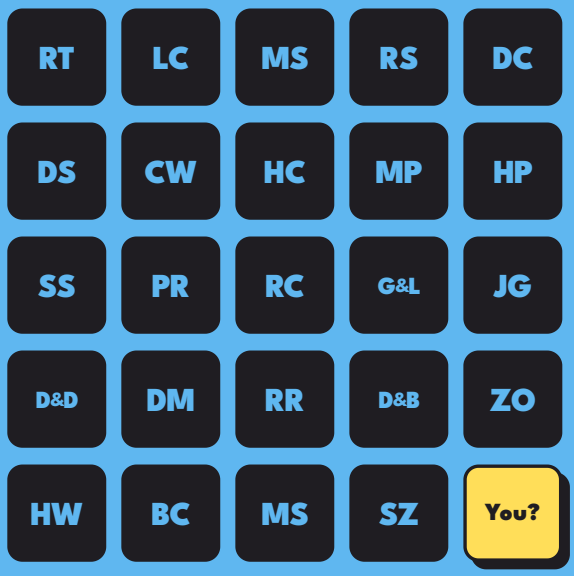
A HUBFORGE MEL BOOK

HUBFORGE.

Giants of Evidence

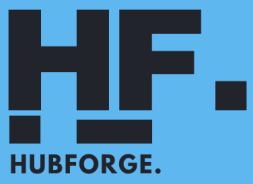
24 lives. 67 big ideas.
One question: how would we know?



The lives and big ideas that built monitoring, evaluation and learning.

Curated by **Shashidhar S A**

A HUBFORGE GLOBAL IMPACT NETWORK FOUNDATION BOOK


HF.
HUBFORGE.

Giants of Evidence

The lives and big ideas that built monitoring, evaluation and learning

THE HUBFORGE MEL SERIES

Books for people who want evidence to change lives



A HUBFORGE MEL BOOK

Giants of Evidence

The lives and big ideas that built monitoring, evaluation and learning.

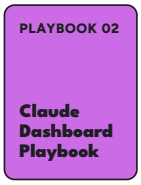
THIS BOOK



PLAYBOOK 01 · DESIGN

Designing Dashboards with AI

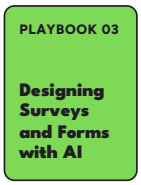
Make dashboards people can use to act. Follow the six-stage DESIGN framework, bring AI in as a collaborator and try each idea with practical exercises.



PLAYBOOK 02 · EVIDENCE

Claude Dashboard Playbook

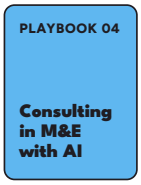
Build evidence dashboards: start with a prompt, connect your data, explore results, trace numbers to their sources and share your work.



PLAYBOOK 03 · BETTER QUESTIONS

Designing Surveys and Forms with AI

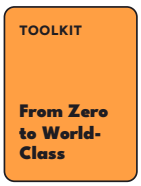
Design surveys and forms with the six-stage SURVEY method, question design tools, a workbook and XLSForms.



PLAYBOOK 04 · CONSULT

Consulting in M&E with AI

Win, run and land monitoring and evaluation assignments with the seven-stage CONSULT method, 36 printable templates and terms of reference checks.



THE M&E CONSULTING TOOLKIT

From Zero to World-Class

36 templates and 36 concept guides: 21 tools across the seven CONSULT stages, nine documents every consultant writes and six tools for your brand and network.

All HubForge playbooks: playbooks.hubforgeglobal.com

A HUBFORGE MEL BOOK

Giants of Evidence

The lives and big ideas that built monitoring, evaluation and learning



Curated by **Shashidhar S A**

HUBFORGE.

HubForge Global Impact Network Foundation · Edition 1 · 2026

COPYRIGHT AND LICENCE

Giants of Evidence

The lives and big ideas that built monitoring, evaluation and learning. Curated by Shashidhar S A. Edition 1, 2026.

ISBN for the print edition: to be assigned. A digital edition of this book is available free from HubForge.

Published by HubForge Global Impact Network Foundation, a registered not-for-profit (Section 8) company in Karnataka, India. **www.hubforgeglobal.com** · **hubforgeglobal@gmail.com**

Copyright © 2026 HubForge Global Impact Network Foundation. The text, design, emblems and other original artwork in this book are licensed under the **Creative Commons Attribution NonCommercial NoDerivatives 4.0 International licence (CC BY-NC-ND 4.0)**. You may copy and share the complete, unaltered book for non-commercial purposes with credit to the publisher. Selling the book, or sharing changed versions of it, needs written permission from the publisher.

Worksheets, templates and reflection pages may be copied, filled in and adapted for personal, classroom and organisational use, provided the source is credited and they are not sold.

Third-party material is not covered by this licence. The five portraits adapted from photographs remain under their own Creative Commons licences, listed in Picture credits on **p. 590**. Names of people, organisations, publications and methods belong to their owners and are used only to identify and describe them.

Please read the Legal notices on **p. 593** before relying on this book. In short: it is general information compiled in good faith from public sources, not professional advice; it may contain errors; the views and lessons are HubForge's own reading; and no person profiled has endorsed it.

HOW TO CITE THIS BOOK

Shashidhar S A (curator). Giants of Evidence: The lives and big ideas that built monitoring, evaluation and learning. Edition 1. HubForge Global Impact Network Foundation, 2026.

THE QUESTION THAT BUILT A FIELD

How would we know?

Every programme is a promise. This book is about the people who taught the world how to check, and about the evidence that lets good intentions become real change.

PROLOGUE

The Question That Built a Field

Every programme is a promise. A school promises that children will learn, a clinic that mothers will be safer, a cash transfer that a family will eat better next month. Monitoring, evaluation and learning exists because promises need checking, and because good intentions alone have never been enough to tell us whether lives actually changed. The field in this book was built around one stubborn question, asked again and again in new settings: how would we know?

The question this book asks

How would we know?

The question took modern form in the 1930s. When around thirty American secondary schools were freed to redesign their curricula, after more than 250 colleges agreed to waive their usual entrance rules, someone had to show credibly what those schools achieved. From 1934 that task fell to Ralph Tyler, who worked from each school's own purposes; the study later tracked 1,475 of their graduates through college. Three decades on, the question grew to the size of governments. An amendment demanded by Senator Robert Kennedy required evaluation of programmes funded under Title I of the Elementary and Secondary Education Act of 1965, and the demand for evaluation that followed helped drive the creation of many of the models still taught today. In 1986 two American societies merged to form the American Evaluation Association. In December 2014 the United Nations General Assembly adopted a resolution on building evaluation capacity in countries themselves.

The answers were never simple, and the people in this book rarely agreed. Some trusted experiments; others trusted the people a programme served. Some said evaluation must deliver a verdict; others said it must be used, or it had failed. Some built methods in American universities; others insisted that evaluation in Africa and South Asia must be rooted in the knowledge and values of

the people evaluated. In our reading, that argument is not a weakness of the field. It is the engine that kept the original question honest as it split into many: did it work, for whom, why, by whose standards, and who gets to decide?

This book tells twenty four life stories of people who carried that question further than anyone before them. They include psychologists and philosophers, a former district officer in Kenya, Nobel prize winning economists whose experiments included an Indian study that, in the Nobel committee's account, led to remedial tutoring benefiting more than five million children, and evaluators who built movements across continents. None of them began as famous. Each began with a problem that would not go away. Read their stories as an invitation: the question is still open, and the next answer may be yours.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340 · D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476 · BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

The question that built a field

How would we know?

Contents

Prologue: The Question That Built a Field — p. 8

How to use this book — p. 11

ACT 1 What Did They Learn? Measuring What Matters — p. 16

RT — Ralph W. Tyler
Chapter 01 · 1930s to 1940s — p. 20

LC — Lee J. Cronbach
Chapter 02 · 1950s to 1980s — p. 38

MS — Michael Scriven
Chapter 03 · 1960s to 2000s — p. 60

RS — Robert E. Stake
Chapter 04 · 1960s to 1990s — p. 82

Great debate: Goals or No Goals? — p. 104

ACT 2 The Experimenting Society — p. 108

DC — Donald T. Campbell
Chapter 05 · 1950s to 1970s — p. 112

DS — Daniel L. Stufflebeam
Chapter 06 · 1960s to 2000s — p. 134

CW — Carol H. Weiss
Chapter 07 · 1960s to 1990s — p. 152

HC — Huey-tsyh Chen
Chapter 08 · 1980s to 2010s — p. 174

Great debate: Proof Here or Promise Elsewhere? — p. 196

ACT 3 Will Anyone Use It? The Use Revolution — p. 200

MP — Michael Quinn Patton
Chapter 09 · 1970s to 2020s — p. 204

HP — Hallie Preskill
Chapter 10 · 1990s to 2010s — p. 226

SS — Sanjeev Sridharan
Chapter 11 · 2010s to 2020s — p. 246

PR — Patricia J. Rogers
Chapter 12 · 2000s to 2020s — p. 268

Great debate: Use or Judgement? — p. 290

ACT 4 Many Voices, Many Truths — p. 292

RC — Robert Chambers
Chapter 13 · 1980s to 2010s — p. 296

G&L — Egon G. Guba and Yvonna S. Lincoln
Chapter 14 · 1980s — p. 318

JG — Jennifer C. Greene
Chapter 15 · 2000s to 2010s — p. 340

D&D — Rick Davies and Jess Dart
Chapter 16 · 1990s to 2000s — p. 362

DM — Donna M. Mertens
Chapter 17 · 1990s to 2020s — p. 384

Great debate: The Paradigm Wars — p. 402

ACT 5 **Evidence for a World in a Hurry** p. 406

RR **Ray C. Rist** p. 410
Chapter 18 · 2000s

D&B **Esther Duflo and Abhijit Banerjee** p. 432
Chapter 19 · 2000s to 2010s

ZO **Zenda Ofir** p. 454
Chapter 20 · 2000s to 2020s

HW **Howard White** p. 476
Chapter 21 · 2000s to 2020s

BC **Bagele M. Chilisa** p. 498
Chapter 22 · 2010s to 2020s

MS **Marco Segone** p. 516
Chapter 23 · 2010s to 2020s

SZ **Sonal Zaveri** p. 536
Chapter 24 · 2000s to 2020s

Great debate: The Randomistas and Their Critics p. 554

The family tree of evaluation p. 558

Epilogue: Your Turn to Ask p. 560

Problem crosswalk p. 562

Write your own story p. 565

Glossary p. 566

Notes and sources p. 569

Picture credits p. 590

An honest note on method p. 591

About HubForge p. 602

About the curator p. 604

Also from HubForge p. 605

Why learn MEL through people?

START HERE

Most of us meet evaluation as a set of finished tools: a logframe template, a theory of change workshop, a sampling formula, a story collection form. They arrive without their history, so they can feel arbitrary, and we use them mechanically.

Every one of those tools began as somebody's answer to a real problem. An evaluator needed to show whether experimental schools still prepared students for college. A field worker in Bangladesh needed to summarise progress without a fixed list of indicators. Economists working with an Indian NGO wanted to know whether a remedial tutoring programme actually helped children learn. When you know the problem behind a tool, you know when to use it, when to bend it and when to put it down.

People also make ideas memorable. It is easier to recall a person's turning point than a definition, and easier to argue with a thinker than with a box on a form. This

book uses 24 stories to help you hold the field in your head as a living conversation, with disagreements, rather than a list of methods.

Inside every chapter

0 Opener
A portrait or emblem, the opening lead, and the line that sums up their contribution.

1 Life and the spark
Where they came from, who shaped them, and what first drew them to evaluation.

2 The breakthrough
The problem in front of them, what they did, and a timeline of the life and work.

3 Impact, debates and why it matters to you
What changed, where people still disagree, the legacy in numbers and the key concepts.

4 Big ideas, complete
Each idea on its own spread: the problem, what it is, how it works, a worked example, when to use it and common misreadings.

5 Put it to work
Steps for Monday, an AI prompt that will not invent evidence, lessons from their path, and a full-page worksheet.

How to use this book

1 Read a story

Each profile opens with a short, factual story in three parts: how the person came to evaluation, the turning point, and what changed because of them. Read one a day or one a week. The order is roughly chronological, but every profile stands alone.

2 Try the tool

Each profile turns the big idea into practice: three steps to use it on Monday, a fill-in template and an AI prompt you can copy. Use them on a live assignment, not a hypothetical one. The prompts are built to stop the AI from inventing evidence; keep it that way.

3 Reflect

End with the reflection question and the links to other figures. The connections show where ideas agree, extend or argue with each other. Write your answer down; over time your notes become your own practice philosophy.

Who this book is for

PATH 1

New to MEL

Students, fellows, programme staff moving into M&E, first-job analysts.

Start with the families of ideas and the timeline to see the shape of the field. Then read Tyler, Scriven, Weiss, Patton, Chambers and Davies and Dart. Use the glossary freely, and try one template per week on whatever you are working on.

RT Tyler 20 | MS Scriven 60 | CW Weiss 152 | MP Patton 204 | RC Chambers 296 | D&D Davies and Dart 362

PATH 2

Practising consultant or MEL lead

Evaluators, MEL managers and consultants who design and deliver studies.

Go straight to the problem crosswalk near the end. Find the problem on your desk this week, read the two or three figures it points to, and use their templates in your inception report. Use the limits sections to anticipate pushback.

PATH 3

Leading evaluation functions and ecosystems

Heads of evaluation, commissioners, funders, network builders and teachers of evaluation.

Read the profiles of institution builders: Stufflebeam, Rist, White, Rogers, Segone, Ofir, Chilisa and Zaveri. Notice how ideas spread through centres, handbooks, training schools and networks. Use the career lessons and reflection questions with your team.

DS Stufflebeam 134 | RR Rist 410 | HW White 476 | PR Rogers 268
MS Segone 516 | ZO Ofir 454 | BC Chilisa 498 | SZ Zaveri 536

PLEASE NOTE

Lessons, advice, templates and prompts are HubForge's own reading of each person's ideas. They are not claims about what the person said or would recommend.

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

SEVEN FAMILIES OF IDEAS

Methods, rigour and causality

How do we know a claim is true? Designs, measures and quality criteria that make evidence trustworthy, from field experiments to qualitative inquiry.

LC Cronbach 38
DC Campbell 112
G&L Guba and Lincoln 318

Value and judgement

Evaluation means judging worth: whose values count, which criteria apply and how to reach a defensible verdict.

MS Scriven 60
RS Stake 82
JG Greene 340

Use and learning

Evaluation that people actually use, built around decisions, intended users, complexity and organisational learning.

DS Stufflebeam 134
MP Patton 204
HP Preskill 226
SS Sridharan 246

Theory and programme logic

Making a programme's purposes and assumptions explicit, from stated objectives to theories of change, and then testing them.

RT Tyler 20
CW Weiss 152
HC Chen 174
PR Rogers 268

Participation, equity and voice

Putting the people a programme serves at the centre of who asks the questions and whose reality counts.

RC Chambers 296
D&D Davies and Dart 362
DM Mertens 384

Made in the Global South

Evaluation grounded in African, South Asian and other Global South knowledge, values, networks and leadership.

ZO Ofir 454
BC Chilisa 498
SZ Zaveri 536

Impact evaluation and evidence systems

Counterfactuals, evidence synthesis, results-based systems and the national and global infrastructure of evaluation.

RR Rist 410
D&B Duflo and Banerjee 432
HW White 476
MS Segone 516

ACT 1 OF 5

Act 1

1

1930s to 1960s

What Did They Learn? Measuring What Matters



Ralph W. Tyler
p. 20



Lee J. Cronbach
p. 38



Michael Scriven
p. 60



Robert E. Stake
p. 82

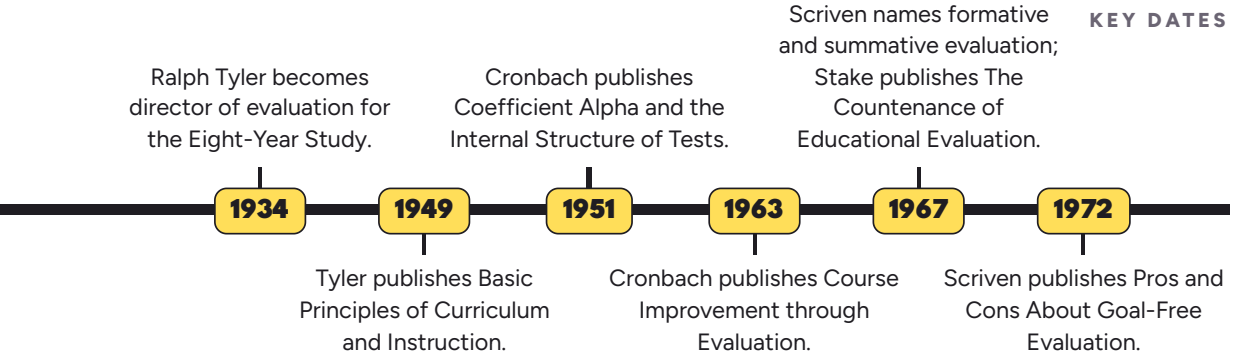
ACT 1 · 1930S TO 1960S

What Did They Learn? Measuring What Matters

In the early 1930s, testing in American schools largely meant checking whether pupils had memorised content, and secondary schools were tightly bound to college entrance requirements. When the Eight-Year Study freed around thirty schools to experiment, Ralph Tyler faced a problem no standard test could solve: each school had its own purposes. His answer, from 1934, was to start from those purposes, define what each would look like in students' work, and build a range of instruments to gather evidence on them. He set out the logic in 1949 in Basic Principles of Curriculum and Instruction, and is widely credited with bringing the word evaluation into the vocabulary of schooling.

His research assistant in 1938, Lee Cronbach, went on to transform measurement itself, with coefficient alpha in 1951 and construct validity, with Paul Meehl, in 1955. In 1963 Cronbach turned back to evaluation and argued that it does most good while a course is still being developed, feeding improvement, rather than only judging a finished product. By the mid 1960s new federal funding was producing ambitious curriculum reforms, yet evaluations often reported little beyond average test scores. In 1967 two answers appeared. Michael Scriven named formative and summative evaluation and argued, against Cronbach, that judging overall worth mattered as much as supporting improvement. Robert Stake argued that a full evaluation must both describe and judge, comparing what was intended with what

KEY DATES



was observed. In 1972 Scriven went further still, proposing that evaluators could set stated goals aside and search for all of a programme's effects.

In our reading, this first act moves from a single question, did pupils reach the objectives, to a harder one: were these the right objectives, and what else happened along the way? Every logframe and every results framework in use today still lives inside that tension.

THE GREAT DEBATE OF THIS ACT

Goals or No Goals? Should a programme be judged against its own stated objectives, or against everything it actually does? p. 104

THE GIANTS OF THIS ACT



Ralph W. Tyler

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

p. 20



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Michael Scriven

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

p. 60



Robert E. Stake

The testing specialist who argued that evaluation should listen to what people around a programme actually care about.

p. 82

Ralph W. Tyler, in one line

He taught educators to ask what they intend and how they will know, and made those questions so natural that we forget someone had to ask them first.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536



Portrait adapted from a photograph; credit on p. 590.

Ralph W. Tyler

1902 to 1994 Active 1930s to 1940s

United States: schools, universities and national policy bodies

Theory and programme logic | Use and learning

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

IN ONE LINE

He taught educators to ask what they intend and how they will know, and made those questions so natural that we forget someone had to ask them first.

In 1934, one of the boldest experiments in American schooling was close to stalling before it had properly begun. More than 250 colleges had agreed to waive their admission rules so that about thirty secondary schools could rebuild their teaching from scratch. There was one problem. Nobody could agree how to judge them. The standard test on offer measured memorised facts, and the schools were trying to teach something else entirely. Their answer arrived in the form of a young testing specialist from Ohio State, barely into his thirties. Ralph Tyler threw out the idea of one test for everyone. He asked each school a simpler question first: what are you actually trying to do? Eight years and 1,475 tracked graduates later, the experiment had its answer. And schools had a new word for what Tyler did: evaluation.

Life CHAPTER 01

Picture a boy of twelve in Nebraska, washing and weighing milk cans at a creamery, and later tasting the cream to judge its quality. That was Ralph Winfred Tyler, born in Chicago on 22 April 1902 and raised on the plains after his family moved west in 1904. He was the sixth of eight children. His father had trained as a doctor and became a minister. Money was so short that, as his Stanford obituary records, the family trapped animals for food and wore clothes from a missionary basket.

He finished his degree at Doane College in Crete, Nebraska, in 1921, aged just 19, and he paid for it partly by working nights as a railway telegraph operator. He taught high school science in Pierre, South Dakota, then moved into training science teachers at the University of Nebraska, where a master's thesis on statistics in testing set his direction. By 1927 he had a PhD from the University of Chicago. Two teachers there

BIG IDEAS

1. **The Tyler rationale**
2. **Objectives-based evaluation with varied evidence**

KEY WORKS

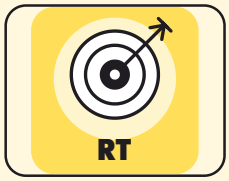
Basic Principles of Curriculum and Instruction
1949

Evaluation volume of the Eight-Year Study
1942 · With Eugene R. Smith

DID YOU KNOW?

As a student at Doane College, Tyler worked nights as a railway telegraph operator, and from the age of 12 he had worked at a creamery, where his jobs eventually included tasting cream.

In his later years he kept up such a heavy consulting schedule that he was reported to fly up to 200,000 air miles a year.



shaped everything that followed. Charles Judd showed him that learning only matters if it travels to new situations, so goals have to be stated broadly enough to travel too. W. W. Charters opened the door to Ohio State.

From there his career reads like a climb. Two years helping teachers in North Carolina redesign their courses. Ohio State from 1929, running achievement testing. Then, in 1934, the call to direct evaluation for the Eight-Year Study. Chicago in 1938, as head of education and chief examiner of the university, and dean of the social sciences from 1948. And in 1953, the founding directorship of a new centre at Stanford where some of the best minds in the behavioural sciences came to think, a post he held until 1967.

Most people would have stopped there. Tyler started a second career. He advised six United States presidents, helped shape the landmark 1965 law that funded schools for poorer children, became the first president of the National Academy of Education and sat on the National Science Board. He took his ideas to Ghana, Indonesia, Ireland, Israel and Sweden, and collected 22 honorary doctorates. He was still working as a visiting scholar when he died of cancer in San Diego on 18 February 1994, aged 91.

The spark

The spark was a worrying number. When Tyler arrived at Ohio State in 1929, the state legislature was alarmed that about half of all new students never came back for a second year. Teachers of botany, zoology and agriculture, struggling with high failure rates, came to him for better tests. He looked at what they were using and saw the problem: the tests only checked whether students had memorised the textbook. So instead of writing a better test, he asked a better question. What do you actually want your students to be able to do? Finding that out, he told them, should be called evaluation.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 01

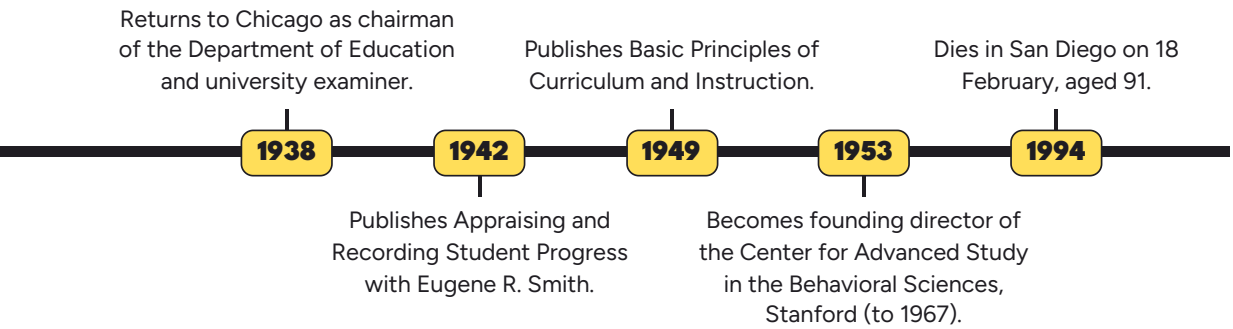
To see why the Eight-Year Study mattered, imagine being a head teacher in America in 1930. You want to teach students to think, to question, to create. But your graduates must still get into college, and colleges only accept students who have covered a fixed list of subjects. Any school that experimented was gambling with its students' futures. So in 1930 the Progressive Education Association made a deal. More than 250 colleges agreed to set aside their usual entry rules for graduates of about thirty experimental schools. The schools were free. But someone had to prove, convincingly, whether freedom worked.

That someone was Tyler. He refused to judge thirty different schools with one standard test. Instead, he and his team worked with each school, wrote down what it was really aiming for, worked out what each aim would look like in a student's actual work, and then built the tools to look for it: tests, yes, but also records, observations and samples of work. Then came the decisive comparison. The study matched 1,475 graduates of the experimental schools with similar students from ordinary schools and followed them through college. The experimental group did about as well in their studies, and were more active in cultural and artistic life. Freedom had not cost them anything. In 1942 Tyler and Eugene R. Smith published how it was done, and in 1949 he boiled the whole approach down into one slim book, Basic Principles of Curriculum and Instruction, which had started life as the syllabus for his own course.

Did you know?

As a student at Doane College, Tyler worked nights as a railway telegraph operator, and from the age of 12 he had worked at a creamery, where his jobs eventually included tasting cream.

TIMELINE, CONTINUED



The problem it solved

Schools had no systematic, defensible way to judge a programme beyond marks on conventional tests.

Impact WHAT CHANGED

That slim book became one of the most influential ever written about teaching. It stayed in print for decades, went through dozens of printings and translations, and in 2000 was named among the books of the century in Educational Leadership. Then, in 1963, the United States Commissioner of Education, Francis Keppel, asked Tyler a simple but enormous question: how could anyone know how good the nation's schools were? Tyler's answer became the National Assessment of Educational Progress. He chaired its planning committee from 1964, the first data were collected in 1969, and it still runs today as the country's report card on what students know.

Tyler's fingerprints are everywhere, usually unnoticed. In our reading, every logframe, results framework and outcome indicator in development work today starts with his question about what a programme intends to achieve and whether it did. Sample-based learning surveys in many countries, which measure how well a whole system is doing rather than ranking individual children, echo the model he helped build. His questions became so ordinary that people now ask them without knowing anyone asked them first. That is perhaps the greatest compliment an idea can receive.

WHERE PEOPLE DISAGREE

Debates

Later curriculum scholars criticised the Tyler rationale as managerial and linear, and linked it to an efficiency-minded tradition that treats schooling like production. Tyler did not formally reply, holding that his critics offered no workable alternative, and he did not fully accept the label of father of behavioural objectives often attached to him. In evaluation, critics argued that judging a programme only against its own objectives can miss unintended effects and cannot say whether the objectives were worth pursuing; Michael Scriven's goal-free evaluation and Robert Stake's responsive evaluation both grew partly from this concern. In our view, the approach also risks narrowing attention to whatever is easiest to measure.

LIMITS, IN OUR READING

- [ ] Judging only against pre-set objectives can miss unintended effects and says little about whether the objectives were worth pursuing; later thinkers, notably Michael Scriven, pushed on exactly this gap.
- [ ] In our view, the approach can tempt teams towards narrow outcomes that are easy to measure.

WHY THIS MATTERS TO YOU

You do not have to be an evaluator to use Tyler. Every time you set a goal, at work, in a project or in your own life, and then ask how you will know you are getting there, you are thinking his way. His sharper lesson is the second half. Big goals, like confidence, curiosity or the courage to negotiate, deserve evidence that fits them, not whatever measurement happens to be easy. When a plan squeezes something rich into a thin number, go back to the purpose, describe what success would look like in real life, and look for it in more than one way. That is when measuring stops being a judgement and starts being help.

1,475

graduates of the Eight-Year Study schools followed into college and compared with matched peers

Eight-Year Study, Education Encyclopedia, StateUniversity.com

6

United States presidents he advised

Stanford University News Service obituary (1994); Encyclopedia of Evaluation, SAGE (2005)

22

honorary doctorates

Stanford University News Service obituary (1994)

THE BIG IDEAS OF THIS CHAPTER

1. **The Tyler rationale** p. 28
   Plan any programme by answering four linked questions in order: what it should achieve, which experiences will get there, how to organise them, and how to tell whether the purposes are being attained.
2. **Objectives-based evaluation with varied evidence** p. 31
   Evaluate by finding out how far a programme's own intended changes in people have actually happened, using whatever kinds of evidence suit each objective rather than a single test.

Key concepts

Tyler rationale
The four-question logic of Basic Principles of Curriculum and Instruction (1949): set purposes, select experiences, organise them and evaluate whether the purposes are attained.

Educational objective
A statement of the change in a learner's behaviour that a programme intends to bring about, written precisely enough to be checked.

Eight-Year Study
A Progressive Education Association study of about thirty secondary schools in the 1930s and early 1940s that tested alternative curricula; Tyler directed its evaluation.

Service study
Tyler's term at Ohio State for applied studies that helped university staff improve teaching, an early model of evaluation in service of practice.

National Assessment of Educational Progress (NAEP)
The continuing national assessment of what American students know and can do, first administered in 1969, whose design Tyler's advice helped launch.

Sources for this chapter: p. 574

GIANTS CONNECTED TO THIS CHAPTER



Michael Scriven
Goal-free evaluation deliberately looks beyond stated objectives to actual effects and needs.
p. 60



Daniel L. Stufflebeam
CIPP widened the frame from outcomes to context, inputs and process, to serve decisions throughout a programme.
p. 134



Carol H. Weiss
Theory-based evaluation asks how activities are supposed to reach the objectives, step by step.
p. 152

KEY WORKS TO READ IN THE ORIGINAL

- **Basic Principles of Curriculum and Instruction** 1949
- **Evaluation volume of the Eight-Year Study** 1942, With Eugene R. Smith

In one line

He taught educators to ask what they intend and how they will know, and made those questions so natural that we forget someone had to ask them first.

1

BIG IDEA 1 OF 2 · TYLER

The Tyler rationale

Plan any programme by answering four linked questions in order: what it should achieve, which experiences will get there, how to organise them, and how to tell whether the purposes are being attained.

PICTURE IT

Think of planning a journey. First you decide where you want to go. Then you choose the route, then the order of the stops, and only then can you check whether you have arrived. Most schools in Tyler's day planned the stops without ever naming the destination. His four questions put the destination first.

The problem

In the 1930s and 1940s American schools planned courses largely as lists of content to cover, and testing usually meant checking recall of textbook facts. At Ohio State, Tyler met a legislature worried that about half of entering students did not return for a second year, and instructors who could not say what their tests were really measuring. Without clear purposes, nobody could judge whether a course worked, improve it sensibly, or explain to families and colleges what pupils had gained.

What it is

The Tyler rationale is the logic of Basic Principles of Curriculum and Instruction (1949), set out as four questions that are also the book's first four chapters. First, what educational purposes should the school seek to attain? Tentative objectives are drawn from three sources: studies of the learners, studies of contemporary life outside school, and suggestions from subject specialists. They are then passed through two screens, the school's philosophy of education and the psychology of learning, leaving a smaller set of important, attainable objectives. Second, which learning experiences are likely to attain them? Third, how can those experiences be organised, with continuity and sequence, for cumulative effect? Fourth, how can we determine whether the purposes are being attained? The underlying logic is a loop: evaluation checks how far the intended changes in learners have occurred, and the findings send planners back to revise objectives and experiences. Tyler said he offered a rationale, a set of questions educators should ask, not a curriculum theory.

How it works

1. **Gather candidate purposes**
   Collect possible objectives from evidence about learners, from studies of life in the community and from subject experts, the three sources named in Basic Principles and traced by William Wraga (2017) to Tyler's Chicago course outline.
2. **Screen them**
   Filter the list through the organisation's stated philosophy and through what is known about how people learn, dropping objectives that conflict with its values or are unrealistic for the learners.
3. **State each objective clearly**
   Write each objective precisely enough that people can agree on whether it has been reached, translating vague words such as understanding into observable actions, as George Madaus (1967) summarised Tyler's approach.
4. **Select learning experiences**
   Choose activities that give learners real chances to practise what each objective describes.
5. **Organise for cumulative effect**
   Sequence the experiences so that each builds on the last across weeks, terms and grades.
6. **Evaluate and revise**
   Gather evidence on how far each purpose is being attained, then use the results to revise objectives, experiences or organisation, the cycle that Tyler's fourth question and final chapter on staff work describe.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO running after-school maths clubs in a district of Odisha wants to redesign its curriculum. Following Tyler, the team gathers evidence from baseline assessments of children, from parents describing the arithmetic they use at markets and in farming, and from teacher educators. It screens the long list against its own commitment to joyful learning and against what children aged eight to ten can realistically master, and keeps four objectives, including that children can solve two-step money problems from daily life. Club sessions are built around market role plays, sequenced from one-step to two-step problems over a term. At the end of the term facilitators record children solving fresh money problems aloud. Because most children stumble on the second step, the team revises the sequence before the next term begins.

Did you know?

In his later years he kept up such a heavy consulting schedule that he was reported to fly up to 200,000 air miles a year.

USE IT WHEN

- You are designing or redesigning a programme or curriculum and need a shared planning logic.
- A team cannot say what a programme is for, or its activities do not match its stated objectives.
- Staff with little evaluation training need a simple, repeatable way to connect purposes, activities and evidence.

SKIP IT WHEN

- The programme is emergent and objectives cannot sensibly be fixed in advance.
- The main question is about unintended effects, or whether the purposes are worth pursuing at all, where goal-free or values-focused approaches fit better.

COMMON MISREADINGS

- Treating the four questions as a rigid one-way production line, when Tyler presented them as a rationale for inquiry, with evaluation feeding back into revision.
- Assuming the objectives must be narrow behavioural targets measured by tests, a label Tyler did not fully accept.

WHY IT WAS GROUNDBREAKING

It turned curriculum planning from an argument about content into a sequence of answerable questions that any school staff could work through. By placing evaluation inside the planning cycle rather than after it, it made judging results part of how a programme is designed, a structure that, in our reading, today's results frameworks still echo.

Origin

The rationale grew out of Tyler's evaluation work for the Eight-Year Study from 1934 and the syllabus for his University of Chicago course Education 360, whose 1947 syllabus and 1948 outline William Wraga examined in 2017. The University of Chicago Press published it as Basic Principles of Curriculum and Instruction in 1949 and reissued it in 2013, still in print, with a foreword by Peter S. Hlebowitsh.

RELATED IDEAS

Objectives-based evaluation with varied evidence

DS Stufflebeam 134 · MS Scriven 60 · CW Weiss 152 · HC Chen 174

2

BIG IDEA 2 OF 2 · TYLER

Objectives-based evaluation with varied evidence

Evaluate by finding out how far a programme's own intended changes in people have actually happened, using whatever kinds of evidence suit each objective rather than a single test.

PICTURE IT

Think of a doctor checking a patient's health. No good doctor relies on one reading. Temperature, pulse, blood tests and a conversation each show something different. Tyler asked evaluators to treat a programme's goals the same way: choose the right check for each goal, instead of forcing everything through one test.

The problem

When the Eight-Year Study began, standard achievement tests mostly checked memorised facts. In 1934 the participating schools rejected a standard test because it did not match what they were trying to teach, and the study's directing committee faced a crisis.

Ambitious aims such as critical thinking, interests or social sensitivity seemed impossible to evidence, so innovative schools risked being judged on narrow measures that ignored what they valued most, with their students' college places at stake.

What it is

Objectives-based evaluation judges a programme by the extent to which its objectives are achieved by the people it serves. In Tyler's version, from his 1942 General Statement on Evaluation as summarised in the Encyclopedia of Evaluation, the procedure has seven steps: formulate the objectives; classify them into major types; define each type in terms of behaviour; identify situations in which learners can be expected to show that behaviour; select and try out promising methods of gathering evidence on each type; choose and refine the most promising methods in the light of those trials; and devise ways of interpreting and using the results. Two principles run through it. Methods must match the outcome, so tests, observation, records, questionnaires and samples of work can all count. And results exist to be used: in the Eight-Year Study, evaluation served records for colleges, continuing appraisal of learning and follow-up after graduation. In a 1981 interview Tyler added that objectives are necessary but not sufficient, because an evaluator must also see what else is happening.

How it works

1. **Formulate objectives**
   Start from what the programme's own staff are trying to achieve, since Tyler held that teachers are best placed to articulate goals.
2. **Classify them**
   Group objectives into major types, such as knowledge, skills, thinking, interests or social attitudes, so that similar evidence can serve similar aims.
3. **Define the behaviour**
   For each type, describe what a person who has reached the objective actually does.
4. **Find the situations**
   Identify where and when people can be expected to show that behaviour, in sessions, at home, at work or in records.
5. **Try out methods**
   Select several promising ways of gathering evidence for each type, such as tests, observation, questionnaires or work samples, and pilot them.
6. **Refine the best**
   Keep and improve the methods that worked in trial, and drop those that proved unworkable or uninformative.
7. **Interpret and use**
   Decide in advance how results will be read and who will act on them, then feed findings back into programme design, as the 1942 procedure's final step requires.

WORKED EXAMPLE · AN ILLUSTRATION

A life skills programme for adolescent girls in Jharkhand states that girls will be able to negotiate decisions about their schooling at home. A test cannot easily show this. Following Tyler's steps, the evaluation team classifies it as a social and communication objective, defines the behaviour as raising the topic with family members and giving reasons, and identifies situations where it might appear: family discussions, school re-enrolment periods and role plays in sessions. It tries out three methods, facilitator observation checklists, short structured interviews with girls and mothers, and school enrolment records, and drops the checklist when facilitators find it unworkable. The findings go back to the programme team before the next cohort starts, and the team adds practice sessions for girls who had not yet raised the topic at home.

In one line

Evaluate by finding out how far a programme's own intended changes in people have actually happened, using whatever kinds of evidence suit each objective rather than a single test.

USE IT WHEN

- Objectives are agreed and the main question is how far they were met.
- Aims such as attitudes, skills or behaviours need evidence beyond test scores.
- A team wants findings it can use before the next cycle of delivery.

SKIP IT WHEN

- Objectives are vague, contested or imposed, so measuring against them would legitimise poor goals.
- Side effects or harms are a major concern; goal-free evaluation is then a needed complement.

COMMON MISREADINGS

- [ ] Reducing it to checking logframe indicators at endline, when Tyler built in trial of methods and use of results throughout.
- [ ] Believing it requires standardised tests, when the Eight-Year Study's point was that methods must fit the objective.

WHY IT WAS GROUNDBREAKING

It showed that ambitious aims could be evidenced without reducing them to recall tests, by matching methods to purposes. The Encyclopedia of Evaluation notes that linking objectives to outcome measures remained the dominant paradigm in evaluation for almost half a century.

Origin

Tyler proposed the word evaluation for this kind of inquiry in his Ohio State work after 1929, when instructors in botany, zoology and agriculture sought his help, as he recalled in a 1981 interview reprinted in 2011. He scaled it up as director of evaluation for the Eight-Year Study from 1934 and published it in Appraising and Recording Student Progress (Smith and Tyler, 1942). Later objectives-based models, such as W. James Popham's, narrowed it to behavioural objectives.

RELATED IDEAS

The Tyler rationale

MS Scriven 60
RS Stake 82
LC Cronbach 38
DS Stufflebeam 134

CONCEPT HIGHLIGHT

Tyler rationale

The four-question logic of Basic Principles of Curriculum and Instruction (1949): set purposes, select experiences, organise them and evaluate whether the purposes are attained.

Put it to work

USE IT ON MONDAY · OBJECTIVES-BASED EVALUATION (THE TYLER RATIONALE)

1. Before you design a single tool, write the programme's stated objectives on one page in the client's own words, and check with the team that this is what they actually mean.

2. For each objective, ask what evidence would show it was reached, and list at least two kinds of evidence beyond a single test or survey score, such as observation, work samples or records.

3. Schedule one moment mid-programme when results are compared with objectives and fed back into design decisions, instead of waiting for the endline.

AI prompt **Copy, fill the brackets, then check every line**

You are helping me apply objectives-based evaluation, in the tradition of Ralph Tyler, to **[programme name]**, which works with **[target group]** in **[location]**. Here are the programme's stated objectives: **[paste objectives]**. Here is the activity list: **[paste activities]**. For each objective: (1) rewrite it so it is observable and time-bound without changing its meaning, and show the original beside your version; (2) suggest two or three types of evidence that could show whether it was reached, mixing quantitative and qualitative sources; (3) flag objectives that are vague, unmeasurable or not linked to any activity. Do not set baselines or targets yourself.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of planning a journey. First you decide where you want to go. Then you choose the route, then the order of the stops, and only then can you check whether you have arrived. Most schools in Tyler's day planned the stops without ever naming the destination. His four questions put the destination first.

YOUR MOVES

Three moves from Tyler's path

1 Start with the problem in front of you

Tyler began with a practical request from struggling colleagues, not a theory, and the answer he built for one university became a method for a nation.

TRY THIS

List the three questions your programme team keeps asking you. Make one of them the core of your next evaluation plan.

2 Make your framework fit on one page

His four questions about purposes, experiences, organisation and evaluation were simple enough to remember, and that simplicity carried them into schools and ministries for decades.

TRY THIS

Take your current tool or framework and redraw its core logic on a single page. If it does not fit, cut until it does.

3 Speak to every level of the system

He moved between classrooms, research offices and government, advising six United States presidents and countries including Ghana and Indonesia, without losing sight of teachers.

TRY THIS

For your next report, write three versions of the key finding: one for a frontline worker, one for a manager, one for a policymaker.

THE BIG IDEAS OF THIS CHAPTER

1 **The Tyler rationale** p. 28

Plan any programme by answering four linked questions in order: what it should achieve, which experiences will get there, how to organise them, and how to tell whether the purposes are being attained.

2 **Objectives-based evaluation with varied evidence** p. 31

Evaluate by finding out how far a programme's own intended changes in people have actually happened, using whatever kinds of evidence suit each objective rather than a single test.

CONCEPT HIGHLIGHT

National Assessment of Educational Progress (NAEP)

The continuing national assessment of what American students know and can do, first administered in 1969, whose design Tyler's advice helped launch.

GIANTS CONNECTED TO THIS CHAPTER



Michael Scriven

Goal-free evaluation deliberately looks beyond stated objectives to actual effects and needs.

p. 60



Daniel L. Stufflebeam

CIPP widened the frame from outcomes to context, inputs and process, to serve decisions throughout a programme.

p. 134



Carol H. Weiss

Theory-based evaluation asks how activities are supposed to reach the objectives, step by step.

p. 152

KEY WORKS TO READ IN THE ORIGINAL

- **Basic Principles of Curriculum and Instruction** 1949
- **Evaluation volume of the Eight-Year Study** 1942, With Eugene R. Smith

CONCEPT HIGHLIGHT

Tyler rationale

The four-question logic of Basic Principles of Curriculum and Instruction (1949): set purposes, select experiences, organise them and evaluate whether the purposes are attained.

CONCEPT HIGHLIGHT

Educational objective

A statement of the change in a learner's behaviour that a programme intends to bring about, written precisely enough to be checked.

WORKSHEET · TRY THIS TEMPLATE

Objectives-to-evidence table

Print this page, or copy the headings into your notes.

1 Stated objective

Copy it exactly from the proposal or logframe. Circle vague verbs such as improve, empower or strengthen.

2 What reaching it looks like

Describe the observable change in plain language: for whom, how much and by when.

3 Evidence sources

Name at least two, mixing tests, observation, records and participants' accounts.

4 Activities linked

Which activities are meant to produce this change? Flag any objective with no activity behind it.

5 Review point

When results will be compared with intent, and who decides what changes as a result.

REFLECT

On your current assignment, which stated objective would you struggle to evidence, and what does that tell you about the objective itself?

CONNECTS TO

MS **Michael Scriven**
Goal-free evaluation deliberately looks beyond stated objectives to actual effects and needs. p. 60

DS **Daniel L. Stufflebeam**
CIPP widened the frame from outcomes to context, inputs and process, to serve decisions throughout a programme. p. 134

CW **Carol H. Weiss**
Theory-based evaluation asks how activities are supposed to reach the objectives, step by step. p. 152



Emblem: an original HubForge design for this figure, not a likeness.

Lee J. Cronbach

1916 to 2001 Active 1950s to 1980s United States

Methods, rigour and causality · Use and learning

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

IN ONE LINE

He taught us that a number is only as trustworthy as the argument behind it, and that evaluation serves everyone who shapes what happens next.

A four-year-old in Fresno, California, is in the grocer's shop comparing the unit prices of goods. A relative notices, and tells a school psychologist. In 1921 the boy is tested and enrolled in Lewis Terman's famous study of gifted children. He finishes high school at 14. At 22 he is working as Ralph Tyler's research assistant on the Eight-Year Study. And at 35 he publishes a short paper on a statistic for judging whether a test is dependable. By 2016 that paper had been cited 26,889 times on Google Scholar. If you have ever filled in a survey that a researcher later checked for reliability, there is a good chance you have met Lee Cronbach's number without knowing his name.

Life CHAPTER 02

Lee Joseph Cronbach was born on 22 April 1916 in Fresno, California, the son of a salesman and a homemaker. His National Academy of Sciences memoir tells how early his talent showed. A relative saw the four-year-old comparing unit prices at the grocer's and told a local school psychologist, a follower of Lewis Terman. The psychologist tested him in 1921, publicised his very high score and enrolled him in Terman's study of gifted children. His mother pressed for him to be moved ahead at school, and he finished Fresno High School at 14.

He studied chemistry and mathematics at Fresno State College, graduating in 1934, then took a teaching credential and, in 1937, a master's degree in education at the University of California, Berkeley. From 1936 to 1938 he taught mathematics and chemistry at Fresno High School. In 1938 he worked as a research assistant to Ralph Tyler on the Eight-Year Study, an experience his memoir describes as having a lasting influence on his later thinking about evaluation. He completed a PhD in educational psychology at the University

BIG IDEAS

1. **Coefficient alpha and generalizability theory**
2. **Construct validity**
3. **Evaluation for the policy-shaping community**

KEY WORKS

Coefficient Alpha and the Internal Structure of Tests
1951

Construct Validity in Psychological Tests
1955 · With Paul E. Meehl

Toward Reform of Program Evaluation
1980 · With Stanford Evaluation Consortium colleagues

of Chicago in 1940, supervised by Guy T. Buswell.

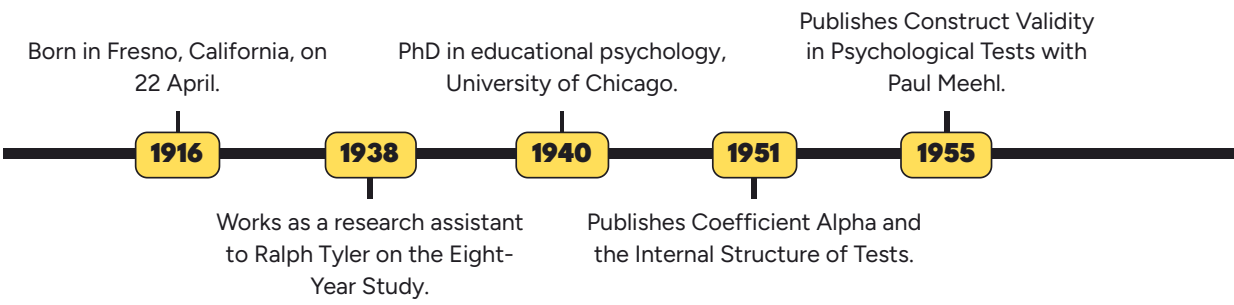
His academic career took him to Washington State University from 1940 to 1946, with wartime leave at a Navy sonar school in San Diego; to Chicago for two years; and to the University of Illinois from 1948 to 1964, where he worked in the Bureau of Educational Research and was among the authors of the proposal that led to the university's evaluation centre, CIRCE, though he left before it opened. In 1964 he moved to Stanford, where he became Vida Jacks Professor of Education and in the 1970s led the Stanford Evaluation Consortium, a group drawing on education, communication and psychology. His textbooks, Essentials of Psychological Testing (1949) and Educational Psychology (1954), were used for decades.

Cronbach was president of the American Psychological Association in 1956 and also led the American Educational Research Association and the Psychometric Society. He was elected to the National Academy of Education in 1965 and the National Academy of Sciences in 1974, and received the APA's Distinguished Scientific Contribution Award in 1973. He married Helen Claresta Bower, and they had five children. He retired from Stanford in 1980 and continued as professor emeritus. He died in Palo Alto on 1 October 2001; a final paper on coefficient alpha, prepared with Richard Shavelson, appeared after his death in 2004.

The spark

Two threads pulled Cronbach into his life's work, and his National Academy of Sciences memoir names both. One was measurement: an early fascination with IQ tests and with Louis Thurstone's scales for measuring attitudes, later sharpened by a dissertation problem at Chicago that left him dissatisfied with how accurately tests measured. The other was evaluation: the year he spent from 1938 as Ralph Tyler's assistant on the Eight-Year Study, which his memoir says shaped his thinking for decades. In our reading, the rest of his career was the work of weaving those two threads together.

TIMELINE · LIFE AND WORK

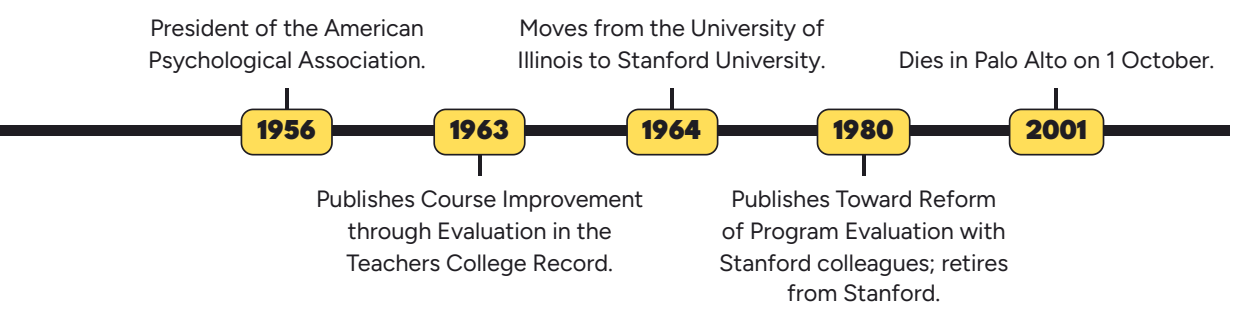


The breakthrough CHAPTER 02

Imagine you have built a test of reading, or of anxiety, or of attitudes. Two questions immediately follow. Does it give consistent results? And does it really measure what you say it measures? In the mid twentieth century, the usual answers were clumsy. To check consistency you gave the test twice, or split it into halves. To check meaning you looked at whether the score predicted something outside it. Cronbach changed both answers. His 1951 paper, Coefficient Alpha and the Internal Structure of Tests, gave researchers one general estimate of reliability from a single sitting. His 1955 paper with Paul Meehl, Construct Validity in Psychological Tests, argued that the meaning of a score has to be built up, like a case in court, from a network of evidence.

His breakthrough for evaluation drew on those early months with Tyler. In 1963, in Course Improvement through Evaluation, published in the Teachers College Record, he made an argument that sounds obvious now and was not then: evaluation is most useful while a course is still being developed, when feedback can still improve it, not only as a final verdict at the end. Over the next two decades he widened this into a full theory of programme evaluation, in Toward Reform of Program Evaluation (1980), written with colleagues from the Stanford Evaluation Consortium, and Designing Evaluations of Educational and Social Programs (1982), written with Karen Shapiro. His conclusion was bold for its time. Evaluation, he argued, is a political activity, and its main audience is not one decision-maker but a whole policy-shaping community.

TIMELINE, CONTINUED



The problem it solved

Researchers lacked a practical reliability check, and evaluations were too often designed for distant verdicts rather than for the people able to act on them.

Impact WHAT CHANGED

Few researchers leave a mark as visible as Cronbach's. Open almost any study in psychology, education, health or the social sciences that uses a questionnaire, and you will probably find coefficient alpha reported in it. In our reading, his work with Meehl underlies the way testing professionals now treat validity as an argument built from evidence rather than a single number. His work with Richard Snow on aptitude-treatment interactions, published in 1977, asked a question every teacher knows: do different methods suit different learners? In evaluation, the Encyclopedia of Evaluation credits him with making the improvement purpose of evaluation legitimate, and with treating evaluation as something that educates a wide community rather than issuing a single verdict. The evaluator Blaine Worthen ranked Toward Reform of Program Evaluation among the most important books ever published in the field.

Our interpretation is that his influence on development practice runs along two tracks. On the measurement side, almost every NGO or government survey that adapts a scale into a new language and reports its reliability is using his statistic, and the more careful ones heed his warnings about what alpha cannot show. On the evaluation side, his question of whether a result will hold for other units, treatments, observations and settings is the same question that arises whenever a programme tested in one district or country is proposed for scale-up elsewhere. Learning-oriented MEL systems that feed findings back to programme teams during implementation also echo, in our reading, his argument for formative evaluation and many smaller studies over one grand verdict.

WHERE PEOPLE DISAGREE

Debates

Cronbach's position on randomised experiments placed him in a long-running debate with Donald Campbell's tradition. Where Campbell prioritised internal validity, Cronbach argued that external validity, whether results transfer, mattered more for policy, and he was sceptical that any single trial could settle a programme's worth because local conditions and later improvements could change the result. Critics in turn have argued that emphasising many local studies can make it harder to reach firm, general conclusions. On measurement, his own memoir notes that alpha is widely misread as proof of unidimensionality or precision, and his final paper, published after his death, revisited alpha and the procedures that succeeded it.

LIMITS, IN OUR READING

- [ ] His National Academy of Sciences memoir notes that alpha is a lower bound on reliability and is often misread as proof that a scale is unidimensional or precise.
- [ ] His scepticism of randomised trials is a contested position, and in our view a strong emphasis on local detail can trade off against findings that generalise.

WHY THIS MATTERS TO YOU

Cronbach's questions belong to anyone who has ever trusted a number. Is this score consistent? Does it mean what it claims to mean? Will this result hold somewhere else? If you report a reliability figure for a survey, you are using his statistic. If you hesitate before trusting a neat index of empowerment or wellbeing, you are heeding his warning. And if you ask whether a pilot that worked in two districts will work in twenty, you are standing on his ground. His lesson is freeing: good evidence is not one magic number. It is a reasoned case that you can build, test and explain to others.

Did you know?

Cronbach was enrolled as a small child in Lewis Terman's famous study of gifted children, and he finished high school at the age of 14.

26,889

Google Scholar citations of the 1951 coefficient alpha paper as of 1 January 2016, by far the most cited article in Psychometrika

Klaas Sijtsma, Invited discussion of Cronbach (1951), Psychometrika (2016), via Tilburg University

3

major professional bodies he led: the American Psychological Association (1956), the American Educational Research Association and the Psychometric Society

Richard J. Shavelson, Biographical Memoir, National Academy of Sciences (2009)

100

Designing Evaluations of Educational and Social Programs was chosen in 2000 among the top 100 education-related books of the twentieth century

Richard J. Shavelson, Biographical Memoir, National Academy of Sciences (2009)

Key concepts

Coefficient alpha (Cronbach's alpha)

A single-administration estimate of internal consistency reliability; it is a lower bound and does not by itself show that a scale is unidimensional.

Construct validity

The degree to which evidence and theory together support interpreting a score as a measure of a particular quality.

Generalizability theory

A framework from 1972 that splits measurement error into sources such as items, raters and occasions, showing which matter most for a given use.

Formative evaluation

Evaluation carried out while a course or programme is still developing, to improve it; Cronbach's 1963 paper helped establish this purpose.

Policy-shaping community

Cronbach's term for the wide audience that evaluation should inform, including decision-makers, advocates, educators, the media and citizens.

UTOS

Units, treatments, observations and settings: Cronbach's framework for asking whether evaluation findings will hold in other contexts.

Sources for this chapter: p. 574

THE BIG IDEAS OF THIS CHAPTER

1. **Coefficient alpha and generalizability theory** p. 46

   Estimate how dependable a measure is from a single administration, then go further and find out which sources of error, such as items, raters or occasions, matter for the decision the scores must support.

2. **Construct validity** p. 49

   A score means what you claim only if a network of theory-based predictions about it keeps being confirmed by evidence; validity is an argument, not a single number.

3. **Evaluation for the policy-shaping community** p. 52

   Evaluation is most useful when it improves programmes while they can still change and informs the whole community that shapes policy, with careful attention to whether results will hold elsewhere.

GIANTS CONNECTED TO THIS CHAPTER



Donald T. Campbell

A contrasting view on how central experiments should be to evaluation.

p. 112



Michael Quinn Patton

Built a whole approach around intended users and intended use.

p. 204



Ralph W. Tyler

The Eight-Year Study influenced Cronbach's turn towards evaluation.

p. 20

KEY WORKS TO READ IN THE ORIGINAL

- **Coefficient Alpha and the Internal Structure of Tests** 1951
- **Construct Validity in Psychological Tests** 1955, With Paul E. Meehl
- **Toward Reform of Program Evaluation** 1980, With Stanford Evaluation Consortium colleagues

1

BIG IDEA 1 OF 3 · CRONBACH

Coefficient alpha and generalizability theory

Estimate how dependable a measure is from a single administration, then go further and find out which sources of error, such as items, raters or occasions, matter for the decision the scores must support.

PICTURE IT

Think of a bathroom scale. If you step on it three times in a minute and get three different weights, you cannot trust any of them. Coefficient alpha asks the same thing of a test: do its questions agree with each other, so that the score is steady enough to rely on?

The problem

Every test, survey scale or observation tool contains error. Before 1951, reliability was commonly estimated by testing people twice or by splitting a test into halves, which was costly, and different splits gave different answers. Researchers also lumped all error together, so they could not tell whether inconsistency came from items, raters or the day of testing. Programmes could therefore make decisions about children, schools or households on scores far less dependable than anyone realised.

What it is

Coefficient alpha, introduced in Cronbach's 1951 paper Coefficient Alpha and the Internal Structure of Tests, estimates how consistently a set of items measures something, from one administration. It equals the mean of all possible split-half coefficients; the older Kuder-Richardson 20 formula is its special case for right-or-wrong items; it works for items with several score levels and is easy to compute. His memoir stresses its limits: alpha is a lower bound and a coefficient of equivalence across items, it does not prove that a scale measures one thing, and it ignores inconsistency across occasions. Generalizability theory, set out in The Dependability of Behavioral Measurements (1972) with Goldine Gleser, H. Nanda and N. Rajaratnam, generalises the idea. The evaluator defines facets, such as items, raters or occasions; a generalizability study uses analysis of variance to estimate how much score variance each contributes; and a decision study works

out how many items, raters or occasions a given use requires. In a single-facet design, the generalizability coefficient takes the same form as alpha.

How it works

1. **Define the intended use**
   Decide whether scores will rank people or judge each against a standard, since generalizability theory treats these as relative and absolute error, as Robert Brennan (2003) explains.
2. **Compute alpha as a first check**
   After piloting, calculate alpha for each scale and inspect items that lower it, remembering that it is a lower bound and says nothing about stability over time.
3. **Name the facets**
   List the conditions that could vary in real use, such as items, enumerators, languages or days of testing.
4. **Run a generalizability study**
   Collect pilot data in which facets are crossed, for example the same children scored by two enumerators on two days, and estimate the variance due to each source.
5. **Plan with a decision study**
   Use the estimates to choose the numbers of items, raters or occasions that give adequate dependability; in a military example in Cronbach's memoir, this led to one well-trained judge and more tasks.
6. **Report what was and was not examined**
   State which sources of error the coefficient covers and which it leaves out, so readers do not overtrust a single number.

WORKED EXAMPLE · AN ILLUSTRATION

A citizen-led learning assessment in rural Karnataka uses volunteer enumerators to score children's oral reading on a ten-item tool. Pilot data give an alpha of 0.88, which looks reassuring, but alpha reflects only consistency across items. Following Cronbach's generalizability theory, the team has a sample of 200 children assessed by two enumerators on two different days. The analysis shows that items contribute little error, differences between enumerators contribute a lot and day-to-day variation a moderate amount. The decision study suggests that stronger enumerator training and a short calibration exercise would improve dependability far more than adding items. The final report presents alpha alongside these results, and states plainly that the scores are fit for comparing districts but not yet for judging individual children.

USE IT WHEN

- You are piloting or adapting a survey scale or assessment into a new language or context.
- Scores depend on human raters or enumerators whose consistency is uncertain.
- You must decide how many items, raters or rounds of data collection are worth paying for.

SKIP IT WHEN

- The tool consists of single factual questions, such as age or household size, where internal consistency is meaningless.
- Items are deliberately varied indicators of different things, so a low alpha is expected and not a fault.

COMMON MISREADINGS

- Treating a high alpha as proof that a scale measures a single construct, which Cronbach's memoir explicitly rejects.
- Reading alpha as precision across time or raters, when it captures only consistency across items on one occasion.

WHY IT WAS GROUNDBREAKING

Alpha gave every researcher a practical, single-administration reliability check, and Klaas Sijtsma (2016) describes the 1951 paper as by far the most cited article in Psychometrika. Generalizability theory went further, showing that reliability is not one fixed property of a test but depends on which sources of error matter for a particular use.

Origin

Alpha appeared in Psychometrika, volume 16, in 1951. The core ideas of generalizability theory were developed in technical reports in 1960 and 1961 and set out fully in The Dependability of Behavioral Measurements (1972). A final paper on alpha, prepared with Richard Shavelson, appeared after Cronbach's death, in 2004, revisiting alpha and the procedures that succeeded it.

RELATED IDEAS

Construct validity

DC Campbell 112

RT Tyler 20

CONCEPT HIGHLIGHT

Coefficient alpha (Cronbach's alpha)

A single-administration estimate of internal consistency reliability; it is a lower bound and does not by itself show that a scale is unidimensional.

2

BIG IDEA 2 OF 3 · CRONBACH

Construct validity

A score means what you claim only if a network of theory-based predictions about it keeps being confirmed by evidence; validity is an argument, not a single number.

PICTURE IT

Think of a detective naming a suspect. No single clue proves the case. Motive, opportunity and witnesses must all point the same way, and each new clue can strengthen or weaken it. Construct validity treats the meaning of a score the same way, as a case built from many pieces of evidence.

The problem

In the early 1950s, a test was usually judged valid if its scores predicted an outside criterion, such as later grades, or if its items sampled a defined body of content. But many qualities that matter most, such as anxiety or reading comprehension, and in development work empowerment or resilience, have no single accepted criterion to check against. Without a way to validate them, researchers either avoided such qualities or claimed to measure them with no real evidence at all.

What it is

Construct validity, set out by Cronbach and Paul Meehl in Construct Validity in Psychological Tests (1955), is the case that a score can be interpreted as a measure of a postulated attribute that has no adequate operational definition or criterion. The American Psychological Association's 1954 Technical Recommendations had listed it beside predictive, concurrent and content validity. Its central device is the nomological network: the interlocking set of laws linking the construct to other constructs and to observable measures. Validating a test means deriving predictions from this network and checking them. Cronbach and Meehl named several kinds of evidence: group differences predicted by theory; correlations and factor analysis with related tests; internal structure, where theory calls for it; change over occasions or after interventions; and studies of the process by which people answer. When a prediction fails, the fault may lie in the test, the theory or the study design, and any revised interpretation needs fresh data. Rationalisation, they insisted, is not validation.

How it works

1. **Define the construct and its network**
   Write down what the construct is and how theory says it should relate to other qualities, behaviours and conditions.
2. **Derive testable predictions**
   Turn the network into concrete expectations, such as groups that should score higher or measures that should correlate.
3. **Check group differences**
   Test whether groups that theory says should differ actually do.
4. **Examine correlations and structure**
   Check that scores relate to measures of related constructs, relate less to unrelated ones, and that items hang together only as far as theory requires.
5. **Study change and process**
   See whether scores shift after relevant interventions, and learn how respondents interpret items, for example through cognitive interviews.
6. **Diagnose failures and revalidate**
   When a prediction fails, decide whether the test, the theory or the study was at fault, revise accordingly and validate the revision on new data, as Cronbach and Meehl advised.

WORKED EXAMPLE · AN ILLUSTRATION

A women's livelihoods programme in Uttar Pradesh develops a 12-item scale of household decision-making agency. The team writes down its network: agency should be higher among women who control some income, should rise after the programme's negotiation training, and should relate moderately to mobility outside the village but only weakly to literacy. Pilot data from 400 women show the expected income difference and a rise after training. But scores correlate strongly with literacy, and cognitive interviews reveal that women with less schooling misread two items worded around written documents. The team treats this as a fault in the test rather than the theory, rewrites the two items in everyday spoken Hindi with local examples, and validates the revised scale on a fresh sample before using it in the evaluation.

CONCEPT HIGHLIGHT

Construct validity

The degree to which evidence and theory together support interpreting a score as a measure of a particular quality.

USE IT WHEN

- You are measuring an abstract quality such as agency, wellbeing, confidence or social cohesion.
- A scale developed elsewhere is being adapted to a new language, culture or population.
- Stakeholders question whether an index really captures what its name claims.

SKIP IT WHEN

- The measure is a direct count or record, such as attendance or enrolment, where a simple accuracy check is enough.
- There is no time or sample for validation, in which case say so plainly rather than claiming validity.

COMMON MISREADINGS

- [ ] Treating construct validity as one statistic, such as a factor loading, rather than an accumulating argument.
- [ ] Assuming a well-named scale measures what its name says, the rationalisation without evidence that Cronbach and Meehl rejected.

WHY IT WAS GROUNDBREAKING

It moved validity from a single correlation with a criterion to a reasoned case about what a score means. It opened rigorous measurement to the abstract qualities, such as agency or wellbeing, that programmes most want to change, and in our reading later testing standards built on it.

Origin

Cronbach and Meehl published the paper in Psychological Bulletin, volume 52, pages 281 to 302, in 1955, after the American Psychological Association's 1954 Technical Recommendations had named construct validity as one of four types. Campbell and Fiske's multitrait-multimethod matrix of 1959 soon gave it a practical test of convergent and discriminant evidence.

RELATED IDEAS

Coefficient alpha and generalizability theory

DC Campbell 112 | BC Chilisa 498

CONCEPT HIGHLIGHT

Generalizability theory

A framework from 1972 that splits measurement error into sources such as items, raters and occasions, showing which matter most for a given use.

3

BIG IDEA 3 OF 3 · CRONBACH

Evaluation for the policy-shaping community

Evaluation is most useful when it improves programmes while they can still change and informs the whole community that shapes policy, with careful attention to whether results will hold elsewhere.

PICTURE IT

Think of a satnav that keeps recalculating while you drive, compared with a report on the journey written after you arrive. Cronbach argued that evaluation helps most like the satnav, while the route can still change, and that its guidance should reach everyone planning the next journey, not just one person at the wheel.

The problem

By the 1960s, evaluations of new curricula and social programmes were often designed as grand head-to-head contests that delivered a single verdict at the end. Cronbach noted in 1963 that such comparisons rarely settled which course was better, because differences between average scores were usually small. Designers who could have used feedback received none until it was too late, and the officials, advocates and citizens debating policy got verdicts they could not interpret for their own contexts.

What it is

In Course Improvement through Evaluation (1963), Cronbach defined evaluation as collecting information to make decisions about an educational programme, whether to improve a course, to plan for individual learners or to judge a system, and argued that its greatest service was to identify where a course needed revision. Over two decades, culminating in Toward Reform of Program Evaluation (1980), written with the Stanford Evaluation Consortium, and Designing Evaluations of Educational and Social Programs (1982), with Karen Shapiro, he built a broader theory. Evaluation is political. Its audience is not one decision-maker but a policy-shaping community of officials, advocacy groups, educators, the media and citizens, and its role is educative. He preferred several smaller, context-rich studies to one horse-race experiment, and feedback during the study rather than only at the end. To judge whether results transfer, he framed studies in terms

of units, treatments, observations and settings, known as UTOS, putting external validity at the centre, in contrast to Campbell's emphasis on internal validity.

How it works

1. **Map the policy-shaping community**
   List everyone who will argue about the programme, not only the funder, and what each needs to understand.
2. **Name the decisions**
   Following the 1963 paper, decide whether the evaluation should inform course improvement, decisions about individuals or judgements about the system.
3. **Describe UTOS**
   Specify the units (who), treatments (which version), observations (which measures) and settings (where) the study will cover, and those to which people will want to generalise.
4. **Spread the effort**
   Instead of one large comparison, plan several smaller studies across contrasting contexts and programme components.
5. **Feed back early**
   Share emerging findings with programme staff during the study so that the programme can improve while it runs.
6. **Report for transfer**
   Explain which features seemed to matter where, and what readers in other settings would need to check, rather than issuing a single verdict.

WORKED EXAMPLE · AN ILLUSTRATION

A state wants to know whether a teacher-coaching model piloted in two districts should go statewide. Instead of one large trial with a single final verdict, the MEL team draws on Cronbach. It lists the policy-shaping community: the education department, teacher unions, district officers, the funder and a newspaper that covers schools. It specifies UTOS: teachers in rural multigrade and urban schools as units, in-person and phone coaching as treatments, classroom observation and pupil assessments as observations, and three contrasting districts as settings. Short studies run in each district, and findings go to coaches every term. The final synthesis explains that phone coaching worked where teachers had first met their coach in person, and tells officials in other districts what to check before adopting it.

CONCEPT HIGHLIGHT

Formative evaluation

Evaluation carried out while a course or programme is still developing, to improve it; Cronbach's 1963 paper helped establish this purpose.

USE IT WHEN

- A pilot is being considered for scale-up in different districts, states or countries.
- Many groups with different interests will use or contest the findings.
- The programme is still developing and can benefit from feedback during the evaluation.

SKIP IT WHEN

- A single, clearly defined decision-maker needs one precise estimate of effect in one setting.
- Resources allow only one small study, so the multi-site strategy is not feasible; then state the limits on transfer explicitly.

COMMON MISREADINGS

- [ ] Reading his scepticism of horse-race experiments as a rejection of rigour, when he recognised both the merits and the limits of randomised field trials.
- [ ] Treating formative feedback as informal conversation, when he saw it as systematic information for decisions.

WHY IT WAS GROUNDBREAKING

It legitimised formative, improvement-oriented evaluation and shifted attention from proving a single effect to informing public debate. It also put the question of whether results transfer, so central to scale-up decisions, at the heart of evaluation design.

Origin

The 1963 paper appeared in the Teachers College Record, volume 64. Cronbach developed the theory with the Stanford Evaluation Consortium, which he led in the 1970s, in Toward Reform of Program Evaluation (Jossey-Bass, 1980), and set out UTOS in Designing Evaluations of Educational and Social Programs (Jossey-Bass, 1982), which was chosen in 2000 among the top 100 education-related books of the twentieth century.

RELATED IDEAS

DC Campbell 112 · RT Tyler 20 · MS Scriven 60 · MP Patton 204 · CW Weiss 152

CONCEPT HIGHLIGHT

Policy-shaping community

Cronbach's term for the wide audience that evaluation should inform, including decision-makers, advocates, educators, the media and citizens.

RELATED GIANTS



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Ralph W. Tyler

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

p. 20



Bagele M. Chilisa

The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

p. 498



Michael Scriven

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

p. 60

CONCEPT HIGHLIGHT

UTOS

Units, treatments, observations and settings: Cronbach's framework for asking whether evaluation findings will hold in other contexts.

THE BIG IDEAS OF THIS CHAPTER

1. **Coefficient alpha and generalizability theory** p. 46

   Estimate how dependable a measure is from a single administration, then go further and find out which sources of error, such as items, raters or occasions, matter for the decision the scores must support.

2. **Construct validity** p. 49

   A score means what you claim only if a network of theory-based predictions about it keeps being confirmed by evidence; validity is an argument, not a single number.

3. **Evaluation for the policy-shaping community** p. 52

   Evaluation is most useful when it improves programmes while they can still change and informs the whole community that shapes policy, with careful attention to whether results will hold elsewhere.

Put it to work

USE IT ON MONDAY · COEFFICIENT ALPHA, AND EVALUATION THAT SERVES LOCAL LEARNING

1. When you adapt a survey scale, pilot it and report alpha, but also check with respondents that the items make sense in their language. A high alpha does not prove the scale measures one thing.

2. Ask which source of measurement error matters most for your use: different enumerators, different days or different tasks. Train for and check the biggest one.

3. Build in at least one formative cycle in which local teams see early findings and can change course, before any summative judgement is made.

AI prompt **Copy, fill the brackets, then check every line**

Help me review the measurement quality of a survey scale used in **[programme]** with **[respondent group]** in **[language and location]**. Here are the items: **[paste items]**. Here is our pilot summary, including any reliability statistics we computed: **[paste pilot notes]**. (1) Identify items that may be ambiguous, double-barrelled or culturally mismatched, and explain why. (2) Explain in plain English what our reported reliability figure can and cannot tell us, noting that coefficient alpha is not proof that a scale measures a single thing. (3) Suggest the likely main sources of measurement error in this setting and practical checks. Do not calculate or estimate any statistic I have not supplied.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a bathroom scale. If you step on it three times in a minute and get three different weights, you cannot trust any of them. Coefficient alpha asks the same thing of a test: do its questions agree with each other, so that the score is steady enough to rely on?

YOUR MOVES

Three moves from Cronbach's path

1 Carry rigour across borders

Cronbach brought the discipline of measurement into evaluation, and the move gave the young field a backbone of careful method.

TRY THIS

Name the strongest skill from your first field or job, and use it deliberately on your current MEL task.

2 Change your mind in public

He revised his views over decades, moving from statistics to the politics of evaluation without abandoning either, and said so with reasons.

TRY THIS

Write a short note on one belief about evaluation you have changed, and why. Share it with your team.

3 Be your own toughest reviewer

He kept refining his best-known idea to the end, preparing a final paper on alpha with Richard Shavelson that appeared after his death.

TRY THIS

Before your next submission, spend one hour attacking your own analysis as a hostile reviewer would.

THE BIG IDEAS OF THIS CHAPTER

1 **Coefficient alpha and generalizability theory** p. 46

Estimate how dependable a measure is from a single administration, then go further and find out which sources of error, such as items, raters or occasions, matter for the decision the scores must support.

2 **Construct validity** p. 49

A score means what you claim only if a network of theory-based predictions about it keeps being confirmed by evidence; validity is an argument, not a single number.

3 **Evaluation for the policy-shaping community** p. 52

Evaluation is most useful when it improves programmes while they can still change and informs the whole community that shapes policy, with careful attention to whether results will hold elsewhere.

CONCEPT HIGHLIGHT

UTOS

Units, treatments, observations and settings: Cronbach's framework for asking whether evaluation findings will hold in other contexts.

GIANTS CONNECTED TO THIS CHAPTER



Donald T. Campbell

A contrasting view on how central experiments should be to evaluation.

p. 112



Michael Quinn Patton

Built a whole approach around intended users and intended use.

p. 204



Ralph W. Tyler

The Eight-Year Study influenced Cronbach's turn towards evaluation.

p. 20

KEY WORKS TO READ IN THE ORIGINAL

- **Coefficient Alpha and the Internal Structure of Tests** 1951
- **Construct Validity in Psychological Tests** 1955, With Paul E. Meehl
- **Toward Reform of Program Evaluation** 1980, With Stanford Evaluation Consortium colleagues

CONCEPT HIGHLIGHT

Coefficient alpha (Cronbach's alpha)

A single-administration estimate of internal consistency reliability; it is a lower bound and does not by itself show that a scale is unidimensional.

CONCEPT HIGHLIGHT

Construct validity

The degree to which evidence and theory together support interpreting a score as a measure of a particular quality.

WORKSHEET · TRY THIS TEMPLATE

Measurement quality check

Print this page, or copy the headings into your notes.

1 Construct

What exactly are we trying to measure, in one sentence?

2 Items and language

Were items translated and back-checked? Note any that confused respondents in piloting.

3 Reliability

Report alpha with its limits; do not present it as proof of a single dimension or of precision.

4 Biggest error source

Enumerator, occasion, task or setting, and what you did about it.

5 Local use

Who in the programme sees these results first, and what could they change?

REFLECT

Who is the evaluation you are designing really for: a distant decision-maker or the people who run the programme every day?

CONNECTS TO

DC **Donald T. Campbell**
A contrasting view on how central experiments should be to evaluation. p. 112

MP **Michael Quinn Patton**
Built a whole approach around intended users and intended use. p. 204

RT **Ralph W. Tyler**
The Eight-Year Study influenced Cronbach's turn towards evaluation. p. 20



Emblem: an original HubForge design for this figure, not a likeness.

Michael Scriven

Died 2023 Active 1960s to 2000s

Born in England, raised in Australia; worked in Australia, the United States and New Zealand

Value and judgement

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

IN ONE LINE

He taught a whole profession that judging value is a discipline to be learned, not a matter of taste.

Two words. That is what a philosopher of science gave evaluation in 1967, in a chapter on judging school curricula: formative and summative. Today they appear in terms of reference, school policies and training courses around the world. Five years later the same man proposed something far more startling. Evaluators, he suggested, could ignore a programme's goals altogether and simply find out what it actually did. Many evaluators greeted the idea with disbelief, as the evaluator Brandon Youker has recorded. The man was Michael Scriven. Across a life that ran from 1928 to 2023, he taught in departments of mathematics, philosophy, psychology, law, education and evaluation, and produced more than 450 publications. His central claim was simple and radical: judging value is not a matter of taste. It can be done well or badly, and it can be learned.

Life CHAPTER 03

Michael John Scriven was born in Britain on 28 March 1928 and grew up in Australia. At the University of Melbourne he took an honours degree in mathematics and then an honours master's degree in the philosophy of mathematical logic, before completing a doctorate in philosophy at Oxford. His first career was in the philosophy of science, where he made his name by finding holes in famous arguments: his papers offered counterexamples to Carl Hempel's influential covering law model of scientific explanation. In 1966 he published Primary Philosophy, an introductory text that took a sceptical line on beliefs such as reincarnation and ghosts.

He taught philosophy at Swarthmore College, Indiana University and the University of California, Berkeley. While at Indiana he helped start the Social Science Education Consortium in 1963, an early

BIG IDEAS

1. **Formative and summative evaluation**
2. **Goal-free evaluation**
3. **The logic of evaluation and the Key Evaluation Checklist**

KEY WORKS

Evaluation Thesaurus
1977 · Later edition 1991

The Logic of Evaluation
1980

Key Evaluation Checklist
2007 · Evaluation Checklist Project

DID YOU KNOW?

The statistician Gene Glass credited Scriven's idea of metaevaluation with leading him to coin the term meta-analysis.

Tributes record that Scriven met Albert Einstein and researched parapsychology; his other enthusiasms included knives, fast cars and rowing.

sign of his interest in how school curricula were designed and judged. By the mid-1960s he was writing about educational evaluation, and over the following decades it became his main field; his colleague Gene Glass dated the point at which evaluation became his principal focus to around 1980. Across his career he taught in departments of mathematics, philosophy, psychology, the history and philosophy of science, law, education and evaluation, and he served on the editorial boards of 42 journals.

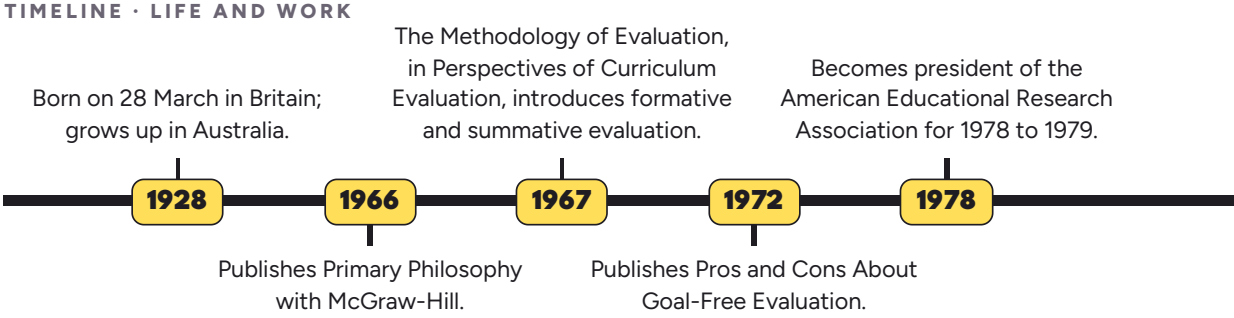
His academic path crossed three continents. He held posts at the University of Western Australia and at Western Michigan University, home of the Evaluation Center, and also taught in New Zealand. He was president of the American Educational Research Association in 1978 to 1979 and later of the American Evaluation Association. In 1997 he joined Claremont Graduate University as a Distinguished Professor, where he helped build its graduate programmes in evaluation and the Claremont Evaluation Center. With E. Jane Davidson he founded the Journal of MultiDisciplinary Evaluation. He wrote or co-wrote more than a dozen books and over 450 publications, and mentored hundreds of students and junior colleagues.

His curiosity ran far beyond the university. Tributes record that he met Albert Einstein, researched parapsychology, designed and oversaw the building of his own house, and knew art and music deeply; his local newspaper added knives, fast cars and rowing to the list. He lived for many years in Inverness, California, with his wife Mary Anne. About fifty years after his Oxford doctorate, the University of Melbourne awarded him an honorary doctorate in education. In 2011 he announced the Faster Forward Fund to support students and projects in professional evaluation. He died on 28 August 2023, aged 95, at a retirement community in San Rafael, California.

The spark

A question was waiting for Scriven. In the early 1960s, American schools were flooded with new courses and materials, and nobody had a rigorous way to say whether any of them were good. Scriven, a philosopher of science, was already involved: at Indiana in 1963 he helped

TIMELINE · LIFE AND WORK



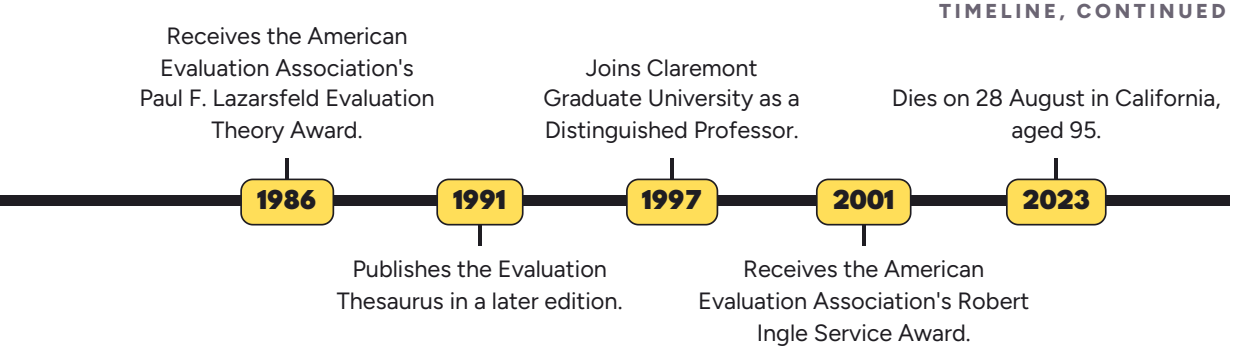
start the Social Science Education Consortium, which worked on how school curricula were designed and judged. Underneath lay a deeper puzzle that suited a philosopher perfectly. Is judging worth even a proper task for science? In our reading, he spent the rest of his life answering yes, and showing how.

The breakthrough CHAPTER 03

In the 1960s, evaluating a programme in the United States usually meant following Ralph Tyler: write down the objectives, then measure whether they were met. A wave of curriculum reform was producing new courses faster than anyone could judge them. In 1963 Lee Cronbach had argued that evaluation helps most while a course is still taking shape. But the field still lacked two things: simple words for the different jobs evaluation can do, and a rigorous account of what it means to judge value rather than merely to measure.

Scriven supplied both. His chapter The Methodology of Evaluation appeared in 1967 in Perspectives of Curriculum Evaluation, the first monograph in the American Educational Research Association's series on the subject, alongside contributions by Tyler and Robert Gagné. There he named formative and summative evaluation, and argued, against Cronbach, that judging overall worth mattered just as much as helping a course improve. Then, in Pros and Cons About Goal-Free Evaluation (1972), he went further. He pointed to the way consumer product testers judge goods such as cars or kettles without asking the manufacturer what it intended, and proposed that evaluators could do the same: set the stated goals aside and look for every effect a programme has.

TIMELINE, CONTINUED



The problem it solved

Official goals biased evaluations and hid unintended effects, while unclear purpose mixed improvement and verdict together.

Impact WHAT CHANGED

Scriven's two words are now so ordinary that it is easy to forget anyone invented them. Terms of reference across the aid and NGO sector routinely ask whether an evaluation is formative or summative, so that its purpose is settled from day one. The pair also travelled into schools, where it shaped the language of formative assessment; in 2024 Janet Clinton and John Hattie argued that this migration blurred his original meaning, which was about the purpose and timing of judgements rather than types of test. And one of his ideas helped spark a method that changed modern science: Gene Glass credited Scriven's concept of metaevaluation with leading him to coin the term meta-analysis.

Goal-free thinking has had a quieter but real influence on development practice. Youker classifies approaches such as most significant change and outcome harvesting, both used widely by NGOs and development agencies, as goal-dismissive forms of goal-free evaluation, because they gather evidence of change before asking whether it matches plans. That classification is his; the developers of those methods did not necessarily start from Scriven. The emphasis on costs and comparisons in the Key Evaluation Checklist anticipated today's interest in cost-inclusive evaluation, which has been applied in international aid. Institutionally, he co-founded the open-access Journal of MultiDisciplinary Evaluation, and his Faster Forward Fund supports students and projects in the field. In our reading, his deepest legacy is the expectation that evaluators state their criteria and defend their judgements openly.

KEY WORKS TO READ IN THE ORIGINAL

- **Evaluation Thesaurus** 1977, Later edition 1991
- **The Logic of Evaluation** 1980
- **Key Evaluation Checklist** 2007, Evaluation Checklist Project

WHERE PEOPLE DISAGREE

Debates

Scriven was deliberately provocative; Claremont notes that many field leaders disagreed with him while respecting him deeply. Goal-free evaluation has been the main target: Youker reports that most evaluators treat it as a rhetorical reminder rather than a workable design, partly because strict blinding to goals is hard to achieve; some argue that the label itself is a misnomer. In our reading, his view that a competent evaluator can establish a programme's worth as a matter of fact sits uneasily with constructivist approaches such as Guba and Lincoln's. The International Evaluation Academy has also observed that his language, examples and ideas were rooted in Western philosophy, a limitation for a field that draws on many ways of knowing, while noting that he valued explicit attention to cultural perspectives.

LIMITS, IN OUR READING

- ☐ The International Evaluation Academy has noted that his perspective was shaped by Western philosophy, which it sees as a limitation for evaluation worldwide.
- ☐ In our view, truly goal-free evaluation is hard in practice, because evaluators usually know the programme's goals before they start.

WHY THIS MATTERS TO YOU

Scriven's ideas reach anyone who has ever had to judge something. Before you judge, ask the formative or summative question: am I trying to help this improve, or to decide its future? Look beyond what was promised to what actually happened, including effects nobody planned, good or bad. And when you reach a verdict, show your reasoning: the criteria you used, how you weighed them, what the alternatives were. His gift is both a permission and a duty. You are allowed to judge. But you owe everyone a clear account of how you did it.

450+

publications, including more than a dozen books he wrote or co-wrote

Remembering Michael Scriven, Claremont Graduate University (2023); Gene V. Glass, Journal of MultiDisciplinary Evaluation (2024)

42

journals on whose editorial boards he served

Gene V. Glass, Even So, Too Soon, Journal of MultiDisciplinary Evaluation (2024)

16

components in the later versions of his Key Evaluation Checklist

Brian Yates and Nadini Persaud, Journal of MultiDisciplinary Evaluation (2024)

Key concepts

Formative evaluation

Evaluation carried out during the development of a programme or product, with the purpose of improving it.

Summative evaluation

Evaluation that judges the overall merit or worth of a programme or product, usually at or near the end, to inform decisions about its future.

Goal-free evaluation

An approach in which the evaluator is kept unaware of, or sets aside, the programme's stated goals in order to find out what it actually does, including unintended effects.

Merit, worth and significance

Scriven's three dimensions of value: roughly, intrinsic quality, value in context including cost, and overall importance. He treated costs as central to judging worth rather than plain merit.

Metaevaluation

The evaluation of an evaluation, applying the same logic of criteria and evidence to judge its quality.

Key Evaluation Checklist

Scriven's multi-part checklist for planning and reviewing evaluations, which in later versions had sixteen components including costs and comparisons.

Sources for this chapter: p. 575

Did you know?

The statistician Gene Glass credited Scriven's idea of metaevaluation with leading him to coin the term meta-analysis.

THE BIG IDEAS OF THIS CHAPTER

1. **Formative and summative evaluation** p. 68

   Evaluation can play different roles, improving something while it can still change (formative) or judging its overall merit and worth to inform a decision about it (summative), and it is the role, not the method, that decides which is which.

2. **Goal-free evaluation** p. 71

   Judge a programme by what it actually does to the people it touches, intended or not, rather than by what its designers hoped it would do.

3. **The logic of evaluation and the Key Evaluation Checklist** p. 74

   Evaluation reaches justified conclusions about merit, worth and significance by combining facts with explicit values, and that logic can be laid out as a checklist any evaluator can follow.

GIANTS CONNECTED TO THIS CHAPTER



Ralph W. Tyler

Goal-free evaluation is a deliberate step beyond judging against stated objectives.

p. 20



Patricia J. Rogers

Patricia Rogers heard him speak on goal-free evaluation at an Australian Evaluation Society event in 1986, early in her own path.

p. 268



Bagele M. Chilisa

African relational approaches ask whose philosophy should define value in the first place.

p. 498

In one line

He taught a whole profession that judging value is a discipline to be learned, not a matter of taste.

1

BIG IDEA 1 OF 3 · SCRIVEN

Formative and summative evaluation

Evaluation can play different roles, improving something while it can still change (formative) or judging its overall merit and worth to inform a decision about it (summative), and it is the role, not the method, that decides which is which.

PICTURE IT

Think of a football coach at half-time and a sports reporter at the final whistle. Both watch the same match. The coach is looking for what to fix while there is still time to change the result. The reporter is deciding how good the performance was. Formative evaluation is the coach; summative is the reporter.

The problem

In the 1960s, American curriculum reform was producing new courses faster than anyone could judge them. Evaluation mostly meant checking at the end whether objectives had been met, and Lee Cronbach had argued in 1963 that evaluation was most useful when it improved a course in development. Commissioners, developers and evaluators had no shared words for what a study was meant to do, so advice for improvement and verdicts on worth were blurred together, and nobody knew whether findings were meant to fix a programme or decide its fate.

What it is

Scriven distinguished two roles that evaluation can play. Formative evaluation is carried out during development or delivery, for the people who can still change the thing being evaluated, so that it gets better. Summative evaluation judges the overall merit, worth or significance of something at or near the end, for those who must decide whether to adopt, continue, expand or end it. The crucial point, which he pressed repeatedly, is that these are roles and not types: the same data, and even the same report, can be formative for one audience and summative for another, as a restaurant review is summative for its readers and formative for the chef. He also insisted that formative work should be as rigorous as summative work, that it can and should reach evaluative conclusions, and that summative evaluation is no less

valuable than formative. In later writing he recognised further roles, such as ascriptive evaluation, and accepted that developmental evaluation could be added to the list.

How it works

1 Name the decision and the decider

Before designing anything, identify who will use the findings and for what: improving the programme while it runs, or deciding on its future.

2 Assign a role to each question

Label each evaluation question formative or summative in the terms of reference. Because Scriven held that role depends on use, a question can serve both, so say so explicitly.

3 Time the evidence to the role

Schedule formative findings early enough for staff to act, for example at the end of a pilot cycle, and summative judgements to arrive before the funding or scale-up decision.

4 Hold both to the same standard

Following Scriven's 1991 critique of the belief that formative work can be informal, as summarised by Clinton and Hattie (2024), apply the same rigour of evidence and criteria in both roles.

5 Reach conclusions in both

Do not stop at a list of suggestions in formative work. State how good the programme is at this stage, which Scriven described as early-warning summative evidence for the team.

6 Report to each audience

Present formative findings to those who can change the design and summative conclusions to decision-makers, noting where the same evidence serves both.

WORKED EXAMPLE · AN ILLUSTRATION

A state education department in India pilots a foundational literacy programme in 200 government schools for one year before deciding whether to extend it statewide. The MEL team writes two sets of questions. Formative questions, for the programme team, ask after each term whether teachers are using the lesson plans, which activities children find hardest and what should change in the next round of training; findings go back within three weeks. Summative questions, for the department's leadership, ask at the end of the year how much children's reading improved compared with similar schools, at what cost per child, and with what effect on teachers' workload. The same classroom observations feed both. Because the purpose of each question was written down at inception, nobody mistakes a term report for the verdict on the scheme.

USE IT WHEN

- A terms of reference mixes improvement questions with accountability questions.
- A pilot must both learn as it goes and inform a later scale-up decision.
- Stakeholders disagree, or have never discussed, what an evaluation is for.

SKIP IT WHEN

- The real need is routine monitoring data, with no evaluative judgement requested.
- You are supporting an innovation with no stable model yet, where a two-way split is too narrow and a developmental role fits better.

COMMON MISREADINGS

- Treating formative and summative as kinds of test or method, a migration into classroom assessment that Scriven regarded as a misreading of his concepts.
- Assuming formative work can be informal, or must avoid negative conclusions; Scriven rejected both ideas.

WHY IT WAS GROUNDBREAKING

It gave commissioners and evaluators a two-word vocabulary for agreeing what a study is for before any data are collected. By separating role from method, it also showed that the same evidence can serve learning and accountability at once, a tension MEL teams still manage every day.

Origin

Scriven introduced the terms in The Methodology of Evaluation, his chapter in Perspectives of Curriculum Evaluation (AERA monograph series, Rand McNally, 1967), partly in response to Cronbach's 1963 article. In 1991 he attacked common fallacies about the pair, in 1993 he described levels of formative work, and in 1996 he stressed that the difference lies in use. In 2012 he added preformative and proformative evaluation in the Journal of MultiDisciplinary Evaluation.

RELATED IDEAS

Goal-free evaluation · LC Cronbach 38 · DS Stufflebeam 134 · MP Patton 204

CONCEPT HIGHLIGHT

Formative evaluation

Evaluation carried out during the development of a programme or product, with the purpose of improving it.

2

BIG IDEA 2 OF 3 · SCRIVEN

Goal-free evaluation

Judge a programme by what it actually does to the people it touches, intended or not, rather than by what its designers hoped it would do.

PICTURE IT

Think of buying a car from a consumer magazine's test rather than from the advert. The advert tells you what the maker intended. The test tells you what the car actually does, including the things nobody mentioned. Goal-free evaluation is the independent road test.

The problem

Since Ralph Tyler, programmes had largely been judged against their own stated objectives. That made evaluation tidy but narrow. Evaluators who read the objectives first tended to look only for the intended outcomes, so side effects, harms and unexpected benefits went unrecorded. Programmes with modest or self-serving goals could pass easily, and the people most affected, whose needs might differ from the programme's targets, had no place in the judgement. In development terms, a scheme could hit every logframe indicator while quietly making some lives worse.

What it is

Goal-free evaluation is an approach in which the evaluator is kept unaware of, or sets aside, a programme's stated goals and objectives, and investigates what the programme actually does. Scriven, who developed the idea while studying how Consumers Union tested products, compared it to a double-blind trial: not knowing the intended outcomes reduces the risk of seeing only what one expects. Brandon Youker sets out four principles: identify relevant effects without reference to goals; identify what occurred without being cued by goals; determine whether the effects can logically be attributed to the programme; and judge each effect as positive, negative or neutral. In Scriven's own version, effects are judged against the assessed needs of the people affected. Variants differ in strictness. Scriven's blinded form often uses a screener who filters goal language out of documents and meetings; Evert Vedung's adaptation sets aside needs as well as goals; reactionary forms arise when goals are simply unknown; and goal-dismissive forms neither seek goals nor give them priority.

How it works

1. **Set up the blind**
   Appoint a screener who handles first meetings with programme staff, removes goal language from the documents passed to the evaluator and acts as liaison, as described in Youker's guidance from the Evaluation Center at Western Michigan University.

2. **Establish needs**
   In Scriven's version, assess the needs of the people the programme affects using independent evidence; in Youker's practice, infer the needs the programme appears to address from what it does.

3. **Search widely for effects**
   Use observation, key informant interviews, documents, surveys and case studies to find out what has changed, for participants and for others, without prompts from targets.

4. **Test attribution**
   For each effect, ask whether it can reasonably be traced to the programme rather than to other causes.

5. **Judge each effect**
   Rate effects as positive, negative or neutral against the needs identified, then reach an overall judgement using explicit criteria.

6. **Reveal goals last, if at all**
   Once findings are fixed, compare them with the stated goals or pass them to a goal-based team; Youker notes that goal-free work often supplements goal-based evaluation.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO gives bicycles to girls entering secondary school in a rural district of Bihar. The funder hires two evaluators. One works from the logframe and confirms that enrolment targets were met. The other works goal-free, briefed only by a screener who forwards schedules and contact lists stripped of objectives. She visits villages, talks with girls, parents, teachers and shopkeepers, and observes school days. She records the expected gains in attendance, and also changes nobody planned to measure: girls running errands for their families, some brothers taking the bicycles over, and safety worries about journeys after dark in winter. Judged against the girls' need for safe, reliable access to school, the scheme looks valuable but leaky. Together the two reports give a fuller judgement than either could alone.

USE IT WHEN

- You suspect important side effects or harms that the logframe ignores.
- Independence from programme staff matters, for example in a high-stakes external review.
- Goals are vague, contested, or were written after the programme began.

SKIP IT WHEN

- The commissioner's core question is whether specific contracted targets were met.
- Budget or time cannot support a screener and wide fieldwork.
- The evaluator already knows the goals well, in which case an openly goal-dismissive variant is more honest than a pretence of blindness.

COMMON MISREADINGS

- Assuming it means programmes should have no goals, when the point is to stop goals from narrowing what the evaluator sees.
- Treating it as a loose, opinion-led exercise, when Scriven's version requires a needs assessment, a test of attribution and explicit criteria.

WHY IT WAS GROUNDBREAKING

It broke the assumption, dominant since Tyler, that a programme's own objectives are the yardstick of its value. It made the search for unintended and negative effects a deliberate design choice, and strengthened the case for evaluators' independence from funders, managers and staff.

Origin

Scriven introduced the idea in Pros and Cons About Goal-Free Evaluation, published in Evaluation Comment in December 1972, after studying how Consumers Union evaluated products. His Evaluation Thesaurus (1991 edition) gave the standard definition. Brandon Youker, who first ran a goal-free evaluation under Scriven's supervision in 2005, set out four principles in 2013 and, in a 2024 retrospective, classified its variants, including goal-dismissive approaches.

RELATED IDEAS

RT Tyler 20 · D&D Davies and Dart 362 · BC Chilisa 498

The logic of evaluation and the Key Evaluation Checklist

Did you know?

Tributes record that Scriven met Albert Einstein and researched parapsychology; his other enthusiasms included knives, fast cars and rowing.

3

BIG IDEA 3 OF 3 · SCRIVEN

The logic of evaluation and the Key Evaluation Checklist

Evaluation reaches justified conclusions about merit, worth and significance by combining facts with explicit values, and that logic can be laid out as a checklist any evaluator can follow.

PICTURE IT

Think of a judge in a baking contest who writes down the criteria before tasting: flavour, texture, appearance, value. Anyone can see how the winner was chosen, and argue with it fairly. The Key Evaluation Checklist makes an evaluator's judgement just as open.

The problem

Through the 1960s and 1970s, many social scientists held that research should be value-free and that judging whether a programme was good was a matter of opinion for politicians. Evaluators therefore reported data and stopped short of conclusions, or slipped in judgements without saying how they had reached them. Commissioners received reports that never answered the question they cared about: is this worth what it costs, compared with what else we could do? Without a stated logic, two evaluators could reach opposite verdicts and neither could show why.

What it is

Scriven defined evaluation as determining merit, worth and significance, and argued that it has a logic of its own rather than one borrowed from a single social science. Merit is roughly intrinsic quality; worth adds value in context, including cost; significance is overall importance. Conclusions come from combining factual premises with value premises drawn from sources such as needs, standards, laws and ethics. The evaluator first chooses the operation the question calls for: ranking, grading against standards, scoring, or apportioning a limited resource. Performance is then profiled on each criterion of merit, with bars, minimum thresholds that rule an option out whatever its other strengths, and weights where some criteria matter more, before being synthesised into an overall judgement. Because every field needs this logic, he called evaluation a transdiscipline, and he defined metaevaluation as evaluation applied to evaluations. The Key Evaluation

Checklist turns the logic into sixteen checkpoints, of which he singled out process, outcomes, costs, comparisons and generalisability as the most important.

How it works

1. **Clarify the evaluand and the task**
   Pin down exactly what is being evaluated and for whom, and whether the task is to rank, grade, score or apportion, the basic operations Scriven distinguishes in The Unruly Logic of Evaluation.
2. **Identify impactees and their needs**
   List everyone affected, not only intended beneficiaries, and establish their needs, which supply many of the value premises.
3. **Make criteria, bars and weights explicit**
   Write down the criteria of merit, set any minimum bars and decide weights; in The Logic and Methodology of Checklists Scriven advises equal weights unless there is strong evidence otherwise, and testing whether conclusions survive changes in weights.
4. **Work through the subevaluations**
   Gather evidence on process, outcomes, costs in money and non-money terms, comparisons with realistic alternatives and generalisability, settling costs before comparisons as the Key Evaluation Checklist specifies.
5. **Synthesise into a judgement**
   Combine the profile into an overall conclusion about merit, worth and significance, using qualitative weight and sum rather than mechanically adding numerical scores, a procedure Scriven called invalid.
6. **Metaevaluate**
   Check the evaluation itself against the same logic before release, ideally with an independent reviewer.

WORKED EXAMPLE · AN ILLUSTRATION

A district administration in Odisha must choose one of three after-school remedial programmes to fund in 300 schools. The MEL lead applies Scriven's logic. She lists the impactees: children, parents, teachers and the volunteers who will run sessions. She sets criteria: learning gains, reach to the lowest-performing children, burden on teachers and cost per child, with a bar that excludes any programme that makes children miss regular lessons. Weights are kept equal. Evidence from pilots is profiled on each criterion. Programme A shows the biggest gains but fails the bar, because its sessions overlap school hours. Programme B is cheapest but reaches few struggling learners. Programme C has moderate gains, strong reach and modest cost. The report recommends C and shows that the conclusion holds when the weights are varied.

USE IT WHEN

- A decision requires choosing among alternatives or allocating a fixed budget.
- Stakeholders need to see exactly how a verdict was reached.
- You are reviewing the quality of an evaluation report.

SKIP IT WHEN

- The task is purely exploratory or descriptive, and no judgement of value is requested.
- The criteria themselves are deeply contested and must be negotiated first, where approaches such as Guba and Lincoln's may need to come before synthesis.
- Time allows only a light review, in which case concentrate on the five core subevaluations rather than all sixteen checkpoints.

COMMON MISREADINGS

- [ ] Treating numerical weight and sum, simply adding weighted scores, as the method, when Scriven rejected it in favour of qualitative weight and sum.
- [ ] Assuming that explicit values make a judgement subjective, when his point was that stated values can be checked and challenged by others.

WHY IT WAS GROUNDBREAKING

It rejected the doctrine that science must be value-free and showed that evaluative conclusions can be reasoned and defended in public. It gave the profession a core logic of its own, and a practical tool that puts costs and comparisons at the heart of every judgement.

Origin

Scriven sketched a logic of evaluation from his 1967 chapter onwards and developed it in the Evaluation Thesaurus (1977, later edition 1991) and The Logic of Evaluation (1980). The Key Evaluation Checklist appeared in a 1989 version with Karen Kirkhart and in later versions, including 2007 and 2015, issued through the Evaluation Center's checklists project; in 2019 he described its sixteen components.

RELATED IDEAS

Goal-free evaluation

DS Stufflebeam 134

G&L Guba and Lincoln 318

HW White 476

RELATED GIANTS



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Daniel L. Stufflebeam

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

p. 134



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Ralph W. Tyler

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

p. 20

CONCEPT HIGHLIGHT

Summative evaluation

Evaluation that judges the overall merit or worth of a programme or product, usually at or near the end, to inform decisions about its future.

CONCEPT HIGHLIGHT

Goal-free evaluation

An approach in which the evaluator is kept unaware of, or sets aside, the programme's stated goals in order to find out what it actually does, including unintended effects.

CONCEPT HIGHLIGHT

Merit, worth and significance

Scriven's three dimensions of value: roughly, intrinsic quality, value in context including cost, and overall importance. He treated costs as central to judging worth rather than plain merit.

Put it to work

USE IT ON MONDAY · GOAL-FREE EVALUATION

1. Before anyone reads the logframe, ask one team member to document what the programme actually does and what changes people report, including unexpected or negative ones.

2. Compare those observed effects with the documented needs of the intended beneficiaries, using needs assessments or community data, not only the programme's targets.

3. In the inception report, state for each evaluation question whether it is formative or summative, so the client knows what kind of answer to expect.

AI prompt **Copy, fill the brackets, then check every line**

Help me run a goal-free scan, in the spirit of Michael Scriven, of **[programme]** in **[location]**. I will give you field observations without the programme's goals: **[paste observations]**. Separately, here is evidence on the needs of the intended beneficiaries: **[paste needs data]**. (1) List the effects the observations describe, including unexpected and negative ones, and the evidence for each. (2) Map each effect against the documented needs. (3) Propose explicit criteria for judging the programme's overall worth, and say which would need more evidence before a judgement is made.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a football coach at half-time and a sports reporter at the final whistle. Both watch the same match. The coach is looking for what to fix while there is still time to change the result. The reporter is deciding how good the performance was. Formative evaluation is the coach; summative is the reporter.

YOUR MOVES

Three moves from Scriven's path

1 Turn your first discipline into an edge

Scriven arrived from mathematics and philosophy, not from education, and an outsider's questions reshaped the field.

TRY THIS

Write down what your original training lets you see that others in MEL miss, and put it to work on one assignment.

2 Name things clearly

Two plain words, formative and summative, did more for everyday practice than many methods. Precise language is a contribution in its own right.

TRY THIS

Find one fuzzy term your team keeps arguing about, and agree a one-line definition everyone will use.

3 Bring a tool with every critique

He questioned the habit of judging programmes only by their goals, and offered goal-free evaluation as a usable alternative.

TRY THIS

Next time you criticise a process, bring a one-page alternative you are willing to pilot.

THE BIG IDEAS OF THIS CHAPTER

1 Formative and summative evaluation p. 68

Evaluation can play different roles, improving something while it can still change (formative) or judging its overall merit and worth to inform a decision about it (summative), and it is the role, not the method, that decides which is which.

2 Goal-free evaluation p. 71

Judge a programme by what it actually does to the people it touches, intended or not, rather than by what its designers hoped it would do.

3 The logic of evaluation and the Key Evaluation Checklist p. 74

Evaluation reaches justified conclusions about merit, worth and significance by combining facts with explicit values, and that logic can be laid out as a checklist any evaluator can follow.

CONCEPT HIGHLIGHT

Key Evaluation Checklist

Scriven's multi-part checklist for planning and reviewing evaluations, which in later versions had sixteen components including costs and comparisons.

GIANTS CONNECTED TO THIS CHAPTER



Ralph W. Tyler

Goal-free evaluation is a deliberate step beyond judging against stated objectives.

p. 20



Patricia J. Rogers

Patricia Rogers heard him speak on goal-free evaluation at an Australian Evaluation Society event in 1986, early in her own path.

p. 268



Bagele M. Chilisa

African relational approaches ask whose philosophy should define value in the first place.

p. 498

KEY WORKS TO READ IN THE ORIGINAL

- **Evaluation Thesaurus** 1977, Later edition 1991
- **The Logic of Evaluation** 1980
- **Key Evaluation Checklist** 2007, Evaluation Checklist Project

CONCEPT HIGHLIGHT

Formative evaluation

Evaluation carried out during the development of a programme or product, with the purpose of improving it.

CONCEPT HIGHLIGHT

Summative evaluation

Evaluation that judges the overall merit or worth of a programme or product, usually at or near the end, to inform decisions about its future.

WORKSHEET · TRY THIS TEMPLATE

Goal-free scan

Print this page, or copy the headings into your notes.

1 What we observed

Activities and changes seen, described plainly, with no reference to targets.

2 Who was affected

Intended participants, others in the community, staff and partners.

3 Unintended effects

Positive and negative, with the evidence for each.

4 Needs it relates to

Which documented needs each effect meets or fails to meet.

5 Criteria for judgement

State the criteria you used to reach an overall judgement of worth.

REFLECT

What has your current programme changed that nobody planned for, and who would know about it?

CONNECTS TO

RT **Ralph W. Tyler**
Goal-free evaluation is a deliberate step beyond judging against stated objectives.
p. 20

PR **Patricia J. Rogers**
Patricia Rogers heard him speak on goal-free evaluation at an Australian Evaluation Society event in 1986, early in her own path.
p. 268

BC **Bagele M. Chilisa**
African relational approaches ask whose philosophy should define value in the first place.
p. 498



Emblem: an original HubForge design for this figure, not a likeness.

Robert E. Stake

Born 1927 Active 1960s to 1990s

United States; Fulbright fellowships in Sweden and Brazil

Value and judgement | Participation, equity and voice

The testing specialist who argued that evaluation should listen to what people around a programme actually care about.

IN ONE LINE

He taught evaluators to listen before they measure, and to show how a programme looks from every place where people stand.

November 1973, a conference in Göteborg, Sweden. A testing expert proposes something that sounds almost like heresy for his profession. An evaluator, he suggests, might spend 30 per cent of the time simply watching a programme, compared with 5 per cent in a conventional design, and build the whole study around the issues that the people involved actually care about. The speaker was Robert Stake, and he knew exactly what tests could do. He held a PhD in psychometrics from Princeton, had worked at the Educational Testing Service and had helped run the Illinois Statewide Testing Program. He called his idea responsive evaluation. It helped turn the whole field towards context, interpretation and the many ways people see the same programme.

Life CHAPTER 04

Robert Earl Stake was born in 1927 in Adams, Nebraska. He took his bachelor's and master's degrees at the University of Nebraska, then a PhD at Princeton University in 1958 in psychometrics, the science of mental measurement. His dissertation, Learning Curve Parameters, Aptitudes and Achievements, was published as a Psychometric Monograph, and he went on to work at the Educational Testing Service. Hold on to that detail. The man who would become one of the best-known critics of narrow testing in evaluation began his career deep inside the testing profession.

After his doctorate he taught at the University of Nebraska, where he became faculty research coordinator in 1962. In 1963 he joined the University of Illinois at Urbana-Champaign, becoming associate director of the Illinois Statewide Testing Program under J. Thomas Hastings. The university's Center for Instructional Research and Curriculum Evaluation, known as CIRCE, was

BIG IDEAS

1. **The countenance model**
2. **Responsive evaluation**
3. **Case study as a way of understanding**

KEY WORKS

The Countenance of Educational Evaluation
1967 · Teachers College Record

The Art of Case Study Research
1995

Qualitative Research: Studying How Things Work
2010

organised in 1964 to continue that testing work and to evaluate curriculum development projects. Stake helped build the centre and later directed it. CIRCE became known internationally for innovative evaluation designs, and its studies ranged over science and mathematics teaching, gifted education, arts education, teacher education for the deaf, environmental education, urban social services and youth sport.

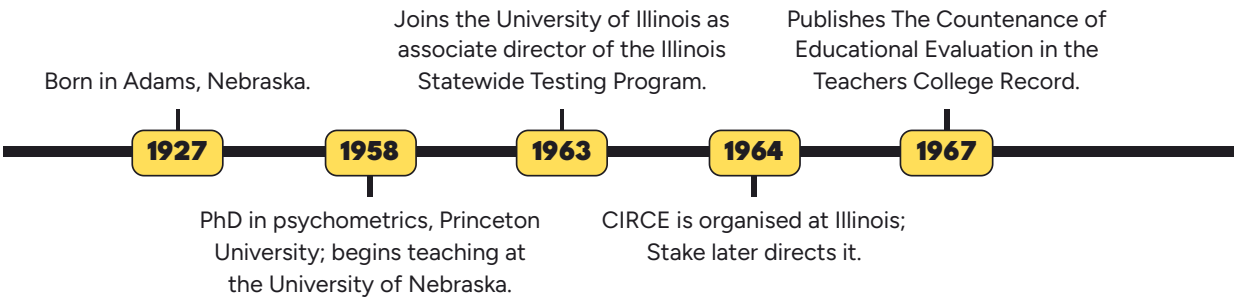
His career was also strongly international. He was a Fulbright fellow in Sweden in 1973 and in Brazil in 1984, and he held visiting posts at institutions including Harvard, the University of British Columbia, Simon Fraser University, the University of East Anglia and the Universidade do Espírito Santo in Brazil. Within the American Educational Research Association he led two divisions, one on curriculum and objectives and one on testing and research design, and he helped organise the Evaluation Research Society and the Evaluation Network, the two bodies that later merged to form the American Evaluation Association.

His books include Case Studies in Science Education, with Jack Easley; Evaluating the Arts in Education; Custom and Cherishing: The Arts in American Elementary Schools, with Liora Bresler and Linda Mabry; Quieting Reform, a metaevaluation of the urban youth programme Cities-in-Schools; The Art of Case Study Research (1995); Standards-Based and Responsive Evaluation (2004); and Qualitative Research: Studying How Things Work (2010). He received the American Evaluation Association's Lazarsfeld Award in 1988, an honorary doctorate from Uppsala University in 1994 and another from the University of Valladolid in Spain in 2009. He became professor emeritus of education at Illinois.

The spark

In the mid 1960s, new federal money was producing ambitious school reforms across the United States, and their evaluations often reported average test scores and almost nothing else. Stake had an unusual vantage point. Working first in a statewide testing programme and then at CIRCE, he sat close to both the tests and the classrooms, and could see what the numbers left out: where each school started, what

TIMELINE · LIFE AND WORK



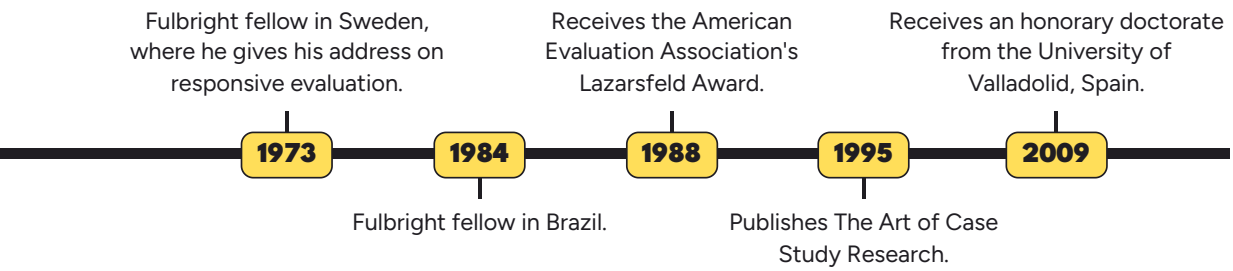
actually happened between teachers and pupils, and how differently people judged the results.

The breakthrough CHAPTER 04

Imagine being told that a new science course worked because average scores rose a few points. You would want to know more. What were these classrooms like before? What did teachers and pupils actually do? Who thought the course was a success, and who did not? In the mid 1960s, most American evaluations could not answer those questions. They judged programmes against their stated objectives, in the tradition of Ralph Tyler, and measured outcomes with standardised tests. Stake, a trained psychometrician, was perfectly placed to see what this left out.

His first major statement, The Countenance of Educational Evaluation, appeared in the Teachers College Record in 1967. A full evaluation, he argued, must both describe and judge a programme. Description should cover three things: antecedents, the conditions beforehand; transactions, what actually goes on; and outcomes, comparing at each stage what was intended with what was observed. A few years later, in a public address given during his 1973 sabbatical in Sweden, Program Evaluation, Particularly Responsive Evaluation, he went further. Instead of answering distant policymakers' questions about average benefits, the evaluator should engage with the people at the site and the issues they care about. Responsive evaluation became one of the founding alternatives to objectives-based and test-centred evaluation.

TIMELINE, CONTINUED



The problem it solved

Evaluations that measured narrow outcomes and ignored what local stakeholders cared about.

Impact WHAT CHANGED

Stake's work helped move evaluation from a largely psychometric enterprise towards one that also values interpretation, context and multiple perspectives. Tineke Abma's encyclopedia account describes responsive evaluation as reframing evaluation from sophisticated analysis for distant policymakers to engagement with practitioners and stakeholders on site, and as part of a wider shift towards hermeneutic and constructivist inquiry. Later approaches built on it: fourth generation evaluation, for example, took the stakeholder's claims, concerns and issues as its starting point, and later writers, including Abma herself and the authors of practitioner guides such as one from the University of Victoria in Canada, have developed responsive evaluation further, with a stronger stress on dialogue, negotiation and power imbalances among stakeholders. His case study typology is standard in qualitative research methods texts.

Institutionally, his role in founding the bodies that became the American Evaluation Association and his long direction of CIRCE gave these ideas a home. Honours from Sweden and Spain and his visiting posts in Brazil, Canada and the United Kingdom reflect a reach well beyond the United States. Our interpretation is that development evaluation carries his imprint wherever terms of reference ask for stakeholder consultations at inception, in-depth case studies of selected sites alongside surveys, or findings fed back in different formats to communities, staff and funders. Participatory and story-based methods used in NGO practice share, in our reading, his starting assumption that a programme's quality looks different from different positions, and that evaluation should show those differences rather than flatten them.

WHERE PEOPLE DISAGREE

Debates

Responsive and case study approaches have drawn questions about how far findings from a few cases can be generalised and how much rests on the evaluator's own judgement and interpretation. Stake's own later writing framed the choice between standards-based and responsive evaluation as a set of trade-offs that each evaluator must make for the situation at hand, rather than a contest one side wins. Later writers in the responsive tradition placed more weight on negotiation among stakeholders and on imbalances of power, since stakeholders' concerns can conflict and the evaluator must still decide whose issues to follow. In our view, being open about that choice is the practical test of a responsive design.

LIMITS, IN OUR READING

- ☐ In our view, case study and responsive designs invite questions about how far findings generalise and how much rests on the evaluator's own judgement.
- ☐ Stakeholder concerns can be many and conflicting, so the evaluator still has to decide whose issues to follow, and should be open about that choice.

WHY THIS MATTERS TO YOU

Stake's lesson reaches anyone who has ever been judged by a number that missed the point. Before you decide whether something worked, ask the people closest to it what matters to them. Spend time watching what actually happens, not only reading reports about it. Look closely at one case in depth, because one story told well can teach what a thousand averages hide. And share what you learn in a form each audience can use, a short story for a community meeting as well as a report for the funder. He gives you permission to treat people's own judgements as evidence, and the discipline to do it well.

Did you know?

Before becoming known for qualitative case study, Stake trained as a psychometrician at Princeton and his dissertation on learning curves was published as a Psychometric Monograph.

30% vs 5%

his estimated share of evaluator time spent observing the programme in a responsive design, compared with a preordinate one

Stake, Program Evaluation, Particularly Responsive Evaluation (1973), reprinted in Journal of MultiDisciplinary Evaluation 7(15), 2011

2

Fulbright fellowships, in Sweden (1973) and Brazil (1984)

Encyclopedia of Evaluation, SAGE (2005)

2

honorary doctorates, from Uppsala University (1994) and the University of Valladolid (2009)

SAGE author biography via Woodslane (2004); Inside Illinois, University of Illinois News Bureau (2009)

Key concepts

Responsive evaluation

An approach that organises an evaluation around the issues and concerns of people connected with a programme and reports their differing value perspectives.

Countenance model

Stake's 1967 framework for describing and judging a programme through its antecedents, transactions and outcomes, comparing intentions with observations.

Antecedents, transactions and outcomes

The conditions before a programme, the activities and encounters within it, and its results.

Congruence and contingency

In the countenance model, congruence is the match between what was intended and what was observed; contingency is the link between antecedents, transactions and outcomes.

Intrinsic, instrumental and collective case study

Stake's three types of case study: studying a case for its own sake, studying it to understand a wider issue, and studying several cases together.

Sources for this chapter: p. 576

KEY WORKS TO READ IN THE ORIGINAL

- **The Countenance of Educational Evaluation** 1967, Teachers College Record
- **The Art of Case Study Research** 1995
- **Qualitative Research: Studying How Things Work** 2010

THE BIG IDEAS OF THIS CHAPTER

1. **The countenance model** p. 90

   A full evaluation both describes and judges a programme, comparing what was intended with what was observed before, during and after it, and treating people's judgements as data.

2. **Responsive evaluation** p. 93

   Organise an evaluation around the issues that matter to the people connected with a programme, spend time observing what actually happens, and report in ways each audience can use, showing their differing judgements.

3. **Case study as a way of understanding** p. 96

   Study a particular bounded case in depth and in context, being clear whether you care about the case itself, use it to understand a wider issue, or study several cases together.

GIANTS CONNECTED TO THIS CHAPTER



Ralph W. Tyler

Responsive evaluation widened the lens beyond objectives and test scores.

p. 20



Egon G. Guba and Yvonna S. Lincoln

Fourth generation evaluation organises the work around stakeholders' claims, concerns and issues.

p. 318



Rick Davies and Jess Dart

Most Significant Change is another route to discovering what people value.

p. 362

In one line

He taught evaluators to listen before they measure, and to show how a programme looks from every place where people stand.

1

BIG IDEA 1 OF 3 · STAKE

The countenance model

A full evaluation both describes and judges a programme, comparing what was intended with what was observed before, during and after it, and treating people's judgements as data.

PICTURE IT

Think of a film review that only gives a star rating. You learn nothing about the plot, the acting or why some viewers loved it and others walked out. Stake's countenance model asks for the whole review: what came before, what happened on screen and how different audiences judged it.

The problem

By the mid 1960s, new federal money was funding ambitious curriculum reforms in the United States, but their evaluations often reported little more than average test scores at the end. They said almost nothing about the conditions schools started from, what actually happened in classrooms, or how teachers, parents and students judged the programme. Decision-makers received a number without a picture, and could not tell whether disappointing results reflected a poor design, poor delivery or difficult starting conditions.

What it is

The countenance model comes from Stake's article The Countenance of Educational Evaluation (1967), which he did not intend as a model but which was widely treated as one. It holds that a full evaluation both describes and judges, with more weight on description, and treats people's judgements as data. Description covers three kinds of information: antecedents, the conditions before the programme, such as learners' backgrounds and teachers' preparation; transactions, the encounters and activities within it; and outcomes. For each, the evaluator records intents and observations, forming a data matrix. Two analyses follow. Congruence asks whether what was intended actually happened. Contingency asks whether antecedents, transactions and outcomes are linked, both logically and in the data, for example whether planned activities could plausibly produce the intended outcomes. Judgement then compares the programme with standards, either absolute standards of excellence

or relative comparison with other programmes. In his 1973 paper on responsive evaluation, Stake described this matrix as having thirteen information categories.

How it works

1. **Record the rationale**
   Write down the programme's philosophy and purposes, which frame the whole matrix.
2. **Fill in intents**
   For antecedents, transactions and outcomes, list what the programme intends, for example who will enrol, what sessions will look like and what learners will gain.
3. **Observe what happens**
   Gather evidence on actual antecedents, transactions and outcomes from records, observation, interviews and assessments.
4. **Analyse congruence**
   Compare intents with observations, cell by cell, to find where delivery departed from design.
5. **Analyse contingency**
   Check whether the links between antecedents, transactions and outcomes hold, both in logic and in the data.
6. **Gather standards and judgements**
   Collect the standards different groups apply, absolute or relative, and their judgements of the programme, and report them as data beside the description.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO runs a teacher mentoring scheme in 60 government primary schools in Telangana. A countenance evaluation records an intended antecedent, mentors with five years' teaching experience, and finds that only half have it. It records an intended transaction, two classroom visits a month with feedback, and observes that visits averaged one, often without feedback, because mentors were also covering vacancies. The intended outcome was better questioning in lessons; observed outcomes show modest change, and only where visits were regular. The contingency analysis therefore points to delivery, not to the mentoring design itself. Finally, the team collects judgements: head teachers value the scheme's support, mentors feel stretched, and the district compares it unfavourably with a cheaper online course. The report presents all three judgements beside the description.

Did you know?

The topics studied at CIRCE, the Illinois centre Stake directed, ranged from teacher education for the deaf to youth sports and Veterans Affairs staff training.

USE IT WHEN

- Outcomes are disappointing and you need to know whether design, delivery or starting conditions are to blame.
- A programme's intended model may differ from what is actually delivered.
- Different stakeholders hold different standards for judging the same programme.

SKIP IT WHEN

- The question is narrowly about the size of one outcome, and process data will not be used.
- Time is so short that a full intents-and-observations matrix cannot be filled; then focus on the cells that matter most for the decision.

COMMON MISREADINGS

- [ ] Treating it as a form to fill in, when Stake did not intend a fixed model and stressed description and judgement as a way of seeing.
- [ ] Collecting only the programme designers' standards, when Stake treated the judgements of many groups as data in their own right.

WHY IT WAS GROUNDBREAKING

It widened evaluation beyond outcomes to include starting conditions, processes and multiple standards of judgement. It made the description of what actually happened a core task of evaluation, and gave evaluators a structure for explaining why results turned out as they did.

Origin

Published as The Countenance of Educational Evaluation in the Teachers College Record in 1967, while Stake was at the University of Illinois, where its evaluation centre, CIRCE, had been organised in 1964. The Encyclopedia of Evaluation notes parallels with the work of Daniel Stufflebeam and Malcolm Provus. Stake carried its data matrix into his responsive approach, presented in 1973.

RELATED IDEAS

Responsive evaluation · RT Tyler 20 · DS Stufflebeam 134 · MS Scriven 60

CONCEPT HIGHLIGHT

Responsive evaluation

An approach that organises an evaluation around the issues and concerns of people connected with a programme and reports their differing value perspectives.

2

BIG IDEA 2 OF 3 · STAKE

Responsive evaluation

Organise an evaluation around the issues that matter to the people connected with a programme, spend time observing what actually happens, and report in ways each audience can use, showing their differing judgements.

PICTURE IT

Think of a good doctor who begins by asking what is bothering you, rather than running only the standard tests. Responsive evaluation starts from the concerns of the people involved and lets them shape what gets examined.

The problem

Even a broad descriptive evaluation could still be designed far from the programme, around questions that distant policymakers set in advance. Teachers, participants and communities rarely saw their own concerns reflected, reports were written in technical language few could use, and formal designs spent much of their effort on instruments and data processing rather than on watching what happened. Where people value programmes differently, a single preset yardstick hid exactly the disagreements decision-makers needed to understand.

What it is

Stake held that an evaluation is responsive if it orients more to a programme's activities than to its intents, responds to the information needs of its audiences, and refers to the different value perspectives of the people involved when reporting success and failure. He contrasted it with preordinate evaluation, which fixes objectives, instruments and analyses in advance. The organising unit is the issue rather than the objective or hypothesis: issues are discovered by talking with students, parents, staff, sponsors and others, and they structure both data collection and discussion. The evaluator enlists a varied group of observers, uses the countenance data matrix, and draws on observation, judgements, case study and tests as fits. Findings are shared through portrayals, from short scripts and logs to narratives, photographs and taped conversations, made understandable to each audience. Stake estimated that a responsive evaluator might spend about 30 per cent of the time observing the programme, against 5 per cent in a preordinate design, and 15 per cent gathering judgements. The events of the study recur and overlap rather than following a fixed order.

How it works

1. **Talk with people first**
   Meet clients, staff, participants and other audiences to learn what they care about and what they need to know.
2. **Identify issues**
   Turn those conversations into a short list of issues, framed as questions, to organise the study, as Stake illustrated with questions about admissions and afternoon attendance from his TCITY evaluation.
3. **Observe the programme**
   Spend substantial time watching activities, using a varied team of observers, and record antecedents, transactions and outcomes.
4. **Gather judgements**
   Collect how different groups judge the programme's quality, and treat those judgements as data.
5. **Prepare portrayals**
   Build case accounts, narratives and other formats that convey what the programme is like, and try them on stand-in audiences before final reporting, the two-stage communication Stake described.
6. **Validate and revisit**
   Check accounts with participants, look for disconfirming evidence and return to earlier steps as new issues emerge, since the events of the study recur.
7. **Report in usable forms**
   Share findings with each audience in its own language and media, agreeing with clients whether a formal final report is needed.

WORKED EXAMPLE · AN ILLUSTRATION

A foundation commissions an evaluation of girls' after-school learning clubs run by an NGO in tribal districts of Chhattisgarh. In the first weeks the evaluator talks with girls, parents, facilitators, head teachers and the funder. Three issues emerge: parents worry about safe travel home after dark, facilitators about attendance falling during harvest, and the funder about learning gains. The study follows all three, with observers drawn from local women's groups and teachers, who walk the routes home, observe sessions across seasons and assess reading. Portrayals include a short photo story for parents' meetings, a summary of facilitator logs and a brief data note for the funder. Each group sees how the others judge the clubs. The resulting decision to hold sessions earlier during harvest months comes from issues the original logframe never contained.

USE IT WHEN

- Stakeholders hold different views of what the programme is for or what success looks like.
- The programme is complex or new and its important issues are not yet known.
- Findings must be used by communities and frontline staff as well as funders.

SKIP IT WHEN

- A funder needs a precise, predefined estimate of impact for a fixed decision, where a preordinate design is more efficient.
- There is no time to engage stakeholders properly, and a token consultation would only create false expectations.

COMMON MISREADINGS

- Equating it with a stakeholder workshop at inception, when Stake meant an evaluation that keeps responding to issues and audiences throughout.
- Assuming it means having no standards, when Stake still required judgement and simply insisted that different value perspectives be shown.

WHY IT WAS GROUNDBREAKING

It shifted evaluation from checking predefined objectives towards understanding a programme's quality from the inside, through the concerns of the people involved. It helped open evaluation to qualitative and naturalistic inquiry, and to the idea that reports must make sense to the people whose programme it is.

Origin

Stake presented Program Evaluation, Particularly Responsive Evaluation at a conference on New Trends in Evaluation in Göteborg, Sweden, in November 1973, during his Fulbright year there. CIRCE issued it as Occasional Paper 5; it was reprinted in Evaluation Models (2002) and the Journal of MultiDisciplinary Evaluation (2011). Later writers, including Tineke Abma, stressed dialogue, negotiation and power, and Stake set it beside standards-based evaluation in a 2004 book.

RELATED IDEAS

The countenance model

Case study as a way of understanding

G&L Guba and Lincoln 318
MP Patton 204
RC Chambers 296
JG Greene 340

In one line

A full evaluation both describes and judges a programme, comparing what was intended with what was observed before, during and after it, and treating people's judgements as data.

3

BIG IDEA 3 OF 3 · STAKE

Case study as a way of understanding

Study a particular bounded case in depth and in context, being clear whether you care about the case itself, use it to understand a wider issue, or study several cases together.

PICTURE IT

Think of a biographer who studies one life in depth. Sometimes the life itself is the point. Sometimes it is a window onto a whole era. Stake's case study types ask you to be clear which you are doing, and why.

The problem

Evaluators and researchers often treated single cases as anecdotes, useful for colour but not as evidence. Yet many evaluation questions concern a particular school, programme or community: how it works, why it struggles, what its people value. Without a clear account of why a case was chosen and how it should be studied, case studies were easy to dismiss and hard to do well, and the rich, particular knowledge that practitioners most need was undervalued in reports and decisions.

What it is

In The Art of Case Study Research (1995), Stake treated a case study as the in-depth study of a particular, bounded case, such as a school, programme or person, in its context, drawing on naturalistic, holistic, ethnographic, phenomenological and biographic methods. He distinguished three purposes. In an intrinsic case study the case itself is of primary interest. In an instrumental case study the case is studied to understand something else, such as an issue or a generalisation. In a collective case study several cases are studied together for that purpose; his Multiple Case Study Analysis (2005) developed this for studying many sites within one programme. The difference lies in purpose, not in the case, and one study can be both intrinsic and instrumental. Within a case, Stake organised inquiry around issues, selected cases to maximise learning, used triangulation to strengthen interpretations, accepted direct interpretation by the researcher as legitimate, and addressed how readers can generalise from one case to their own situations.

How it works

1. **Clarify purpose**
   Decide whether the case matters for itself, as an instrument for understanding an issue, or as one of a collection.
2. **Bound the case**
   Define what lies inside the case, its people, place and time, and what counts as context.
3. **Select to maximise learning**
   Choose the cases you can learn most from, which may not be the most typical, a point The Art of Case Study Research makes about case selection.
4. **Frame issues**
   Organise the inquiry around a few issue questions rather than a long list of variables.
5. **Gather and triangulate**
   Use observation, interviews and documents, and check key interpretations against other sources and against participants' views.
6. **Interpret and write for the reader**
   Combine direct interpretation with pattern analysis, and give readers enough of the case to judge what applies to their own settings.

WORKED EXAMPLE · AN ILLUSTRATION

A state department asks why one government residential school for tribal girls in Odisha keeps far more students through class ten than similar schools. Studied only for the department's interest in that school, it would be an intrinsic case. The MEL team instead frames it instrumentally: what can this school teach about how residential schools handle the move from home languages to the state language? It bounds the case to one academic year, frames three issues, including how teachers use home languages in the first months, and triangulates classroom observation, interviews with girls and wardens, and attendance records. Because the department wants broader lessons, the team adds three contrasting schools, making the design collective. The report describes each school richly enough that officials can judge what might transfer.

CONCEPT HIGHLIGHT

Countenance model

Stake's 1967 framework for describing and judging a programme through its antecedents, transactions and outcomes, comparing intentions with observations.

USE IT WHEN

- You need to understand how and why something works in a particular place, not only how much.
- A survey shows striking variation between sites and you need to explain it.
- A funder or government wants lessons from a positive or negative outlier.

SKIP IT WHEN

- The decision needs a statistically generalisable estimate across a population.
- Cases are chosen only to showcase success, which turns case study into promotion.

COMMON MISREADINGS

- Treating any short success story as a case study, without bounding the case, framing issues or triangulating evidence.
- Assuming that case studies cannot support generalisation at all, when Stake addressed how readers generalise from rich cases to their own situations.

WHY IT WAS GROUNDBREAKING

It gave evaluators and qualitative researchers a clear vocabulary for why and how to study single cases. It made case study a respected design within evaluation rather than an afterthought, and its three types are now standard in research methods teaching.

Origin

Stake set out the typology in The Art of Case Study Research (Sage, 1995), illustrated with a case from Harper School. He extended collective case work in Multiple Case Study Analysis (Guilford, 2005), with three case reports from a transnational early childhood programme, and in Qualitative Research: Studying How Things Work (2010). His earlier Case Studies in Science Education, with Jack Easley, had already applied case methods to schooling.

RELATED IDEAS

Responsive evaluation

G&L Guba and Lincoln 318

D&D Davies and Dart 362

BC Chilisa 498

CONCEPT HIGHLIGHT

Antecedents, transactions and outcomes

The conditions before a programme, the activities and encounters within it, and its results.

RELATED GIANTS



Ralph W. Tyler

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

p. 20



Daniel L. Stufflebeam

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

p. 134



Michael Scriven

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

p. 60



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318

CONCEPT HIGHLIGHT

Congruence and contingency

In the countenance model, congruence is the match between what was intended and what was observed; contingency is the link between antecedents, transactions and outcomes.

CONCEPT HIGHLIGHT

Intrinsic, instrumental and collective case study

Stake's three types of case study: studying a case for its own sake, studying it to understand a wider issue, and studying several cases together.

In one line

Organise an evaluation around the issues that matter to the people connected with a programme, spend time observing what actually happens, and report in ways each audience can use, showing their differing judgements.

Put it to work

USE IT ON MONDAY · RESPONSIVE EVALUATION

1

In the inception phase, talk to a handful of people from each stakeholder group and ask what worries them and what they hope to learn. List the issues before you fix the evaluation questions.

2

Choose one or two cases, such as a school, a block or a self-help group, to study in depth through observation and interviews alongside your wider data.

3

Report back in formats each group can use: an oral briefing for field staff, a visual summary for communities and a written report for the funder.

AI prompt **Copy, fill the brackets, then check every line**

I am designing a responsive evaluation, in the spirit of Robert Stake, for **[programme]** in **[location]**. Stakeholder groups are **[list groups]**. Here are my notes from initial conversations: **[paste notes]**. (1) Group the concerns and issues each stakeholder group raised, using only what is in my notes. (2) Show where groups' issues conflict. (3) Suggest three or four issues to organise the study around and one site for an in-depth case study, with reasons drawn from the notes. (4) Propose a reporting format for each group. If the notes are too thin to decide, say so rather than filling the gap.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a film review that only gives a star rating. You learn nothing about the plot, the acting or why some viewers loved it and others walked out. Stake's countenance model asks for the whole review: what came before, what happened on screen and how different audiences judged it.

YOUR MOVES Three moves from Stake's path

1 Master a craft before you reform it

Stake knew testing from the inside, so his critique of narrow measurement carried weight that an outsider's never could.

TRY THIS
Choose the method you are most tempted to dismiss, and learn it well enough to teach it before you criticise it.

2 Give your ideas a home

He directed CIRCE and helped organise the two bodies that later merged into the American Evaluation Association. Institutions let his ideas travel.

TRY THIS
Start or join one community of practice, even a monthly call of five evaluators, and keep it alive for a year.

3 Let other countries change your eyes

Fulbright years in Sweden and Brazil and visiting posts in Canada and the United Kingdom gave him, in our reading, a lesson that quality looks different from different places.

TRY THIS
Ask a peer from another country or state to review your evaluation criteria, and note what they would judge differently.

THE BIG IDEAS OF THIS CHAPTER

1 **The countenance model** p. 90

A full evaluation both describes and judges a programme, comparing what was intended with what was observed before, during and after it, and treating people's judgements as data.

2 **Responsive evaluation** p. 93

Organise an evaluation around the issues that matter to the people connected with a programme, spend time observing what actually happens, and report in ways each audience can use, showing their differing judgements.

3 **Case study as a way of understanding** p. 96

Study a particular bounded case in depth and in context, being clear whether you care about the case itself, use it to understand a wider issue, or study several cases together.

CONCEPT HIGHLIGHT

Intrinsic, instrumental and collective case study

Stake's three types of case study: studying a case for its own sake, studying it to understand a wider issue, and studying several cases together.

GIANTS CONNECTED TO THIS CHAPTER



Ralph W. Tyler

Responsive evaluation widened the lens beyond objectives and test scores.

p. 20



Egon G. Guba and Yvonna S. Lincoln

Fourth generation evaluation organises the work around stakeholders' claims, concerns and issues.

p. 318



Rick Davies and Jess Dart

Most Significant Change is another route to discovering what people value.

p. 362

KEY WORKS TO READ IN THE ORIGINAL

- **The Countenance of Educational Evaluation** 1967, Teachers College Record
- **The Art of Case Study Research** 1995
- **Qualitative Research: Studying How Things Work** 2010

CONCEPT HIGHLIGHT

Responsive evaluation

An approach that organises an evaluation around the issues and concerns of people connected with a programme and reports their differing value perspectives.

WORKSHEET · TRY THIS TEMPLATE

Stakeholder issues map

Print this page, or copy the headings into your notes.

1 Stakeholder group

Be specific: frontline teachers, not just staff.

2 Their main concerns

In their terms, from conversations, not from the proposal.

3 Evidence they would trust

Numbers, observation, stories, peer accounts or official records.

4 Format they would use

Meeting, poster, voice note, dashboard or report, and in which language.

5 Priority and trade-offs

Whose issues the evaluation will follow, and why. Say what is being left out.

REFLECT

Whose concerns are shaping your current evaluation questions, and whose are missing?

CONNECTS TO

RT **Ralph W. Tyler**
Responsive evaluation widened the lens beyond objectives and test scores. p. 20

G&L **Egon G. Guba and Yvonna S. Lincoln**
Fourth generation evaluation organises the work around stakeholders' claims, concerns and issues. p. 318

D&D **Rick Davies and Jess Dart**
Most Significant Change is another route to discovering what people value. p. 362

THE GREAT DEBATES · ACT 1

Goals or No Goals?

THE QUESTION

Should a programme be judged against its own stated objectives, or against everything it actually does?

RT Tyler 20 | MS Scriven 60 | RS Stake 82

SIDE A

Start from objectives (Tyler)

Tyler's approach began with purposes. In the Eight-Year Study, rather than impose one standard test on very different schools, he and his staff worked from each school's own stated aims, clarified what each would look like in students' work, and built varied instruments to gather evidence on them. He distilled this into the 1949 Basic Principles of Curriculum and Instruction. The strength of the approach is fairness and clarity: a programme is held to what it set out to do, in terms its own staff understand, and evidence is gathered on more than test scores. In our reading, every results framework that asks a programme to define its intended outcomes and then measure them descends from this logic.

READ THE ORIGINALS

- **Eight-Year Study** Education Encyclopedia, StateUniversity.com
- **Scriven's Goal-Free Evaluation (Youker)** Journal of MultiDisciplinary Evaluation 20(47), 2024
- **A Tree: Planted and Growing (Christie)** Journal of MultiDisciplinary Evaluation 20(48), 2024

SIDE B

Look for all effects (Scriven, with Stake)

Critics argued that judging a programme only against its own objectives can miss unintended effects and cannot say whether the objectives were worth pursuing. In Pros and Cons About Goal-Free Evaluation (1972), Scriven drew on the way consumer product testers judge goods without asking manufacturers what they intended, and proposed that evaluators set stated goals aside and search for all of a programme's effects. Robert Stake, from a different angle, argued in 1967 that a full evaluation must both describe and judge, comparing intentions with what was actually observed, and later developed responsive evaluation around the issues people at a site care about. In our reading, their shared case is that goals can become blinkers.

WHAT HAPPENED

Goal-free evaluation never became the norm. Brandon Youker, who has written extensively on the approach, reports that most evaluators treat it as a rhetorical reminder rather than a workable design, partly because strict blinding to goals is hard to achieve. Yet its central worry spread. In Christie's account of the second edition of the theory tree, goal-free evaluation sits on the objectivist side of the valuing branch, and Stake's responsive evaluation underlies most theories on the subjectivist side. Most Significant Change was designed partly because predefined indicators risked missing the unexpected. Tyler did not formally reply to his critics.

Where you stand

In your last evaluation, what was the most important effect that nobody had written into the logframe?

Could you design one part of your next evaluation in which the field team is not told the programme's objectives? What would you gain and lose?

Your position

Which side do you find yourself on today, and what evidence would change your mind?

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

THE GIANTS IN THIS DEBATE



Ralph W. Tyler

The science teacher who taught evaluation to begin with one plain question: what were we trying to achieve?

p. 20



Michael Scriven

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

p. 60



Robert E. Stake

The testing specialist who argued that evaluation should listen to what people around a programme actually care about.

p. 82

ACT 2 OF 5

2

1960s to 1990s

The Experimenting Society



Donald T. Campbell
p. 112



Daniel L. Stufflebeam
p. 134



Carol H. Weiss
p. 152



Huey-tsyh Chen
p. 174

ACT 2 · 1960S TO 1990S

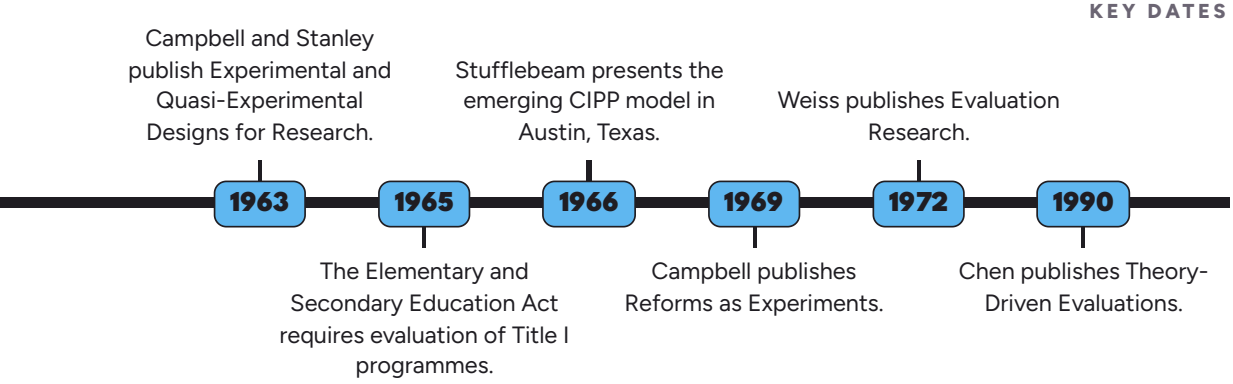
The Experimenting Society

In the 1960s evaluation left the classroom and entered government. The Elementary and Secondary Education Act of 1965 brought federal money to American schools, and an amendment demanded by Senator Robert Kennedy required evaluation of its Title I programmes. Local educators were often unprepared, and that demand, the Encyclopedia of Evaluation records, helped drive the development of many evaluation models. The tools to hand had been built for testing pupils, not for judging whole policies.

Donald Campbell supplied a discipline for judging causes in the field. In 1963, with Julian Stanley, he catalogued sixteen designs and listed eight threats to internal validity and four to external validity. In 1969, in Reforms as Experiments, he turned the toolkit on public policy, showing how a Connecticut claim about a speeding crackdown ignored trends and ordinary year-to-year swings. Daniel Stufflebeam, a test developer suddenly advising on federal evaluation, built the CIPP model, first presented in Austin in 1966, which treated evaluation as a service to decisions at every stage of a programme's life. Carol Weiss, whose 1972 textbook was among the first in the field, asked why findings seemed to be ignored, and in 1979 described seven ways research enters policy, including the slow enlightenment of debate. Huey-tsyh Chen, working with Peter Rossi, argued in 1983 that experiments treated programmes as black boxes, and in 1990 set out theory-driven evaluation.

In 1991 Shadish, Cook and Leviton grouped Campbell with Scriven as theorists who sought to bring truth to social problem solving, and Weiss with Stake and Joseph Wholey as theorists who emphasised use and pragmatism. In our reading, this act answered the first by scaling up its ambition: from

KEY DATES



whether pupils learned to whether a society could learn from its own reforms.

THE GREAT DEBATE OF THIS ACT

Proof Here or Promise Elsewhere? Should evaluation first prove that a programme caused an effect, or show whether the effect will travel? p. 196

THE GIANTS OF THIS ACT



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Daniel L. Stufflebeam

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

p. 134



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174

Huey-tsyh Chen, in one line

Huey-tsyh Chen opened the black box of programme evaluation and showed practitioners how to read the machinery inside.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

05



Emblem: an original HubForge design for this figure, not a likeness.

Donald T. Campbell

1916 to 1996 Active 1950s to 1970s United States; fieldwork in East Africa

Methods, rigour and causality **Impact evaluation and evidence systems**

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

IN ONE LINE

He gave evaluators a way to be honest about cause when the world will not hold still for an experiment.

In 1955, Connecticut counted a record 324 deaths on its roads. Governor Abraham Ribicoff ordered a fierce crackdown on speeding, and in 1956 the toll fell to 284. The governor announced 40 lives saved, a cut of 12.3 per cent. Case closed? Not for Donald Campbell. In a 1969 paper now widely credited with helping launch programme evaluation as a field, he showed that a drop after a record year is exactly what normal ups and downs alone might produce. Only a longer run of years, compared with neighbouring states, could reveal a real effect, and it proved modest. Campbell, a psychologist with more than 200 papers to his name, spent his career on one stubborn question: how do we know it was the programme that made the difference?

Life CHAPTER 05

Donald Thomas Campbell was born on 20 November 1916 in Grass Lake, Michigan. His father was an agronomist, and his work took the family first to a cattle ranch in Wyoming and later to California, where he became an agricultural extension agent. After finishing high school in 1934, the future pioneer of research design spent a year working on a turkey ranch in Victorville, California, before enrolling at San Bernardino Valley Union Junior College in 1935.

He transferred to the University of California, Berkeley, and graduated in 1939 at the top of his class, sharing the honour with his sister, Fayette. His doctoral studies in psychology at Berkeley were interrupted by service in the US Naval Reserve during the Second World War, and he completed his PhD in 1947. He then taught at Ohio State University from 1947 to 1950 and at the University of Chicago from 1950 to 1953, before settling at Northwestern University, where he spent 26 years and became the core of

BIG IDEAS

1. **Quasi-experiments and threats to validity**
2. **The experimenting society**
3. **Campbell's law: the corruption of indicators**

KEY WORKS

Experimental and Quasi-Experimental Designs for Research
1963 · With Julian C. Stanley

Reforms as Experiments
1969

Quasi-Experimentation: Design and Analysis Issues for Field Settings
1979 · With Thomas D. Cook

its social psychology doctoral programme, training several generations of students.

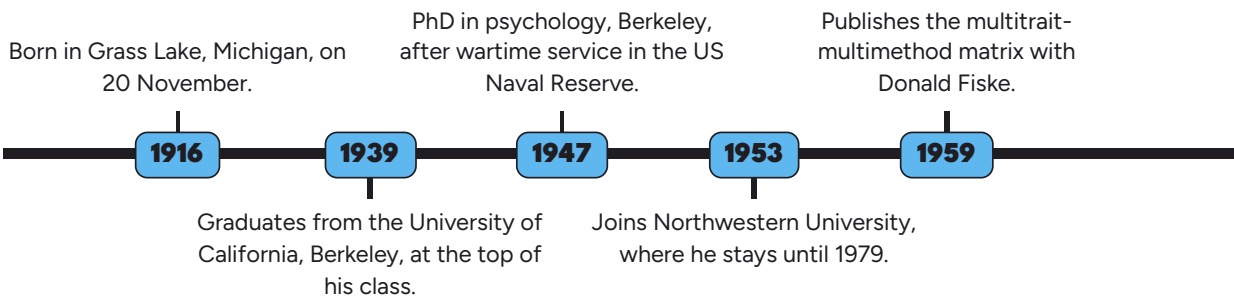
Campbell's range was unusually wide. His research ran from perceptual contrast effects to cross-cultural studies of ethnocentrism and relations between groups, and in 1964 he joined the ethnographer Robert A. LeVine in fieldwork among the Gusii of East Africa. He developed a model of knowledge growth built on blind variation and selective retention, which linked perception, learning and the passing on of culture, and he became a leading figure in evolutionary epistemology and, later, the sociology of science. Across these interests he published more than 200 scholarly papers. Colleagues remembered an intimidating intellect softened by self-deprecating humour, and a mentor who encouraged dissent rather than conformity.

In 1979 he moved to Syracuse University's Maxwell School as New York State Board of Regents Albert Schweitzer Professor, and in 1982 to Lehigh University as University Professor of Social Relations, Psychology and Education. He was elected to the National Academy of Sciences and the American Academy of Arts and Sciences in 1973, and received honorary doctorates from Michigan, Chicago, Southern California, Northwestern and Oslo, among others. He semi-retired from Lehigh in 1994; the dinner marking the occasion drew colleagues from psychology, biology, philosophy, sociology, education, anthropology and law. His wife, Barbara Frankel, was an anthropologist on the Lehigh faculty. Campbell died on 6 May 1996 in Bethlehem, Pennsylvania, after complications following surgery, a few months before his 80th birthday.

The spark

By the 1950s, scientists knew how to run controlled experiments in a laboratory. But the questions that mattered to schools and governments were being answered with simple before-and-after comparisons, and nobody could say how much those comparisons proved. Campbell and his co-author Julian Stanley later hailed William McCall's 1923 book How to Experiment in Education as an undervalued classic that had seen the problem first. In our reading, Campbell's unusually wide interests, from how the

TIMELINE · LIFE AND WORK



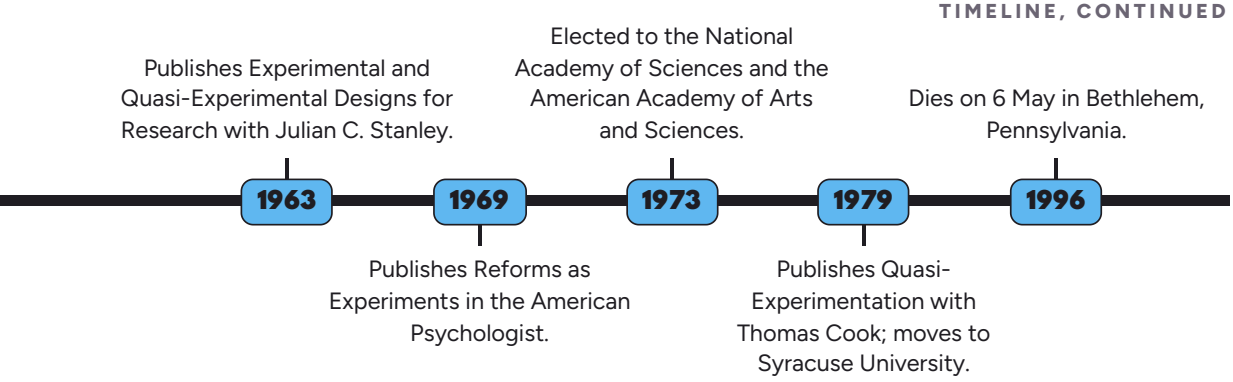
eye perceives to how cultures pass on knowledge, made him alert to how easily people fool themselves.

The breakthrough CHAPTER 05

Here is the problem Campbell faced. A school introduces a new reading scheme, and a year later scores are up. Did the scheme work? Perhaps. Or perhaps the children simply grew older, or the test was easier, or the best pupils were picked for the scheme, or the scores were unusually low to begin with and were bound to bounce back. In a laboratory you can rule these out by randomly assigning who gets the treatment. In real schools, cities and states, that is often impossible. By the mid twentieth century, nobody had a shared, systematic way to say what a real-world comparison could and could not prove.

The key text appeared in 1963: Experimental and Quasi-Experimental Designs for Research, written with Julian C. Stanley as a chapter in N. L. Gage's Handbook of Research on Teaching and soon reissued as a short book. It catalogued sixteen designs, from the weakest one-shot case study to true experiments and quasi-experiments such as the nonequivalent control group, the time series and the regression-discontinuity design, and it listed eight threats to internal validity and four to external validity. In 1969, in Reforms as Experiments in the American Psychologist, Campbell turned this toolkit on public policy. He showed, for example, that a Connecticut governor's claim that a 1955 speeding crackdown had cut road deaths ignored long-term trends, normal year-to-year swings and a natural fall after a record year, although a longer series with comparison states did suggest some real effect.

TIMELINE, CONTINUED



The problem it solved

Field researchers had no disciplined way to judge programme effects outside the laboratory.

Impact WHAT CHANGED

Campbell's 1969 paper is widely credited with helping launch programme evaluation as a field, and the vocabulary he and Stanley introduced, of internal and external validity and of history, maturation and selection as threats, became the common language of evaluation methods teaching. He extended it with Thomas Cook in Quasi-Experimentation: Design and Analysis Issues for Field Settings (1979). Some of his designs travelled far beyond psychology: the regression-discontinuity design, first used by Thistlethwaite and Campbell in 1960 to study the effect of National Merit Scholarship certificates, was revived by economists in the 1990s and 2000s, including in work on Mexico's PROGRESA cash-transfer programme.

His name lives on institutionally in the Campbell Collaboration, which grew out of meetings at University College London in 1999 and at the University of Pennsylvania in February 2000, was modelled on the Cochrane Collaboration and was named after him because of Reforms as Experiments. It produces systematic reviews of social interventions, and its international development coordinating group, run with 3ie, focuses on reviews relevant to low- and middle-income countries in areas such as nutrition, education, social protection and fragile states. Our interpretation is that when development evaluators today defend a difference-in-differences or matched comparison design, or warn that a payment-by-results indicator may be gamed, they are working with tools and cautions Campbell supplied.

KEY WORKS TO READ IN THE ORIGINAL

- **Experimental and Quasi-Experimental Designs for Research** 1963, With Julian C. Stanley
- **Reforms as Experiments** 1969
- **Quasi-Experimentation: Design and Analysis Issues for Field Settings** 1979, With Thomas D. Cook

WHERE PEOPLE DISAGREE

Debates

Campbell's emphasis on internal validity drew sustained debate. Lee Cronbach argued that the more pressing question for policy is usually external validity, whether a programme will work in another place and with other people, and developed his own framework of units, treatments, observations and settings in response. Qualitative and theory-based evaluators have argued that a focus on whether a programme caused an effect can leave aside how and why it worked. Others note that quasi-experiments still leave doubt when groups differ in ways nobody measured. His own model of the experimenting society, notably, depended on exactly this kind of open criticism and reanalysis.

LIMITS, IN OUR READING

- ☐ Quasi-experimental designs can still leave real doubt when comparison groups differ in ways nobody measured.
- ☐ A focus on whether a programme caused an effect can underweight how and why it worked, which theory-based and qualitative approaches set out to explain.

WHY THIS MATTERS TO YOU

Campbell gives everyone a habit worth keeping for life: before you believe that something caused a change, ask what else could explain it. Would it have happened anyway? Was the starting point unusual? Were the people who took part different from those who did not? When you choose a comparison group, look at five years of data instead of two, or use the cut-off by which a scheme was allocated, you are using his designs. And when you warn that paying people per registration will inflate registrations, you are applying his law. He offers a way to be rigorous when a perfect trial is impossible, which in real life is most of the time.

16

research designs catalogued and compared in Experimental and Quasi-Experimental Designs for Research

Campbell and Stanley, Experimental and Quasi-Experimental Designs for Research (1963)

200+

scholarly papers published

APS Observer (1996); Lehigh University Libraries memorial

17

books dedicated to him by colleagues and former students, at last count

APS Observer (1996); Lehigh University Libraries memorial

Key concepts

Internal validity
The degree to which a study rules out rival explanations, so that an observed change can be attributed to the programme rather than to something else.

External validity
The degree to which a finding can be generalised to other people, settings, times and versions of the programme.

Nonequivalent control group design
A quasi-experiment comparing a treated group with a similar untreated group that was not formed by random assignment, usually with measures before and after.

Interrupted time series
A design that looks for a break in a long run of repeated measurements at the point a reform began, which helps separate a real effect from trends and normal fluctuation.

Regression-discontinuity design
A design, introduced by Donald Thistlethwaite and Campbell in 1960, that compares cases just above and just below a cut-off used to allocate a benefit.

Multitrait-multimethod matrix
A 1959 method, with Donald Fiske, for checking a measure's validity by comparing several traits each measured by several methods.

Sources for this chapter: p. 576

Did you know?

By the time of his death, at least 17 books had been dedicated to Campbell by colleagues and former students.

THE BIG IDEAS OF THIS CHAPTER

1. **Quasi-experiments and threats to validity** p. 120

   When you cannot randomise, you can still make a credible causal claim by choosing a design that rules out, one by one, the named rival explanations for an observed change.

2. **The experimenting society** p. 123

   Governments should treat reforms as trials, announcing the problem rather than promising a solution, testing responses with the strongest feasible design and keeping or dropping them on honest evidence.

3. **Campbell's law: the corruption of indicators** p. 126

   The more a quantitative indicator is used to make decisions that reward or punish, the more it will be corrupted and the more it will distort the very activity it was meant to monitor.

GIANTS CONNECTED TO THIS CHAPTER



Lee J. Cronbach

A fellow measurement giant who took a more sceptical view of relying mainly on randomised comparisons.

p. 38



Esther Duflo and Abhijit Banerjee

Randomised evaluations carried experimental logic into anti-poverty programmes at scale.

p. 432



Howard White

Theory-based impact evaluation pairs a credible counterfactual with a tested causal chain.

p. 476

In one line

He gave evaluators a way to be honest about cause when the world will not hold still for an experiment.

1

BIG IDEA 1 OF 3 · CAMPBELL

Quasi-experiments and threats to validity

When you cannot randomise, you can still make a credible causal claim by choosing a design that rules out, one by one, the named rival explanations for an observed change.

PICTURE IT

Think of a detective with a list of usual suspects. Before naming the culprit, she rules out each alternative in turn: the person with an alibi, the one who was out of town. Campbell's threats to validity are the usual suspects for any claimed effect, and a quasi-experiment is the investigation that rules them out.

The problem

Much of what matters in education and social policy happens in schools, districts and whole states, where random assignment is often impossible. In the mid twentieth century, claims of programme success usually rested on simple before-and-after comparisons, and nobody had a shared way to say what such comparisons could and could not prove. Officials could take credit for changes caused by something else, while effective programmes could be dismissed, with real consequences for public money and for the people programmes served.

What it is

A quasi-experiment compares groups or periods that were not formed by random assignment. Campbell and Julian Stanley's 1963 text catalogued sixteen designs in three families: three pre-experimental designs, such as the one-shot case study; three true experiments; and ten quasi-experiments, including the time series, the nonequivalent control group, the multiple time series and regression-discontinuity analysis. Each design is judged by which threats it controls. Internal validity asks whether the programme, rather than something else, made the difference; its eight threats are history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality and the interaction of selection with maturation. External validity asks to whom and where a result can be generalised, and faces four threats, including the reactive effect of testing and the reactive effects of experimental arrangements. Campbell and Stanley treated internal validity as the indispensable

minimum, while insisting that external validity also mattered. The logic is elimination: a design earns trust by making specific rival explanations implausible.

How it works

1. **State the claim and its rivals**
   Write down the causal claim, then list which of the named threats could produce the same pattern, using Campbell and Stanley's checklist.
2. **Find a comparison**
   Look for a comparable untreated group, such as neighbouring blocks or a later cohort, to form a nonequivalent control group.
3. **Lengthen the time series**
   Collect as many measurements as possible before and after the change, so that trends and normal ups and downs become visible, as Campbell did with Connecticut's road deaths in 1969.
4. **Exploit allocation rules**
   Where a benefit is allocated by a cut-off score, compare cases just either side of it, the regression-discontinuity design that Thistlethwaite and Campbell introduced in 1960.
5. **Check measures and dropouts**
   Confirm that instruments, scorers and record-keeping did not change, and compare who left each group.
6. **Judge each threat explicitly**
   For every threat, record whether the design rules it out, makes it less plausible or leaves it open, and report the remaining doubts honestly.

WORKED EXAMPLE · AN ILLUSTRATION

A state introduces a remedial reading scheme only in blocks whose average scores were lowest in the previous year's assessment, and the next year those blocks improve. Applying Campbell, the MEL team notes that statistical regression alone could produce a rise, because extremely low scores tend to drift back towards the average. History is another rival: a new textbook arrived statewide the same year.

The team assembles five years of block scores to see the earlier trend, adds blocks that just missed the cut-off as a comparison, and checks that the assessment and its scoring did not change. The scheme blocks improve more than blocks just above the threshold, beyond the earlier trend. The report states which threats this rules out and which, such as differences in local leadership, remain open.

Did you know?

Campbell graduated from Berkeley in 1939 sharing top honours with his own sister, Fayette.

USE IT WHEN

- Random assignment is impossible or unethical, but a comparison group or long data series is available.
- A programme was allocated by a clear cut-off, such as a score, poverty rank or date.
- You need to review or defend the causal claims in someone else's evaluation.

SKIP IT WHEN

- The question is how or why a programme works for different people, where theory-based or qualitative designs are better suited.
- No credible comparison or pre-programme data exist, so a design would only dress up a simple before-and-after comparison.

COMMON MISREADINGS

- [ ] Treating the threat list as a box-ticking exercise rather than reasoning about which threats are plausible in this setting.
- [ ] Assuming a quasi-experiment is either as good as a randomised trial or worthless, when Campbell graded designs by the specific threats each one controls.

WHY IT WAS GROUNDBREAKING

It turned vague doubts about causal claims into a named, shared checklist that any evaluator could apply and argue about. It made credible causal reasoning possible outside the laboratory, and its vocabulary of internal and external validity became the common language of evaluation methods teaching.

Origin

First set out in Experimental and Quasi-Experimental Designs for Research, by Campbell and Julian C. Stanley, a chapter in N. L. Gage's Handbook of Research on Teaching (1963), soon reissued as a short book. It acknowledged William McCall's How to Experiment in Education (1923) as an undervalued forerunner. Campbell and Thomas Cook expanded the framework in Quasi-Experimentation: Design and Analysis Issues for Field Settings (1979).

RELATED IDEAS

The experimenting society · LC Cronbach 38 · D&B Duflo and Banerjee 432 · HW White 476 · CW Weiss 152

2

BIG IDEA 2 OF 3 · CAMPBELL

The experimenting society

Governments should treat reforms as trials, announcing the problem rather than promising a solution, testing responses with the strongest feasible design and keeping or dropping them on honest evidence.

PICTURE IT

Think of a sensible cook trying a new recipe. She does not promise her family it will be delicious. She tries it, tastes honestly and keeps it only if it works. Campbell asked governments to treat reforms the same way.

The problem

Governments launch reforms with confident promises. Once an official has staked a reputation on success, an honest evaluation becomes a threat, so weak programmes are defended with testimonials and selective data, and successful ones cannot be told apart from lucky ones. In 1969 Campbell argued that this was mainly a political failure rather than a technical one. Without a different stance from administrators, even the best designs would go unused, and citizens would keep paying for reforms nobody knew worked.

What it is

The experimenting society is Campbell's vision of a polity that treats reforms as trials to be tested and adjusted, set out in Reforms as Experiments (1969) and in Methods for the Experimenting Society, prepared for talks in 1971. Its core contrast is between the trapped administrator, who has promised in advance that a reform will work and so cannot tolerate failure, and the experimental administrator, who stresses how important the problem is and commits to trying one solution and, if it fails, another. Campbell paired this stance with designs suited to real reforms, ordered roughly from weaker to stronger: the interrupted time series; the control series, adding comparison units; regression discontinuity; and randomised experiments using lotteries for scarce places, staged roll-out or pilot areas with controls. He urged consistent record-keeping, several measures of benefits and side effects, public audit of records, and evaluation of policies rather than of individual officials. The society he sketched would be honest, non-dogmatic, open to reanalysis by outsiders and decentralised.

How it works

1. **Announce the problem, not the solution**
   State the seriousness of the problem and the intention to test responses, so that no one's reputation depends on a single programme, the stance of Campbell's experimental administrator.
2. **Choose the strongest feasible design**
   Where places are scarce, allocate by lottery; where roll-out must be gradual, stage it so that later areas act as comparisons; otherwise use a time series with control units.
3. **Keep record-keeping stable**
   Hold definitions and recording practices constant before and after the reform, since Campbell warned that changes in bookkeeping can mimic or hide effects.
4. **Measure benefits and side effects**
   Track several outcome measures, including possible harms, and keep records on those who did not receive the programme.
5. **Evaluate the policy, not the person**
   Present results as a test of the reform rather than a judgement on individual officials, which Campbell argued reduces the incentive to sabotage measurement.
6. **Publish and invite reanalysis**
   Release results and data whatever they show, let outside critics reanalyse them, and keep, change or drop the reform accordingly.

WORKED EXAMPLE · AN ILLUSTRATION

A state government in India wants to expand a scheme giving bicycles to girls entering secondary school, but the first year's budget covers only some districts. Instead of announcing that the scheme will end dropout, the department announces that girls leaving school after class eight is a serious problem and that it will test responses. It stages the roll-out so that districts join in a randomly drawn order, keeps enrolment definitions unchanged, and tracks attendance, enrolment and reported safety on the journey to school. Results are published after two years, and the data are shared with university researchers for reanalysis. If later districts show the same gains when their turn comes, the case grows stronger; if not, the department tries a different response without anyone losing face.

CONCEPT HIGHLIGHT

Internal validity

The degree to which a study rules out rival explanations, so that an observed change can be attributed to the programme rather than to something else.

USE IT WHEN

- A government or large NGO is about to roll out a reform in phases or with a limited budget.
- Leaders are willing to frame a new scheme as a test rather than a promise.
- You need to argue for publishing results and data, including disappointing ones.

SKIP IT WHEN

- The decision has already been taken and the pilot is a formality, where an honest test is unlikely to change anything.
- Withholding or delaying the programme would cause clear harm and no fair allocation rule exists.

COMMON MISREADINGS

- [ ] Reading it as a demand that every policy be settled by randomised trial, when Campbell offered a ladder of designs and stressed openness, criticism and decentralised trial as much as method.
- [ ] Using pilots as a public relations step towards a decision already taken, the very trap Campbell described.

WHY IT WAS GROUNDBREAKING

It linked research methods to the politics of how governments learn, explaining why so many reforms were never honestly evaluated: the incentives worked against candour. It also offered a practical way out, combining a change in how reforms are announced with designs that fit how reforms are actually rolled out.

Origin

The argument first appeared in Reforms as Experiments, American Psychologist, volume 24, 1969, pages 409 to 429. Campbell developed the political vision in Methods for the Experimenting Society, prepared for presentations in 1971 and published in revised form in 1988 in Methodology and Epistemology for Social Science, edited by E. S. Overman. The Campbell Collaboration, founded in 2000, took his name because of this work.

RELATED IDEAS

Quasi-experiments and threats to validity

Campbell's law: the corruption of indicators

D&B Duflo and Banerjee 432

HW White 476

RR Rist 410

In one line

When you cannot randomise, you can still make a credible causal claim by choosing a design that rules out, one by one, the named rival explanations for an observed change.

3

BIG IDEA 3 OF 3 · CAMPBELL

Campbell's law: the corruption of indicators

The more a quantitative indicator is used to make decisions that reward or punish, the more it will be corrupted and the more it will distort the very activity it was meant to monitor.

The problem

As governments began to manage by numbers in the 1970s, quantitative indicators were increasingly tied to budgets, promotions and contracts. Planners assumed that a number used for decisions would remain an honest reflection of reality. Campbell saw evidence that it did not: crime statistics, employment office placements and factory output targets all bent under pressure. For development programmes paid by results, the risk is that the indicator improves while the lives it was meant to reflect do not.

What it is

Campbell's law is the pessimistic generalisation Campbell stated in Assessing the Impact of Planned Social Change (1976): the more a quantitative social indicator is used for social decision-making, the more it is subject to pressures to corrupt it, and the more it distorts the social processes it was meant to monitor. It has two parts. The indicator itself is corrupted, through underrecording, reclassification or gaming; and the underlying activity is distorted, as people work to the number rather than the purpose. His examples included police underrecording complaints and downgrading offences to improve clearance rates, employment offices taking the easiest clients when judged by placements, contractors paid on test gains teaching the specific test items, and production targets by weight yielding only the heaviest goods. He compared indicators to votes: both attract corruption once real stakes attach. In the same paper he argued for combining qualitative, common-sense knowledge with quantitative data and for evaluating policies rather than individuals.

In one line

The more a quantitative indicator is used to make decisions that reward or punish, the more it will be corrupted and the more it will distort the very activity it was meant to monitor.

How it works

1. **Map the stakes**
   For each indicator, note who is rewarded or penalised by it, since Campbell tied corruption pressure to how far a number is used for decisions.
2. **Predict the gaming**
   Ask how someone could move the number without moving the reality, using his examples of underrecording, choosing easy cases and teaching to the test as prompts.
3. **Use several indicators**
   Pair each high-stakes number with others that are hard to game at the same time, including measures of quality and of side effects.
4. **Add qualitative checks**
   Triangulate the numbers with observation, case studies and participants' accounts, as Campbell urged in the 1976 paper.
5. **Audit independently**
   Verify a sample of records against reality through independent spot checks.
6. **Judge systems, not individuals**
   Where possible, use the indicator to judge policies or systems rather than individual staff, and review any indicator that shows signs of distortion.

WORKED EXAMPLE · AN ILLUSTRATION

A nutrition programme in several districts of Maharashtra pays community workers a bonus based on the number of children weighed each month. Within a year reported coverage approaches 100 per cent, yet independent visits find many entries copied from earlier months, and counselling for mothers of underweight children, the activity the weighing was meant to trigger, has become rare. Applying Campbell's law, the MEL team keeps weighing coverage but stops tying bonuses to it alone, adds a quarterly independent re-weighing of a sample, tracks whether underweight children received a counselling visit, and holds case reviews with workers about why the numbers drifted. Coverage falls slightly on paper, while the share of underweight children receiving follow-up rises, and the data become trustworthy enough to guide decisions.

CONCEPT HIGHLIGHT

External validity

The degree to which a finding can be generalised to other people, settings, times and versions of the programme.

USE IT WHEN

- Indicators will be tied to payments, bonuses, rankings or funding decisions.
- Reported figures look too good, or improve suddenly without matching change on the ground.
- You are designing a results framework or payment-by-results contract.

SKIP IT WHEN

- An indicator is used only for low-stakes learning within a team, where pressure to corrupt it is small; even then, revisit the risk if stakes rise.
- It is used as a reason to avoid measurement altogether, which Campbell never advised.

COMMON MISREADINGS

- Concluding that numbers should be abandoned, when Campbell kept advocating quantitative evaluation and asked for numbers to be combined with qualitative knowledge and protected from corruption.
- Assuming corruption means dishonest individuals, when the pressure is structural and arises wherever real stakes attach to a number.

WHY IT WAS GROUNDBREAKING

It warned, years before results-based management became common, that measuring for accountability changes the behaviour being measured. It gave indicator designers a law to plan around rather than a surprise to discover too late.

Origin

Campbell presented an early version at a social psychology conference in Visegrad, Hungary, in May 1974, and published it in December 1976 as Assessing the Impact of Planned Social Change, Occasional Paper 8 of the Public Affairs Center at Dartmouth College, with support from the Russell Sage Foundation and the National Science Foundation. The Journal of MultiDisciplinary Evaluation reprinted it in 2011.

RELATED IDEAS

The experimenting society

RC Chambers 296 · D&D Davies and Dart 362 · RR Rist 410

CONCEPT HIGHLIGHT

Nonequivalent control group design

A quasi-experiment comparing a treated group with a similar untreated group that was not formed by random assignment, usually with measures before and after.

RELATED GIANTS



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Esther Duflo and Abhijit Banerjee

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

p. 432



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152

CONCEPT HIGHLIGHT

Interrupted time series

A design that looks for a break in a long run of repeated measurements at the point a reform began, which helps separate a real effect from trends and normal fluctuation.

CONCEPT HIGHLIGHT

Regression-discontinuity design

A design, introduced by Donald Thistlethwaite and Campbell in 1960, that compares cases just above and just below a cut-off used to allocate a benefit.

CONCEPT HIGHLIGHT

Multitrait-multimethod matrix

A 1959 method, with Donald Fiske, for checking a measure's validity by comparing several traits each measured by several methods.

Put it to work

USE IT ON MONDAY · QUASI-EXPERIMENTAL DESIGN AND THREATS TO VALIDITY

1. When the terms of reference ask for impact but randomisation is impossible, list the plausible rival explanations for any change you expect to see, such as other programmes, seasonal effects, who chose to join and who dropped out, before you choose a design.

2. Pick the design that rules out the most plausible rivals within the budget: a comparison group, repeated measures before and after, or a cut-off rule. Write down which threats remain.

3. In the report, add a short section that names the remaining threats and how much they could change the conclusion. Being explicit about weaknesses makes the claim stronger, not weaker.

AI prompt **Copy, fill the brackets, then check every line**

Act as a methods reviewer using Donald Campbell's threats-to-validity thinking. Our programme is **[programme description]**, delivered to **[participants]** in **[setting]**. We want to claim that it caused **[outcome]**. Our planned design is **[design, for example before-and-after with a non-random comparison group]**. (1) List the most plausible rival explanations for the change we might observe. (2) Rate each as high, medium or low plausibility using only the context I describe, and explain why. (3) Suggest practical design or data steps that could address each one within a budget of **[budget]**. Do not produce effect sizes or sample sizes.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

CONCEPT HIGHLIGHT

Multitrait-multimethod matrix

A 1959 method, with Donald Fiske, for checking a measure's validity by comparing several traits each measured by several methods.

YOUR MOVES **Three moves from Campbell's path**

1 Name your weaknesses before critics do

Campbell made listing the threats to validity of his own designs a standard practice, and that openness is what made his arguments trusted by experts.

TRY THIS
Add a limitations box to the first page of your next evaluation design, and list the three biggest ways it could mislead.

2 Read far outside your lane

He worked across psychology, anthropology, including fieldwork among the Gusii of East Africa in 1964, and the philosophy of science, and that breadth sharpened his methods.

TRY THIS
Pick one book this quarter from a field you never read, such as anthropology, economics or design, and note one idea you can borrow.

3 Grow people, not only papers

Over 26 years at Northwestern he trained generations of doctoral students and encouraged dissent. At least 17 books were dedicated to him.

TRY THIS
Offer one junior colleague a regular half hour each month to review their work and argue with your ideas.

Picture it

Think of a detective with a list of usual suspects. Before naming the culprit, she rules out each alternative in turn: the person with an alibi, the one who was out of town. Campbell's threats to validity are the usual suspects for any claimed effect, and a quasi-experiment is the investigation that rules them out.

THE BIG IDEAS OF THIS CHAPTER

1. **Quasi-experiments and threats to validity** p. 120

   When you cannot randomise, you can still make a credible causal claim by choosing a design that rules out, one by one, the named rival explanations for an observed change.

2. **The experimenting society** p. 123

   Governments should treat reforms as trials, announcing the problem rather than promising a solution, testing responses with the strongest feasible design and keeping or dropping them on honest evidence.

3. **Campbell's law: the corruption of indicators** p. 126

   The more a quantitative indicator is used to make decisions that reward or punish, the more it will be corrupted and the more it will distort the very activity it was meant to monitor.

GIANTS CONNECTED TO THIS CHAPTER



Lee J. Cronbach

A fellow measurement giant who took a more sceptical view of relying mainly on randomised comparisons.

p. 38



Esther Duflo and Abhijit Banerjee

Randomised evaluations carried experimental logic into anti-poverty programmes at scale.

p. 432



Howard White

Theory-based impact evaluation pairs a credible counterfactual with a tested causal chain.

p. 476

KEY WORKS TO READ IN THE ORIGINAL

- **Experimental and Quasi-Experimental Designs for Research** 1963, With Julian C. Stanley
- **Reforms as Experiments** 1969
- **Quasi-Experimentation: Design and Analysis Issues for Field Settings** 1979, With Thomas D. Cook

WORKSHEET · TRY THIS TEMPLATE

Threats-to-validity register

Print this page, or copy the headings into your notes.

1 Claim we want to make

One sentence: the programme caused this change for this group.

2 Rival explanation

Something other than the programme that could produce the same change.

3 How plausible here

High, medium or low, with a reason grounded in this context.

4 What addresses it

Comparison group, baseline, timing, administrative records or qualitative checks.

5 What remains open

Threats you could not rule out, stated plainly for the reader.

REFLECT

What is the most plausible rival explanation for the success your client hopes to report, and have you tested it?

CONNECTS TO

LC **Lee J. Cronbach**
A fellow measurement giant who took a more sceptical view of relying mainly on randomised comparisons. p. 38

D&B **Esther Duflo and Abhijit Banerjee**
Randomised evaluations carried experimental logic into anti-poverty programmes at scale. p. 432

HW **Howard White**
Theory-based impact evaluation pairs a credible counterfactual with a tested causal chain. p. 476

06



Emblem: an original HubForge design for this figure, not a likeness.

Daniel L. Stufflebeam

1936 to 2017 Active 1960s to 2000s

United States; evaluation work and lectures in more than 20 countries

Use and learning · Impact evaluation and evidence systems

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

IN ONE LINE

He made evaluation a companion to every stage of a programme's life, and then held evaluators to standards of their own.

Chicago, 1961. A young music teacher from Iowa is working as a substitute, sent from school to school, more than 40 of them. As he later recalled, he came away believing he might be one of very few people who knew how poorly many of those schools, especially in the inner city, were performing. Four years later, a new federal law poured money into American schools and demanded that it be evaluated. Nobody quite knew how. Daniel Stufflebeam, still in his twenties and directing test development at Ohio State, was chosen as chief evaluation adviser to Nolan Estes, who led the federal work. Out of that collision came CIPP, one of the first evaluation models and still one of the most widely used in the world.

Life CHAPTER 06

Daniel Leroy Stufflebeam was born on 19 September 1936 in Waverly, Iowa, the son of Leroy and Melva Mae Stufflebeam. He went to Wartburg College in his home town, where he wrestled on the college team, and then to the University of Iowa, graduating in music education in 1958. That December he married Carolyn Teresa Joseph in Chicago; they were married for 59 years and had three children. He served in the US Army Reserve from 1960 to 1968, and his first jobs were in schools, as a high school band director and wrestling coach. Few careers in evaluation have started further from a statistics office.

A spell of substitute teaching in Chicago in 1961, which took him into more than 40 schools, proved formative. A fellowship then took him to Purdue University, where he earned a master's degree in 1962 and a PhD in 1964.

In 1963 he joined Ohio State University to lead test development, with the GED high school equivalency tests among his main projects. As federal funding began to

BIG IDEAS

1. **The CIPP model: Context, Input, Process, Product**
2. **Standards for evaluation and metaevaluation**

KEY WORKS

Educational Evaluation and Decision Making · With the Phi Delta Kappa National Study Committee on Evaluation

The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability 2017 · With Guili Zhang

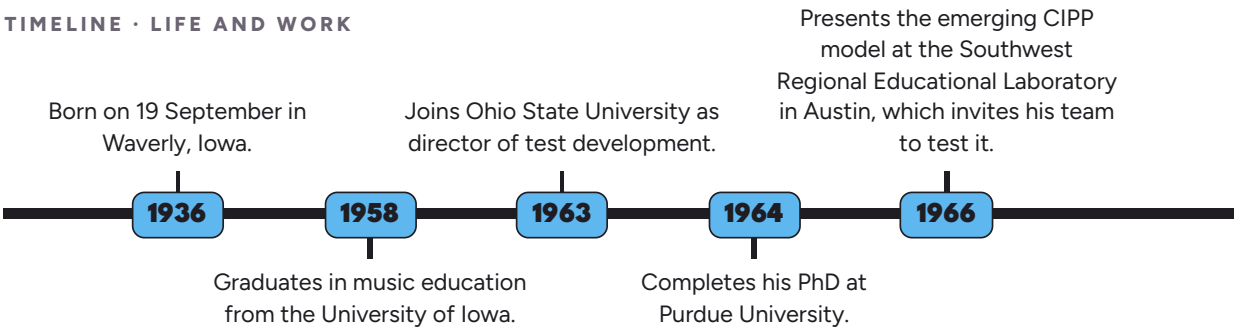
require school districts to submit sound evaluation plans, the head of Ohio State's School of Education, John Ramseyer, saw a duty to help, and Stufflebeam's testing unit grew into the Ohio State Evaluation Center, with Stufflebeam as founding director in the mid-1960s. The centre sat within the Bureau of Educational Research and Service directed by Egon Guba. In 1965 Nolan Estes, who led the federal work on the new Elementary and Secondary Education Act, chose Stufflebeam as his chief evaluation adviser. In 1973 Stufflebeam moved the centre to Western Michigan University in Kalamazoo and led it until 2002, growing it from three staff to about two dozen researchers. He also founded the federally funded Center for Research on Educational Accountability and Teacher Evaluation.

He carried out evaluation work for the US Department of Education, the US Marine Corps and foundations including MacArthur, Kellogg and Alger, worked for the United Nations and the World Bank, and lectured or gave technical assistance in more than 20 countries, among them Australia, the Philippines, Colombia and Spain. He was credited with around 20 books. He retired from Western Michigan University in 2007, later working as a civil servant at the Volpe National Transportation Systems Center, and kept writing: his final book on the CIPP model, with Guili Zhang, appeared at the end of 2016 with a 2017 date. He died of heart failure in Kalamazoo on 23 July 2017, aged 80.

The spark

Stufflebeam himself traced his life's work to those months as a substitute teacher in Chicago in 1961. Moving through more than 40 schools, he later recalled, he concluded that he might be one of very few people who knew how badly many of them were doing, and he wondered how such a situation could be accepted. He linked that experience to his lifelong interest in context evaluation: the careful study of needs and conditions before anyone designs a solution. A few years later, at Ohio State, federal funding rules turned his question into a profession.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 06

In 1965, the Elementary and Secondary Education Act sent a wave of federal money into American schools, and with it a demand that every funded project be evaluated. Picture a school district official receiving that demand. The only tool most people knew was testing students against objectives at the end of a programme. That could not tell the official whether the programme suited local needs, whether its plan was sound, or how it was going while there was still time to fix it. In our reading, Stufflebeam, a test developer suddenly advising on federal evaluation, was perfectly placed to see the gap between what tests offered and what decision-makers needed.

His answer was the CIPP model. In 1966 he presented the emerging framework at the Southwest Regional Educational Laboratory in Austin, Texas, whose director, Edwin Hindsman, invited the Ohio State team to install and test it there; Stufflebeam, Egon Guba and Robert Hammond travelled to Austin each month for the work. The model was refined through a Phi Delta Kappa national study committee on evaluation, which Stufflebeam chaired, and set out in the committee's book Educational Evaluation and Decision Making, published by F. E. Peacock in 1971. Its central move was to treat evaluation as a service to decision-makers at every stage of a programme's life, from deciding what to aim for to deciding whether to continue, rather than as a verdict delivered only at the end.

Did you know?

Before evaluation, Stufflebeam trained in music education and worked as a high school band director and wrestling coach.

1971 — Educational Evaluation and Decision Making, from the Phi Delta Kappa committee he chaired, sets out the CIPP model.

1973 — Moves the Evaluation Center to Western Michigan University.

1981 — The Joint Committee publishes its first standards, Standards for Evaluations of Educational Programs, Projects, and Materials.

1985 — Receives the American Evaluation Association's Paul F. Lazarsfeld Evaluation Theory Award.

2017 — TIMELINE, CONTINUED The CIPP Evaluation Model, with Guili Zhang, is published; he dies on 23 July in Kalamazoo, aged 80.

The problem it solved

Evaluation was largely end-of-programme testing, which gave decision-makers little to work with while there was still time to improve.

Impact WHAT CHANGED

CIPP is among the most durable models in evaluation. Western Michigan University describes it as one of the first evaluation models and still one of the most widely used. Its reach is visible far from Ohio and Michigan. A 2025 systematic review by Indonesian researchers found CIPP widely adopted in the evaluation of educational programmes, especially in Asia between 2019 and 2025, valuing it as comprehensive and flexible while noting the burden of mixed-methods analysis and a shortage of skilled evaluators. A 2026 study in Nepal used it to evaluate a teacher education programme serving students from remote and disadvantaged areas. In our reading, many NGO MEL plans follow the CIPP shape without naming it: needs assessment, design appraisal, implementation monitoring and results evaluation.

His institutional legacy matters as much as the model. The Evaluation Center he founded became a lasting base for evaluation research and training, and its evaluation checklists project, which draws on his guidelines for checklist development, makes practical tools freely available online. The Joint Committee standards set a precedent for professional norms that has been adapted in other countries. He took his approach abroad himself, through work for the United Nations and the World Bank and lectures and technical assistance in more than 20 countries. Michael Quinn Patton, writing a memorial tribute in 2017, singled out the standards as a lasting contribution and linked their attention to context analysis to Stufflebeam's experience as a Chicago substitute teacher.

WHERE PEOPLE DISAGREE

Debates

Stufflebeam's work attracted less public controversy than that of some contemporaries, but its orientation has been questioned. Because CIPP is built to serve decision-makers, evaluators in the participatory and constructivist traditions have, in our reading, seen it as giving programme participants a smaller voice than those who fund and manage programmes. Its comprehensiveness can also be a burden: the 2025 Indonesian review pointed to the complexity of analysing mixed-methods data and a shortage of skilled evaluators. Used mechanically, the four boxes can become a filing system rather than a way of thinking. His 2001 conclusion that only nine of 22 approaches merited continued use was itself a pointed act of metaevaluation.

LIMITS, IN OUR READING

- ☐ In our view, a model built around decision-makers can give less attention to participants' voices than participatory and constructivist approaches do.
- ☐ Applied mechanically, the four boxes can become a checklist rather than a way of thinking about one specific programme.

WHY THIS MATTERS TO YOU

Stufflebeam's idea is useful far beyond evaluation: good evidence should help at every stage of a decision, not just at the end. Before you start anything, study the need. Before you commit, check the plan. While you work, watch how it is going. At the end, judge the results. That four-step habit, CIPP, sits inside almost every NGO and government MEL plan today. His standards give you something just as valuable: a way to defend your own work and to judge whether someone else's evidence deserves trust. Use them to push back on an impossible timeline, an unethical consent process or a report nobody will read.

20+

countries in which he lectured or gave technical assistance

Remembering Daniel L. Stufflebeam, The Evaluation Center, Western Michigan University (2017)

29

years leading the Evaluation Center at Western Michigan University, from 1973 to 2002

Remembering Daniel L. Stufflebeam, The Evaluation Center, Western Michigan University (2017)

13

years as founding chair of the Joint Committee on Standards for Educational Evaluation

Michael Quinn Patton, Journal of MultiDisciplinary Evaluation (2017); SAGE Encyclopedia of Evaluation (2005)

Key concepts

Context evaluation
Assessment of needs, problems, assets and opportunities in a setting, used to decide what a programme should aim to achieve.

Input evaluation
Comparison of alternative strategies, plans and resources, used to choose and justify a programme design.

Process evaluation
Monitoring and documenting how a programme is being implemented, so that it can be adjusted while it runs.

Product evaluation
Identifying and judging a programme's outcomes, intended and unintended, to inform decisions about its future.

Program Evaluation Standards
The Joint Committee's standards for judging the quality of evaluations, grouped under utility, feasibility, propriety and accuracy.

Metaevaluation
The evaluation of an evaluation or of an evaluation approach, for example against professional standards.

Sources for this chapter: p. 577

THE BIG IDEAS OF THIS CHAPTER

1. **The CIPP model: Context, Input, Process, Product** p. 142
   Evaluate a programme's context, inputs, process and products so that evidence serves decisions at every stage of its life, for improvement as well as accountability.

2. **Standards for evaluation and metaevaluation** p. 145
   Evaluations should themselves be judged against shared, public standards of utility, feasibility, propriety and accuracy.

KEY WORKS TO READ IN THE ORIGINAL

- **Educational Evaluation and Decision Making** With the Phi Delta Kappa National Study Committee on Evaluation
- **The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability** 2017, With Guili Zhang

GIANTS CONNECTED TO THIS CHAPTER



Egon G. Guba and Yvonna S. Lincoln

Egon Guba collaborated with Stufflebeam at Ohio State in the 1960s before taking a constructivist turn.

p. 318



Michael Quinn Patton

Both put use at the centre; Patton focuses on named intended users.

p. 204



Ray C. Rist

The Ten Steps took management-oriented M&E into government systems.

p. 410

In one line

He made evaluation a companion to every stage of a programme's life, and then held evaluators to standards of their own.

1

BIG IDEA 1 OF 2 · STUFFLEBEAM

The CIPP model: Context, Input, Process, Product

Evaluate a programme's context, inputs, process and products so that evidence serves decisions at every stage of its life, for improvement as well as accountability.

PICTURE IT

Think of building a house. First you study the site and what the family needs (context). Then you check the plans and the budget (input). While the walls go up, you inspect the work (process). When it is finished, you ask whether it is a good home (product). CIPP asks the same four questions of any programme.

The problem

The Elementary and Secondary Education Act of 1965 suddenly required American school districts to evaluate federally funded projects. The tools to hand were tests of student achievement at the end of a programme, measured against stated objectives. Such tests could not tell administrators whether a project fitted local needs, whether its design was the best available, or how it was going while there was still time to change course. Districts risked producing verdicts too late to help, while money and children's time went into projects nobody could improve.

What it is

CIPP is a comprehensive framework for evaluations that guide and judge programmes, projects, institutions and systems. Context evaluation assesses needs, problems, assets and opportunities in a setting, to help define goals and priorities. Input evaluation identifies and assesses alternative strategies, plans, budgets and staffing, to help choose a sound design. Process evaluation monitors, documents and assesses implementation, to guide adjustments and later explain results. Product evaluation identifies and judges outcomes, intended and unintended, short and long term. In Stufflebeam's 2007 checklist, product evaluation is divided into impact, whether the right beneficiaries were reached; effectiveness, the quality and significance of outcomes; sustainability, whether gains are institutionalised and continued; and transportability, whether the approach could be adapted elsewhere. The four core studies can be used in sequence or singly, either prospectively to guide decisions or retrospectively to account for what happened. Underlying the

model is the view that evaluation's main purpose is improvement, with accountability alongside, and that evaluations should themselves meet professional standards.

How it works

1. **Agree the contract**
   Set out the evaluation's purposes, audiences, questions, budget, reporting arrangements and standards in advance, the first component of Stufflebeam's 2007 CIPP Evaluation Model Checklist.
2. **Context evaluation**
   Assess the needs, assets and problems of the target population and setting, to confirm or sharpen the programme's goals.
3. **Input evaluation**
   Identify and compare alternative strategies and plans for feasibility, cost and likely effect, and help choose or refine one.
4. **Process evaluation**
   Monitor and document implementation, give staff regular feedback and record adaptations, so that problems can be fixed and results later explained.
5. **Product evaluation in four parts**
   Assess impact on the intended beneficiaries, the effectiveness of outcomes, the sustainability of gains and the transportability of the approach to other settings.
6. **Metaevaluation**
   Check the evaluation against professional standards, such as those of the Joint Committee, ideally with an independent reviewer.
7. **Final synthesis report**
   Bring the findings together for decision-makers and other audiences, serving both improvement and accountability.

WORKED EXAMPLE · AN ILLUSTRATION

A state department of school education plans a bridge course for children who fell behind during school closures. A context study uses learning assessment data and district consultations to find which grades and blocks have the largest gaps and how much teacher time is available. An input study compares two options, a summer camp run by volunteers and a daily in-school period taught by regular teachers, on cost, feasibility and evidence of effect; the department chooses the in-school period. Process evaluation uses monthly classroom visits to track whether the period is actually held and how teachers adapt materials, and feeds problems back to block officers. Product evaluation at the end of the year asks whether the children furthest behind were reached, how much they gained, whether schools plan to continue, and whether the approach could work in neighbouring states.

USE IT WHEN

- Designing an MEL system for a new programme, from needs assessment onwards.
- Decision-makers need evidence at several points, not just at the end.
- A programme must serve both improvement and accountability.

SKIP IT WHEN

- You need a quick, narrow answer to a single question.
- The main purpose is to centre participants' own definitions of value, where participatory or constructivist approaches fit better.
- Resources cannot support the studies chosen, and the boxes would be filled with token data.

COMMON MISREADINGS

- [ ] Treating the four parts as a fixed sequence that must always be completed, when they can be used singly or in combination as decisions require.
- [ ] Equating process evaluation with counting activities, when it is meant to assess the quality of implementation and help explain results.

WHY IT WAS GROUNDBREAKING

It turned evaluation from an end-of-programme verdict into a management tool used across a programme's whole life. It was one of the first academic evaluation models to gain wide use and remains one of the most widely applied, from American school districts to education programmes across Asia.

Origin

Stufflebeam presented the emerging model at the Southwest Regional Educational Laboratory in Austin in 1966, where it was field-tested with Egon Guba and Robert Hammond. It was set out in Educational Evaluation and Decision Making (Phi Delta Kappa committee, F. E. Peacock, 1971), extended with four product subparts in his CIPP checklist for the Evaluation Center, including a 2007 version, and given its fullest statement in The CIPP Evaluation Model, with Guili Zhang (Guilford, 2017).

RELATED IDEAS

Standards for evaluation and metaevaluation

- G&L Guba and Lincoln 318
- MP Patton 204
- RR Rist 410

In one line

Evaluations should themselves be judged against shared, public standards of utility, feasibility, propriety and accuracy.

2

BIG IDEA 2 OF 2 · STUFFLEBEAM

Standards for evaluation and metaevaluation

Evaluations should themselves be judged against shared, public standards of utility, feasibility, propriety and accuracy.

PICTURE IT

Think of food safety inspectors. Before you trust a restaurant's own claim that its kitchen is clean, you want someone to check it against public standards everyone agreed in advance. Metaevaluation does that for evaluations, judging them against shared standards of usefulness, practicality, fairness and accuracy.

The problem

In the 1970s, evaluation was spreading fast but had no agreed definition of good practice. Commissioners could not easily tell a sound study from a weak one, evaluators worked to private norms, and ethical lapses, wasted budgets and unusable reports had no shared yardstick against which they could be named. A young field that judged other people's programmes had no public way of being judged itself, which threatened its credibility with the governments, schools and communities that relied on it.

What it is

The Program Evaluation Standards are principles, developed by the Joint Committee on Standards for Educational Evaluation, for designing and judging evaluations. Stufflebeam chaired the committee for its first 13 years and was principal author of its original standards. They are grouped under four attributes. Utility standards ask whether an evaluation serves the information needs of its intended users. Feasibility standards ask whether it is realistic, prudent and cost-effective. Propriety standards ask whether it is conducted legally and ethically, with due regard for the welfare of those involved and affected. Accuracy standards ask whether it produces technically sound, dependable information. The committee was deliberately mixed, eventually 17 members split equally between evaluation specialists and practitioners, so that the standards reflected the needs of users as well as experts. Metaevaluation applies such standards to evaluation plans, reports or whole approaches. Stufflebeam himself rated 22 evaluation approaches against them in 2001 and concluded that only nine merited continued use.

How it works

1. **Choose the standards at inception**
   Agree with the commissioner that the evaluation will be designed and judged against the Program Evaluation Standards or a national adaptation of them.
2. **Design to the four attributes**
   For each design choice, ask whether it serves users, is practical and affordable, is ethical and fair to those affected, and will produce sound information.
3. **Build in formative metaevaluation**
   Review the plan and early instruments against the standards, preferably with an independent reviewer, and fix weaknesses before fieldwork.
4. **Document as you go**
   Keep records of decisions, data sources and ethical safeguards so that the evaluation can later be examined.
5. **Conduct summative metaevaluation**
   At the end, rate the evaluation against the standards and report its strengths and weaknesses alongside the findings, as the metaevaluation component of Stufflebeam's 2007 CIPP checklist provides.
6. **Use the standards to choose approaches**
   When selecting a model or a firm, compare options against the standards, as Stufflebeam did for 22 approaches in 2001.

WORKED EXAMPLE · AN ILLUSTRATION

A foundation in Nairobi commissions an evaluation of a youth employment programme in three Kenyan counties. Before fieldwork, its MEL officer asks an independent reviewer to rate the inception report against the four attributes. On utility, the questions match the board's scale-up decision, but county governments, also key users, have not been consulted. On feasibility, a survey of 3,000 households is planned within six weeks, during planting season. On propriety, the consent form exists only in English. On accuracy, the comparison group comes from a county with a much stronger job market. The reviewer's short memo leads to county consultations, a smaller survey spread over ten weeks, consent forms in Swahili and local languages, and a better matched comparison. The final report includes a one-page metaevaluation so the board knows how far to trust it.

CONCEPT HIGHLIGHT

Context evaluation

Assessment of needs, problems, assets and opportunities in a setting, used to decide what a programme should aim to achieve.

USE IT WHEN

- Reviewing inception or final reports for quality.
- Writing terms of reference that set clear expectations for evaluators.
- Choosing between evaluation approaches, teams or firms.

SKIP IT WHEN

- Very light internal reviews, where a full metaevaluation would cost more than the evaluation itself.
- Settings that have their own adapted national or regional standards, which should be used in place of the original American set.

COMMON MISREADINGS

- Treating the standards as a compliance list ticked at the end, rather than design guidance used from the start.
- Attending only to accuracy, when utility, feasibility and propriety are equally part of a good evaluation.

WHY IT WAS GROUNDBREAKING

It gave a young field public, shared criteria for good practice, an essential step towards becoming a profession. It made evaluators accountable for usefulness and ethics, not only technical accuracy, and the standards have been adapted for use in other countries.

Origin

The Joint Committee, chaired by Stufflebeam for its first 13 years, published Standards for Evaluations of Educational Programs, Projects, and Materials in 1981 and revised them in 1994 as The Program Evaluation Standards; a third edition followed in 2011. His comparative metaevaluation of 22 approaches appeared in Evaluation Models (New Directions for Evaluation 89, 2001), and metaevaluation became a standard component of his CIPP checklists.

RELATED IDEAS

The CIPP model: Context, Input, Process, Product

MS Scriven 60 MP Patton 204 DM Mertens 384

CONCEPT HIGHLIGHT

Input evaluation

Comparison of alternative strategies, plans and resources, used to choose and justify a programme design.

Put it to work

USE IT ON MONDAY · THE CIPP MODEL: CONTEXT, INPUT, PROCESS, PRODUCT

1. Sort every question in the terms of reference into the four CIPP boxes. Empty boxes show you what the client has not thought to ask.

2. For a new programme, propose context and input evaluation before launch: a needs check and a comparison of plan options, not just a baseline.

3. Agree two or three decision points with the client when process findings will be reviewed, so evaluation feeds decisions during the programme.

AI prompt **Copy, fill the brackets, then check every line**

Using Daniel Stufflebeam's CIPP model (Context, Input, Process, Product), help me structure an evaluation plan for **[programme]** in **[location]**, currently at **[stage: design, early implementation, mid-term or end]**. Here are the terms of reference questions: **[paste questions]**. Here is what I know about the programme: **[paste summary]**. (1) Sort each question into Context, Input, Process or Product, and show which boxes are empty. (2) For each box, suggest two evaluation questions, likely data sources and the decision the findings would inform. (3) Recommend which boxes matter most at this stage, and why. Mark every assumption you make about the programme as ASSUMPTION.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of building a house. First you study the site and what the family needs (context). Then you check the plans and the budget (input). While the walls go up, you inspect the work (process). When it is finished, you ask whether it is a good home (product). CIPP asks the same four questions of any programme.

YOUR MOVES Three moves from Stufflebeam's path

1 Let ground-level work give you the question

Stufflebeam began as a band director and wrestling coach, and classroom work in Chicago shaped his whole agenda.

TRY THIS
Spend a day in the field this month, watching the programme as a participant would, and write the question it raises.

2 Move fast when policy opens a gap

When the 1965 federal demand for evaluation arrived, schools lacked the capacity to meet it. He built the CIPP model for exactly that gap.

TRY THIS
Track one new policy, funder rule or reporting requirement in your sector, and draft the tool people will need to meet it.

3 Build what others can build on

He invested in an evaluation centre, a standards committee and a free checklists project. Institutions multiplied one person's ideas.

TRY THIS
Turn one thing you do well into a free checklist or template, and share it publicly.

THE BIG IDEAS OF THIS CHAPTER

1. **The CIPP model: Context, Input, Process, Product** p. 142
   Evaluate a programme's context, inputs, process and products so that evidence serves decisions at every stage of its life, for improvement as well as accountability.
2. **Standards for evaluation and metaevaluation** p. 145
   Evaluations should themselves be judged against shared, public standards of utility, feasibility, propriety and accuracy.

CONCEPT HIGHLIGHT

Metaevaluation

The evaluation of an evaluation or of an evaluation approach, for example against professional standards.

GIANTS CONNECTED TO THIS CHAPTER



Egon G. Guba and Yvonna S. Lincoln

Egon Guba collaborated with Stufflebeam at Ohio State in the 1960s before taking a constructivist turn.

p. 318



Michael Quinn Patton

Both put use at the centre; Patton focuses on named intended users.

p. 204



Ray C. Rist

The Ten Steps took management-oriented M&E into government systems.

p. 410

KEY WORKS TO READ IN THE ORIGINAL

- **Educational Evaluation and Decision Making** With the Phi Delta Kappa National Study Committee on Evaluation
- **The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability** 2017, With Guili Zhang

CONCEPT HIGHLIGHT

Context evaluation

Assessment of needs, problems, assets and opportunities in a setting, used to decide what a programme should aim to achieve.

CONCEPT HIGHLIGHT

Input evaluation

Comparison of alternative strategies, plans and resources, used to choose and justify a programme design.

WORKSHEET · TRY THIS TEMPLATE

CIPP planning grid

Print this page, or copy the headings into your notes.

1 Context

Which needs, assets and problems justify this programme here and now?

2 Input

Which strategies, plans and budgets were considered, and why was this one chosen?

3 Process

Is it being delivered as planned? What is being adapted, and why?

4 Product

What results, intended and unintended, for whom?

5 Decision informed

For each box: who decides what, and by when the findings are needed.

REFLECT

Which decision will your next evaluation actually inform, and will the findings arrive before that decision is made?

CONNECTS TO

G&L **Egon G. Guba and Yvonna S. Lincoln** Egon Guba collaborated with Stufflebeam at Ohio State in the 1960s before taking a constructivist turn. p. 318

MP **Michael Quinn Patton** Both put use at the centre; Patton focuses on named intended users. p. 204

RR **Ray C. Rist** The Ten Steps took management-oriented M&E into government systems. p. 410

07



Emblem: an original HubForge design for this figure, not a likeness.

Carol H. Weiss

Died 2013 Active 1960s to 1990s

United States; consulted in Africa, Asia, Australia and Europe

Theory and programme logic · Use and learning

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

IN ONE LINE

She taught evaluators to ask not only whether a programme works, but what it assumes and where its evidence will travel.

If you have ever drawn boxes and arrows to show how a project is supposed to create change, you have used an idea that Carol Weiss helped popularise. In 1972, when programme evaluation was still a young craft, she published one of its first comprehensive textbooks. It sold hundreds of thousands of copies and was translated into Spanish, German and Thai. In 1979 she took on a fear that haunted her colleagues, that their reports were simply ignored, and showed that research is used in at least seven different ways, some of them slow and almost invisible. And in 1995 she urged evaluators to trace the small steps by which a programme is meant to work. Today, foundations, governments and development agencies commonly ask for exactly that: a theory of change.

Life CHAPTER 07

Carol Hirschon Weiss was born and raised in New York. She studied government at Cornell University and then took a master's degree and a doctorate in sociology at Columbia University. Harvard's tribute dates her doctorate to 1977, by which time her best-known textbook had already been in print for five years. At Columbia she worked at the Bureau of Applied Social Research, and it was there that she began to study a question that would occupy her for the rest of her life: how is social research actually used in public policy?

Her path into evaluation began in the 1960s, when she evaluated a domestic Peace Corps programme in central Harlem, at a time when new federal social and education programmes were creating demand for evaluation. In 1972 she published Evaluation Research: Methods of Assessing Program Effectiveness, one of the first comprehensive textbooks in the field. It sold hundreds of thousands of copies and was translated

BIG IDEAS

1. **Theory-based evaluation**
2. **The many meanings of research use**
3. **The politics of evaluation**

KEY WORKS

Evaluation Research: Methods of Assessing Program Effectiveness
1972

The Many Meanings of Research Utilization
1979

Evaluation: Methods for Studying Programs and Policies
1998

DID YOU KNOW?

After retiring from Harvard, Weiss became a docent at the Boston Museum of Science.

Her 1972 textbook Evaluation Research sold hundreds of thousands of copies and was translated into Spanish, German and Thai.

into Spanish, German and Thai, and it already discussed, with a diagram, the idea of evaluating the theory behind a programme.

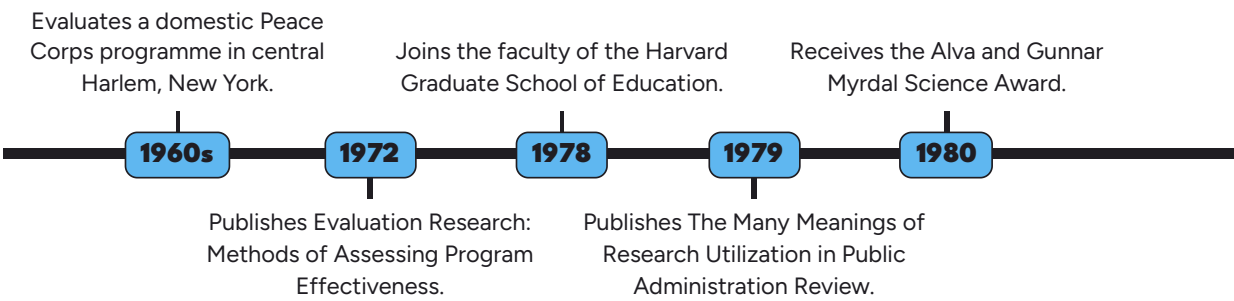
In 1978 she joined the Harvard Graduate School of Education, where she spent the rest of her career, and in 1999 she was named the Beatrice B. Whiting Professor of Education. She held fellowships at the Center for Advanced Study in the Behavioral Sciences and the Brookings Institution, served as a senior fellow at the US Department of Education and as a congressional fellow, and sat on more than ten National Academies committees. Her consulting clients included the US General Accounting Office, the World Bank, the Rockefeller Foundation and Canada's International Development Research Centre, and she advised both US and international agencies and governments. In 1980 she received the Alva and Gunnar Myrdal Science Award. In all she published 11 books and more than 100 articles on evaluation, the use of research in policymaking, and the influence of ideology, interests and information on institutions. Her doctoral advisees included Sharon Rallis, later president of the American Evaluation Association.

Weiss retired from Harvard in the mid-2000s; Harvard gives the year as 2005 and her obituary as 2006. In retirement she became a docent at the Boston Museum of Science. She was married to Malcolm A. Weiss for 64 years, and they had three children. She died on 8 January 2013, aged 86, of complications of heart disease and cancer.

The spark

One question held Carol Weiss for her whole career: what actually happens to research after it is delivered? She began studying it at Columbia University's Bureau of Applied Social Research. Her first big evaluation, of a domestic Peace Corps programme in central Harlem in the 1960s, came just as new federal social programmes were creating a hunger for evaluators. Again and again, her work shows a sociologist who refused to stop at the final report, and kept asking where the evidence went next.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 07

Imagine spending a year on an evaluation, delivering a careful report, and then watching nothing change. By the 1970s many American evaluators felt exactly that. Decision-makers rarely did what reports recommended, and many concluded that research was simply ignored. A second puzzle made things worse. As programmes grew more ambitious, especially community initiatives working across health, housing, schools and jobs at once, evaluations could say whether outcomes had changed but not why. Programme designers were left holding a verdict with nothing to learn from it.

Weiss addressed the first puzzle in The Many Meanings of Research Utilization, published in Public Administration Review in 1979. Drawing on her studies of how officials actually encounter research, she set out seven distinct ways in which research can be used, and argued that one of the most important was also the slowest: the gradual enlightenment of policy debate. She addressed the second in 1995, in a chapter written through her work on the evaluation steering committee of the Aspen Institute's roundtable on community initiatives, published in New Approaches to Evaluating Community Initiatives. Its title, Nothing as Practical as Good Theory, echoed a maxim usually attributed to the psychologist Kurt Lewin. She argued that complex programmes were hard to evaluate because their underlying assumptions were poorly articulated, and proposed that evaluators trace the mini-steps linking activities to long-term goals.

Did you know?

After retiring from Harvard, Weiss became a docent at the Boston Museum of Science.

TIMELINE, CONTINUED

1995 — Publishes Nothing as Practical as Good Theory in the Aspen Institute volume New Approaches to Evaluating Community Initiatives.

1998 — Publishes Evaluation: Methods for Studying Programs and Policies, the second edition of her textbook.

1999 — Named Beatrice B. Whiting Professor of Education at Harvard.

2013 — Dies on 8 January, aged 86.

The problem it solved

Evaluations of complex community programmes could say whether outcomes occurred but not why.

Impact WHAT CHANGED

Weiss is one of the few evaluation theorists whose ideas have become everyday management practice. The Center for Theory of Change credits her with popularising the term theory of change through the Aspen Institute work, and an AEA365 retrospective notes that foundations, government agencies, nonprofits and international development organisations now commonly require a theory of change in funding proposals. The same retrospective counts her 1995 chapter among the most influential works in the history of programme evaluation. In our reading, the later family of theory-driven approaches, which seek to open the black box between activities and outcomes, builds directly on the ground she prepared.

Her work on use proved equally durable. By separating problem-solving, political, tactical and enlightenment uses, she gave evaluators and research funders a way to describe influence that does not depend on a single decision changing overnight, and her 1998 American Journal of Evaluation article revisited what the field had learned about use. Kathryn Newcomer chose Evaluation Research for The Oxford Handbook of Classics in Public Policy and Administration in 2015, and its translations into Spanish, German and Thai carried her approach well beyond the United States. Her advisory work for the World Bank and Canada's International Development Research Centre brought her into development evaluation directly. In our view, every MEL plan that includes a learning and dissemination strategy, rather than ending with the final report, reflects her lesson that use has to be planned.

KEY WORKS TO READ IN THE ORIGINAL

- **Evaluation Research: Methods of Assessing Program Effectiveness** 1972
- **The Many Meanings of Research Utilization** 1979
- **Evaluation: Methods for Studying Programs and Policies** 1998

WHERE PEOPLE DISAGREE

Debates

Weiss was candid about the limits of her own approach. In 1998 she noted that theory-based evaluation is hard to apply where programmes have no explicit or implicit theory, are vague in form, or change substantially over time, all common in community initiatives. Her account of enlightenment also had a darker side that she herself identified: diffuse influence can spread weak generalisations, and advocates of almost any position can find research to support them. In our view, theories of change can slide into compliance documents drawn for funders and never tested, a risk she did not create but which her popularity has amplified. Evaluators who prioritise direct use by named users, a position associated with Michael Quinn Patton, place more weight on designing for immediate decisions.

LIMITS, IN OUR READING

- [ ] Weiss herself noted, in a 1998 interview, that theory-based evaluation is hard where programmes lack explicit theories or change over time.
- [ ] In our view, theory of change work slides into box-ticking when the diagram is drawn for the funder and never tested.

WHY THIS MATTERS TO YOU

Weiss's lessons work for anyone trying to make change happen. First, write down how you think your plan will work, step by step, and test the weakest step before you bet everything on it. Second, do not judge your influence only by whether someone acted on your advice the next morning; evidence often changes how people think slowly and indirectly, long before it changes any single decision. Third, accept that every decision about evidence sits inside politics and interests, and plan for that instead of pretending it is not there. If you have ever felt your work vanish into a drawer, she offers both comfort and a strategy.

In one line

She taught evaluators to ask not only whether a programme works, but what it assumes and where its evidence will travel.

11

books, alongside more than 100 articles on evaluation and the use of research in policy

HGSE remembers Professor Carol Weiss, Harvard Graduate School of Education (2013)

3

languages, Spanish, German and Thai, into which her 1972 textbook was translated, selling hundreds of thousands of copies

HGSE remembers Professor Carol Weiss, Harvard Graduate School of Education (2013)

7

models of research use set out in The Many Meanings of Research Utilization (1979)

The Many Meanings of Research Utilization, Public Administration Review (1979), as summarised by AcaWiki

Key concepts

Theory-based evaluation

Evaluation that makes explicit the assumptions linking a programme's activities to its outcomes and tests the intermediate steps, not only the final result.

Theory of change

The set of assumptions about the mini-steps through which a programme is expected to bring about its long-term goals; Weiss is credited with popularising the term.

Enlightenment use

The indirect influence of research, as its ideas and perspectives gradually shape how people understand a problem.

Tactical use

Use of research for purposes unrelated to its content, such as showing that action is under way, deflecting criticism or enhancing an agency's prestige.

Political model of use

Research used as ammunition by people whose positions on an issue are already fixed.

Sources for this chapter: p. 578

THE BIG IDEAS OF THIS CHAPTER

1. **Theory-based evaluation** p. 160

   Make explicit the chain of assumptions by which a programme is supposed to produce change, then test each link, so that the evaluation can explain why results did or did not appear.

2. **The many meanings of research use** p. 163

   Research and evaluation influence policy in at least seven different ways, and the slow enlightenment of how people think about a problem can matter more than any single decision.

3. **The politics of evaluation** p. 166

   Evaluation is political from start to finish, so evaluators must understand the interests around a programme and protect the conditions for credible work.

GIANTS CONNECTED TO THIS CHAPTER



Huey-tsyh Chen

Theory-driven evaluation gave the idea a detailed structure of change and action models.

p. 174



Patricia J. Rogers

Purposeful program theory asks what a theory of change is for before drawing it.

p. 268



Michael Quinn Patton

Both studied why findings go unused; Patton designed the whole evaluation around use.

p. 204

1

BIG IDEA 1 OF 3 · WEISS

Theory-based evaluation

Make explicit the chain of assumptions by which a programme is supposed to produce change, then test each link, so that the evaluation can explain why results did or did not appear.

PICTURE IT

Think of a row of dominoes. If the last one does not fall, you want to know exactly which domino failed to knock over the next. Theory-based evaluation lines up a programme's dominoes, its chain of assumptions, and checks each one.

The problem

By the early 1990s, American foundations were backing comprehensive community initiatives that worked on housing, jobs, schools and health at once. Conventional evaluation could report whether neighbourhood outcomes moved, but such programmes rarely allowed clean comparison groups, and a bare verdict told designers nothing about what to change. When results disappointed, nobody could say whether the idea was wrong or had simply been badly delivered. Weiss saw that a deeper difficulty was that the programmes' own assumptions about how change would happen had never been spelled out.

What it is

Theory-based evaluation starts from the premise that every programme embodies a theory, stated or implicit, about how its activities will lead to its goals. The evaluator works with staff and other stakeholders to articulate that theory as a sequence of mini-steps, each resting on assumptions that must hold for the next step to follow, and then collects evidence at each link. Weiss separated two parts of the theory: implementation theory, about whether activities are carried out as planned, and programme theory, about the mechanisms through which those activities change people's knowledge, behaviour or circumstances. The distinction lets the evaluator tell implementation failure, where the programme was not delivered as intended, from theory failure, where it was delivered but the expected mechanism did not operate. Where stakeholders hold competing theories, the evaluation can follow more than one. Weiss also held that articulating the theory is valuable in itself, because it pushes a programme's people to agree on what they are doing and why.

How it works

1. **Surface the theories**
   Interview planners, staff and participants and read programme documents to learn how each believes the activities will lead to the long-term goals, as Weiss proposed in her 1995 Aspen Institute chapter.
2. **Lay out the mini-steps**
   Write the chain of intermediate changes from activities to outcomes, separating implementation steps from the mechanisms of change in participants.
3. **Name the assumptions**
   For each link, state what must be true for the next step to follow, such as people attending, understanding, trusting the advice and being able to act on it.
4. **Choose the links to test**
   Prioritise the steps that are most central to the theory or least supported by existing evidence, since few studies can measure everything.
5. **Collect evidence along the chain**
   Gather data on each chosen link over time, using whatever methods fit, so that progress or breakdown can be located.
6. **Locate where the chain held or broke**
   Distinguish implementation failure from theory failure and report which assumptions held, so that designers know what to change.
7. **Feed it back**
   Share the articulated theory and emerging findings with staff while the programme runs, since making the theory explicit is itself useful to them.

WORKED EXAMPLE · AN ILLUSTRATION

A state government in India trains community health workers to counsel mothers on infant feeding, expecting fewer underweight children within two years. An evaluator following Weiss meets officials, trainers and workers and maps the mini-steps: workers attend training; they learn the messages; they visit homes regularly; mothers understand and trust the advice; families can afford the recommended foods; feeding practices change; weight improves.

Data collection follows each step. Attendance and knowledge scores are high and visits happen, but mothers in poorer hamlets say they cannot often buy eggs or pulses, and feeding changes little there. The evaluation can now say that counselling was delivered and understood, but that one assumption, affordability, broke. The programme links families to the supplementary nutrition scheme instead of retraining workers.

USE IT WHEN

- Programmes are complex or multi-component and final outcomes take years to appear.
- Designers need to know where to adjust, not just whether to continue.
- A rigorous comparison group is not available, but the causal chain can be traced step by step.

SKIP IT WHEN

- The programme has no coherent theory, is vague in form or changes substantially over time, limits Weiss herself noted in 1998.
- The only question is whether a well-understood intervention reached its targets.
- A theory of change would be drawn only to satisfy a funder, with no plan or budget to test it.

COMMON MISREADINGS

- [ ] Treating the theory of change diagram as the deliverable, when the point is to collect evidence on each link.
- [ ] Mapping only implementation steps, such as trainings held, and never testing the mechanisms that are supposed to change people's behaviour.

WHY IT WAS GROUNDBREAKING

It moved evaluation from asking only whether a programme worked to asking how and why, which is what designers need in order to improve it. It laid the foundations for the theory of change, now a routine requirement of foundations, government agencies and international development organisations.

Origin

Weiss's 1972 textbook Evaluation Research already discussed, with a diagram, evaluating the theory behind a programme. The fuller statement came in Nothing as Practical as Good Theory, her 1995 chapter in the Aspen Institute volume New Approaches to Evaluating Community Initiatives, which popularised the term theory of change.

She reviewed the approach's past, present and future in 1997 and built it into Evaluation: Methods for Studying Programs and Policies, the 1998 second edition of her textbook.

RELATED IDEAS

HC Chen 174 · PR Rogers 268 · HW White 476 · SS Sridharan 246

Did you know?

Her 1972 textbook Evaluation Research sold hundreds of thousands of copies and was translated into Spanish, German and Thai.

2

BIG IDEA 2 OF 3 · WEISS

The many meanings of research use

Research and evaluation influence policy in at least seven different ways, and the slow enlightenment of how people think about a problem can matter more than any single decision.

PICTURE IT

Think of how attitudes to seatbelts or smoking changed. No single study made people change overnight. Evidence slowly seeped into newspapers, conversations and laws until thinking shifted. Weiss called that slow seeping the enlightenment use of research, and showed it often matters most.

The problem

By the late 1970s, years of federal investment in social research and evaluation in the United States had left a widespread sense of disappointment. Decision-makers rarely did what reports recommended, and many researchers concluded that their work was simply ignored. Funders wondered whether evaluation was worth paying for. Part of the trouble, Weiss argued, was a narrow picture of use: if only direct, immediate adoption of findings counted, most real influence would be invisible, and evaluators would keep designing their work for a kind of use that rarely happens.

What it is

In The Many Meanings of Research Utilization (1979), Weiss described seven models of how research enters policy. In the knowledge-driven model, basic research leads to applied research and then to practical application. In the problem-solving model, a study is commissioned to settle a specific decision. In the interactive model, research is one input among many in a search for solutions involving officials, practitioners and others. In the political model, research becomes ammunition for people whose positions are already fixed. In the tactical model, the fact that research is under way is used, to delay action, deflect criticism or show that something is being done, regardless of its content. In the enlightenment model, concepts and generalisations from many studies seep into the thinking of officials, journalists and informed publics over years, changing how problems are framed. Finally, research is part of the intellectual enterprise of society, shaped by and shaping wider currents of

thought. Weiss treated the problem-solving model as the common expectation but an optimistic one, and enlightenment as one of the most important routes.

How it works

1. **Map the policy arena**
   Before design, identify who will encounter the findings, what positions they already hold and which decisions, if any, are pending.
2. **Anticipate the likely kinds of use**
   Judge whether problem-solving, interactive, political, tactical or enlightenment use is most plausible, and be candid about any tactical purposes a commissioner may have.
3. **Design for direct use where decisions exist**
   Time findings to known decision points and shape questions around them.
4. **Plan for enlightenment**
   Build ways for ideas to travel over years, such as short briefs, media pieces, convenings and syntheses with other studies; Weiss noted in a 1998 interview that policymakers often learn of research through newspapers and magazines.
5. **Track influence broadly**
   When reviewing what an evaluation achieved, look for changes in language, framing and agendas, not only for adopted recommendations.
6. **Guard against misuse**
   Watch for selective citation and the spread of weak generalisations, a darker side of diffuse influence that Weiss herself identified.

WORKED EXAMPLE · AN ILLUSTRATION

An evaluation of a state's mid-day meal kitchens finds that centralised kitchens cut costs but reduce local hiring of cooks from marginalised communities. The next budget ignores the report, and the team feels it has failed. Using Weiss's models, they look again. The education department used the study tactically, citing it in the legislature to show oversight. A civil society alliance used it politically, as ammunition for local hiring.

Over four years, the idea that kitchen design affects livelihoods as well as nutrition turns up in newspaper columns, a legislative committee report and a review of national guidelines: enlightenment. For its next evaluation, the team writes a communication plan for each route, with briefings timed to the budget cycle and short pieces for journalists.

USE IT WHEN

- Planning dissemination for an evaluation with policy audiences.
- Assessing the influence of a body of evaluations over several years.
- A commissioner's reasons for an evaluation seem partly tactical or political.

SKIP IT WHEN

- A single, well-defined operational decision is all that matters and the user is known; a utilisation-focused design may fit better.
- You need to attribute a specific policy change precisely to one study, which the models do not do.

COMMON MISREADINGS

- In our reading, using enlightenment as an excuse for findings nobody reads; it is a route to plan for, not a consolation prize.
- Treating political and tactical use as always illegitimate, when Weiss described them as normal features of policymaking that evaluators need to understand.

WHY IT WAS GROUNDBREAKING

It explained why evaluators so often felt ignored, and showed that apparent non-use could conceal real, slow influence. It made the pathways and politics of use something to study and design for, rather than an afterthought.

Origin

The seven models appeared in The Many Meanings of Research Utilization, Public Administration Review, 1979, drawing on Weiss's studies of how officials actually encounter research. She developed the analysis in later work on research and policy, received the Alva and Gunnar Myrdal Science Award in 1980, and revisited what the field had learned about the use of evaluation in a 1998 article in the American Journal of Evaluation.

RELATED IDEAS

MP Patton 204 · HP Preskill 226 · RR Rist 410 · The politics of evaluation

CONCEPT HIGHLIGHT

Theory-based evaluation

Evaluation that makes explicit the assumptions linking a programme's activities to its outcomes and tests the intermediate steps, not only the final result.

3

BIG IDEA 3 OF 3 · WEISS

The politics of evaluation

Evaluation is political from start to finish, so evaluators must understand the interests around a programme and protect the conditions for credible work.

PICTURE IT

Think of a referee in a heated final. The match is full of passionate interests, and the referee's authority depends on being seen as fair. Weiss saw evaluators in the same position: always inside politics, so they must guard the conditions that make their judgements credible.

The problem

Early evaluators often imagined themselves as neutral technicians who measured effects and let the facts speak. In practice, programmes were created by political bargains, evaluations were commissioned by people with interests at stake, and findings arrived in debates shaped by ideology and budgets.

Evaluators who ignored this were surprised when reports were shelved, distorted or attacked, and some accepted assignments that could never produce credible answers, such as a verdict on a flagship scheme with no time, no data and no comparison.

What it is

Weiss argued that evaluation is never a purely technical exercise, and that politics enters it in several ways. The programmes being evaluated are themselves products of political decisions, proposed, funded and defended by coalitions with a stake in their survival. Evaluations are commissioned by people with interests, which shape the questions asked and the time and money available. Findings then land in an arena where they compete with ideology, interests and other information for the attention of decision-makers. And evaluators, by choosing questions, criteria and comparisons, take positions whether they intend to or not, so they need to be aware of their own political role. From this she drew practical lessons. Evaluators should understand the interests around a programme, expect findings to inform rather than settle decisions, communicate in forms that reach real audiences, including through the press, and decline studies when the conditions for credible work, such as adequate time, data, comparisons or funding, are missing. In our reading, her stance is realistic rather than cynical.

How it works

1. **Map stakeholders and stakes**
   Identify who created, funds, runs and opposes the programme, and what each stands to gain or lose from different findings.
2. **Test the conditions for credibility**
   Check whether there is enough time, data, comparison and funding for a credible study; in her 1998 interview Weiss advised declining studies when these are missing.
3. **Negotiate scope and independence**
   Where conditions are weak, renegotiate questions, timing or methods, and agree in writing who controls and publishes the report.
4. **Make your own choices visible**
   State which questions, criteria and comparisons you chose and why, recognising that these choices carry values.
5. **Communicate for the arena**
   Prepare findings in forms that meet journalistic standards and reach officials through the channels they actually use, not only in technical reports.
6. **Prepare for contest**
   Anticipate how different groups may use or attack the findings, and make sure the evidence for each key claim is clear and defensible.

WORKED EXAMPLE · AN ILLUSTRATION

Six months before a state election, a department asks an evaluation firm for an impact assessment of its flagship skilling scheme, to be delivered in eight weeks. There is no baseline and no comparison group. Following Weiss, the evaluation lead maps the stakes: the department wants a positive headline, opposition parties want a scandal, and trainees want jobs. She judges that a credible impact verdict is impossible in the time available. Rather than simply walking away, she proposes a narrower process study of placement records and trainees' experiences, with a plain statement of what it can and cannot show, and an impact design for after the election. The contract records that the full report will be published, and the team prepares a two-page brief for journalists.

USE IT WHEN

- Evaluating high-profile or contested programmes.
- Commissioners have strong stakes in particular findings.
- The timing coincides with budget, election or scale-up decisions.

SKIP IT WHEN

- Small internal learning reviews where the stakes are low and a light touch is enough, though awareness still helps.
- Whenever political analysis is becoming a reason to avoid uncomfortable findings rather than to protect them.

COMMON MISREADINGS

- [ ] Concluding that because evaluation is political, rigour matters less; in our reading, rigour is what gives findings a chance in a contested arena.
- [ ] Treating political awareness as licence to tailor findings to the client, the opposite of her advice to decline studies that cannot be done credibly.

WHY IT WAS GROUNDBREAKING

It punctured the image of the evaluator as a neutral technician standing outside politics. It gave the field a more realistic basis for designing studies that can survive, and matter, in real policy settings.

Origin

Weiss's concern with politics runs through her work from the 1970s, including her early essay Where Politics and Evaluation Research Meet (1973), her 1979 analysis of political and tactical use, and her later studies of how ideology, interests and information shape institutions. In a 1998 interview with the Harvard Family Research Project she distilled practical lessons about communicating through the press and declining compromised studies.

RELATED IDEAS

MP Patton 204 · MS Segone 516 · SZ Zaveri 536 · RR Rist 410

RELATED GIANTS



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174



Patricia J. Rogers

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

p. 268



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476



Sanjeev Sridharan

The evaluator of complex health programmes who argues that designs should fit the programme, and that evidence lives in ecologies, not silos.

p. 246

CONCEPT HIGHLIGHT

Theory of change

The set of assumptions about the mini-steps through which a programme is expected to bring about its long-term goals; Weiss is credited with popularising the term.

CONCEPT HIGHLIGHT

Enlightenment use

The indirect influence of research, as its ideas and perspectives gradually shape how people understand a problem.

CONCEPT HIGHLIGHT

Tactical use

Use of research for purposes unrelated to its content, such as showing that action is under way, deflecting criticism or enhancing an agency's prestige.

CONCEPT HIGHLIGHT

Political model of use

Research used as ammunition by people whose positions on an issue are already fixed.

In one line

Make explicit the chain of assumptions by which a programme is supposed to produce change, then test each link, so that the evaluation can explain why results did or did not appear.

Put it to work

USE IT ON MONDAY · THEORY-BASED EVALUATION

1. In your inception workshop, ask the team to describe each step between activity and outcome, then ask what has to be true for each step to happen. Write those assumptions down.

2. Choose the two or three most important or shakiest assumptions and design data collection to test them, rather than measuring only the final outcome.

3. Plan for slow use as well as quick use: share findings through briefings and short notes over months, not only in a final report.

AI prompt **Copy, fill the brackets, then check every line**

Help me apply theory-based evaluation, in the tradition of Carol Weiss, to **[programme]** in **[location]**. Here is the programme's theory of change or description: **[paste]**. (1) Lay out the chain of steps from activities to long-term outcomes as the document describes it. (2) For each step, list the assumptions that must hold, labelling which are stated in the document and which are implied. (3) Identify the three assumptions that look most important or least supported, and suggest how to test each with data we could realistically collect. (4) Suggest how findings could be shared over time with **[audience]**.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a row of dominoes. If the last one does not fall, you want to know exactly which domino failed to knock over the next. Theory-based evaluation lines up a programme's dominoes, its chain of assumptions, and checks each one.

YOUR MOVES

Three moves from Weiss's path

1 Plan for use from day one

Weiss studied how evidence is actually used, and showed that findings often shape thinking slowly rather than decisions directly.

TRY THIS

In your next inception report, name who will use the findings, when and how, before you design a single tool.

2 Ask what the programme assumes

Her theory-based evaluation made a programme's hidden assumptions visible, which can be worth as much to a client as the final verdict.

TRY THIS

Run a one-hour session with your client listing every assumption between activity and outcome, then test the weakest one.

3 Stay with one big question

She stayed with how research is used for decades, while advising the World Bank and Canada's IDRC and teaching at Harvard.

TRY THIS

Write the one question you want your career to answer, and pin it where you work.

THE BIG IDEAS OF THIS CHAPTER

1 **Theory-based evaluation** p. 160

Make explicit the chain of assumptions by which a programme is supposed to produce change, then test each link, so that the evaluation can explain why results did or did not appear.

2 **The many meanings of research use** p. 163

Research and evaluation influence policy in at least seven different ways, and the slow enlightenment of how people think about a problem can matter more than any single decision.

3 **The politics of evaluation** p. 166

Evaluation is political from start to finish, so evaluators must understand the interests around a programme and protect the conditions for credible work.

CONCEPT HIGHLIGHT

Political model of use

Research used as ammunition by people whose positions on an issue are already fixed.

GIANTS CONNECTED TO THIS CHAPTER



Huey-tsyh Chen

Theory-driven evaluation gave the idea a detailed structure of change and action models.

p. 174



Patricia J. Rogers

Purposeful program theory asks what a theory of change is for before drawing it.

p. 268



Michael Quinn Patton

Both studied why findings go unused; Patton designed the whole evaluation around use.

p. 204

KEY WORKS TO READ IN THE ORIGINAL

- **Evaluation Research: Methods of Assessing Program Effectiveness** 1972
- **The Many Meanings of Research Utilization** 1979
- **Evaluation: Methods for Studying Programs and Policies** 1998

CONCEPT HIGHLIGHT

Theory-based evaluation

Evaluation that makes explicit the assumptions linking a programme's activities to its outcomes and tests the intermediate steps, not only the final result.

CONCEPT HIGHLIGHT

Theory of change

The set of assumptions about the mini-steps through which a programme is expected to bring about its long-term goals; Weiss is credited with popularising the term.

WORKSHEET · TRY THIS TEMPLATE

Assumptions ladder

Print this page, or copy the headings into your notes.

1 Activity

What the programme does, in one line.

2 Expected next step

The immediate change the activity should produce.

3 Assumption

What must be true for that step to happen: people attend, understand, can act, are allowed to act.

4 How we will test it

Data source and timing for checking the assumption.

5 If it fails

What the programme would change, and who would decide.

REFLECT

Which assumption in your programme's theory of change has nobody actually checked?

CONNECTS TO

HC **Huey-tsyh Chen**
Theory-driven evaluation gave the idea a detailed structure of change and action models. p. 174

PR **Patricia J. Rogers**
Purposeful program theory asks what a theory of change is for before drawing it. p. 268

MP **Michael Quinn Patton**
Both studied why findings go unused; Patton designed the whole evaluation around use. p. 204

08



Emblem: an original HubForge design for this figure, not a likeness.

Huey-tsyh Chen

Active 1980s to 2010s Born in Taiwan; worked in the United States

Theory and programme logic

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

IN ONE LINE

Huey-tsyh Chen opened the black box of programme evaluation and showed practitioners how to read the machinery inside.

For years, the gold standard for judging a social programme was an experiment that could answer one question: did it work? It was much worse at a second, equally important question: why? In 1980 a young sociologist trained in Taipei and at the University of Massachusetts Amherst, Huey-tsyh Chen, challenged that blind spot in the journal Social Forces, writing with his mentor Peter Rossi. Every programme rests on a theory of how it will help, they argued, and evaluations that ignore it treat programmes as black boxes. A decade later his book Theory-Driven Evaluations set out the approach in full. Then, in 1997, he took it out of the university and into the US Centers for Disease Control and Prevention, where he led the design of a national evaluation system for HIV prevention programmes.

Life CHAPTER 08

Huey-tsyh Chen, who also publishes as Huey T. Chen, was born and raised in Taiwan. He took a bachelor's degree in sociology in Taipei and then moved to the United States for doctoral study, earning a PhD in sociology at the University of Massachusetts Amherst. There he worked with Peter H. Rossi, one of the most influential figures in American programme evaluation. Chen later wrote about that apprenticeship in a 2007 article in the American Journal of Evaluation on his mentorship under Rossi.

His early career was a partnership with Rossi. In 1980 the two published an article in Social Forces proposing a multi-goal, theory-driven approach to evaluation, presented as a model linking basic and applied social science. In 1983, when Chen was at Johns Hopkins University, they followed it with Evaluating with Sense in Evaluation Review. He later taught at the University of

BIG IDEAS

1. **Theory-driven evaluation: normative and causal theory**
2. **The change model and the action model**
3. **The bottom-up approach to integrative validity**

KEY WORKS

The multi-goal, theory-driven approach to evaluation
1980 · With Peter H. Rossi, Social Forces

Theory-Driven Evaluations
1990

Practical Program Evaluation
2005 · Second edition 2015

Akron, and in 1990 SAGE published his book Theory-Driven Evaluations, which his publisher describes as a landmark in programme evaluation. In 1992 he and Rossi co-edited Using Theory to Improve Program and Policy Evaluations for Praeger, and in 1993 the American Evaluation Association gave him its Paul F. Lazarsfeld award for contributions to evaluation theory.

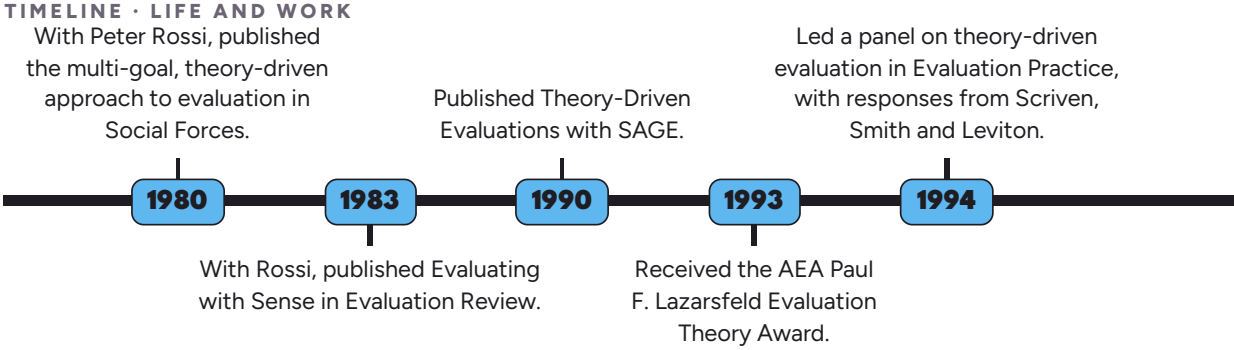
In 1997 Chen left Akron for the US Centers for Disease Control and Prevention, where he became chief of an evaluation branch and a senior evaluation scientist. He led the design and implementation of a national evaluation system for HIV prevention programmes funded by the CDC and run by health departments and community organisations, and described the work in a 2001 journal article. The CDC gave him its Senior Biomedical Research Service Award in 1998. In 2002 he became a professor in the School of Public Health at the University of Alabama at Birmingham, where he produced Practical Program Evaluation (SAGE, released in late 2004 and often dated 2005) and its second edition in 2015.

In 2013 he joined Mercer University as Professor of Public Health and Director of the Center for Evaluation and Applied Research. Mercer credits him with more than 70 peer-reviewed articles and lists evaluations of a juvenile delinquency prevention and treatment policy in Taiwan and of an HIV prevention and care initiative in China. He also serves on the editorial advisory board of Evaluation and Program Planning. In his later work he turned to validity, proposing in 2010 a bottom-up approach to integrative validity, and in 2024 he co-wrote an article on programme plan evaluation in the American Journal of Evaluation.

The spark

As a doctoral student at the University of Massachusetts Amherst, he worked beside Peter Rossi, one of the most influential figures in American programme evaluation, at a time when experiments dominated the field. Chen and Rossi saw what those designs left out: they said little about how a programme was actually delivered, which the two identified as often the main reason programmes failed. That

TIMELINE · LIFE AND WORK



gap, between a verdict and an explanation, became the problem of his career.

The breakthrough CHAPTER 08

Imagine a trial shows that a new health programme made no difference. What should you do next? Scrap it? Perhaps the idea was wrong. But perhaps the idea was right and the clinics never received the supplies, or the staff were never trained. A standard experiment cannot tell those two stories apart. In 1983 Chen and Rossi argued that this tradition treated programmes as black boxes: it could show whether outcomes differed between groups, but not why, and it said little about implementation. They proposed that a clear model of how a programme is supposed to work could strengthen experiments and offset the weaknesses of less rigorous designs.

Chen's 1990 book Theory-Driven Evaluations, described in secondary accounts as the first comprehensive treatment of the approach, consolidated these arguments. He distinguished normative theory, what a programme should be and do, from causal theory, how it is expected to produce its effects, and argued that evaluation needs both. A 1994 panel in Evaluation Practice, organised around his paper on the need, difficulties and options for theory-driven evaluation, brought critics and supporters into public debate. In Practical Program Evaluation, a CDC reviewer noted in 2006, he went further and used an action model and a change model in place of the familiar logic model.

TIMELINE, CONTINUED

1997 — Moved from the University of Akron to the CDC as chief of an evaluation branch.

2001 — Published his account of the CDC national HIV prevention evaluation system.

2002 — Became a professor at the University of Alabama at Birmingham School of Public Health.

2010 — Proposed the bottom-up approach to integrative validity in Evaluation and Program Planning.

2013 — Joined Mercer University as Professor of Public Health and Director of the Center for Evaluation and Applied Research.

The problem it solved

Black-box evaluations that reported outcomes without explaining the mechanism, and could not separate bad ideas from bad delivery.

Impact WHAT CHANGED

Chen's most tangible impact came through public health. At the CDC from 1997 he led the building of a national evaluation system for HIV prevention programmes run by health departments, based on evaluation guidance that defined standard data elements for different evaluation tasks; his 2001 account argued its lessons would apply to any organisation building a large evaluation system. His textbook Practical Program Evaluation was reviewed in the CDC journal Preventing Chronic Disease in 2006, where the reviewer recommended its diagrams, examples and taxonomy for evaluation teaching. Beyond the United States, Mercer University lists his evaluations of a juvenile delinquency prevention and treatment policy in Taiwan and an HIV prevention and care initiative in China.

His influence also runs through the wider family of theory-based approaches that development agencies now use. In our reading, the routine expectation that NGO and donor programmes state their causal assumptions in a theory of change or results framework, and test them, owes part of its intellectual footing to the theory-driven tradition Chen built with Rossi, alongside Carol Weiss and others. His insistence on separating delivery from mechanism remains a practical tool for evaluators in low-resource settings, where implementation gaps are common. His later work on viable validity, in our assessment, speaks to the same audience by asking whether programmes can work with ordinary staff and budgets.

KEY WORKS TO READ IN THE ORIGINAL

- **The multi-goal, theory-driven approach to evaluation** 1980, With Peter H. Rossi, Social Forces
- **Theory-Driven Evaluations** 1990
- **Practical Program Evaluation** 2005, Second edition 2015

WHERE PEOPLE DISAGREE

Debates

Theory-driven evaluation was contested from early on. In 1994 Evaluation Practice published a panel built around Chen's paper on the need, difficulties and options for the approach, with responses from Michael Scriven, Nick Smith and Laura Leviton; Scriven's piece was titled The Fine Line between Evaluation and Explanation, and his later paper on minimalist theory signalled, in our reading, his view that evaluation needs less theory than Chen proposed. E. Jane Davidson argued in 2006 that both advocates and sceptics of programme theory held serious misconceptions. A 2006 CDC review of Practical Program Evaluation praised the book but noted unfamiliar terminology, the burden of replacing the logic model, and limited attention to evaluation use and intended users.

LIMITS, IN OUR READING

- ☐ Critics have said the approach leaned heavily on statistical methods such as experiments, quasi-experiments and structural equation modelling, giving qualitative methods less weight.
- ☐ Its terminology overlaps with other approaches, so theory-driven, theory-based and similar labels are often used interchangeably, which can confuse clients.

WHY THIS MATTERS TO YOU

Chen's central question is useful whenever something you care about fails: was it a bad idea, or a good idea badly delivered? The answer changes everything. A bad idea should be dropped; a good idea badly delivered should be fixed. If you have ever drawn a theory of change or a results framework, you are working on ground he helped clear. And his later work adds a practical test before anyone spends on a costly trial: can this programme work with the people, budgets and conditions that actually exist? Asking that early can save years.

70+

peer-reviewed articles

Mercer University College of Health Professions, Huey Chen biography

6

basic types of theory-driven evaluation set out in Theory-Driven Evaluations (1990)

SAGE Publishing, Theory-Driven Evaluations table of contents

3

editions of Practical Program Evaluation, from late 2004 to the third edition listed for 2026

SAGE Publishing, Practical Program Evaluation third edition listing; CDC Preventing Chronic Disease review (2006)

Key concepts

Black-box evaluation

An evaluation that measures whether outcomes changed without examining the mechanisms or implementation that connect a programme to those outcomes.

Programme theory

The set of assumptions, stated or implicit, about how a programme should be delivered and how it is expected to produce its intended effects.

Normative theory

In Chen's usage, the part of programme theory that specifies what the programme should be: its goals, treatment and implementation.

Causal theory

In Chen's usage, the part of programme theory that specifies how the treatment is expected to change intermediate factors and outcomes.

Action model and change model

Chen's paired models: the action model sets out who delivers what, to whom, how and in what context; the change model sets out how the intervention shifts determinants that lead to outcomes.

Viable validity

A standard proposed by Chen within his integrative validity model, built on the bottom-up approach and presented as suited to health promotion and social betterment programmes; as we read it, it concerns whether a programme works under real-world conditions.

Sources for this chapter: p. 578

Did you know?

In a 1996 issue of Evaluation Practice, Chen wrote both an opening typology and the closing epilogue for a debate on formative and summative evaluation that also featured Michael Quinn Patton, Joseph Wholey and Michael Scriven.

THE BIG IDEAS OF THIS CHAPTER

1. **Theory-driven evaluation: normative and causal theory** p. 182

   Make a programme's theory explicit, both what it should look like and how it is meant to cause change, and evaluate against that theory so results can be explained, not just reported.

2. **The change model and the action model** p. 185

   Split programme theory into a change model, how the intervention shifts determinants and outcomes, and an action model, what must be in place to deliver it, so evaluators can tell a bad idea from bad delivery.

3. **The bottom-up approach to integrative validity** p. 188

   Before testing whether a programme works under ideal conditions, check whether it can work in the real world, so that evaluation serves both science and practice.

GIANTS CONNECTED TO THIS CHAPTER



Carol H. Weiss

Theory-based evaluation shares the aim of opening the black box.

p. 152



Patricia J. Rogers

Purposeful program theory shows how to fit programme theory to its purpose.

p. 268



Howard White

Theory-based impact evaluation joins programme theory to counterfactual designs.

p. 476

In one line

Huey-tsyh Chen opened the black box of programme evaluation and showed practitioners how to read the machinery inside.

1

BIG IDEA 1 OF 3 · CHEN

Theory-driven evaluation: normative and causal theory

Make a programme's theory explicit, both what it should look like and how it is meant to cause change, and evaluate against that theory so results can be explained, not just reported.

PICTURE IT

Think of a recipe and a cook. A recipe has a theory: these ingredients, in this order, make this dish. Chen asked evaluators to write down both what the dish should be and why the recipe should produce it, so that when the dish disappoints, they can tell where things went wrong.

The problem

Since Donald Campbell and Julian Stanley's influential 1966 work on research design, impact assessment had been dominated by experiments. Randomised and quasi-experimental studies could show whether outcomes differed between groups, but treated the programme as a black box. When a programme failed, nobody could say whether the idea was wrong, the delivery was weak or the context was unsuitable. Decision-makers were left with a verdict but no guidance, and promising ideas could be abandoned over fixable implementation problems.

What it is

Theory-driven evaluation assesses a programme against its programme theory: the stated or implicit assumptions about how it should be organised and why it should work. In Theory-Driven Evaluations (SAGE, 1990) Chen divided programme theory into two parts. Normative theory describes what the programme should be: its goals and outcomes, the treatment to be delivered and the implementation environment it needs. Causal theory describes how the treatment is expected to produce its effects. From these he derived six basic types of evaluation. Normative outcome, normative treatment and normative implementation environment evaluations examine whether the programme is what it should be. Impact, intervening mechanism and generalisation evaluations examine whether and how it causes change, and whether results would hold elsewhere. Composite types combine them, such as normative treatment and impact evaluation. The logic is that a programme

can fail at any link, so the evaluator examines the links that matter to stakeholders, strengthening experimental designs rather than replacing them.

How it works

1. **Elicit the programme theory**
   Draw out the assumptions of designers and stakeholders, and relevant research, about what the programme should do and how it should work; Chen's 1990 book gives a chapter to constructing programme theory.
2. **Separate normative from causal**
   Write the normative theory (goals, treatment, implementation environment) and the causal theory (impact, intervening mechanisms, conditions for generalisation) as distinct statements.
3. **Choose the evaluation types needed**
   Match stakeholders' questions to Chen's types, for example a normative treatment evaluation if delivery is in doubt, or an intervening mechanism evaluation if the causal link is.
4. **Check fidelity to the normative theory**
   Compare the treatment and environment actually delivered with what the theory requires.
5. **Test the causal links**
   Measure intermediate factors as well as outcomes, so analysis can show whether the treatment changed them and whether they changed the outcomes.
6. **Explain and advise**
   Locate where the chain held or broke, and report what that implies for redesign, delivery or scale.

WORKED EXAMPLE · AN ILLUSTRATION

A district in Madhya Pradesh enriches its school mid-day meals, adding eggs or milk three days a week, to raise test scores in grades one to five. The normative theory says enriched meals will be served on schedule in all 400 schools, with cooks trained and supplies delivered monthly. The causal theory says better nutrition will improve attention and attendance, which will raise learning. After a year, scores are flat. A black-box study would stop there. This evaluation also collected supply records, observed meals and checked attendance registers. It finds that enriched meals were served on fewer than half the planned days because supply contracts lapsed, a failure of the normative treatment. Where meals were served, attendance rose. The advice is to fix procurement, not abandon the idea.

USE IT WHEN

- Stakeholders need to know why results occurred, not only whether they did.
- A programme is to be scaled or adapted and its causal mechanism must be understood.
- Earlier outcome studies produced puzzling or null results.

SKIP IT WHEN

- ○ The only question is a simple compliance check or output count.
- ○ Time and budget do not allow intermediate factors to be measured.
- ○ The programme is so new and emergent that no theory can yet be stated, where developmental evaluation may fit better.

COMMON MISREADINGS

- ☐ Treating a drawn diagram of the programme theory as the evaluation itself, without collecting data that test its links.
- ☐ Assuming theory-driven means anti-experimental, when Chen and Rossi presented it as a way to strengthen experimental and quasi-experimental designs.

WHY IT WAS GROUNDBREAKING

It challenged the dominance of black-box experiments by insisting that evaluation should explain why programmes succeed or fail. It gave a formal vocabulary for linking social science theory to practical programme assessment, which in our reading laid groundwork for the theory-based approaches development agencies use today.

Origin

Chen and Peter Rossi proposed a multi-goal, theory-driven approach in Social Forces in 1980 and developed it in Evaluating with Sense (Evaluation Review, 1983). Chen's Theory-Driven Evaluations (SAGE, 1990) gave the first book-length treatment, and he and Rossi co-edited Using Theory to Improve Program and Policy Evaluations (Praeger, 1992). A 1994 Evaluation Practice panel debated it, and his later textbook recast it as action and change models.

RELATED IDEAS

The change model and the action model

CW Weiss 152 · DC Campbell 112 · PR Rogers 268 · HW White 476

CONCEPT HIGHLIGHT

Black-box evaluation

An evaluation that measures whether outcomes changed without examining the mechanisms or implementation that connect a programme to those outcomes.

2

BIG IDEA 2 OF 3 · CHEN

The change model and the action model

Split programme theory into a change model, how the intervention shifts determinants and outcomes, and an action model, what must be in place to deliver it, so evaluators can tell a bad idea from bad delivery.

PICTURE IT

Think of a cake that fails to rise. Either the recipe was wrong, or the oven was faulty and the baker skipped a step. The change model checks the recipe; the action model checks the kitchen. Only by checking both can you know which to fix.

The problem

Logic models spread widely, but many drew a single chain from inputs to impact that blurred two different kinds of claim: claims about delivery and claims about causation. When a programme disappointed, managers often could not tell whether to fix the delivery system or rethink the intervention. In public health, where Chen led evaluation work at the CDC from 1997, that confusion could mean wasted budgets and communities left without effective prevention.

What it is

In Practical Program Evaluation (SAGE, released late 2004 and often dated 2005), Chen reorganised programme theory into two linked models. The change model is the causal part. It has three elements: the intervention; the determinants, the intermediate or mediating factors the intervention is meant to shift; and the outcomes that should follow. The action model is the operational part: what must be arranged for the intervention to reach people. It covers the implementing organisation and its partners, the delivery protocols, the context in which delivery takes place and the target population to be reached. A failure can sit in either model. A sound intervention poorly delivered is an action model failure; a well-delivered intervention that does not move its determinants is a change model failure. The SAGE Encyclopedia of Evaluation describes matching evaluation types: theory-driven process evaluation tests the action model against actual implementation, intervening mechanism evaluation tests the change model, and moderating mechanism evaluation examines factors that condition the effect.

How it works

1. **Draw the change model**
   State the intervention, the determinants it must shift and the outcomes expected, drawing on research and stakeholder knowledge.
2. **Draw the action model**
   Set out who delivers, through which organisation and partners, under which protocols, in what context and to which target population.
3. **Check the fit**
   Ask whether the action model, as resourced, can plausibly deliver the intervention that the change model needs.
4. **Evaluate delivery**
   Use a theory-driven process evaluation to see whether the target group was reached and protocols were followed, as described in the SAGE Encyclopedia of Evaluation.
5. **Test the mechanism**
   Measure determinants as well as outcomes in an intervening mechanism evaluation, and look for moderators that change the effect for different groups.
6. **Report two diagnoses**
   Tell decision-makers separately whether to fix delivery, rethink the intervention or both; SAGE's listing for the 2026 third edition presents this schema alongside six stages of a programme's life cycle.

WORKED EXAMPLE · AN ILLUSTRATION

A state AIDS control society in India funds community organisations to raise HIV testing among migrant construction workers. The change model says peer outreach will increase knowledge and reduce fear of stigma, the determinants, which will raise testing and linkage to treatment, the outcomes. The action model says three community organisations will deploy 30 trained peer educators at worksites, using a referral protocol with government testing centres. Monitoring shows testing barely rises. Process evaluation reveals an action model failure: the centres close at five in the evening, before shifts end. Where one centre extended its hours, testing rose, yet few workers who tested positive started treatment, because fear of losing work was untouched. That second finding is a change model problem, calling for new determinants such as job protection.

CONCEPT HIGHLIGHT

Programme theory

The set of assumptions, stated or implicit, about how a programme should be delivered and how it is expected to produce its intended effects.

USE IT WHEN

- A programme has a complex delivery chain running through partners or government systems.
- Results disappoint and managers need to know what to fix.
- A scale-up is being planned and the delivery requirements must be specified.

SKIP IT WHEN

- The intervention is very simple, with a single delivery step.
- A funder mandates a particular logic model format and there is no room for a parallel framework, although the two can usually be mapped onto each other.
- An innovation is so early that neither model is yet stable.

COMMON MISREADINGS

- Drawing only the change model and neglecting the action model, which is where many failures of delivery hide.
- Treating the two models as paperwork for proposals instead of a guide to which data to collect and how to diagnose results.

WHY IT WAS GROUNDBREAKING

It gave evaluators a structured way to distinguish design failure from implementation failure, which is what decision-makers most need to know. It tied programme theory to the organisational realities of delivery, where, in our reading, many programmes in low-resource settings actually break.

Origin

Chen presented the paired models in Practical Program Evaluation (SAGE, released late 2004, often dated 2005), building on his 1990 typology. A 2006 review in the CDC journal Preventing Chronic Disease noted that he offered them in place of the logic model. A second edition followed in 2015, and SAGE lists a third edition, Foundations of the Field and Current Advancements, for September 2026 that keeps the action model and change model schema.

RELATED IDEAS

Theory-driven evaluation: normative and causal theory

CW Weiss 152 · DS Stufflebeam 134 · PR Rogers 268

CONCEPT HIGHLIGHT

Normative theory

In Chen's usage, the part of programme theory that specifies what the programme should be: its goals, treatment and implementation.

3

BIG IDEA 3 OF 3 · CHEN

The bottom-up approach to integrative validity

Before testing whether a programme works under ideal conditions, check whether it can work in the real world, so that evaluation serves both science and practice.

PICTURE IT

Think of testing a new car only on a perfect track with a professional driver. It may perform beautifully, and still fail on potholed roads with ordinary drivers. Chen's bottom-up approach starts with the potholed road: can it work in real conditions first?

The problem

The Campbellian tradition gave internal validity, confidence that the programme caused the effect, the highest priority. Efficacy studies under tightly controlled conditions could produce interventions that ordinary agencies were unable to deliver with their own staff and budgets. Practitioners saw evaluations that were rigorous but of little use to them, and the gap between academics and practitioners widened. Communities waited for proven programmes that never became workable services. Chen asked whether the sequence of evaluation itself was part of the problem.

What it is

In a 2010 article in Evaluation and Program Planning, Chen argued that the traditional Campbellian approach works from the top down: researchers set the priorities, internal validity comes first, and stakeholders' concerns follow later, if at all. He held that this emphasis can make evaluations less useful and widen the gap between academics and practitioners. His alternative, set within an integrative validity model, works from the bottom up. It starts with stakeholders' evaluation views and concerns, and makes external validity, whether results hold in real settings, practical to pursue. At its base is viable validity, which we read as the question of whether a programme is practical, affordable and acceptable, and able to work in real-world conditions with ordinary resources. Rigorous testing of effects, and of whether they hold in other settings, follows once viability is established. Chen presented viable validity as a fitting standard for health promotion and social betterment programmes, and the approach as a way to satisfy both scientific and practical demands.

How it works

1. **Start with stakeholders**
   Ask implementers, managers and communities what they need to know and what would make the programme usable for them, the starting point of Chen's 2010 article.
2. **Assess viability**
   Examine whether the programme can be delivered by ordinary staff, within existing budgets, and is acceptable to the people it serves.
3. **Adapt or stop**
   Use the viability findings to strengthen the programme, or to stop before investing in costly trials of something that cannot be run.
4. **Test effects rigorously**
   Once the programme is viable, evaluate its outcomes with designs strong enough to support causal claims.
5. **Examine generalisability**
   Study whether effects hold across other sites, populations and conditions before recommending scale-up.

WORKED EXAMPLE · AN ILLUSTRATION

A state health department in Odisha is offered an adolescent anaemia programme that worked in a well-funded university trial with dedicated nurses. Following a bottom-up sequence, the evaluator first spends four months in two blocks testing viability. Can existing auxiliary nurse midwives and teachers deliver weekly iron tablets and nutrition sessions alongside their other duties? Do girls and parents accept the tablets, given side effects? The viability data show that sessions fit the school timetable, but that supply gaps and nausea reduce uptake. The programme is adapted, with buffer stocks and guidance on taking tablets after meals. Only then does the department fund a cluster-randomised effectiveness study across 60 schools, followed by checks in a tribal district with different conditions.

USE IT WHEN

- A programme designed under ideal conditions is about to be tested at scale in ordinary public systems.
- Delivery will depend on government or NGO frontline workers with limited time and resources.
- Acceptability to communities is uncertain, as in many health promotion programmes.

SKIP IT WHEN

- An intervention is already proven viable and the open question is purely causal.
- Regulation requires a fixed efficacy trial sequence, as for new medicines.
- Stakeholders cannot be engaged early enough to shape the viability questions.

COMMON MISREADINGS

- [ ] Taking bottom-up to mean rigour is optional, when Chen keeps rigorous testing of effects in the sequence.
- [ ] Treating viability as a quick satisfaction survey after a pilot, rather than a serious assessment of practicality, affordability and acceptability.

WHY IT WAS GROUNDBREAKING

It questioned, from within a tradition that prized rigour, the assumption that internal validity must always come first. It offered a sequence that aims to make evaluation both scientifically credible and useful to the agencies that must run programmes.

Origin

Chen set out the bottom-up approach to integrative validity in Evaluation and Program Planning in 2010 (volume 33, issue 3), extending the concern with real-world conditions that ran through his theory-driven work. SAGE's listing for the 2026 third edition of Practical Program Evaluation includes a holistic effectuality perspective for real-world outcome evaluation, which in our reading continues the same line of thought.

RELATED IDEAS

DC Campbell 112 · LC Cronbach 38

Theory-driven evaluation: normative and causal theory

D&B Duflo and Banerjee 432

RELATED GIANTS



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Patricia J. Rogers

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

p. 268



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476

CONCEPT HIGHLIGHT

Causal theory

In Chen's usage, the part of programme theory that specifies how the treatment is expected to change intermediate factors and outcomes.

CONCEPT HIGHLIGHT

Action model and change model

Chen's paired models: the action model sets out who delivers what, to whom, how and in what context; the change model sets out how the intervention shifts determinants that lead to outcomes.

CONCEPT HIGHLIGHT

Viable validity

A standard proposed by Chen within his integrative validity model, built on the bottom-up approach and presented as suited to health promotion and social betterment programmes; as we read it, it concerns whether a programme works under real-world conditions.

MORE RELATED GIANTS



Daniel L. Stufflebeam

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

p. 134



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Esther Duflo and Abhijit Banerjee

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

p. 432

Put it to work

USE IT ON MONDAY · THEORY-DRIVEN EVALUATION: THE CHANGE MODEL AND THE ACTION MODEL

1. Draw two diagrams for the programme, not one: the change model (intervention, intermediate factors, outcomes) and the action model (who delivers, to whom, how, in what context).

2. When outcomes disappoint, check the action model first: did the programme reach the intended group, through a capable organisation, following its protocols?

3. Report separately on implementation and on causal links, so the client can see whether to fix delivery or rethink the idea.

AI prompt **Copy, fill the brackets, then check every line**

Using Huey-tsyh Chen's theory-driven evaluation framework, help me separate the change model from the action model for **[programme]** in **[location]**. Here is the programme description: **[paste]**. Here are the results we have so far: **[paste findings or write NONE YET]**. (1) Draft the change model: intervention, intermediate factors and outcomes. (2) Draft the action model: target population, implementing organisation, partners, protocols and context. (3) For each element, suggest an evaluation question and a data source. (4) If results are provided, list the possible explanations for them under design failure and implementation failure, without choosing between them unless the evidence supports it.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a recipe and a cook. A recipe has a theory: these ingredients, in this order, make this dish. Chen asked evaluators to write down both what the dish should be and why the recipe should produce it, so that when the dish disappoints, they can tell where things went wrong.

YOUR MOVES Three moves from Chen's path

1 Find a mentor and build with them

Chen's partnership with Peter Rossi produced the papers of 1980 and 1983 that framed his career, and he later wrote the book that consolidated it.

TRY THIS

Ask a senior evaluator you admire to co-author a short piece with you, offering to do the first draft.

2 Test your ideas inside a big system

He moved from the University of Akron to the CDC in 1997 and back to academia in 2002, and his ideas returned sharper.

TRY THIS

Seek one assignment inside a government or large agency system, where your framework meets real scale.

3 Keep revising your best work

He kept refining his framework for more than four decades, from 1980 to a third edition of his textbook listed for 2026.

TRY THIS

Reopen one tool you built more than a year ago and publish a version 2 with what practice has taught you.

THE BIG IDEAS OF THIS CHAPTER

1 **Theory-driven evaluation: normative and causal theory** p. 182

Make a programme's theory explicit, both what it should look like and how it is meant to cause change, and evaluate against that theory so results can be explained, not just reported.

2 **The change model and the action model** p. 185

Split programme theory into a change model, how the intervention shifts determinants and outcomes, and an action model, what must be in place to deliver it, so evaluators can tell a bad idea from bad delivery.

3 **The bottom-up approach to integrative validity** p. 188

Before testing whether a programme works under ideal conditions, check whether it can work in the real world, so that evaluation serves both science and practice.

KEY WORKS TO READ IN THE ORIGINAL

- **The multi-goal, theory-driven approach to evaluation** 1980, With Peter H. Rossi, Social Forces
- **Theory-Driven Evaluations** 1990
- **Practical Program Evaluation** 2005, Second edition 2015

CONCEPT HIGHLIGHT

Viable validity

A standard proposed by Chen within his integrative validity model, built on the bottom-up approach and presented as suited to health promotion and social betterment programmes; as we read it, it concerns whether a programme works under real-world conditions.

GIANTS CONNECTED TO THIS CHAPTER



Carol H. Weiss

Theory-based evaluation shares the aim of opening the black box.

p. 152



Patricia J. Rogers

Purposeful program theory shows how to fit programme theory to its purpose.

p. 268



Howard White

Theory-based impact evaluation joins programme theory to counterfactual designs.

p. 476

CONCEPT HIGHLIGHT

Black-box evaluation

An evaluation that measures whether outcomes changed without examining the mechanisms or implementation that connect a programme to those outcomes.

CONCEPT HIGHLIGHT

Programme theory

The set of assumptions, stated or implicit, about how a programme should be delivered and how it is expected to produce its intended effects.

WORKSHEET · TRY THIS TEMPLATE

Change model and action model sheet

Print this page, or copy the headings into your notes.

1 Intervention

What is actually delivered, in concrete terms.

2 Intermediate factors

What must change first, such as knowledge, access or norms, for outcomes to follow.

3 Outcomes

Short and longer term, for whom.

4 Target population and reach

Who should be reached, and who actually is.

5 Implementing organisation and partners

Capacity, roles and incentives.

6 Protocols and context

How delivery should happen, and what in the setting helps or hinders it.

REFLECT

If your programme's results disappoint, would you be able to tell whether the idea was wrong or the delivery was?

CONNECTS TO

CW **Carol H. Weiss**
Theory-based evaluation shares the aim of opening the black box. p. 152

PR **Patricia J. Rogers**
Purposeful program theory shows how to fit programme theory to its purpose. p. 268

HW **Howard White**
Theory-based impact evaluation joins programme theory to counterfactual designs. p. 476

THE GREAT DEBATES · ACT 2

Proof Here or Promise Elsewhere?

THE QUESTION

Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

DC Campbell 112 | LC Cronbach 38 | HC Chen 174

SIDE A

Internal validity first (Campbell)

Campbell's starting point was that a claim of programme success means little if rival explanations have not been ruled out. In 1963, with Julian Stanley, he catalogued sixteen designs and eight threats to internal validity, giving field researchers a shared language for what a comparison could and could not prove. In 1969, in Reforms as Experiments, he urged governments to treat reforms as experiments, and showed how a governor's claim about a speeding crackdown ignored trends and ordinary year-to-year swings. The Encyclopedia of Evaluation records that Campbell continued to prioritise internal validity. In our reading, his case is simple and powerful: before asking whether a result will transfer, be sure the result is real.

SIDE B

External validity and many studies (Cronbach)

Cronbach argued that the question policymakers most need answered is whether a programme that succeeded in one place will succeed in another, with different people and conditions. In Designing Evaluations of Educational and Social Programs (1982) he set out his units, treatments, observations and settings framework to address exactly that. According to the Encyclopedia of Evaluation, he opposed relying on one large experiment, which he likened to a horse race, and favoured many smaller studies, each looking closely at an important case or component. He saw evaluation as political and its main audience as a wide policy-shaping community. In our reading, his case is that a clean answer to a narrow question can still mislead a busy decision-maker.

WHAT HAPPENED

Neither side simply won. In 1991 Shadish, Cook and Leviton placed Campbell among the stage one theorists who sought to bring truth to social problem solving, and Cronbach among the stage three theorists who tried to integrate what came before. In Evaluation Roots (2004), Thomas Cook wrote a chapter on how both men shaped his own theory of causal generalisation. Chen, who had criticised black box experiments with Peter Rossi in 1983, proposed in 2010 a bottom-up approach to integrative validity. In our reading, the same argument resurfaced in development economics after 2010, when critics asked whether trial results transfer across settings.

Where you stand

Think of a pilot in your organisation that showed good results. What would you need to know before recommending it for a different district or state?

If you had budget for one large study or five small ones, which would you choose for your next evaluation, and why?

Your position

Which side do you find yourself on today, and what evidence would change your mind?

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340 · D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476 · BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

THE GIANTS IN THIS DEBATE



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174

ACT 3 OF 5



1970s to 2010s

Will Anyone Use It? The Use Revolution



Michael Quinn Patton
p. 204



Hallie Preskill
p. 226



Sanjeev Sridharan
p. 246



Patricia J. Rogers
p. 268

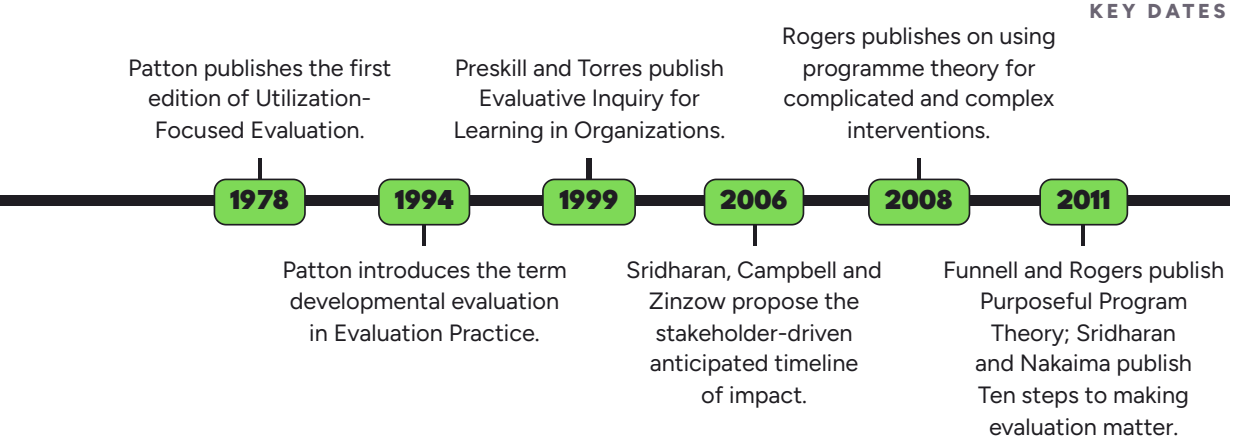
ACT 3 · 1970S TO 2010S

Will Anyone Use It? The Use Revolution

By the 1970s programme evaluation had grown quickly in the United States, and an uncomfortable question had become pressing: did any of it change decisions? Michael Quinn Patton treated this as a research problem. A 1977 field study by Patton and colleagues interviewed about 60 project officers, decision-makers and evaluators linked to 20 health programme evaluations, and pointed to what he called the personal factor: findings were most likely to be used when an identifiable person cared about the evaluation and what it would find. In 1978 the first edition of Utilization-Focused Evaluation argued that an evaluation should be judged by its actual use. In 1994 he introduced the term developmental evaluation for work alongside innovations still taking shape.

Others carried the idea inside organisations and into complexity. In 1999 Hallie Preskill and Rosalie Torres described evaluative inquiry, in which an organisation's own members continually examine what they do as part of daily work. Sanjeev Sridharan and colleagues proposed in 2006 that stakeholders set out when a programme should first be expected to show impact, and in 2011 he and Nakaima set out ten steps to make evaluation matter, arguing that a programme's complexity should shape its evaluation's design. Patricia Rogers, introduced to program logic at an evaluation conference in Canberra in 1987, argued in 2008 that evaluators should classify aspects of an intervention as simple, complicated or complex, and in 2011, with Sue Funnell, presented programme theory as a family of tools chosen to fit purpose.

KEY DATES



In Alkin and Christie's evaluation theory tree, Stufflebeam, Patton and Preskill sit on the use branch. In our reading, this act answered the experimenting society with a practical rejoinder: a rigorous finding that no one acts on has helped no one, so the evaluator's craft must include the people and organisations who will use what is learned.

THE GREAT DEBATE OF THIS ACT

Use or Judgement? Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth? p. 290

THE GIANTS OF THIS ACT



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Hallie Preskill

The educator who reframed evaluation as a way for organisations to learn, starting from what already works.

p. 226



Sanjeev Sridharan

The evaluator of complex health programmes who argues that designs should fit the programme, and that evidence lives in ecologies, not silos.

p. 246



Patricia J. Rogers

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

p. 268

Jennifer C. Greene, in one line

Jennifer Greene taught evaluators to listen across differences: between methods, between disciplines, and between the powerful and the unheard.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

09



Portrait adapted from a photograph; credit on p. 590.

Michael Quinn Patton

Born 1945 Active 1970s to 2020s United States; keynotes and consultancy worldwide

Use and learning

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

IN ONE LINE

Michael Quinn Patton taught a whole profession that an evaluation is only as good as the difference it makes in someone's hands.

In the 1970s, programme evaluation in the United States was booming, and a sociologist at the University of Minnesota decided to test an uncomfortable question: did anyone actually use it? Michael Quinn Patton and his colleagues interviewed about 60 project officers, decision-makers and evaluators connected with 20 health programme evaluations. The answer they found was not a technique or a statistic. It was a person. Evaluations were used when someone identifiable cared about them. In 1978 Patton built a book on that finding. Utilization-Focused Evaluation reached its fifth edition in 2021, 43 years after the first, and from the same root grew developmental evaluation, principles-focused evaluation and Blue Marble Evaluation. One stubborn question had travelled from a Minnesota campus to the launch of an African evaluation society in Nairobi, and on to the scale of the whole planet.

BIG IDEAS

1. **Utilization-focused evaluation**
2. **Developmental evaluation**
3. **Blue Marble Evaluation**

KEY WORKS

Utilization-Focused Evaluation
1978 · First edition; fifth edition 2021 with Charmagne E. Campbell-Patton

Getting to Maybe
2006 · With Frances Westley and Brenda Zimmerman

Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use
2010 or 2011 · Sources differ on the year

Life CHAPTER 09

Michael Quinn Patton was born in 1945 in Pewee Valley, Kentucky. He took a bachelor's degree in anthropology and sociology at the University of Cincinnati, then a master's degree in rural sociology and a PhD in sociology at the University of Wisconsin, Madison. His work in evaluation began in the 1970s, when the systematic evaluation of public and non-profit programmes was still a young field in the United States.

In 1973 he joined the faculty of the University of Minnesota, where he remained for 18 years, until 1991. During that time he spent five years as director of the Minnesota Center for Social Research and ten years working with the Minnesota Extension Service. In the 1970s he and colleagues studied what made evaluations get used, and that research fed directly

into the first edition of Utilization-Focused Evaluation in 1978. The American Evaluation Association lists him as the 1984 recipient of its Alva and Gunnar Myrdal Evaluation Practice Award, and in 1988 he served as the association's president, the year its annual conference in New Orleans took Evaluation and Politics as its theme.

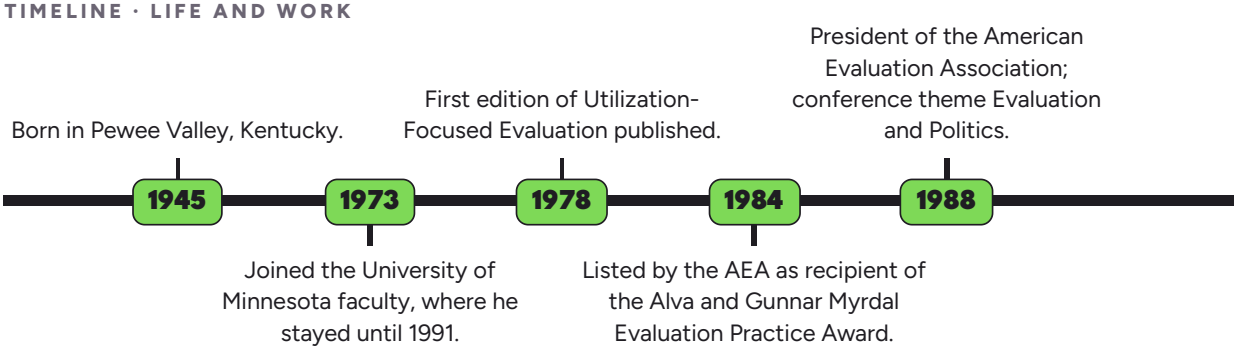
In 1991 he left the university for independent practice, founding Utilization-Focused Evaluation, an organisational development and evaluation practice based in Saint Paul, Minnesota. He also taught at the Union Institute and University Graduate College. His audience became international: the SAGE Encyclopedia of Evaluation records that he gave the keynote at the launch of the African Evaluation Society in Nairobi in 1999 and at the European Evaluation Society in 2000, and he co-chaired the 2005 international evaluation conference in Toronto, held jointly by the American and Canadian associations. His writing ranged from the methods text Qualitative Research and Evaluation Methods, in its fourth edition by 2015, to Grand Canyon Celebration, a work of creative nonfiction about a father and son journey that was a finalist for Minnesota Book of the Year.

Over five decades he kept carrying one question, how evaluation can be made useful, into new settings. Developmental evaluation, first named in a 1994 article and set out at book length in 2010, addressed innovation under complexity. Principles-Focused Evaluation followed in 2018 and Blue Marble Evaluation at the end of 2019, alongside a fifth edition of Utilization-Focused Evaluation in 2021, co-written with Charmagne Campbell-Patton. The American Evaluation Association gave him its Paul F. Lazarsfeld Evaluation Theory Award in 1997 and its Research on Evaluation Award in 2017, and he received the Lester F. Ward Award for contributions to applied sociology in 2001.

The spark

Patton's spark was a research problem. In the 1970s, reports that nobody read were a familiar frustration in the young field of evaluation. Most people complained about it. Patton and his colleagues studied it, asking what made some evaluations get used while others gathered dust. He later wrote that his idea

TIMELINE · LIFE AND WORK



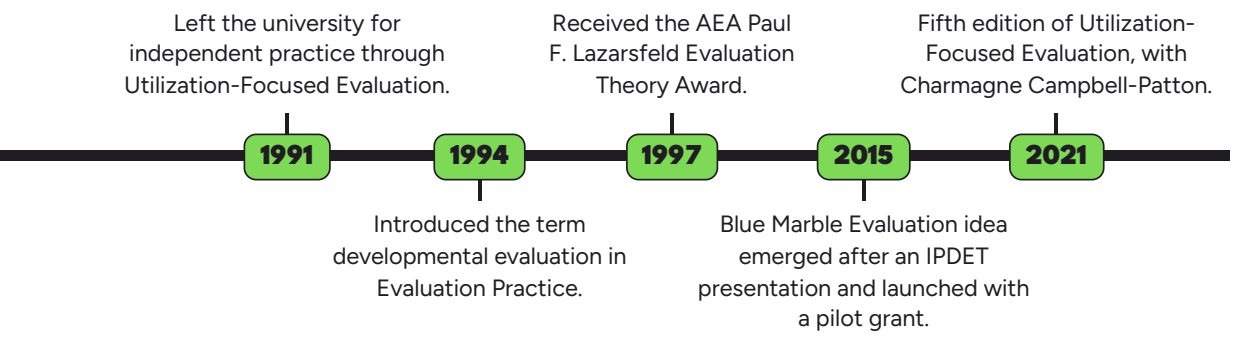
of the personal factor first emerged from that work. Once the evidence pointed to people rather than procedures, he spent the following decades designing evaluation around them.

The breakthrough CHAPTER 09

Picture two evaluations of similar quality. One changes how a programme is run. The other is filed and forgotten. What makes the difference? By the 1970s, as evaluation grew quickly in the United States, that question had become pressing. Patton and his colleagues treated it as a research problem rather than a complaint. A 1986 review by J. Bradley Cousins and Kenneth Leithwood summarises their retrospective study, dated 1977: about 60 project officers, decision-makers and evaluators connected with 20 health programme evaluations were interviewed. Leadership, caring and interest among the people involved were linked to greater use of the findings.

Patton called this the personal factor: evaluations were most likely to be used when an identifiable person or group, other than the evaluator, cared about the evaluation and what it would find. Instead of treating it as a curiosity, he built an approach on it. The first edition of Utilization-Focused Evaluation, published in 1978, argued that an evaluation should be judged by its actual use, and that the evaluator should therefore organise the whole process around specific primary intended users and the uses they intend. The book was revised repeatedly, with a fourth edition in 2008 and a fifth, with Charmagne Campbell-Patton, in 2021.

TIMELINE, CONTINUED



The problem it solved

Well-made evaluation reports that sat unread on shelves.

Impact WHAT CHANGED

Patton's ideas have shaped evaluation practice well beyond the United States. In our reading, the now common habit of naming intended users and intended uses in evaluation terms of reference descends from utilization-focused evaluation. More concretely, his keynote at the launch of an African evaluation society in Nairobi in 1999 placed him early in the growth of evaluation associations in the Global South. In 2013 a primer on utilization-focused evaluation was published in Penang, Malaysia, under the IDRC-supported project Developing Evaluation Capacity in Information and Communication for Development. Its authors included Ricardo Ramírez, Dal Brodhead and Sonal Zaveri, and it drew on five experiences of applying the approach with research projects on information technology for development.

Developmental evaluation soon gathered early adopters: a 2011 University of Waterloo thesis by Mark Cabaj interviewed 18 of them and mentions a 2005 to 2006 training programme with Patton linked to the J.W. McConnell Family Foundation. Blue Marble Evaluation has moved his work into planetary questions. A 2021 webinar hosted by CECAN, the UK centre for evaluating complexity, reported a worldwide network of more than 600 people learning the approach, some capturing the knowledge of women farmers, Indigenous peoples, smallholders and youth ahead of the UN Food Systems Summit. In our assessment, his influence lies as much in how evaluators frame their role as in any single method.

KEY WORKS TO READ IN THE ORIGINAL

- **Utilization-Focused Evaluation** 1978, First edition; fifth edition 2021 with Charmagne E. Campbell-Patton
- **Getting to Maybe** 2006, With Frances Westley and Brenda Zimmerman
- **Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use** 2010 or 2011, Sources differ on the year

WHERE PEOPLE DISAGREE

Debates

Patton's widening of evaluation's purposes has drawn principled disagreement. In a 1996 issue of Evaluation Practice, he argued that the world of evaluation was larger than the formative and summative distinction, citing developmental and knowledge-generating work, while Michael Scriven defended that distinction as one reasonable way of classifying evaluation. A recurring concern in the field is that focusing on primary intended users can tilt an evaluation towards what powerful users want to know. Cabaj's 2011 study of developmental evaluation found practitioners wrestling with how far evaluators should engage in development itself, whether to judge, to whom they are accountable, and how to balance rigour with rapid feedback.

LIMITS, IN OUR READING

- ☐ A widely discussed tension: focusing on primary intended users can narrow an evaluation to what powerful users want to know.
- ☐ Developmental evaluation can blur the line between evaluator and programme team, which raises questions about independence.

WHY THIS MATTERS TO YOU

Patton's question works for any report, proposal or piece of advice you will ever produce: who will use this, and for what? Ask it before you start, not after you finish. Find the people who genuinely care about the answer, involve them early and shape the work around the decisions they face. That is how a report becomes a tool instead of a ritual. His later ideas go further: when you are building something new in an unpredictable world, bring evidence into the team as you go, and when a problem is global, think at the scale of the planet. In our reading, his lesson is simple and freeing: work succeeds when people act on it, and you can design for that.

Did you know?

Besides evaluation texts, Patton wrote Grand Canyon Celebration: A Father-Son Journey of Discovery, a work of creative nonfiction that was a finalist for Minnesota Book of the Year.

5

editions of Utilization-Focused Evaluation, from 1978 to 2021

SAGE Encyclopedia of Evaluation (2005); Utilization-Focused Evaluation, Our Team; BRAC University Library catalogue (third edition, 1997)

17

steps in the utilization-focused evaluation checklist

Essentials of Utilization-Focused Evaluation, SAGE (2011), checklist chapter

12

in-depth developmental evaluation exemplars in the 2015 companion volume

Guilford Press, Developmental Evaluation Exemplars (2015)

600+

people in the worldwide Blue Marble Evaluation learning network reported in 2021

CECAN webinar (2021)

Key concepts

Primary intended users

The specific people, identified by name or role, who have a stake in an evaluation and the capacity to act on its findings, and around whom a utilization-focused evaluation is organised.

Personal factor

Patton's name for the finding that evaluations are more likely to be used when an identifiable person or group, other than the evaluator, cares about the evaluation and its results.

Intended use

The agreed purpose an evaluation is meant to serve for its users, such as programme improvement, a decision, accountability or knowledge building.

Developmental evaluation

An approach that supports innovation in complex, changing settings by embedding evaluative thinking and rapid feedback in the development process.

Principles-focused evaluation

Evaluation that treats guiding principles, rather than a standard intervention, as the object of judgement, asking whether principles are meaningful, followed and leading to desired results.

GLOCAL

A Blue Marble operating principle that links local action and evaluation to global patterns and trends, and the reverse.

Sources for this chapter: p. 579

THE BIG IDEAS OF THIS CHAPTER

1. **Utilization-focused evaluation** p. 212
   Judge an evaluation by whether specific people actually use it, and design every step around those people and the uses they intend.
2. **Developmental evaluation** p. 215
   When an innovation is still taking shape in a complex setting, embed evaluative thinking in the team so it can adapt on evidence in real time.
3. **Blue Marble Evaluation** p. 218
   Evaluate global problems at the scale at which they operate, connecting local action to the health of the whole planet.

GIANTS CONNECTED TO THIS CHAPTER



Carol H. Weiss

Her work on how research is used, often slowly, complements his focus on direct, intended use.

p. 152



Hallie Preskill

Evaluative inquiry extends use into organisational learning.

p. 226



Sonal Zaveri

Real-time evaluation shares developmental evaluation's commitment to feedback as events unfold.

p. 536

In one line

Michael Quinn Patton taught a whole profession that an evaluation is only as good as the difference it makes in someone's hands.

1

BIG IDEA 1 OF 3 · PATTON

Utilization-focused evaluation

Judge an evaluation by whether specific people actually use it, and design every step around those people and the uses they intend.

PICTURE IT

Think of a tailor. A tailor does not make a suit and then look for someone it fits. The tailor measures the person first. Utilization-focused evaluation is tailoring: it starts with the specific people who will use the findings and shapes every step to fit them.

The problem

By the 1970s programme evaluation in the United States had grown fast, yet a worrying pattern had emerged: technically sound reports often changed nothing. Findings arrived after decisions had been taken, answered questions nobody was asking, or were written for an abstract audience. For programmes, this meant money spent on evidence that never reached the people who could fix a failing service or expand a good one. For communities, it meant the same mistakes repeated. The field needed to know what made evaluation matter.

What it is

Utilization-focused evaluation (U-FE) is a decision framework rather than a single method or model. Its organising premise is that evaluations should be judged by their utility and actual use, so the evaluator facilitates the whole process with use in mind. It rests on two anchors. Primary intended users are specific, named people, not vague audiences, who care about the findings and can act on them; Patton grounded this in the personal factor, his research finding that use was most likely when an identifiable person cared about the evaluation.

Intended uses are what those users will do with the findings, such as improving a programme, making a summative judgement, ensuring accountability, supporting development or generating knowledge. Process use, the learning people gain simply by taking part, is a further benefit to plan for. Methods follow from these choices: quantitative, qualitative or mixed designs all fit if users find them credible and suited to the use. The evaluator's own effectiveness is judged by use too.

How it works

1. **Assess readiness**
   Check the commitment of the organisation and its funders, and your own competence and fit as evaluator; Essentials of Utilization-Focused Evaluation (SAGE, 2011) makes these the first two of its 17 steps.

2. **Identify and engage primary intended users**
   Find people who are interested, knowledgeable, credible and available, and form them into a working group that stays involved throughout.

3. **Analyse the situation together**
   Review past evaluation experience, politics, resources, deadlines, leadership support and risks with the users, and revisit this analysis as conditions change.

4. **Agree priority uses and questions**
   Choose among purposes such as formative, summative, accountability, developmental or knowledge generation, then let users select the few questions that matter most and link each to a use.

5. **Negotiate credible methods**
   Select methods that users will find credible and that are practical, ethical and affordable, and explain in plain terms any methods debates that affect the choice.

6. **Simulate use**
   Before collecting data, show users fabricated findings and rehearse what they would do with them; if no action would follow, rethink the questions or design.

7. **Interpret, report and follow up**
   Organise findings around the priority questions, involve users in interpretation and recommendations, deliver on time, then follow up and review how the evaluation was actually used.

WORKED EXAMPLE · AN ILLUSTRATION

A state education department in India commissions an evaluation of a foundational literacy programme running in 300 government schools. The evaluator asks who will act on the findings and names four people: the state programme director, two district education officers and the funder's programme lead. At the first meeting they explain that the real decision is whether to extend the programme to six more districts in the next budget cycle, due in February.

The evaluator therefore drops a planned three-year outcome study and designs a five-month study of implementation fidelity, cost per child and early reading gains across contrasting districts. Before fieldwork, the group reviews mock tables filled with invented numbers and agrees which results would change their decision. Findings are presented at a working session in January, not in a final report in June.

USE IT WHEN

- A clear decision or improvement process is waiting on the evidence.
- Identifiable people with authority to act are willing to engage over the life of the evaluation.
- Earlier evaluations in the organisation have gone unread or unused.

SKIP IT WHEN

- Nobody with the power to act is available or interested, so the approach has no anchor.
- The evaluation must be strictly independent of programme managers, unless independence can be built into how users are chosen and engaged.
- The main purpose is broad public knowledge with no particular user in view.

COMMON MISREADINGS

- Treating all stakeholders as primary intended users, which dilutes the approach into a long consultation list with no one accountable for use.
- Assuming a focus on use means lower rigour, when the approach asks for methods users find credible, which often demands more rigour, not less.

WHY IT WAS GROUNDBREAKING

It moved the test of a good evaluation from technical quality alone to whether real people used it. It rested on empirical research about use rather than exhortation, and it gave evaluators a concrete role as facilitators of decisions.

Origin

Patton set out the approach in the first edition of Utilization-Focused Evaluation (SAGE, 1978), drawing on 1970s research with colleagues into the use of health programme evaluations. A third edition, subtitled The New Century Text, followed in 1997, a fourth in 2008 and a fifth, with Charmagne Campbell-Patton, in 2021. Essentials of Utilization-Focused Evaluation (2011) condensed it into a 17-step checklist, and a 2013 IDRC-supported primer adapted it for development research in Asia.

RELATED IDEAS

Developmental evaluation · Process use

CW Weiss 152 · HP Preskill 226 · SZ Zaveri 536

In one line

Evaluate global problems at the scale at which they operate, connecting local action to the health of the whole planet.

2

BIG IDEA 2 OF 3 · PATTON

Developmental evaluation

When an innovation is still taking shape in a complex setting, embed evaluative thinking in the team so it can adapt on evidence in real time.

PICTURE IT

Think of a satnav on a road that is still being built. A fixed map would be useless, so the satnav recalculates as new roads open. Developmental evaluation travels inside an innovation team, helping it adjust its route as it learns.

The problem

Formative evaluation improves a model and summative evaluation judges it, but both assume a fairly stable model with settled goals. Social innovators often have neither: they are experimenting, their environment shifts, and success is still being defined. Forcing such work into fixed logframes and baseline and endline designs can freeze an idea too early, treat adaptation as deviation and deliver a verdict long after it could help. Innovators needed evaluation that kept pace with their learning.

What it is

Developmental evaluation (DE) supports the development of innovations and their adaptation in complex, dynamic conditions. Rather than testing a fixed model, the evaluator, usually working inside or alongside the team, helps innovators frame questions, surface assumptions, gather and interpret data quickly and decide what to change. Patton's 2010 book describes several purposes, among them the ongoing development of a programme, adapting effective principles to a new local context, generating innovations and taking them to scale, and rapid response in a crisis.

Developmental Evaluation Exemplars (2015) names eight essential principles: a developmental purpose; evaluation rigour; a utilisation focus; an innovation niche, meaning DE applies where something new is being developed; a complexity perspective; systems thinking; co-creation of the innovation and its evaluation; and timely feedback. The logic is that in complex situations cause and effect often become clear only as things unfold, so evaluation must help people learn as they go while keeping the discipline of testing ideas against evidence.

How it works

1. **Check the niche**
   Confirm that the initiative is genuinely innovating in a complex setting; where a stable model already exists, formative or summative evaluation fits better.
2. **Join the team**
   Agree a role close to the innovators, often attending planning meetings, and settle early how judgement and independence will be handled.
3. **Make assumptions visible**
   Map the hypotheses, values and systems the innovation rests on, using systems thinking to set boundaries and bring in different perspectives.
4. **Gather data in short cycles**
   Use bricolage, a term in Patton's 2010 book for assembling whatever methods fit the question at hand, from quick surveys to interviews, observation and administrative data.
5. **Hold sense-making sessions**
   Bring findings back promptly and help the team interpret them and decide what to keep, drop or change, in line with the principle of timely feedback.
6. **Document the development**
   Record decisions, the evidence behind them and how the innovation evolved, so learning is not lost and later accountability is possible.
7. **Revisit the purpose**
   As the innovation stabilises, consider whether to move to formative or summative evaluation.

WORKED EXAMPLE · AN ILLUSTRATION

A Karnataka NGO is testing ways to bring adolescent girls who have left school back into learning. It has three untested ideas: community learning camps, mentoring by college students and a mobile phone curriculum. An embedded evaluator joins the core team. Each month she pulls enrolment and attendance records, conducts a handful of phone interviews with girls and mothers, and observes two camps.

At a monthly two-hour session the team reviews the evidence. By month four the data show that phone lessons reach few girls because the handsets belong to fathers, while camps held near home draw steady attendance. The team drops the phone stream, redesigns mentoring around the camps and records why. After eighteen months, with a more stable model, the NGO commissions a formative evaluation.

USE IT WHEN

- An innovation is being developed in a complex, fast-changing situation where goals are still emerging.
- The team is willing to adapt in response to evidence and to have an evaluator in the room.
- A crisis response, or the adaptation of proven principles to a very different context, needs rapid learning.

SKIP IT WHEN

- A stable, well-specified model is ready for a summative test of its effects.
- The funder requires only fixed indicators and an independent end-of-project verdict.
- There is no budget, time or appetite for frequent sense-making with the team.

COMMON MISREADINGS

- Using developmental evaluation as a label for having no evaluation plan, or for loose monitoring without rigour.
- Assuming it avoids judgement, when the evaluator still asks hard evaluative questions about what is and is not working.

WHY IT WAS GROUNDBREAKING

It gave evaluation a legitimate role in the early, uncertain life of innovations, where fixed indicators and end-of-project verdicts fit poorly. It recast the evaluator as a partner in learning while keeping the discipline of testing ideas against evidence.

Origin

Patton first used the term in a 1994 article in Evaluation Practice. He linked it to complexity and social innovation in Getting to Maybe (2006), written with Frances Westley and Brenda Zimmerman, and gave it book length in Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use (Guilford, June 2010). Developmental Evaluation Exemplars (Guilford, 2015), with Kate McKegg and Nan Wehipeihana, added twelve cases and the eight essential principles.

RELATED IDEAS

Utilization-focused evaluation

Blue Marble Evaluation

PR Rogers 268

SS Sridharan 246

Complexity-aware evaluation

CONCEPT HIGHLIGHT

Intended use

The agreed purpose an evaluation is meant to serve for its users, such as programme improvement, a decision, accountability or knowledge building.

3

BIG IDEA 3 OF 3 · PATTON

Blue Marble Evaluation

Evaluate global problems at the scale at which they operate, connecting local action to the health of the whole planet.

PICTURE IT

Think of the famous photograph of Earth from space, a single blue marble with no borders. Rivers, climate and markets cross every boundary. Blue Marble Evaluation asks evaluators to see local projects as part of that one connected system.

The problem

Evaluation grew up around single projects, single funders and national borders. Yet climate change, pandemics, migration and food systems cross every boundary. A project can meet all its indicators while the river basin, forest or climate it depends on keeps deteriorating. Separate national evaluations of linked efforts rarely add up to a judgement on whether the world is moving towards sustainability. Practitioners lacked a framework for judging transformation at the scale of the problems themselves.

What it is

Blue Marble Evaluation, named after the photograph of the whole Earth taken on 7 December 1972 during the last Apollo mission, is a principles-based approach for evaluating efforts at global systems transformation. Patton's 2019 book organises it as principles in three tiers. Four overarching principles frame the approach: global thinking, treating the world as an interconnected whole; the Anthropocene as context, recognising that human activity now shapes planetary systems; transformative engagement, judging whether change is transformational rather than merely incremental; and integration, applying the principles together. Operating principles guide practice, including GLOCAL, which links local and global levels in both directions; cross-silos, which works across sectors and disciplines; time being of the essence, which brings urgency into evaluation; and bricolage methods, which assemble whatever methods fit. A third set addresses how to judge global systems transformation itself. The underlying logic is principles-focused: in a turbulent world, guiding principles travel across contexts better than fixed models do.

How it works

1. **Set a wider frame**
   Place the intervention in its global and Anthropocene context, asking which planetary systems and global trends it touches; these steps are our ordering of the principles in Patton's 2019 book.
2. **Trace GLOCAL links**
   Map how local actions connect to regional and global patterns, and how global forces shape local results.
3. **Work across silos**
   Bring in perspectives from other sectors, disciplines and countries working on the same system, instead of evaluating each project alone.
4. **Judge transformation, not only outputs**
   Ask whether the combined efforts are shifting the system at the depth and speed the problem requires.
5. **Assemble methods by bricolage**
   Combine methods suited to the scale, such as systems mapping, satellite or administrative data, case studies and community knowledge.
6. **Report with urgency**
   Fit timelines to the pace of the problem, so that findings reach decision-makers while they can still change course.

WORKED EXAMPLE · AN ILLUSTRATION

A consortium of NGOs and a state government runs groundwater recharge projects across 40 villages in a drought-prone district of the Deccan plateau. Each project already reports on check dams built and wells recharged. A Blue Marble lens asks wider questions. Are gains in one village drawing water from its neighbours? How do new sugarcane contracts from distant markets, and shifting monsoon patterns, affect the shared aquifer? The evaluation team brings together hydrologists, farmers' groups and the agriculture department, combines satellite data on cropping with village water budgets, and judges the effort on whether the aquifer as a whole is stabilising. Findings are shared before the next sowing season, so that cropping advisories can change in time.

CONCEPT HIGHLIGHT

Primary intended users

The specific people, identified by name or role, who have a stake in an evaluation and the capacity to act on its findings, and around whom a utilization-focused evaluation is organised.

USE IT WHEN

- An initiative addresses cross-border or planetary problems such as climate, water, food systems or migration.
- Funders or networks want to judge whether many projects add up to systems change.
- A programme's success depends on ecological or economic systems beyond its own boundary.

SKIP IT WHEN

- The question is tightly bounded, such as whether a clinic's waiting times improved.
- No partners beyond the single project can be engaged.
- The commissioner needs only a quick, standard compliance review.

COMMON MISREADINGS

- Thinking it applies only to global organisations, when the GLOCAL principle is meant for local projects too.
- Treating it as a new method, when it is a set of principles that guide the choice and combination of methods.

WHY IT WAS GROUNDBREAKING

It asked evaluators to treat the planet and its interlinked systems as a unit of judgement, not only the project or the nation. It pushed sustainability and long time horizons into the core of evaluative reasoning.

Origin

According to Patton's Utilization-Focused Evaluation organisation, the idea emerged after a presentation he gave at IPDET in mid-2015 and was launched with a pilot grant later that year. It built on his Principles-Focused Evaluation (2018). Guilford published Blue Marble Evaluation: Premises and Principles at the end of 2019, and a 2021 CECAN webinar reported a worldwide learning network of more than 600 people.

RELATED IDEAS

Developmental evaluation

Principles-focused evaluation

ZO Ofir 454

PR Rogers 268

CONCEPT HIGHLIGHT

Personal factor

Patton's name for the finding that evaluations are more likely to be used when an identifiable person or group, other than the evaluator, cares about the evaluation and its results.

RELATED GIANTS



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Hallie Preskill

The educator who reframed evaluation as a way for organisations to learn, starting from what already works.

p. 226



Sonal Zaveri

The Indian evaluator who carried participatory, feminist practice from grassroots work into regional networks and real-time evaluation.

p. 536



Patricia J. Rogers

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

p. 268

CONCEPT HIGHLIGHT

Developmental evaluation

An approach that supports innovation in complex, changing settings by embedding evaluative thinking and rapid feedback in the development process.

CONCEPT HIGHLIGHT

Principles-focused evaluation

Evaluation that treats guiding principles, rather than a standard intervention, as the object of judgement, asking whether principles are meaningful, followed and leading to desired results.

CONCEPT HIGHLIGHT

GLOCAL

A Blue Marble operating principle that links local action and evaluation to global patterns and trends, and the reverse.

Put it to work

USE IT ON MONDAY · UTILIZATION-FOCUSED EVALUATION

1. Name the primary intended users as people, not organisations: the district coordinator, the programme director, the funder's portfolio lead.

2. Ask each of them what they would do differently depending on what the evaluation finds. If nobody can answer, rethink the questions before collecting any data.

3. Before fieldwork, show users a mock report with empty tables and charts, and ask whether those results, once filled in, would help them decide anything.

AI prompt **Copy, fill the brackets, then check every line**

Help me design a utilization-focused evaluation, in the tradition of Michael Quinn Patton, for **[programme]** commissioned by **[organisation]**. The people who may use the findings are **[list names or roles and what I know about each]**. The current draft evaluation questions are **[paste]**. (1) For each potential user, suggest what uses the evaluation might serve and what decisions it could inform, flagged as hypotheses to check with them. (2) Draft interview questions I can ask each user to confirm intended uses. (3) Show which draft evaluation questions do not clearly connect to any use. (4) Suggest reporting formats and timing for each user.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a tailor. A tailor does not make a suit and then look for someone it fits. The tailor measures the person first. Utilization-focused evaluation is tailoring: it starts with the specific people who will use the findings and shapes every step to fit them.

YOUR MOVES Three moves from Patton's path

1 Own one signature question

Patton kept returning to one question, whether evaluations get used, from the first edition of Utilization-Focused Evaluation in 1978 to the fifth in 2021.

TRY THIS

Finish this sentence and use it on your profile: the question I help organisations answer is (your question).

2 Gather evidence about your own practice

He treated a frustration of his field as a research problem, studying 20 health programme evaluations before proposing a solution.

TRY THIS

Look back at your last five evaluations and track which recommendations were acted on. Let the pattern change how you work.

3 Go independent without going small

Leaving a university post in 1991 for independent practice did not narrow his reach; keynotes in Nairobi in 1999 and at the European Evaluation Society in 2000 followed.

TRY THIS

Share one lesson from your practice publicly each month, in a post, talk or short paper.

THE BIG IDEAS OF THIS CHAPTER

1 **Utilization-focused evaluation** p. 212

Judge an evaluation by whether specific people actually use it, and design every step around those people and the uses they intend.

2 **Developmental evaluation** p. 215

When an innovation is still taking shape in a complex setting, embed evaluative thinking in the team so it can adapt on evidence in real time.

3 **Blue Marble Evaluation** p. 218

Evaluate global problems at the scale at which they operate, connecting local action to the health of the whole planet.

CONCEPT HIGHLIGHT

GLOCAL

A Blue Marble operating principle that links local action and evaluation to global patterns and trends, and the reverse.

GIANTS CONNECTED TO THIS CHAPTER



Carol H. Weiss

Her work on how research is used, often slowly, complements his focus on direct, intended use.

p. 152



Hallie Preskill

Evaluative inquiry extends use into organisational learning.

p. 226



Sonal Zaveri

Real-time evaluation shares developmental evaluation's commitment to feedback as events unfold.

p. 536

KEY WORKS TO READ IN THE ORIGINAL

- **Utilization-Focused Evaluation** 1978, First edition; fifth edition 2021 with Charmagne E. Campbell-Patton
- **Getting to Maybe** 2006, With Frances Westley and Brenda Zimmerman
- **Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use** 2010 or 2011, Sources differ on the year

CONCEPT HIGHLIGHT

Primary intended users

The specific people, identified by name or role, who have a stake in an evaluation and the capacity to act on its findings, and around whom a utilization-focused evaluation is organised.

CONCEPT HIGHLIGHT

Personal factor

Patton's name for the finding that evaluations are more likely to be used when an identifiable person or group, other than the evaluator, cares about the evaluation and its results.

WORKSHEET · TRY THIS TEMPLATE

Intended use agreement

Print this page, or copy the headings into your notes.

1 Primary intended users

Names and roles of the few people who will act on findings.

2 Intended uses

For each user: improve, decide, account, learn or something else.

3 Decisions and timing

Which decisions, and the date findings are needed by.

4 Questions that matter to them

Three to five, in their words, agreed with them.

5 Reporting moment and format

When and how they want findings: workshop, memo, dashboard.

6 Follow-up

Who checks, and when, whether the findings were used.

REFLECT

If your report vanished tomorrow, who would notice, and what would they be unable to do?

CONNECTS TO

Carol H. Weiss (CW)
Her work on how research is used, often slowly, complements his focus on direct, intended use. p. 152

Hallie Preskill (HP)
Evaluative inquiry extends use into organisational learning. p. 226

Sonal Zaveri (SZ)
Real-time evaluation shares developmental evaluation's commitment to feedback as events unfold. p. 536

10



Emblem: an original HubForge design for this figure, not a likeness.

Hallie Preskill

Active 1990s to 2010s

United States; health, nonprofit, education, foundation, government and corporate sectors

Use and learning

The educator who reframed evaluation as a way for organisations to learn, starting from what already works.

IN ONE LINE

Hallie Preskill turned evaluation from a verdict delivered at the end into a habit of learning practised every day.

Most organisations treat evaluation like a visit to the dentist: it happens now and then, often when something has gone wrong, it is done by an outsider, and everyone is relieved when it is over. In 1999 Hallie Preskill and Rosalie Torres put that problem into words. Evaluation in most organisations, they wrote, was episodic, event-driven, often triggered by crisis, carried out by outsiders and disconnected from daily work. Their alternative, evaluative inquiry, asked organisations to make learning from evidence part of ordinary work. Preskill came to this from an unusual direction: a degree in Spanish, a master's in elementary and special education, and work in private industry on training and evaluation before her doctorate. She went on to lead the American Evaluation Association in 2007, and in 2009 carried her ideas into philanthropy at FSG.

Life CHAPTER 10

Hallie Preskill was born in New York City in 1953. Her education took an unusual route into evaluation. She earned a BA in Spanish at California State University, Fullerton, an MS in elementary and special education at Long Island University, a Certificate of Advanced Study in Evaluation and Policy Research at the University of Vermont, and a PhD in education at the University of Illinois at Urbana-Champaign. Before her academic career she worked in private industry in training and evaluation.

She was an associate professor in the School of Education at the University of St. Thomas and, from 1994, a professor in the College of Education at the University of New Mexico. Over more than 20 years in universities she taught graduate courses in programme evaluation, training design and development, organisational learning, appreciative inquiry and consulting. Her research centred on

BIG IDEAS

1. **Evaluative inquiry for organisational learning**
2. **Appreciative inquiry in evaluation**
3. **Evaluation capacity building**

KEY WORKS

Evaluative Inquiry for Learning in Organizations
1999 · With Rosalie T. Torres

Evaluation in Organizations: A Systematic Approach to Enhancing Learning, Performance and Change
2001 · With Darlene Russ-Eft; second edition 2009

Reframing Evaluation through Appreciative Inquiry
2006 · With Tessie Catsambas

evaluation capacity building, the transfer of learning from training, evaluation use, and evaluation as a catalyst for learning by individuals, teams and organisations. Her books from this period include Evaluative Inquiry for Learning in Organizations (1999, with Rosalie Torres), Evaluation in Organizations (2001 and 2009, with Darlene Russ-Eft), Evaluation Strategies for Communicating and Reporting (2005, with Torres and M. Piontek) and Reframing Evaluation through Appreciative Inquiry (2006, with Tessie Catsambas).

Recognition followed. The American Evaluation Association gave her its Alva and Gunnar Myrdal award for evaluation practice in 2002, an award its list shows as shared with James Altschuld; the University of Illinois named her a distinguished alumna in 2004; and she served as AEA president in 2007. In 2009 she left academia for FSG, a nonprofit strategy, evaluation and research consultancy, as a managing director leading its Strategic Learning and Evaluation practice; FSG has since listed her as a senior adviser. Her clients there have included the Kresge, MasterCard, Knight, Packard, Robert Wood Johnson, W.K. Kellogg and Bill and Melinda Gates foundations.

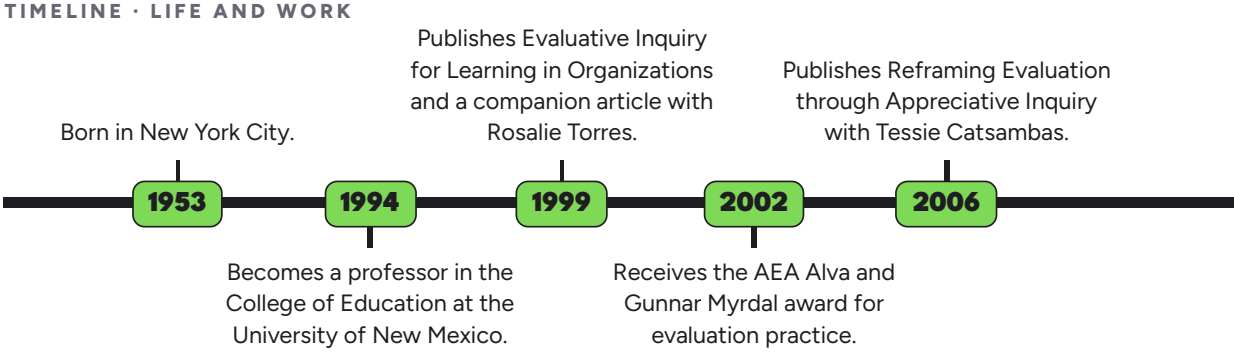
The spark

Her early work in private industry, on training and evaluation, sat where two fields met. Evaluation had methods for gathering and judging evidence. Organisational learning had ideas about how people and teams change the way they work. In our reading, the two had grown up largely apart through the 1990s. Her research on how learning transfers from training into practice, on how evaluations get used and on evaluation as a spark for learning set out to bring them together.

The breakthrough CHAPTER 10

Think about how evaluation usually happened in the 1990s. A study was commissioned near the end of a project. Outsiders arrived, collected data and left. A report was delivered, and the organisation moved on. In our summary

TIMELINE · LIFE AND WORK



of the context, the lessons rarely became part of how people worked the next day. Evaluators had methods for gathering and judging evidence; writers on organisational learning described how teams change their practice. What was missing was a way to make careful inquiry part of how an organisation learns, rather than something done to it from outside at the end.

Preskill and Rosalie Torres joined them in 1999, in an article in Evaluation 5(1), Building capacity for organizational learning through evaluative inquiry, and in the SAGE book Evaluative Inquiry for Learning in Organizations. They described evaluative inquiry as a practice in which members of an organisation, with stakeholders and clients, continually examine what they do and how they do it, built into daily work and aimed at learning and improved performance. In June 2006, with Tessie Catsambas, she extended the argument in Reframing Evaluation through Appreciative Inquiry, presented by its publisher as the first book to apply that strengths-based change approach to evaluation.

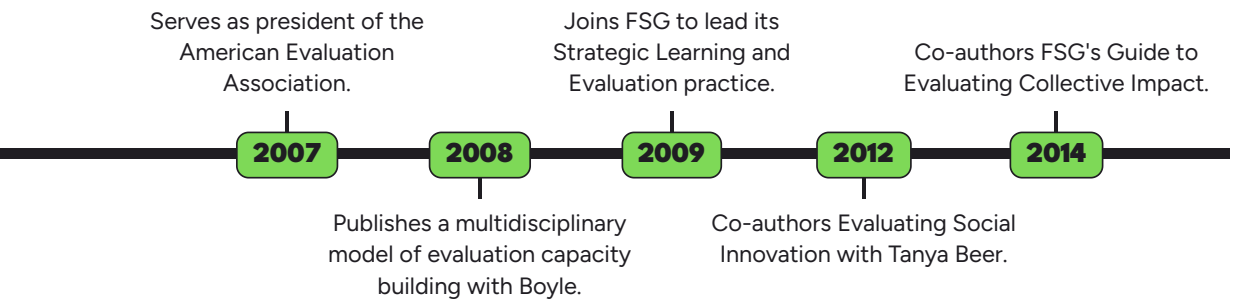
The problem it solved

Evaluations produced reports but little organisational learning, and deficit-focused evaluations left stakeholders defensive and less willing to use findings.

Impact WHAT CHANGED

Preskill's ideas shaped how many foundations and nonprofits in the United States think about evaluation. At FSG she applied them to philanthropy. In 2012, with Tanya Beer, she wrote Evaluating Social Innovation for the Center for

TIMELINE, CONTINUED



Evaluation Innovation, arguing that traditional formative and summative evaluation can slow innovative initiatives and that developmental evaluation better supports adaptation. In 2014, with Marcie Parkhurst and Jennifer Splansky Juster, she wrote FSG's Guide to Evaluating Collective Impact, stressing that design depends on an initiative's maturity and context and that continuous learning is essential.

Her clients have included the MasterCard Foundation, and her appreciative inquiry book draws on international case examples. Our interpretation is that her influence reaches MEL practitioners everywhere through the now common idea that the point of evaluation is what an organisation learns and does next. Learning reviews, after-action sessions and MEL teams framed as learning partners all echo evaluative inquiry, and her activity books remain practical resources for trainers building evaluation skills in NGOs and government.

WHERE PEOPLE DISAGREE

Debates

The publisher's description of Evaluative Inquiry for Learning in Organizations notes that it concentrates on single organisations rather than large multi-site programmes, a limit for those evaluating national schemes. Preskill and Beer's 2012 brief itself acknowledged tensions between innovation, accountability, rigour and adaptation. In our reading, a strengths-based frame can sit uneasily with accountability evaluations that must report shortfalls clearly, and learning-oriented evaluation depends on an organisational culture willing to question itself, which outsiders cannot simply install.

LIMITS, IN OUR READING

- ☐ Her evaluative inquiry book is aimed at single organisations and is not designed for large multi-site programmes, according to its publisher's description.
- ☐ In our view, a strengths-based frame can sit uneasily with accountability evaluations that must report shortfalls clearly.

WHY THIS MATTERS TO YOU

Preskill's ideas work in any team, not only in evaluation. Instead of saving all the reflection for one big review at the end, build small moments of learning into ordinary work: ask what happened, why and what you will do differently. When people feel defensive, start with what is working at its best, because that builds the energy needed to change. And whenever you help with an evaluation, aim to leave skills behind, so the team can keep learning after you have gone. These habits turn evaluation from a report people fear into a practice people use.

6

books authored or co-authored, from Evaluative Inquiry for Learning in Organizations (1999) to Becoming the Change (2009)

FSG biography

89

structured activities in the 2015 second edition of Building Evaluation Capacity, up from 72 in 2005

SAGE (2015); FSG biography

25+

years providing evaluation, organisational learning and training services across health, nonprofit, education, foundation, government and corporate sectors

FSG biography

Key concepts

Evaluative inquiry

Ongoing inquiry by members of an organisation, with stakeholders, into what they do and how, built into daily work and aimed at learning and better performance.

Dialogue and reflection

Two of the four learning processes of evaluative inquiry: structured conversation to build shared understanding, and stepping back to consider experience and its meaning.

Appreciative inquiry

An approach to organisational change that focuses on strengths and what works rather than on weaknesses and problems.

Evaluation capacity building

Deliberate work to strengthen an organisation's ability to design, conduct and use evaluation.

Developmental evaluation

An approach that supports continuous adaptation in innovative initiatives, as an alternative to fixed formative and summative designs.

Strategic learning

Using evaluation and data to inform an organisation's ongoing strategic decisions rather than only to judge past performance.

Sources for this chapter: p. 580

Did you know?

Her first degree was a BA in Spanish, followed by a master's in elementary and special education, before she trained in evaluation.

1

BIG IDEA 1 OF 3 · PRESKILL

Evaluative inquiry for organisational learning

Make evaluation an ongoing, collaborative practice inside the organisation, built into daily work, so that people continually learn from what they do and change it.

PICTURE IT

Think of the difference between a once-a-year health check and a fitness tracker you glance at every day. The annual check tells you something important, once. The tracker makes noticing and adjusting part of daily life. Evaluative inquiry is the tracker built into how an organisation works.

The problem

In the 1990s evaluation was usually episodic and event-driven: triggered by a perceived crisis, carried out by outside consultants or a detached unit, focused on products and summative outcomes, and delivered as a report. Organisations had plenty of data but little timely, meaningful interpretation, and the step of applying learning was often neglected. Writers on organisational learning described how teams change, but rarely connected this to systematic inquiry. Staff experienced evaluation as something done to them, and lessons rarely reached practice.

What it is

Evaluative inquiry for organisational learning, set out by Preskill and Rosalie Torres in 1999, is an ongoing process in which members of an organisation, often with trained evaluators, study their processes and outcomes as part of daily work in order to learn and improve performance. It also questions the value of what the organisation does and why. It rests on four learning processes that overlap rather than follow in sequence: dialogue; reflection; asking questions; and identifying and clarifying values, beliefs, assumptions and knowledge.

These run through three phases: focusing the inquiry, carrying out the inquiry, and applying learning. The whole depends on an organisational infrastructure with four elements: a culture of trust that tolerates risk and learns from mistakes; leadership that models learning; communication that shares information openly and for learning rather than power; and systems and structures that encourage collaboration across units. Inquiry can focus on any unit, from a team to the whole organisation.

How it works

1. **Check the infrastructure**
   Assess culture, leadership, communication, and systems and structures, and address the barriers most likely to block learning.
2. **Focus the inquiry**
   With the team, define the issue, identify stakeholders, agree guiding questions and clarify intended users and uses.
3. **Carry out the inquiry**
   Design the study, collect and analyse data and interpret findings in context, with dialogue and reflection running throughout.
4. **Surface assumptions**
   Use dialogue and questions to bring values, beliefs and assumptions into the open, including issues people usually avoid discussing.
5. **Apply learning**
   Weigh options for action, write and carry out an action plan and monitor progress; Preskill and Torres note that this phase is often neglected.
6. **Make it continuous**
   Build the cycle into routine meetings and planning so that inquiry becomes part of how the organisation works.

WORKED EXAMPLE · AN ILLUSTRATION

A Bengaluru-based NGO runs after-school learning centres in 60 urban settlements. Instead of waiting for an end-line report, its MEL lead sets up evaluative inquiry with programme staff. In the focusing phase, the team chooses one question: why attendance at some centres falls sharply after the first three months. In the inquiry phase, field coordinators review attendance data, visit centres with high and low attendance and talk with children and parents. In monthly dialogue sessions, the team surfaces an assumption no one had questioned: that evening timings suit all families. In the applying phase, three centres trial earlier timings and a parent volunteer system, and the team tracks the results. Leadership commits to sharing failures openly, and the inquiry becomes a standing item in the monthly review.

CONCEPT HIGHLIGHT

Evaluative inquiry

Ongoing inquiry by members of an organisation, with stakeholders, into what they do and how, built into daily work and aimed at learning and better performance.

USE IT WHEN

- An organisation wants staff to use evidence routinely, not only at end-line.
- An internal MEL team is supporting programme improvement.
- Leaders are willing to have their own assumptions questioned.

SKIP IT WHEN

- An independent, summative judgement is required for accountability.
- The culture punishes mistakes and leaders will not support open inquiry, unless that infrastructure is addressed first.
- Large multi-site schemes spanning many organisations, which the book's publisher notes it does not focus on.

COMMON MISREADINGS

- [ ] Equating it with holding more meetings, without systematic data collection and analysis.
- [ ] Assuming learning happens once findings are shared, when the applying learning phase requires planned action and follow-up.

WHY IT WAS GROUNDBREAKING

It relocated evaluation from the end of a project to the centre of how an organisation works. It gave evaluators a theory of learning to match their methods of inquiry, and gave organisations a way to own evaluation rather than endure it.

Origin

Preskill and Torres published the framework in 1999, in Building capacity for organizational learning through evaluative inquiry, Evaluation 5(1), pages 42 to 60, and in the SAGE book Evaluative Inquiry for Learning in Organizations. Preskill extended the learning focus in Evaluation in Organizations with Darlene Russ-Eft (2001, second edition 2009), and from 2009 carried it into philanthropy through FSG's strategic learning and evaluation practice.

RELATED IDEAS

MP Patton 204 · LC Cronbach 38 · RR Rist 410

CONCEPT HIGHLIGHT

Dialogue and reflection

Two of the four learning processes of evaluative inquiry: structured conversation to build shared understanding, and stepping back to consider experience and its meaning.

2

BIG IDEA 2 OF 3 · PRESKILL

Appreciative inquiry in evaluation

Start an evaluation from what works at its best and what people value, using the 4-I cycle of inquire, imagine, innovate and implement, to build the energy needed to use findings.

PICTURE IT

Think of a sports coach reviewing a match. One coach starts with every mistake. Another starts by replaying the team's best moves and asking how to do more of them. Both learn, but the second team walks out ready to try. Appreciative inquiry is the second coach.

The problem

Many evaluation approaches begin from deficits: what went wrong, what is missing, who failed. Staff and partners often met evaluators defensively, held back stories and resisted findings that felt like blame. In complex, changing programmes, hunting for problems alone also missed the conditions behind success, which are often what managers most need to replicate. Evaluators lacked a structured way to engage stakeholders around strengths without losing rigour, and organisations lost the motivation that taking part in an evaluation could create.

What it is

Appreciative inquiry is an organisational development and change approach that focuses on strengths rather than weaknesses. In Reframing Evaluation through Appreciative Inquiry (SAGE, 2006), Preskill and Tessie Tzavaras Catsambas applied it to evaluation, in what the publisher presented as the first book to do so. They showed how it can be used to focus an evaluation, design interviews and surveys, develop evaluation systems and build evaluation capacity, with case examples from the United States and international settings.

Its process is a four-phase cycle. Preskill and Catsambas relabelled the widely used 4-D model of Discover, Dream, Design and Destiny as the 4-I model: Inquire, identifying peak moments and the conditions behind them; Imagine, building a shared picture of a preferred future; Innovate, deciding how structures, culture and practices should change, often written as possibility statements; and Implement, acting, monitoring and adjusting. The aim is not to ignore problems but to reduce defensiveness and increase use.

How it works

1. **Reframe the focus**
   Turn deficit questions into affirmative topics, for example from why attendance is low to when attendance has been highest and why.
2. **Inquire**
   Hold paired interviews in which participants describe a peak experience of the programme and the conditions that made it possible, then identify common themes.
3. **Imagine**
   In small groups, build a shared picture of what the programme would look like if those peak moments were the norm.
4. **Innovate**
   Write possibility statements describing how structures, practices and culture should change to realise that picture.
5. **Implement**
   Choose which statements to act on, plan actions and monitor progress, using appreciative evaluation tools to guide course corrections.
6. **Keep accountability in view**
   As our addition, include direct questions about shortfalls and concerns so that the evaluation still reports clearly what is not working.

WORKED EXAMPLE · AN ILLUSTRATION

A state health mission wants to understand why some primary health centres in Maharashtra sustain high immunisation coverage while others struggle. Rather than starting with an audit of failures, the evaluator uses the 4-I model. In the Inquire phase, auxiliary nurse midwives and ASHA workers from stronger and weaker centres pair up and interview each other about a time immunisation sessions went especially well. Themes emerge: reliable vaccine delivery, a medical officer who joined outreach days, and reminders passed through women's groups. In Imagine and Innovate workshops, staff draft possibility statements, such as every session being confirmed with the community a day ahead. The final report presents these conditions alongside administrative data on coverage and a frank section on stock-outs and staff vacancies, so the focus on strengths does not hide problems.

USE IT WHEN

- Stakeholders are defensive after earlier deficit-focused evaluations.
- The aim is to identify and replicate the conditions behind success.
- Building an evaluation system or capacity with staff who need to own it.

SKIP IT WHEN

- Investigating suspected misconduct, harm or serious failure, where direct, problem-focused inquiry is essential.
- Accountability evaluations that must report shortfalls prominently, unless problem-focused questions are added.

COMMON MISREADINGS

- ☐ Treating it as positive thinking that leaves problems unreported, when its aim is to reduce defensiveness, not to hide shortfalls.
- ☐ Adding one appreciative question to an otherwise standard survey and calling the whole evaluation appreciative.

WHY IT WAS GROUNDBREAKING

According to its publisher, it brought a change method from organisational development into evaluation practice in book form for the first time. It reframed stakeholder engagement as a source of energy for change rather than a compliance step. In our reading, it also showed that studying success carefully is a rigorous way to learn about the conditions behind it.

Origin

In 2003 Preskill and Coghlan edited Using Appreciative Inquiry in Evaluation, number 100 of New Directions for Evaluation. In June 2006 Preskill and Catsambas published Reframing Evaluation through Appreciative Inquiry, a 192-page SAGE book, relabelling the 4-D cycle associated with Cooperrider and Whitney as 4-I. A 2015 UNITAR guide to appreciative inquiry, developed largely by Preskill and Arani Grindle, sets out both sets of labels for trainers.

RELATED IDEAS

D&D Davies and Dart 362 · RC Chambers 296 · MP Patton 204

RELATED GIANTS



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Ray C. Rist

The World Bank evaluator who turned results-based M&E into ten steps a government could follow, and helped build a school to teach them.

p. 410



Rick Davies and Jess Dart

Two practitioners who turned the stories people tell about change into a disciplined monitoring method.

p. 362

3

BIG IDEA 3 OF 3 · PRESKILL

Evaluation capacity building

Treat an organisation's ability to do and use evaluation as something to design deliberately, with clear objectives, chosen teaching and learning strategies and the conditions needed for practice to last.

PICTURE IT

Think of teaching someone to fish rather than handing them a fish. An evaluation done for an organisation feeds it once. Evaluation capacity building leaves behind the skills, systems and habits for it to keep evaluating itself.

The problem

By the 2000s many organisations ran evaluation workshops, hired consultants and spoke of building a culture of evaluation, but without any overarching model of how capacity is built. Training rarely transferred to daily work, skills left with the consultant, and no one could say which approaches worked. For NGOs and public agencies this meant repeated spending on capacity that did not stick, and continued dependence on outsiders for every evaluation they needed.

What it is

Evaluation capacity building (ECB), as defined by Preskill and S. Boyle in the American Journal of Evaluation 29 (2008), is the intentional design and implementation of teaching and learning strategies that help individuals, groups and organisations learn what constitutes effective, useful and professional evaluation practice. Its goal is sustainable evaluation practice, in which members routinely ask questions that matter, collect, analyse and interpret data, and use findings for decisions and action. Their multidisciplinary model identifies ten teaching and learning strategies, ranging from internships to coaching or mentoring and including communities of practice. It stresses transfer of learning: people need leadership support, incentives, resources and opportunities to apply what they learn in their daily work. Sustainability also needs systems, processes, policies and plans that embed evaluation in how the organisation pursues its mission and strategic goals. The model was designed both to guide practice and to support research on ECB.

How it works

1. **Set ECB objectives**
   Decide what knowledge, skills and attitudes people need, and for which evaluation tasks.
2. **Assess the conditions for transfer**
   Examine the leadership support, culture, incentives, resources and systems that will help or hinder people applying what they learn.
3. **Choose strategies**
   Select a mix from the range of teaching and learning strategies in the 2008 model, from internships to communities of practice and coaching or mentoring, drawing on structured activities such as those in Building Evaluation Capacity.
4. **Design for transfer**
   Plan how people will apply new skills in daily work, with time, incentives and visible support from leaders.
5. **Embed in systems**
   Build evaluation into policies, plans and processes so that practice outlasts individual staff and consultants.
6. **Evaluate the capacity building itself**
   Track whether people use the skills and whether evaluation practice is sustained, adding to the evidence on what works.

WORKED EXAMPLE · AN ILLUSTRATION

A network of 20 small NGOs in Rajasthan wants to evaluate its own programmes rather than depend on consultants. Its coordinator designs capacity building around the 2008 model. Objectives are specific: each NGO should be able to write evaluation questions, design a simple survey and run a learning review. Strategies are mixed: a short training using structured activities, a community of practice that meets monthly online, and coaching by an experienced evaluator as each NGO plans a real evaluation. For transfer, each NGO's director commits staff time and discusses results at board meetings. To embed the practice, the network adds an evaluation plan to its standard project template. After a year, the coordinator reviews which NGOs completed evaluations without outside help and which strategies participants found most useful.

CONCEPT HIGHLIGHT

Appreciative inquiry

An approach to organisational change that focuses on strengths and what works rather than on weaknesses and problems.

USE IT WHEN

- An organisation aims to reduce its dependence on external evaluators.
- Training for MEL staff has not changed day-to-day practice.
- Networks or government departments are planning multi-year investments in evaluation capacity.

SKIP IT WHEN

- A one-off evaluation where no one will need the skills afterwards.
- Leaders will not commit time, incentives or resources, so transfer cannot happen.

COMMON MISREADINGS

- Equating capacity building with running workshops, when the model stresses transfer of learning and embedded systems.
- Measuring success by the number of people trained rather than by sustained evaluation practice.

WHY IT WAS GROUNDBREAKING

It moved capacity building from a vague aspiration to a designed and researchable process. It named the conditions, from leadership support to embedded systems, that decide whether training turns into practice. Its companion activity books put practical teaching tools into the hands of trainers worldwide.

Origin

Preskill's ECB work grew alongside her teaching: Building Evaluation Capacity: 72 Activities for Teaching and Training, with Darlene Russ-Eft, appeared in 2005, and a second edition with 89 activities in 2015. In September 2008 Preskill and Boyle published A multidisciplinary model of evaluation capacity building in the American Journal of Evaluation 29, pages 443 to 459. Later ECB research, including a 2021 New Directions for Evaluation paper by Hargraves and colleagues, builds on their definition.

RELATED IDEAS

RR Rist 410 · MP Patton 204 · LC Cronbach 38

CONCEPT HIGHLIGHT

Evaluation capacity building

Deliberate work to strengthen an organisation's ability to design, conduct and use evaluation.

CONCEPT HIGHLIGHT

Developmental evaluation

An approach that supports continuous adaptation in innovative initiatives, as an alternative to fixed formative and summative designs.

CONCEPT HIGHLIGHT

Strategic learning

Using evaluation and data to inform an organisation's ongoing strategic decisions rather than only to judge past performance.

THE BIG IDEAS OF THIS CHAPTER

1 **Evaluative inquiry for organisational learning** p. 232

Make evaluation an ongoing, collaborative practice inside the organisation, built into daily work, so that people continually learn from what they do and change it.

2 **Appreciative inquiry in evaluation** p. 235

Start an evaluation from what works at its best and what people value, using the 4-I cycle of inquire, imagine, innovate and implement, to build the energy needed to use findings.

3 **Evaluation capacity building** p. 238

Treat an organisation's ability to do and use evaluation as something to design deliberately, with clear objectives, chosen teaching and learning strategies and the conditions needed for practice to last.

Did you know?

Before entering academia she worked in private industry on training and evaluation.

Put it to work

USE IT ON MONDAY · EVALUATIVE INQUIRY FOR ORGANISATIONAL LEARNING

1. After each round of data, hold a short learning session with the team using four prompts: what happened, why, so what and now what. Record the decisions.

2. Add appreciative questions to your interview guides, such as asking people to describe a time the programme worked at its best and what made that possible, alongside questions on problems.

3. Pair programme staff with the evaluation team for at least one task, such as tool design or analysis, so capacity stays in the organisation when you leave.

AI prompt **Copy, fill the brackets, then check every line**

Help me design a learning-oriented evaluation process, drawing on Hallie Preskill's evaluative inquiry and appreciative inquiry, for **[organisation or programme]**. The team is **[describe size, roles and experience with evaluation]**. Our key learning questions are **[paste]**. (1) Draft an interview guide that combines appreciative questions with direct questions about shortfalls. (2) Design a 90-minute learning session the team can run after each round of data. (3) Suggest two ways to build the team's own evaluation capacity during the assignment. Keep the plan realistic for **[time and budget]**.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of the difference between a once-a-year health check and a fitness tracker you glance at every day. The annual check tells you something important, once. The tracker makes noticing and adjusting part of daily life. Evaluative inquiry is the tracker built into how an organisation works.

YOUR MOVES

Three moves from Preskill's path

1 Treat your unusual route as an asset

A BA in Spanish, training in special education and work in private industry gave Preskill a view of learning that many evaluators lacked.

TRY THIS
Write a two-line story of your path that explains why it makes you a better evaluator, and use it when you introduce yourself.

2 Borrow from neighbouring fields

She brought appreciative inquiry from organisational development into evaluation, and helped many organisations learn from what works.

TRY THIS
Try one technique from design, coaching or organisational development in your next workshop.

3 Evaluate for what happens next

She linked evaluation to learning and change, and moved from a professorship to FSG in 2009, two years after serving as AEA president.

TRY THIS
End your next evaluation with a learning session where the team decides what they will do differently, not just a report.

THE BIG IDEAS OF THIS CHAPTER

1 **Evaluative inquiry for organisational learning** p. 232

Make evaluation an ongoing, collaborative practice inside the organisation, built into daily work, so that people continually learn from what they do and change it.

2 **Appreciative inquiry in evaluation** p. 235

Start an evaluation from what works at its best and what people value, using the 4-I cycle of inquire, imagine, innovate and implement, to build the energy needed to use findings.

3 **Evaluation capacity building** p. 238

Treat an organisation's ability to do and use evaluation as something to design deliberately, with clear objectives, chosen teaching and learning strategies and the conditions needed for practice to last.

CONCEPT HIGHLIGHT

Strategic learning

Using evaluation and data to inform an organisation's ongoing strategic decisions rather than only to judge past performance.

GIANTS CONNECTED TO THIS CHAPTER



Michael Quinn Patton

Utilization-focused evaluation shares her focus on findings that people act on.

p. 204



Lee J. Cronbach

An early argument that evaluation should serve local learning and improvement.

p. 38



Ray C. Rist

Evaluation capacity building at the level of whole countries and systems.

p. 410

KEY WORKS TO READ IN THE ORIGINAL

- **Evaluative Inquiry for Learning in Organizations** 1999, With Rosalie T. Torres
- **Evaluation in Organizations: A Systematic Approach to Enhancing Learning, Performance and Change** 2001, With Darlene Russ-Eft; second edition 2009
- **Reframing Evaluation through Appreciative Inquiry** 2006, With Tessie Catsambas

CONCEPT HIGHLIGHT

Evaluative inquiry

Ongoing inquiry by members of an organisation, with stakeholders, into what they do and how, built into daily work and aimed at learning and better performance.

CONCEPT HIGHLIGHT

Dialogue and reflection

Two of the four learning processes of evaluative inquiry: structured conversation to build shared understanding, and stepping back to consider experience and its meaning.

WORKSHEET · TRY THIS TEMPLATE

Appreciative evaluation interview guide

Print this page, or copy the headings into your notes.

1 Best experience

Invite a specific story of the programme at its best.

2 What made it possible

People, conditions, decisions and resources behind that moment.

3 What people value most

About the work, the team and the change it creates.

4 Wishes for the future

What would make more of those moments happen.

5 Shortfalls and concerns

Ask directly about what is not working, so accountability is not lost.

6 Learning to act on

What the team will do differently, by when.

REFLECT

What does your client's organisation actually do with findings in the week after a report lands?

CONNECTS TO

MP **Michael Quinn Patton** Utilization-focused evaluation shares her focus on findings that people act on. p. 204

LC **Lee J. Cronbach** An early argument that evaluation should serve local learning and improvement. p. 38

RR **Ray C. Rist** Evaluation capacity building at the level of whole countries and systems. p. 410

11



Emblem: an original HubForge design for this figure, not a likeness.

Sanjeev Sridharan

Active 2010s to 2020s

Worked in Scotland, Canada, India and the United States; capacity building in China, Chile, India, Sri Lanka and Scotland

Use and learning | Impact evaluation and evidence systems

The evaluator of complex health programmes who argues that designs should fit the programme, and that evidence lives in ecologies, not silos.

IN ONE LINE

Sanjeev Sridharan asked evaluators to respect complexity, and to design every study so that it reaches the people who decide.

In 1993, at the University of California, Irvine, Sanjeev Sridharan completed a doctorate on one of humanity's oldest catastrophes: the dynamics of famine. A famine is never a single event. It builds over time, hits different people differently and depends on everything around it. That interest in how big social problems unfold would travel with him across three continents: to the University of Edinburgh, where he headed an evaluation programme; to Toronto, as founding director of an evaluation centre for complex health interventions; and to the Gates Foundation's India Country Office, where he served as Country Lead for learning systems and systems evaluation. Along the way he kept asking questions most evaluation designs skip. When should change appear? For whom might it differ? And how will the findings reach the people who decide?

Life CHAPTER 11

In 1993 Sanjeev Sridharan completed a PhD in Social Ecology at the University of California, Irvine, with a dissertation titled Theoretical and Empirical Essays on the Dynamics of Famine, chaired by Richard McCleary. The topic is a reminder that his later work on health and social programmes grew out of an interest in how large social problems unfold over time.

His career then ran through three countries. At the University of Edinburgh he was head of the Evaluation Program and a Senior Research Fellow in the Research Unit in Health, Behaviour and Change. In 2006, with Bernadette Campbell and Heidi Zinzow, he published a method for asking stakeholders when a programme should first be expected to show impact. He moved to Toronto as founding director of the Evaluation Centre for Complex Health Interventions at the Li Ka

BIG IDEAS

1. **The anticipated timeline of impact**
2. **Ten steps to making evaluation matter**
3. **An ecology of evidence**

KEY WORKS

Ten steps to making evaluation matter
2011 · With April Nakaima, Evaluation and Program Planning

Valuing and embracing complexity
2020 · With April Nakaima, Evaluation and Program Planning

From Silos to Ecologies: Towards an Ecology of Evidence to Navigate Solutions
2026 · Co-edited with Frans L. Leeuw; release listed for 2026

Shing Knowledge Institute of St. Michael's Hospital, and associate professor in the University of Toronto's Department of Health Policy, Management and Evaluation. From there he published, with April Nakaima, Ten steps to making evaluation matter (2011) and wrote a guide to evaluating place-based initiatives for Policy Horizons Canada.

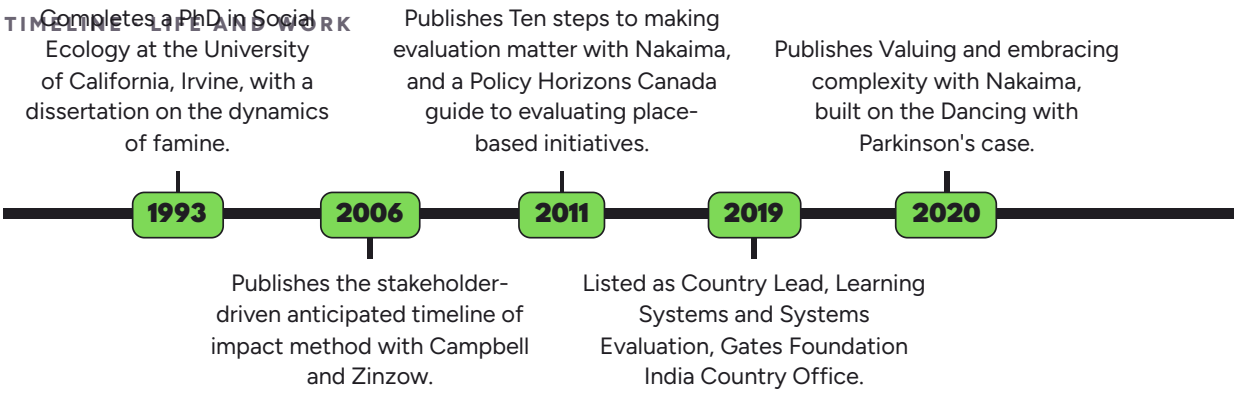
By March 2019 he was Country Lead for Learning Systems and Systems Evaluation at the Bill and Melinda Gates Foundation's India Country Office, a title still listed in October 2021. He has led evaluation capacity-building projects in China, Chile, Scotland, India and Sri Lanka, been an associate editor of the American Journal of Evaluation, and sat on the boards of the Canadian Journal of Program Evaluation, New Directions for Evaluation and Evaluation and Program Planning.

His 2026 publisher biography describes him as a professor specialising in health policy evaluation at the University of Hawai'i at Mānoa and an adviser to the evaluation functions of the UN Office of Internal Oversight Services and the Pan American Health Organization, with research focused on sustainability, equity and new evaluation approaches for brain health. Recent books include A New Blueprint for Brain Health (Springer, 2025), co-edited with Jordan Antflick, Kaela Scott and April Nakaima, and From Silos to Ecologies, co-edited with Frans Leeuw for Routledge, with release listed for September 2026.

The spark

His doctorate studied the dynamics of famine, in our reading a subject in which timing and context shape everything. Years later, in a 2019 talk, he described the kind of programme that would occupy him: many parts, heavily shaped by context, changing over time and often starting without a clear theory of how it would work. His 2011 paper with April Nakaima listed the many reasons evaluations of such programmes so often failed to matter.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 11

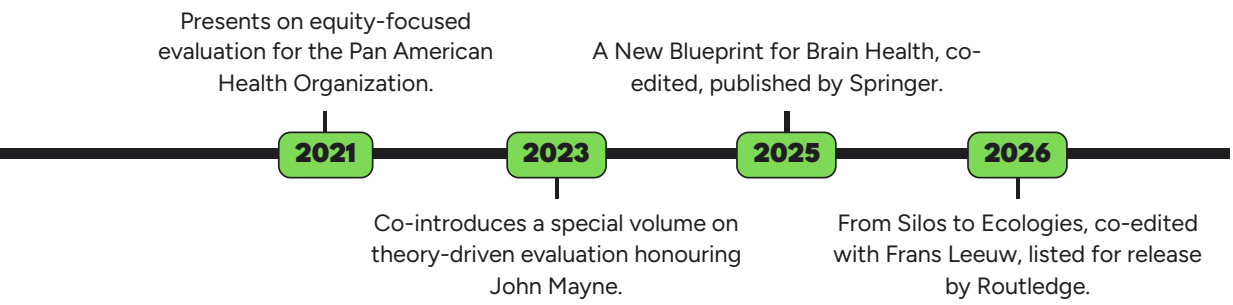
Imagine a health programme is evaluated after two years and shows no change. Has it failed? Perhaps. Or perhaps nobody asked, before the evaluation began, how long change should realistically take. Health and social programmes rarely behave like the single, stable treatments that standard evaluation designs assume. In a 2019 talk at the University of California, San Francisco, Sridharan described complex interventions as having many components, depending heavily on context, changing over time and often starting with an unclear programme theory. His 2011 paper with Nakaima criticised formulaic evaluation designs and argued that evaluations should be built to influence policy and practice.

His work tackled these gaps in stages. In June 2006, in the American Journal of Evaluation 27(2), Sridharan, Campbell and Zinzow proposed a stakeholder-driven anticipated timeline of impact, illustrated with a comprehensive community initiative and built with concept mapping. In 2011, in Evaluation and Program Planning 34(2), Sridharan and Nakaima set out ten steps to make evaluation matter, arguing that the complexity of a programme should shape the design of its evaluation, and drawing heavily on a realist approach that asks what it is about a programme that makes it work.

The problem it solved

Evaluations that were technically sound but ignored, or poorly matched to multi-component, context-dependent programmes.

TIMELINE, CONTINUED



Impact WHAT CHANGED

Sridharan's influence runs mainly through capacity building and editorial work rather than a single branded method. He has led evaluation capacity-building projects in China, Chile, Scotland, India and Sri Lanka; presented on equity-focused evaluation for the Pan American Health Organization in 2021; and spoken at the Asian Development Bank's Asian Evaluation Week. The 2023 special volume he co-introduced drew on evaluations in Scotland, India, Canada and the United States, and his 2026 co-edited book includes case studies from Canada, China and India on maternal health, brain health, poverty reduction and community-led innovation.

His years as Country Lead for learning systems and systems evaluation at the Gates Foundation's India office placed him close to large-scale health and development work in India. Our interpretation is that his most useful legacy for MEL practitioners is a set of questions that discipline design: when should change appear, for whom might it differ, how will findings reach decisions, and what other evidence should be read alongside a study. These questions sit comfortably with the learning-oriented MEL increasingly expected by funders in South Asia.

WHERE PEOPLE DISAGREE

Debates

In our reading, the approach has some general limits. Designs tailored to each programme's complexity can be slower and costlier to plan than standard templates, and they depend heavily on the evaluator's judgement. Stakeholder timelines of impact may reflect hopes or political pressures rather than evidence. Realist and ecological approaches can make it harder to give funders a single clear verdict. Several of his key outputs are briefs, special issues and edited volumes, so evidence of their measured effect on practice is limited.

LIMITS, IN OUR READING

- ☐ In our view, stakeholder timelines of impact can reflect hopes or political pressure as much as evidence, so they need testing. In our view, context-specific designs can be harder to apply quickly than checklists.
- ☐ Several key outputs are practitioner briefs and edited volumes, so evidence of their measured effects on practice is limited.

WHY THIS MATTERS TO YOU

Sridharan's questions help anyone judging something complicated. Before you declare that something has failed, ask when it should realistically have shown results. Before you design a study, ask which decisions it must inform, and who makes them. And instead of waiting for one perfect study to settle a hard question, bring several kinds of evidence together: routine data, the knowledge of people on the ground and published research. Each habit protects good programmes from being judged too early, and makes you the person in the room who designs evidence to matter.

10

steps in the 2011 framework for making evaluation matter, written with April Nakaima

Sridharan and Nakaima, Evaluation and Program Planning 34(2), 2011

5

countries in which he has led evaluation capacity-building projects: China, Chile, Scotland, India and Sri Lanka

Asian Evaluation Week speaker biography, Asian Development Bank

9

case papers from Scotland, India, Canada and the United States in the 2023 special volume he co-introduced

Evaluation and Program Planning 97 (2023), IDEAS/RePEc record

14

chapters in From Silos to Ecologies, co-edited with Frans Leeuw, including cases from Canada, China and India

Routledge book listing (2026)

Did you know?

His doctoral dissertation, completed in 1993, was a set of theoretical and empirical essays on the dynamics of famine.

Key concepts

Complex intervention
A programme with many interacting components, strong dependence on context, change over time and often an unclear programme theory at the outset.

Anticipated timeline of impact
Stakeholders' explicit expectations of when a programme should first affect each outcome, used to time and interpret an evaluation.

Realist evaluation
An approach that asks what it is about a programme that makes it work, for whom and in what circumstances, rather than only whether it works.

Learning framework
A plan setting out what decisions an evaluation should inform, who makes them and how lessons will be fed back and updated.

Pathways of evaluation influence
The routes by which findings reach and change decisions, policies or practice, which can themselves be traced and evaluated.

Epistemic fluency
The ability to work across different kinds of knowledge and evidence and to judge what each can and cannot show.

Sources for this chapter: p. 580

THE BIG IDEAS OF THIS CHAPTER

1. **The anticipated timeline of impact** p. 254
   Before judging a programme, agree with the people closest to it when each outcome should realistically begin to change, and use that timeline to time and interpret the evaluation.

2. **Ten steps to making evaluation matter** p. 257
   Design an evaluation around the programme's real complexity and around the decisions it must influence, rather than around a standard formula.

3. **An ecology of evidence** p. 260
   For complex social problems, cultivate many forms of evidence that work together over time, instead of waiting for one study at the top of a hierarchy to settle the question.

GIANTS CONNECTED TO THIS CHAPTER



Patricia J. Rogers

Matching evaluation design to simple, complicated and complex challenges.

p. 268



Michael Quinn Patton

Developmental evaluation for complex, changing situations.

p. 204



Howard White

Evidence synthesis as a way beyond reliance on single studies.

p. 476

KEY WORKS TO READ IN THE ORIGINAL

- **Ten steps to making evaluation matter** 2011, With April Nakaima, Evaluation and Program Planning
- **Valuing and embracing complexity** 2020, With April Nakaima, Evaluation and Program Planning
- **From Silos to Ecologies: Towards an Ecology of Evidence to Navigate Solutions** 2026, Co-edited with Frans L. Leeuw; release listed for 2026

In one line

Sanjeev Sridharan asked evaluators to respect complexity, and to design every study so that it reaches the people who decide.

1

BIG IDEA 1 OF 3 · SRIDHARAN

The anticipated timeline of impact

Before judging a programme, agree with the people closest to it when each outcome should realistically begin to change, and use that timeline to time and interpret the evaluation.

PICTURE IT

Think of planting a mango tree. Nobody judges it a failure because it has no fruit after one year; everyone knows roughly when to expect flowers and fruit. The anticipated timeline of impact asks the people closest to a programme to agree, in advance, when each change should appear.

The problem

Evaluations are usually timed by funding cycles and reporting deadlines, not by how change actually unfolds. A community programme might be judged after two years on outcomes that could take six years to appear, or declared a success on early gains that later fade.

Logic models show what should change but not when. The result can be unfair verdicts: good programmes closed too early, weak ones extended on hope, and communities losing services because an evaluation asked the right question at the wrong time.

What it is

The anticipated timeline of impact is an explicit statement of when a programme's outcomes are expected to begin to change, and in what pattern. Sridharan, Bernadette Campbell and Heidi Zinzow proposed a stakeholder-driven method for building it in the American Journal of Evaluation 27(2) in 2006, demonstrated on a comprehensive community initiative using concept mapping. It has two parts. The timeline sets realistic expectations for when each outcome should first shift. The trajectory describes the shape of change, which may be gradual or sharp, or may even get worse before it gets better. The 2011 Ten steps paper makes it step four and recommends drawing mainly on stakeholders closest to the programme, who tend to hold more realistic views of community change than funders or academics. Disagreements among stakeholders are documented rather than smoothed away. The logic is simple: a verdict means little unless it is measured against when change should have appeared.

How it works

1. **List the outcomes**
   Take each outcome from the programme theory, from early changes in practice to long-term impact.
2. **Ask those closest to the work**
   Ask frontline staff, community members, managers and funders when each outcome should first change; the 2011 paper gives most weight to those nearest the programme.
3. **Structure the views**
   Use a method such as concept mapping, as in the 2006 paper, to sort and rate expectations and show where groups agree or differ.
4. **Sketch the trajectory**
   For each outcome, agree whether change should be gradual, sudden, or dip before improving, and note the inertia and delays that could slow it.
5. **Ground it in experience**
   Check expectations against evidence and past experience, since the 2011 Policy Horizons Canada guide asks that timelines rest on realistic experience rather than aspiration.
6. **Time and judge the evaluation by it**
   Schedule data collection for when change is expected, judge results against the agreed timeline, and revise the timeline as learning accumulates.

WORKED EXAMPLE · AN ILLUSTRATION

A district programme in Odisha trains government primary teachers in activity-based mathematics. The funder wants an end-line test after 18 months. The evaluator first brings together teachers, cluster resource coordinators, block officials and the funder to build a timeline. Teachers expect classroom routines to change within one term; coordinators expect test scores to move only in the second year, perhaps dipping briefly while children adjust to new methods; officials expect effects on transition to upper primary only after four years. Concept mapping shows the funder's expectations are the most optimistic. The evaluation is redesigned with classroom observation at six months, learning assessments at 12 and 24 months, and a later tracking study of transition. When scores barely move at 12 months, the team reads this against the timeline rather than declaring failure.

CONCEPT HIGHLIGHT

Complex intervention

A programme with many interacting components, strong dependence on context, change over time and often an unclear programme theory at the outset.

USE IT WHEN

- Outcomes are long-term or community-level and likely to lag well behind activities.
- The funder's reporting deadline may not match how change unfolds.
- Stakeholders disagree about what should have happened by now.

SKIP IT WHEN

- The intervention has a short, stable and well-evidenced window of effect.
- The work is a quick process check that makes no judgement about outcomes.

COMMON MISREADINGS

- [ ] Treating the timeline as a promise or target rather than an expectation to test and revise.
- [ ] Using it to excuse any null result indefinitely, rather than as a basis for judgement agreed in advance.

WHY IT WAS GROUNDBREAKING

It made time an explicit part of programme theory instead of an accident of funding cycles. It gave the people closest to a programme a structured say in what counts as a fair test of it. In our reading, it protects programmes from premature verdicts and funders from false comfort.

Origin

The method first appeared in June 2006 in the American Journal of Evaluation 27(2), by Sridharan, Campbell and Zinzow. Sridharan carried it into the fourth of the Ten steps to making evaluation matter with April Nakaima (Evaluation and Program Planning 34(2), 2011), adding the idea of an anticipated performance trajectory, and into the fourth question of his 2011 guide to evaluating place-based initiatives for Policy Horizons Canada.

RELATED IDEAS

PR Rogers 268 · MP Patton 204 · HC Chen 174 · CW Weiss 152

CONCEPT HIGHLIGHT

Anticipated timeline of impact

Stakeholders' explicit expectations of when a programme should first affect each outcome, used to time and interpret an evaluation.

2

BIG IDEA 2 OF 3 · SRIDHARAN

Ten steps to making evaluation matter

Design an evaluation around the programme's real complexity and around the decisions it must influence, rather than around a standard formula.

PICTURE IT

Think of a tailor-made suit versus one bought off the rack. A standard evaluation design is off the rack. Sridharan's ten steps fit the design to the real shape of the programme, and to the decisions it must serve.

The problem

Many evaluations of health and social programmes were technically sound but changed little. Sridharan and Nakaima traced the causes: unclear purposes, a divide between programme planning and evaluation, evaluation treated as form-filling for compliance, designs that measured well but said little about causation, and programmes treated as static when they kept changing. Commissioners hired evaluators for narrow tasks rather than broader learning. For programme staff and the people they served, this meant evaluations that arrived, were filed and left the next round of decisions untouched.

What it is

Ten steps to making evaluation matter, by Sridharan and April Nakaima (Evaluation and Program Planning 34(2), 2011), is a framework for planning evaluations that influence policy and practice. Its steps are: understand the programme and the stability of its components; make the programme theory explicit; integrate a synthesis of existing evidence into that theory; develop the anticipated timeline of impact; develop a learning framework; discuss how the evaluation will be used and exert influence; build a comprehensive design that fits the programme's complexity; capture emergent, dynamic learning during implementation; relate the evaluation to a framework of spread; and consider the relationship between performance and sustainability. Some steps are conventional, others less so. The logic, drawn heavily from realist evaluation, is that design should follow the programme's complexity, and that an evaluation is itself an intervention whose influence must be planned. The steps are meant to structure dialogue between planners and evaluators.

How it works

1. **Understand the programme**
   Map its components, how they relate and whether they are stable over time, before choosing any design.
2. **Make the theory explicit and test it against evidence**
   Write out the if-then assumptions, then use realist synthesis to mark which links are well supported and which remain uncertain.
3. **Set the timeline and learning framework**
   Agree when outcomes should change and state at the outset what can be learned at policy, organisational, process and individual levels.
4. **Plan for influence**
   Treat the evaluation as an intervention, naming the decisions it should shape and how its influence will be traced.
5. **Fit the design to complexity**
   Choose a design that addresses causation, context and mechanisms and takes account of threats to validity.
6. **Learn as the programme changes**
   Use methods such as concept mapping, network analysis or system dynamics to track emergent change during implementation.
7. **Plan for spread and sustainability**
   Decide what knowledge should spread and who is responsible for spreading it, and what besides effectiveness will decide whether the programme lasts.

WORKED EXAMPLE · AN ILLUSTRATION

A state health department in India is piloting home visits by community health workers for people with hypertension in three districts. The evaluation team works through the ten steps at inception. It finds that the screening component is stable but the follow-up model differs by district. A rapid realist synthesis shows strong evidence for screening and weak evidence for adherence support through home visits, so that link becomes the evaluation's focus. The learning framework names three decisions: whether to standardise the follow-up model, whether to stock medicines at sub-centres, and whether to extend to more districts. An influence plan schedules briefings before each budget cycle. A spread question asks what district officials would need in order to adopt the model, and a sustainability question asks what happens when pilot funding ends.

Did you know?

One of his main case studies on evaluating complexity was a Toronto organisation that runs dance classes for people with Parkinson's disease.

USE IT WHEN

- Multi-component health or social programmes whose parts vary by place or change over time.
- Pilots that may be scaled up and need evidence on spread and sustainability.
- Commissioners want findings to feed specific, upcoming decisions.

SKIP IT WHEN

- Simple, stable interventions where a standard design answers the question well.
- Very small evaluations where working through the full sequence would cost more than the decisions are worth.

COMMON MISREADINGS

- Treating the ten steps as a rigid checklist, when they are meant to structure dialogue between planners and evaluators.
- Assuming that fitting complexity means giving up rigour, when the framework calls for a comprehensive design that addresses threats to validity.

WHY IT WAS GROUNDBREAKING

It treated an evaluation's influence on decisions as something to design for, not hope for. It brought realist questions, evidence synthesis, spread and sustainability into one planning sequence that programme planners and evaluators can work through together.

Origin

The framework appeared in 2011 in Evaluation and Program Planning 34(2), drawing among other work on the evaluation of Have a Heart Paisley, a Scottish demonstration programme on which Sridharan and colleagues had published in 2008. In 2020, in Valuing and embracing complexity (Evaluation and Program Planning 80), Sridharan and Nakaima used the Dancing with Parkinson's case in Toronto to set out the skills that complexity demands of evaluators.

RELATED IDEAS

PR Rogers 268 · MP Patton 204 · HW White 476 · HC Chen 174

CONCEPT HIGHLIGHT

Realist evaluation

An approach that asks what it is about a programme that makes it work, for whom and in what circumstances, rather than only whether it works.

3

BIG IDEA 3 OF 3 · SRIDHARAN

An ecology of evidence

For complex social problems, cultivate many forms of evidence that work together over time, instead of waiting for one study at the top of a hierarchy to settle the question.

PICTURE IT

Think of how a forest stays healthy. No single tree does it; soil, water, insects and trees work together over time. An ecology of evidence works the same way, drawing many kinds of evidence together instead of relying on one study at the top.

The problem

Evidence hierarchies taught decision-makers to rank studies by design and trust the top. For complex problems such as maternal health or brain health, this could mean waiting years for a definitive study while discarding routine data, implementation experience and community knowledge. Evaluation systems themselves could become bureaucratic, focused on compliance and at risk of capture by the interests they served. Programmes that needed to adapt quickly were left with either a single verdict or nothing usable at all.

What it is

An ecology of evidence is the term Sridharan has developed, notably with Frans Leeuw, for a system in which different kinds of evidence, from trials and evaluations to routine data, implementation experience and community knowledge, work together to help people navigate complex problems. The 2023 special volume he introduced with April Nakaima, honouring the theory-driven evaluator John Mayne, linked theories of change to these broader ecologies, stressed the epistemic fluency evaluators need to learn across kinds of knowledge, and noted that knowledge of programme mechanisms is incomplete at the start. The Routledge volume From Silos to Ecologies, listed for September 2026, draws on realist evaluation and systems thinking. It calls for three shifts: from routine monitoring towards deeper explanation; from bureaucratic systems towards learning systems that are not centrally orchestrated; and from isolated interventions towards solutions that reflect community realities. Knowledge is treated as fallible, adaptive and rooted in local context.

How it works

1. **Start from the problem and its terrain**
   Define the real-world problem and its context first, rather than starting from a preferred method, as the 2026 volume recommends.
2. **State how solutions are thought to work**
   Write the theory of change as hypotheses about mechanisms, accepting that it will be incomplete at first.
3. **Map the available evidence**
   List the trials, evaluations, routine data, implementation records and community accounts that bear on each part of the theory.
4. **Judge what each source can show**
   Use epistemic fluency to decide what each kind of evidence can and cannot establish, rather than ranking by design alone.
5. **Combine and test**
   Bring the sources together to test the theory, looking for convergence, contradiction and gaps.
6. **Build learning loops**
   Set regular points at which the evidence is revisited and the theory and programme adapted, so that the system learns rather than merely reports.

WORKED EXAMPLE · AN ILLUSTRATION

A state maternal health team wants to reduce deaths in childbirth in its remote districts. Rather than waiting for a single trial, it assembles an ecology of evidence. Systematic reviews show what has worked for emergency obstetric referral elsewhere. Routine facility data reveal where women arrive late. Implementation records show which ambulance routes fail in the rainy season. Accounts from community health workers and village councils explain why some families delay seeking care. The team writes its theory of change as hypotheses and meets every quarter to test them against each new stream of data. When facility data and community accounts disagree, it investigates rather than choosing one. Over two years the programme adapts its referral model district by district and documents what it learned.

CONCEPT HIGHLIGHT

Learning framework

A plan setting out what decisions an evaluation should inform, who makes them and how lessons will be fed back and updated.

USE IT WHEN

- Complex problems where no single study can settle the question.
- Settings with rich routine and community data but few rigorous studies.
- Programmes that must adapt as they go and learn from each round of data.

SKIP IT WHEN

- A narrow, well-defined question that one well-designed study can answer.
- Accountability settings that require a single independent verdict by a fixed date.

COMMON MISREADINGS

- Taking ecology to mean that any evidence is as good as any other, when the approach demands judgement about what each source can show.
- Treating it as a reason to avoid rigorous studies, which remain an important part of the ecology.

WHY IT WAS GROUNDBREAKING

It reframes evidence as a system to be cultivated rather than a ladder to be climbed. It connects evaluation directly to how organisations and communities learn over time. In our reading, it gives practitioners in settings rich in routine data but poor in rigorous studies a principled way to use what they have.

Origin

The idea took shape across Sridharan's later work: the 2023 special volume of Evaluation and Program Planning 97 with April Nakaima, whose nine case papers came from Scotland, India, Canada and the United States, and From Silos to Ecologies, co-edited with Frans Leeuw for Routledge's Comparative Policy Evaluation series. Its chapters include a maternal health case from Meghalaya and a critique of evidence hierarchies by Patricia Rogers and Thomas Aston.

RELATED IDEAS

HW White 476 · PR Rogers 268 · MP Patton 204 · HC Chen 174

CONCEPT HIGHLIGHT

Pathways of evaluation influence

The routes by which findings reach and change decisions, policies or practice, which can themselves be traced and evaluated.

RELATED GIANTS



Patricia J. Rogers

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

p. 268



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152

CONCEPT HIGHLIGHT

Epistemic fluency

The ability to work across different kinds of knowledge and evidence and to judge what each can and cannot show.

THE BIG IDEAS OF THIS CHAPTER

1. **The anticipated timeline of impact** p. 254
   Before judging a programme, agree with the people closest to it when each outcome should realistically begin to change, and use that timeline to time and interpret the evaluation.
2. **Ten steps to making evaluation matter** p. 257
   Design an evaluation around the programme's real complexity and around the decisions it must influence, rather than around a standard formula.
3. **An ecology of evidence** p. 260
   For complex social problems, cultivate many forms of evidence that work together over time, instead of waiting for one study at the top of a hierarchy to settle the question.

Put it to work

USE IT ON MONDAY · EVALUATION DESIGNED TO FIT COMPLEXITY, WITHIN AN ECOLOGY OF EVIDENCE

1. At inception, describe the programme's complexity honestly: how many components, how much they depend on context, and how much is expected to emerge. Let that shape the design rather than reusing last year's template.

2. Write a short learning framework: the decisions findings should inform, who makes them, and how you will trace whether the evaluation actually influenced them.

3. Add questions about spread and sustainability: what would it take for this to continue or grow, and for whom does it work less well?

AI prompt **Copy, fill the brackets, then check every line**

Help me fit an evaluation design to the complexity of **[programme]** in **[location]**, drawing on Sanjeev Sridharan's work on evaluating complex interventions and realist questions of what works, for whom and in what context. Here is the programme description: **[paste]**. Here is the current draft design: **[paste]**. (1) Describe the programme's complexity: components, interactions, context dependence and emergence. (2) Identify where the draft design treats the programme as simpler than it is. (3) Suggest realist-style questions and a learning framework linking findings to decisions. (4) Suggest questions on spread and sustainability. Do not assert how the programme works in practice; frame mechanisms as hypotheses to test.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of planting a mango tree. Nobody judges it a failure because it has no fruit after one year; everyone knows roughly when to expect flowers and fruit. The anticipated timeline of impact asks the people closest to a programme to agree, in advance, when each change should appear.

YOUR MOVES

Three moves from Sridharan's path

1 Specialise so people know when to call

From Edinburgh to Toronto, Sridharan built a centre and a publication trail around one hard problem: evaluating complex health interventions.

TRY THIS
Choose the niche you want to be known for, and make your next three pieces of work sit inside it.

2 Work close to implementation

He moved between universities, a hospital research institute and a large foundation's India office, which sharpened the questions his research asked.

TRY THIS
Spend time with an implementation team before you finalise your evaluation questions.

3 Turn partnerships into frameworks

His long collaboration with April Nakaima, from 2011 to 2023, turned practical lessons into frameworks others could use.

TRY THIS
Pick one long-term collaborator and agree to write up a shared lesson together this year.

THE BIG IDEAS OF THIS CHAPTER

1. **The anticipated timeline of impact** p. 254
   Before judging a programme, agree with the people closest to it when each outcome should realistically begin to change, and use that timeline to time and interpret the evaluation.
2. **Ten steps to making evaluation matter** p. 257
   Design an evaluation around the programme's real complexity and around the decisions it must influence, rather than around a standard formula.
3. **An ecology of evidence** p. 260
   For complex social problems, cultivate many forms of evidence that work together over time, instead of waiting for one study at the top of a hierarchy to settle the question.

CONCEPT HIGHLIGHT

Epistemic fluency

The ability to work across different kinds of knowledge and evidence and to judge what each can and cannot show.

GIANTS CONNECTED TO THIS CHAPTER



Patricia J. Rogers

Matching evaluation design to simple, complicated and complex challenges.

p. 268



Michael Quinn Patton

Developmental evaluation for complex, changing situations.

p. 204



Howard White

Evidence synthesis as a way beyond reliance on single studies.

p. 476

KEY WORKS TO READ IN THE ORIGINAL

- **Ten steps to making evaluation matter** 2011, With April Nakaima, Evaluation and Program Planning
- **Valuing and embracing complexity** 2020, With April Nakaima, Evaluation and Program Planning
- **From Silos to Ecologies: Towards an Ecology of Evidence to Navigate Solutions** 2026, Co-edited with Frans L. Leeuw; release listed for 2026

CONCEPT HIGHLIGHT

Complex intervention

A programme with many interacting components, strong dependence on context, change over time and often an unclear programme theory at the outset.

CONCEPT HIGHLIGHT

Anticipated timeline of impact

Stakeholders' explicit expectations of when a programme should first affect each outcome, used to time and interpret an evaluation.

WORKSHEET · TRY THIS TEMPLATE

Complexity fit check

Print this page, or copy the headings into your notes.

1 Components and interactions

List components and note where they depend on each other.

2 Context dependence

Where results are likely to differ by place, group or timing.

3 Mechanisms

What it is about the programme that might work, for whom, in what context.

4 Pathways to influence

Which decisions findings feed, and how you will trace that.

5 Spread and sustainability

Questions about scale-up and continuation.

6 Forms of evidence

The mix of studies, data and knowledge you will bring together.

REFLECT

Is your evaluation design shaped by your programme's complexity, or by the template you used last time?

CONNECTS TO

PR **Patricia J. Rogers** Matching evaluation design to simple, complicated and complex challenges. p. 268

MP **Michael Quinn Patton** Developmental evaluation for complex, changing situations. p. 204

HW **Howard White** Evidence synthesis as a way beyond reliance on single studies. p. 476

12



Emblem: an original HubForge design for this figure, not a likeness.

Patricia J. Rogers

Active 2000s to 2020s

Australia; professional development delivered across Asia, Africa, Europe and the Americas

Theory and programme logic | Use and learning

The Australian evaluator who taught that a logic model is a tool with a purpose, and built a free library of methods for everyone.

IN ONE LINE

Patricia Rogers showed evaluators that every model needs a purpose, and then gave the whole world a free map of the work.

Some careers are changed by a single book. Patricia Rogers, by her own account, was changed by two conferences. In 1986, working in Sydney and about to move to a local government job in Melbourne, she attended the Australian Evaluation Society's conference, where a keynote by Michael Scriven stirred debate about evaluating without reference to stated goals. A year later, in Canberra, she was introduced to program logic. She went on to turn those encounters into tools for a whole profession: a 2008 article that taught evaluators to separate the complicated from the complex, a 2011 book with Sue Funnell on purposeful program theory, and BetterEvaluation, a free platform whose Rainbow Framework arranged more than 200 options into seven clusters of tasks. Practitioners anywhere with an internet connection can now use the map she helped draw.

BIG IDEAS

1. **Complicated and complex aspects of interventions**
2. **Purposeful program theory**
3. **The Rainbow Framework**

KEY WORKS

Purposeful Program Theory: Effective Use of Theories of Change and Logic Models
2011 · With Sue C. Funnell

Developing a research agenda for impact evaluation in development
2014 · With Greet Peersman, IDS Bulletin 45

When and how to develop an impact-oriented monitoring and evaluation system
2016 · With Peersman, Guijt, Hearn, Pasanen and Buffardi, ODI

Life CHAPTER 12

Patricia J. Rogers is an Australian evaluator based in Melbourne. Her own account of how she entered the field, given to Anthea Rutter for the Australian Evaluation Society in 2019, begins in 1986. She was working in Sydney and about to move back to Melbourne to work in local government when she attended the society's Sydney conference, opened by a Michael Scriven keynote, and met the debates around goal-free evaluation. At the 1987 conference in Canberra she was introduced to program logic, the line of work that later led to a long collaboration with Sue Funnell, a former president of the society.

She went on to earn a PhD in evaluation at the University of Melbourne and to hold a post-doctoral fellowship in evaluation at Harvard University. She taught at Phillip Institute, which became part of RMIT University, for about three decades, and names Jerry

Winston, whom she met there, as a major influence. As Professor of Public Sector Evaluation at RMIT she worked in its Collaboration for Interdisciplinary Research, Consulting and Learning in Evaluation. In 2000 she co-edited a New Directions for Evaluation issue on program theory, with Timothy Hacsi, Anthony Petrosino and Tracy Huebner.

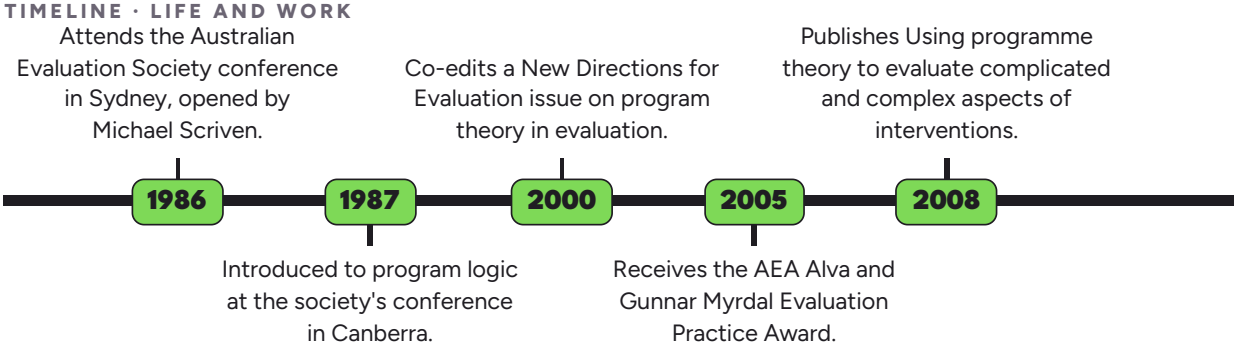
BetterEvaluation, which describes itself as having begun in 2009 as an interdisciplinary public good project, became the centre of her later career. She took leave from RMIT to lead the evidence and evaluation hub at the Australia and New Zealand School of Government, which then became BetterEvaluation, an independent not-for-profit, supported in its first year by an ANZSOG grant. She was its founder and chief executive; it is now part of the Global Evaluation Initiative. She now works as an independent consultant and is part of the core team of the Footprint Evaluation Initiative.

Her honours include the American Evaluation Association's Alva and Gunnar Myrdal Evaluation Practice Award in 2005, the Australian Evaluation Society's Outstanding Contribution to Evaluation Award and a fellowship of that society. She has been a visiting professor at the University of the Witwatersrand and has delivered professional development in Australia, New Zealand, the United Kingdom, the United States, Canada, Japan, South Africa, Malaysia, Nepal and Brazil.

The spark

Rogers has told the story herself, in a 2019 interview for the Australian Evaluation Society. Two conferences lit the way: Sydney in 1986, where goal-free evaluation was debated, and Canberra in 1987, where she met program logic. She has described being drawn to evaluation models and to ideas that could make a practical difference. Later, a simple distinction from the work of Brenda Zimmerman, between problems that are simple, complicated and complex, gave her a way to tell different kinds of evaluation challenge apart.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 12

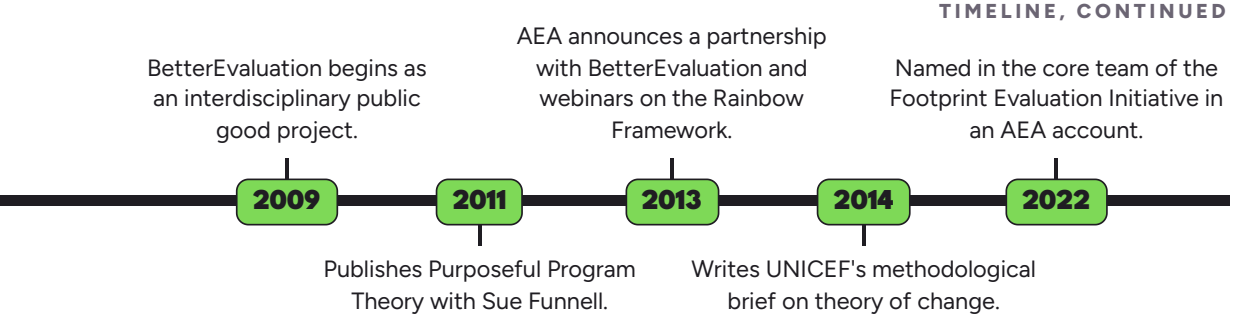
Picture the diagram that sits in thousands of funding proposals: a neat row of boxes and arrows running from inputs to activities, outputs and outcomes. In our summary, by the 2000s such logic models were everywhere in public and development programmes, often because funders asked for them. The trouble was that real programmes rarely ran in a straight line. Many involved several agencies, parallel pathways that all had to work, effects that differed by place, feedback loops and goals that only took shape along the way. Forcing them into one tidy chain could overstate what a programme achieved and produce indicators that measured the wrong things.

Rogers addressed this in a January 2008 article in Evaluation 14(1), Using programme theory to evaluate complicated and complex aspects of interventions. Drawing on the simple, complicated and complex distinction drawn by Glouberman and Zimmerman, she argued that evaluators should classify aspects of an intervention, not whole interventions, and represent each aspect in a way that fits. Her book with Sue Funnell, Purposeful Program Theory, published by Jossey-Bass in March 2011, developed the wider argument: program theory is a family of tools to be chosen according to purpose and situation, not one prescribed method.

The problem it solved

Logic models used as one formula to satisfy funders, and linear designs applied to complex interventions.

TIMELINE, CONTINUED



Impact WHAT CHANGED

Rogers's widest reach comes through BetterEvaluation, an open-access platform now part of the Global Evaluation Initiative, which supports governments in strengthening monitoring and evaluation. Its task-based structure gives practitioners without access to university libraries a route to methods and guidance. Her 2014 theory of change brief appeared in UNICEF's methodological series on impact evaluation, and her 2008 article is hosted, among other places, on the website of India's National Health Systems Resource Centre, an indication that the ideas circulate in Global South health systems work. Her teaching in South Africa, Malaysia, Nepal and Brazil, and her visiting professorship at Witwatersrand, extend that reach.

Through the Footprint Evaluation Initiative she has helped push environmental sustainability onto the evaluation agenda; a 2022 account noted that most evaluations still fail to consider it. Our interpretation is that her main contribution to MEL culture is a habit of mind: ask what a theory of change is for before drawing it, sort the simple from the complicated and complex, and check the full set of evaluation tasks before choosing methods. Many practitioners who never cite her now work in exactly this way.

WHERE PEOPLE DISAGREE

Debates

In our reading, her frameworks have some practical limits. A menu of more than 200 options organised by task does not by itself tell a novice which option fits a given context; judgement still has to come from the evaluator. The simple, complicated and complex distinction can become a label applied loosely rather than an analysis. Her own 2008 article anticipates some limits, warning that logic models used too broadly overstate causal contribution, but tailored, evolving models can be harder for funders to hold programmes accountable against.

LIMITS, IN OUR READING

- ☐ In our view, a menu of options organised by task does not by itself tell a novice which option fits their context; judgement still has to come from the evaluator.
- ☐ In our view, theory of change tools can still be drawn for funders and never used, however purposeful the method.

WHY THIS MATTERS TO YOU

Rogers offers a question worth asking before you draw any plan: what is it for, and who will use it? Then look closely at what you are dealing with. Some parts of any big effort are simple, like following a recipe. Some are complicated, like sending a rocket to the moon, needing expertise and coordination. Some are complex, like raising a child, where every situation is different and you learn as you go. Each needs a different kind of plan and a different kind of evidence. And through BetterEvaluation, anyone, in a small NGO or a district office, can plan an evaluation as thoroughly as a large agency, for free.

200+

evaluation options arranged in the Rainbow Framework when the AEA partnership was announced

AEA365, April 2013

7

clusters of evaluation tasks in the Rainbow Framework

AEA365, April 2013; The Change Agency (2015)

5

aspects of interventions, three complicated and two complex, identified in her 2008 article

Rogers, Evaluation 14(1), 2008

30+

years working as an evaluator, adviser and researcher with governments, UN agencies, development banks, NGOs and philanthropy

Asian Evaluation Week speaker biography, ADB (2023)

Did you know?

Her path into evaluation began with a single conference: the 1986 Australian Evaluation Society meeting in Sydney, when she was about to start a job in local government.

Key concepts

Program theory
An explicit model of how an intervention is expected to produce its outcomes, used to guide design, monitoring and evaluation.

Theory of change
A description of how activities produce a chain of results that lead to intended long-term impacts, at any level from an event to an organisation.

Simple, complicated, complex
A distinction between tested recipes, problems with many coordinated parts, and situations where outcomes are uncertain and emerge during the work.

Emergent outcomes
Results whose specific form, and the means to reach them, develop during implementation rather than being fixed in advance.

Goal displacement
When measuring indicators becomes the aim in itself, pulling effort away from the outcomes the indicators were meant to track.

Footprint evaluation
A collaborative effort premised on the view that all evaluations should consider environmental sustainability, even when it is not a stated goal.

Sources for this chapter: p. 581

THE BIG IDEAS OF THIS CHAPTER

1 **Complicated and complex aspects of interventions** p. 276
Do not ask whether a whole programme is simple or complex; identify which aspects are complicated or complex, and represent and evaluate each in a way that fits.

2 **Purposeful program theory** p. 279
Decide what a theory of change is for, and who will use it, before deciding how to develop and draw it.

3 **The Rainbow Framework** p. 282
Organise the whole of evaluation by the tasks evaluators must do, with a range of options for each task, so that no task is forgotten and every method choice is deliberate.

GIANTS CONNECTED TO THIS CHAPTER



Michael Scriven

His 1986 talk on goal-free evaluation was among her first encounters with the field.

p. 60



Carol H. Weiss

Theory-based evaluation laid the foundations of program theory practice.

p. 152



Sanjeev Sridharan

Shares a commitment to fitting evaluation design to complexity.

p. 246

KEY WORKS TO READ IN THE ORIGINAL

- **Purposeful Program Theory: Effective Use of Theories of Change and Logic Models** 2011, With Sue C. Funnell
- **Developing a research agenda for impact evaluation in development** 2014, With Greet Peersman, IDS Bulletin 45
- **When and how to develop an impact-oriented monitoring and evaluation system** 2016, With Peersman, Guijt, Hearn, Pasanen and Buffardi, ODI

In one line

Patricia Rogers showed evaluators that every model needs a purpose, and then gave the whole world a free map of the work.

1

BIG IDEA 1 OF 3 · ROGERS

Complicated and complex aspects of interventions

Do not ask whether a whole programme is simple or complex; identify which aspects are complicated or complex, and represent and evaluate each in a way that fits.

PICTURE IT

Think of the difference between following a recipe, sending a rocket to the moon and raising a child. The first is simple, the second complicated, the third complex. Most programmes contain all three, and Rogers asks evaluators to treat each part in the way it deserves.

The problem

By the 2000s funders routinely asked for logic models, and the standard format was a single linear chain of boxes and arrows. Real interventions often involved several agencies, parallel pathways that all had to work, mechanisms that differed by context, feedback loops and goals that emerged during implementation. Forcing all this into one tidy diagram overstated a programme's causal contribution, encouraged goal displacement when the model was used for accountability, and generated performance indicators that could reward counterproductive behaviour.

What it is

In Using programme theory to evaluate complicated and complex aspects of interventions (Evaluation 14(1), 2008), Rogers drew on the simple, complicated and complex distinction of Glouberman and Zimmerman to identify five aspects of interventions that need special treatment. Three are complicated. Governance: several agencies or jurisdictions deliver the intervention. Simultaneous causal strands: several pathways must all work for success. Alternative causal strands: the programme works through different mechanisms in different contexts. Two are complex. Recursive causality with disproportionate effects: feedback loops and tipping points, where small changes can have large effects. Emergent outcomes: specific outcomes, and the pathways to them, that develop during implementation. Each calls for a different representation and evaluation response. The skill, Rogers argued, lies in knowing which parts can be simplified and which cannot, rather than either oversimplifying everything or producing a model too tangled to use.

How it works

1. **Break the intervention into aspects**
   Look separately at governance, causal strands, mechanisms, feedback and how settled the goals are.
2. **Handle multi-agency governance**
   Use one shared logic model, report data separately for each site or agency and then aggregate, treating the main challenges as negotiation and logistics.
3. **Show simultaneous strands as jointly needed**
   Draw each strand with arrows that make clear all are essential, and collect data on every strand so none is neglected.
4. **Document alternative strands in context**
   Use tables to show which mechanism is expected in which context, and compare similar initiatives across settings.
5. **Annotate feedback loops and tipping points**
   Mark them on the diagram and track early signs of change throughout implementation.
6. **Let emergent models evolve**
   Use a series of revised logic models, developmental evaluation, or a shared set of guiding principles in place of a fixed causal model.
7. **Use models carefully for accountability**
   Do not derive performance targets mechanically from models of complex work; rely on discussion, reflection and participatory learning instead, as the 2008 article advises.

WORKED EXAMPLE · AN ILLUSTRATION

A state nutrition mission in India combines take-home rations, growth monitoring, home visits by health and anganwadi workers, and village campaigns planned by local women's groups. The MEL team sorts its aspects. Ration distribution is relatively simple and suits standard output tracking. Joint delivery by the health and the women and child development departments is complicated governance, so the team agrees one shared model with data reported by department. Home visits and rations are simultaneous strands, both needed, so both are monitored. Village campaigns are emergent: each women's group decides its own focus, so instead of fixed indicators the team agrees guiding principles and revises a light logic model every six months. It deliberately avoids district targets for campaign outcomes, which could push groups towards easy, visible activities.

In one line

Decide what a theory of change is for, and who will use it, before deciding how to develop and draw it.

USE IT WHEN

- Multi-agency or multi-component programmes.
- Programmes in which some elements are well understood and others are being invented as they go.
- Designing indicators that will be used for accountability or performance management.

SKIP IT WHEN

- ○ A single, well-understood intervention delivered by one agency in a stable setting.
- ○ A quick communication diagram for a lay audience, where the full analysis would add detail no one will use.

COMMON MISREADINGS

- ☐ Labelling a whole programme complex to justify avoiding any model or measure.
- ☐ Treating complicated and complex as synonyms, when complicated parts can be planned and coordinated while complex ones need emergent, adaptive approaches.

WHY IT WAS GROUNDBREAKING

It replaced a blunt question, whether a programme is complex, with a sharper one: which parts are, and what follows for evaluation. It gave evaluators concrete ways to represent each case instead of a single template. It also warned, early and clearly, against generating performance measures mechanically from models of complex work.

Origin

The framework appeared in January 2008 in Evaluation 14(1), pages 29 to 48, building on Glouberman and Zimmerman's 2002 discussion paper for Canada's Commission on the Future of Health Care. Rogers had first met program logic at the Australian Evaluation Society's 1987 conference in Canberra. With Sue Funnell she developed the distinctions further in Purposeful Program Theory (2011), which gives a chapter to the nature of the situation and the intervention.

RELATED IDEAS

SS Sridharan 246 · MP Patton 204 · CW Weiss 152 · HC Chen 174

CONCEPT HIGHLIGHT

Program theory

An explicit model of how an intervention is expected to produce its outcomes, used to guide design, monitoring and evaluation.

2

BIG IDEA 2 OF 3 · ROGERS

Purposeful program theory

Decide what a theory of change is for, and who will use it, before deciding how to develop and draw it.

PICTURE IT

Think of a map. A hiking map, a road map and a metro map of the same city all look different, because each serves a different traveller. A theory of change is a map too, and Rogers asks you to decide who will use it before you draw it.

The problem

Theories of change and logic models had become standard requirements, but were often drawn once to satisfy a funder and then ignored. One template was applied to every purpose and every kind of intervention. Diagrams were either too simple to guide evaluation or too cluttered to communicate. Funnell and Rogers saw both promise and risk: a good programme theory could focus monitoring and evaluation and support learning, while a careless one could mislead decisions about what to measure and about what had caused change.

What it is

Purposeful program theory, set out by Sue Funnell and Patricia Rogers in Purposeful Program Theory: Effective Use of Theories of Change and Logic Models (Jossey-Bass, 2011), treats program theory as an explicit model of how an intervention contributes to a chain of intermediate results and then to its intended or observed outcomes. It has two components. The theory of change describes the central processes that drive change for people, groups or communities, whether drawn from research or from tacit understanding. The theory of action describes how the intervention is designed to set those processes in motion. An outcomes chain links immediate and intermediate outcomes to final impacts. The purposeful part is the sequence: first scope intended uses and users, and assess whether the situation and intervention are simple, complicated or complex; then develop, represent and critique the theory to fit; then use it for monitoring, causal inference, synthesis and reporting.

How it works

1. **Scope intended uses**
   Decide who will use the theory and for what, such as design, monitoring, evaluation or communication, as Part Two of the 2011 book sets out.
2. **Assess the situation and intervention**
   Judge which aspects are simple, complicated or complex, since this shapes how detailed and flexible the theory must be.
3. **Choose development methods**
   Combine deductive sources such as policy and research, inductive observation of the intervention, and the mental models of stakeholders.
4. **Build the theory of change**
   Work through a situation analysis, define the programme's scope and develop the outcomes chain.
5. **Build the theory of action**
   Use a program theory matrix to set out how the intervention will set each part of the theory of change in motion.
6. **Represent and critique**
   Draw the theory in a form suited to its users, then test it for internal coherence and external validity.
7. **Use and revisit**
   Use it to plan monitoring and evaluation, examine causal inference and report findings, and revise it as circumstances change.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO running digital literacy courses for rural women in Karnataka needs a theory of change for three purposes: a funding proposal, a monitoring system and an end-line evaluation. Following a purpose-first approach, its MEL lead produces three linked versions rather than one. For the proposal, a one-page outcomes chain shows the path from training to confident phone use to access to government schemes and income. For monitoring, a program theory matrix sets out what the course does to achieve each outcome, with indicators and data sources. For the evaluation, a detailed version spells out the assumption behind each link, such as whether family members allow women to keep and use a phone, and notes alternative explanations for any change. The team agrees to revisit the theory after every two course cycles.

CONCEPT HIGHLIGHT

Theory of change

A description of how activities produce a chain of results that lead to intended long-term impacts, at any level from an event to an organisation.

USE IT WHEN

- Writing or revising a theory of change for any purpose.
- A programme's existing logic model is not being used by anyone.
- Different audiences need different levels of detail from the same theory.

SKIP IT WHEN

- Stakeholders cannot be involved at all, since the approach relies on agreeing uses and drawing out mental models.
- A decision is needed within days and an existing, adequate theory of change already serves it.

COMMON MISREADINGS

- Thinking purposeful means less rigorous, when the book includes critique, causal inference and synthesis.
- Treating the theory of change and the theory of action as the same thing, so that what the programme does is confused with how change happens.

WHY IT WAS GROUNDBREAKING

It challenged the habit of drawing a logic model for a donor and then ignoring it. It made the purpose and users of the model the first design decision, and it offered a family of approaches rather than one true way.

Origin

Rogers was introduced to program logic at the Australian Evaluation Society's 1987 Canberra conference, through the work of Sue Funnell and colleagues. In 2000 she co-edited a New Directions for Evaluation issue on program theory. Funnell and Rogers published Purposeful Program Theory with Jossey-Bass in 2011, in five parts and 16 chapters. Rogers's 2014 UNICEF methodological brief on theory of change carried the approach into impact evaluation guidance.

RELATED IDEAS

CW Weiss 152 · HC Chen 174 · MP Patton 204 · SS Sridharan 246

CONCEPT HIGHLIGHT

Simple, complicated, complex

A distinction between tested recipes, problems with many coordinated parts, and situations where outcomes are uncertain and emerge during the work.

3

BIG IDEA 3 OF 3 · ROGERS

The Rainbow Framework

Organise the whole of evaluation by the tasks evaluators must do, with a range of options for each task, so that no task is forgotten and every method choice is deliberate.

PICTURE IT

Think of a pilot's pre-flight checklist: it does not fly the plane, but nothing essential is forgotten. The Rainbow Framework is that checklist for evaluation.

The problem

Evaluation guidance was scattered across textbooks, journals and agency manuals, much of it behind paywalls and organised by method or school rather than by what an evaluator actually has to do. Practitioners, especially outside universities and in the Global South, tended to use the few methods they knew. Many planned data collection carefully while neglecting tasks such as framing boundaries, understanding causes or supporting use. Evaluations then suffered from gaps no one noticed until it was too late.

What it is

The Rainbow Framework is BetterEvaluation's way of organising evaluation methods and processes by the tasks commonly undertaken in monitoring and evaluation. When the American Evaluation Association announced its partnership with BetterEvaluation in April 2013, it described the framework as arranging more than 200 options into seven clusters of tasks. The clusters are: manage an evaluation; define what is to be evaluated; frame the boundaries of an evaluation; describe activities, outcomes, impacts and context; understand causes of outcomes and impacts; synthesise data from one or more evaluations; and report and support use of findings. Each cluster contains specific tasks, and each task links to options, methods or processes, with guidance on when each is appropriate. The underlying logic is that method choice should follow from the task and the situation, not from habit, and that a complete plan addresses every cluster. The framework drew on contributions from evaluators in many countries and is freely available.

How it works

1. **Start with management**
   Use the manage cluster to agree who will be involved, how decisions will be made, and what resources and standards apply.
2. **Define and frame**
   Describe what is being evaluated, often with a theory of change, then set the boundaries: purposes, key questions and how success will be judged.
3. **Describe**
   Choose options for collecting and analysing data on activities, outcomes, impacts and context.
4. **Understand causes**
   Select approaches to causal inference that suit the situation, from comparison groups to checking results against the theory of change.
5. **Synthesise**
   Decide how data will be combined into an overall judgement, within one evaluation or across several.
6. **Report and support use**
   Plan products, timing and follow-up so that findings reach and help the people who will act on them.
7. **Review the whole plan**
   Check the plan against every cluster to spot neglected tasks, and consult the options BetterEvaluation lists for each task.

WORKED EXAMPLE · AN ILLUSTRATION

A small education NGO in Telangana is planning its first external evaluation of a reading programme. Its draft terms of reference list a baseline and end-line test, a sample of 40 schools and interviews with teachers. The MEL officer checks the draft against the seven clusters of the Rainbow Framework. Describe is well covered. Manage is thin: no one has decided who signs off findings or how the community will be involved. Frame is missing: the team has not agreed whether the evaluation is for improving the programme or for the funder's renewal decision. Understand causes is absent: there is no comparison group and no alternative approach to attribution. Report and support use is a single line about a final report. The revised plan adds a reference group, a clear purpose, a contribution analysis and two learning workshops.

CONCEPT HIGHLIGHT

Emergent outcomes

Results whose specific form, and the means to reach them, develop during implementation rather than being fixed in advance.

USE IT WHEN

- Drafting or reviewing terms of reference and evaluation plans.
- Training new evaluators to see the whole evaluation process.
- Looking for alternatives to the methods a team habitually uses.

SKIP IT WHEN

- ○ You need detailed technical instruction in one method, which the linked resources rather than the framework itself provide.
- ○ An experienced team has already designed a sound evaluation, and a full checklist pass would add little.

COMMON MISREADINGS

- ☐ Treating the clusters as a strict sequence, when in practice they overlap and recur throughout an evaluation.
- ☐ Assuming that choosing more options makes a better evaluation, rather than choosing the options that fit the situation.

WHY IT WAS GROUNDBREAKING

It turned scattered, method-by-method advice into a single task map of the whole evaluation. Because it is free and open, it put that map in the hands of practitioners everywhere, including those without university libraries. In our reading, it also made neglected tasks, such as framing and supporting use, as visible as data collection.

Origin

BetterEvaluation describes itself as having begun in 2009 as an interdisciplinary public good project. It grew from the evidence and evaluation hub Rogers led at the Australia and New Zealand School of Government, with Rogers as founder and chief executive. In April 2013 the AEA announced a partnership and eight free Coffee Break webinars on the framework, two presented by Rogers. BetterEvaluation is now part of the Global Evaluation Initiative.

RELATED IDEAS

MP Patton 204 · MS Scriven 60 · DS Stufflebeam 134

CONCEPT HIGHLIGHT

Goal displacement

When measuring indicators becomes the aim in itself, pulling effort away from the outcomes the indicators were meant to track.

RELATED GIANTS



Sanjeev Sridharan

The evaluator of complex health programmes who argues that designs should fit the programme, and that evidence lives in ecologies, not silos.

p. 246



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174

CONCEPT HIGHLIGHT

Footprint evaluation

A collaborative effort premised on the view that all evaluations should consider environmental sustainability, even when it is not a stated goal.

THE BIG IDEAS OF THIS CHAPTER

1. **Complicated and complex aspects of interventions** p. 276
   Do not ask whether a whole programme is simple or complex; identify which aspects are complicated or complex, and represent and evaluate each in a way that fits.
2. **Purposeful program theory** p. 279
   Decide what a theory of change is for, and who will use it, before deciding how to develop and draw it.
3. **The Rainbow Framework** p. 282
   Organise the whole of evaluation by the tasks evaluators must do, with a range of options for each task, so that no task is forgotten and every method choice is deliberate.

Put it to work

USE IT ON MONDAY · PURPOSEFUL PROGRAM THEORY

1. Before drawing any theory of change, write one line on what it is for: designing, monitoring, evaluating or communicating. Let that purpose decide how detailed it needs to be.

2. Sort the programme's components into simple, complicated and complex, and choose evaluation methods to match. Do not force a linear design onto the complex parts.

3. Use BetterEvaluation's task-based framework as a checklist for your evaluation plan, to spot tasks you have not yet planned for.

AI prompt **Copy, fill the brackets, then check every line**

Help me develop a theory of change, following the purposeful program theory approach of Sue Funnell and Patricia Rogers, for **[programme]** in **[location]**. Its main purpose is **[design, monitoring, evaluation or communication]**, and its main audience is **[audience]**. Here is the programme description: **[paste]**. (1) Suggest the right level of detail for this purpose and audience. (2) Draft the outcome chain and key assumptions in text form. (3) Classify components as simple, complicated or complex, with reasons, and suggest evaluation approaches that fit each. (4) Propose when and how the theory of change should be updated.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of the difference between following a recipe, sending a rocket to the moon and raising a child. The first is simple, the second complicated, the third complex. Most programmes contain all three, and Rogers asks evaluators to treat each part in the way it deserves.

YOUR MOVES

Three moves from Rogers's path

1 Show up to your community early

Two Australian Evaluation Society conferences, in 1986 and 1987, set the direction of Rogers's work.

TRY THIS

Register for one evaluation conference or network meeting this year, and go with one question to ask five people.

2 Keep your ideas grounded

She taught for about three decades at what became RMIT University while working with practitioners on real evaluation tasks.

TRY THIS

Pair every training or paper you produce with one live assignment where you test it.

3 Turn know-how into shared infrastructure

She took leave from a professorship to lead the hub that became BetterEvaluation, a free resource used by evaluators worldwide.

TRY THIS

Document one method you know well as a free, step-by-step guide others can reuse.

THE BIG IDEAS OF THIS CHAPTER

1 **Complicated and complex aspects of interventions** p. 276

Do not ask whether a whole programme is simple or complex; identify which aspects are complicated or complex, and represent and evaluate each in a way that fits.

2 **Purposeful program theory** p. 279

Decide what a theory of change is for, and who will use it, before deciding how to develop and draw it.

3 **The Rainbow Framework** p. 282

Organise the whole of evaluation by the tasks evaluators must do, with a range of options for each task, so that no task is forgotten and every method choice is deliberate.

CONCEPT HIGHLIGHT

Footprint evaluation

A collaborative effort premised on the view that all evaluations should consider environmental sustainability, even when it is not a stated goal.

GIANTS CONNECTED TO THIS CHAPTER



Michael Scriven

His 1986 talk on goal-free evaluation was among her first encounters with the field.

p. 60



Carol H. Weiss

Theory-based evaluation laid the foundations of program theory practice.

p. 152



Sanjeev Sridharan

Shares a commitment to fitting evaluation design to complexity.

p. 246

KEY WORKS TO READ IN THE ORIGINAL

- **Purposeful Program Theory: Effective Use of Theories of Change and Logic Models** 2011, With Sue C. Funnell
- **Developing a research agenda for impact evaluation in development** 2014, With Greet Peersman, IDS Bulletin 45
- **When and how to develop an impact-oriented monitoring and evaluation system** 2016, With Peersman, Guijt, Hearn, Pasanen and Buffardi, ODI

CONCEPT HIGHLIGHT

Program theory

An explicit model of how an intervention is expected to produce its outcomes, used to guide design, monitoring and evaluation.

CONCEPT HIGHLIGHT

Theory of change

A description of how activities produce a chain of results that lead to intended long-term impacts, at any level from an event to an organisation.

WORKSHEET · TRY THIS TEMPLATE

Purpose-first theory of change brief

Print this page, or copy the headings into your notes.

1 Purpose

Design, monitoring, evaluation, communication or several, in priority order.

2 Audience

Who will read and use the diagram.

3 Level of detail

One-page overview, nested diagrams or detailed causal pathways.

4 Simple, complicated, complex

Which parts are which, and what that means for evaluation.

5 Key assumptions to show

The few that matter most for this purpose.

6 Use and update plan

When it will be revisited, and by whom.

REFLECT

Who is your current theory of change diagram really for, and would they recognise their programme in it?

CONNECTS TO

Michael Scriven (MS)
His 1986 talk on goal-free evaluation was among her first encounters with the field. p. 60

Carol H. Weiss (CW)
Theory-based evaluation laid the foundations of program theory practice. p. 152

Sanjeev Sridharan (SS)
Shares a commitment to fitting evaluation design to complexity. p. 246

THE GREAT DEBATES · ACT 3

Use or Judgement?

THE QUESTION

Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

MP Patton 204 MS Scriven 60 CW Weiss 152

SIDE A

Judge evaluation by its use (Patton)

Patton built his approach on evidence. A 1977 field study by Patton and colleagues found that findings were most likely to be used when an identifiable person cared about the evaluation, which he called the personal factor. Utilization-Focused Evaluation (1978) argued that evaluations should therefore be judged by their actual use and organised around specific primary intended users and their intended uses. In a 1996 issue of Evaluation Practice he argued that the world of evaluation was larger than the formative and summative distinction, pointing to developmental and knowledge-generating work. In our reading, his case is that an evaluation nobody uses is a cost without a benefit, however rigorous it is.

SIDE B

Evaluation means judging merit and worth (Scriven)

Scriven saw evaluation as the determination of value, and held that a competent evaluator can establish a programme's worth as a matter of fact rather than as a matter of taste. It was Scriven who in 1967 named formative and summative evaluation and insisted that judging overall worth was as important as supporting improvement. In the same 1996 issue of Evaluation Practice, he replied to Patton and others by defending the formative and summative distinction as one reasonable way to classify evaluation. A recurring concern in the field is that a focus on primary intended users can tilt an evaluation towards what powerful users want to know.

WHAT HAPPENED

Both positions became pillars of the field. In Alkin and Christie's evaluation theory tree, Patton's work sits on the use branch and Scriven's at the base of the valuing branch, and the 2023 third edition of Evaluation Roots still gives both their own chapters. Patton went on to develop developmental evaluation, whose early adopters, in Mark Cabaj's 2011 study, wrestled with exactly this tension: whether to judge, and to whom they are accountable. Weiss's earlier work had already shown that use itself takes many forms, from direct application to slow enlightenment.

Where you stand

Who are the primary intended users of the evaluation you are working on now, and have you asked them what they will do with the findings?

When a funder and a community want different things from an evaluation, whose question comes first for you?

Your position

Which side do you find yourself on today, and what evidence would change your mind?

ACT 4 OF 5

4

1970s to 2010s

Many Voices, Many Truths



Robert Chambers

p. 296



Egon G. Guba and Yvonna S. Lincoln

p. 318



Jennifer C. Greene

p. 340



Rick Davies and Jess Dart

p. 362



Donna M. Mertens

p. 384

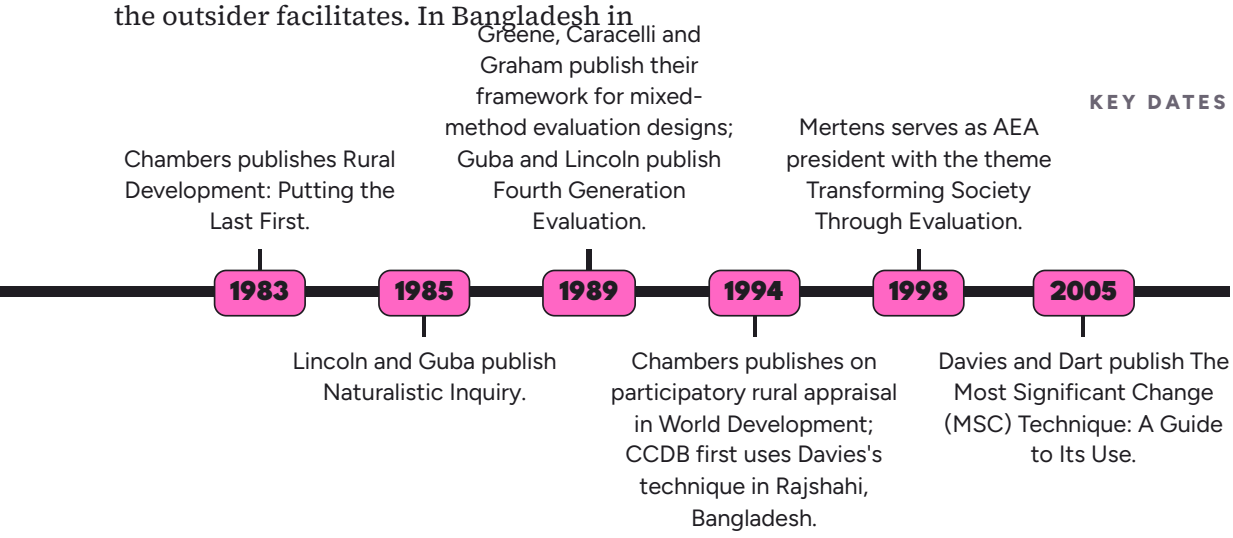
ACT 4 · 1970S TO 2010S

Many Voices, Many Truths

From the 1970s, the Encyclopedia of Evaluation records, evaluators fought what they themselves described as their most bitter debate, between quantitative and qualitative ways of knowing. It was methodological and philosophical, but also political, because it touched who defines good evidence. Judged by the conventional tests of validity and objectivity, a study built on interviews in one setting looked soft. In 1985, in Naturalistic Inquiry, Yvonna Lincoln and Egon Guba argued that such work rests on different assumptions and should be judged by its own criteria of trustworthiness; in 1989 Fourth Generation Evaluation organised a whole approach around negotiation among stakeholders. The same year N. L. Gage wrote about the paradigm wars in research on teaching, and Jennifer Greene, Valerie Caracelli and Wendy Graham reviewed 57 mixed-method evaluations to set out five purposes for combining methods.

Far from American campuses, a parallel challenge was growing. Robert Chambers showed in 1983 how short official visits to villages were biased towards roadsides, better off people and the dry season, and in 1994 he documented participatory rural appraisal, in which rural people map and analyse their own realities while the outsider facilitates. In Bangladesh in 1994 the Christian Commission for Development in Bangladesh first used the technique Rick Davies designed, which Jess Dart later trialled in Australia and which became Most Significant Change: a way to collect stories of change and choose openly among them. In 1998 Donna Mertens led the American

KEY DATES



Evaluation Association under the theme of transforming society through evaluation, and went on to build a transformative paradigm that starts from power and justice.

In our reading, this act answered the use revolution by asking whose use, and whose truth. It widened the field's circle of legitimate knowers, from experts to participants, and made the evaluator's own values part of the question.

THE GREAT DEBATE OF THIS ACT

The Paradigm Wars Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined? p. 402

THE GIANTS OF THIS ACT



Robert Chambers

The former colonial administrator who spent a career asking development professionals to learn from the people they usually see last.

p. 296



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318



Jennifer C. Greene

The methodologist who argued that evaluators are never neutral, and that mixing methods should mean mixing ways of seeing.

p. 340



Rick Davies and Jess Dart

Two practitioners who turned the stories people tell about change into a disciplined monitoring method.

p. 362

Bagele M. Chilisa, in one line

Bagele Chilisa gave the communities that research once studied a say in deciding what counts as knowledge.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

13



Portrait adapted from a photograph; credit on p. 590.

Robert Chambers

Active 1980s to 2010s

Britain; worked mainly in Kenya and South Asia, including India, Bangladesh and Nepal

Participation, equity and voice | Made in the Global South

The former colonial administrator who spent a career asking development professionals to learn from the people they usually see last.

IN ONE LINE

Robert Chambers taught a profession to ask whom it had not met, and then to hand them the chalk.

In 1958 a young Cambridge graduate began work as a District Officer in Samburu, in northern Kenya. A quarter of a century later he published a book with an uncomfortable message for his own profession: development experts were systematically failing to see the poorest people they claimed to serve. Rural Development: Putting the Last First appeared in 1983, and its English edition went through 14 printings. By 1994, when Robert Chambers described participatory rural appraisal in the journal World Development, the approach was spreading fast: one programme alone had run about 30 field training workshops in 15 countries, and Kenya's soil and water conservation branch had adopted it in more than 40 districts. Yet his own institute had at first turned him down for a professorship, because he had published too few journal articles.

Life CHAPTER 13

Robert Chambers came to development through administration rather than theory. He studied at Cambridge, starting in natural sciences and switching to history after National Service. From 1958 he served as a District Officer in Kenya, first for about two and a half years in Samburu district and then for a year in Nyeri. He then spent about five years training administrators at the Kenya Institute of Administration and the East African Staff College. A doctorate on land settlement schemes produced his 1969 book Settlement Schemes in Tropical Africa. Long before he wrote about other people's blind spots, he had spent years inside the system that produced them.

The Institute of Development Studies at the University of Sussex became his base from 1969 and he has been a Fellow there since 1972, with long periods abroad. India shaped him as

BIG IDEAS

1. **Putting the last first: offsetting the biases of rural development tourism**
2. **Participatory rural appraisal (PRA)**
3. **Behaviour and attitudes before methods**

KEY WORKS

Rural Development: Putting the Last First
1983

Whose Reality Counts? Putting the First Last
1997

Handbook on Community-Led Total Sanitation
2008 · With Kamal Kar

much as Kenya did. He worked there on Green Revolution research in South India and Sri Lanka in 1973 and 1974, with the Ford Foundation from 1981 to 1984, and in Hyderabad from 1989 to 1991, the years when participatory rural appraisal was emerging. A 2026 reflection published by the IDS Participatory Methods team notes that he was at first turned down for a professorship at IDS because he had too few journal articles, a detail that sits oddly beside his later reputation as one of the most cited writers in development studies.

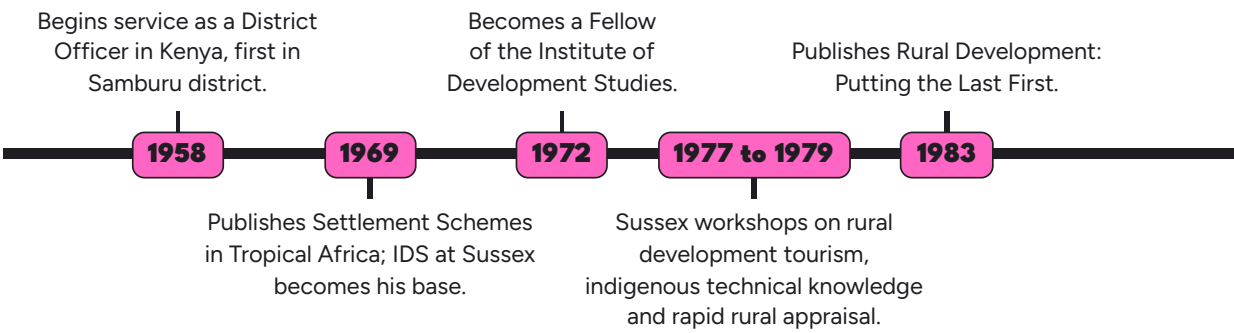
From the late 1970s he helped convene the Sussex workshops on what he called rural development tourism, indigenous technical knowledge and rapid rural appraisal. His 1983 book Rural Development: Putting the Last First, which the International Institute of Social Studies reports went through 14 printings of its English edition, argued that outsiders systematically miss the poorest people. In 1994 he set out the origins and practice of participatory rural appraisal in World Development, and in 1997 he published Whose Reality Counts? Putting the First Last. Later books, including Revolutions in Development Inquiry (2008), Provocations for Development (2012) and Can We Know Better? (2017), kept pressing the same question about whose knowledge counts.

His interests moved with the work. He took part in the World Bank's Consultations with the Poor in 1999 and 2000, wrote with Kamal Kar the 2008 Handbook on Community-Led Total Sanitation with support from Plan International UK and IDS, and became involved with the IDS Sanitation Learning Hub, including rapid action learning linked to India's sanitation campaign between 2014 and 2019. He has received honorary doctorates from the Universities of Sussex, Edinburgh and East Anglia, and from Erasmus University Rotterdam in 2013. In 2011 colleagues published a festschrift, Revolutionizing Development, edited by Andrea Cornwall and Ian Scoones. He describes his own way of working as a kind of opportunistic nomadism: moving into a new field, and leaving once its insights become routine.

The spark

Chambers's spark was a long exposure to a gap: the difference between what officials knew and what rural life was

TIMELINE · LIFE AND WORK



really like. He saw it as an administrator and trainer in Kenya, in doctoral work on settlement schemes, and in Green Revolution research in South India and Sri Lanka in 1973 and 1974. By the late 1970s the problem had a name. The Sussex workshops he helped convene from 1977 called it rural development tourism, and asked how outsiders could learn faster and more fairly.

The breakthrough CHAPTER 13

Imagine a senior official visiting villages to understand rural poverty. The car keeps to tarmac roads. The visit happens in the dry season, when travel is easy. The people who come forward to talk are mostly men, the better off and those already using services. The official sees the project sites the programme wants to show. Each choice is reasonable. Together they hide the poorest people almost completely. Chambers named this rural development tourism and listed its biases in Putting the Last First (1983): towards places near towns and roads, towards project sites, towards better off people, men and service users, towards the dry season and towards each visitor's own specialism. The alternative, long questionnaire surveys, was slow and costly, and often reported findings after decisions had been taken.

Rapid rural appraisal, discussed at Sussex workshops from 1977 to 1979 and at an international conference at Khon Kaen University in Thailand in 1985, offered quicker and cheaper learning. The decisive shift came between 1988 and 1991. In Kenya, the National Environment Secretariat and Clark University produced a village resource management plan at Mbusanyi in Machakos District in 1988, and in India, the Aga Khan Rural Support Programme, MYRADA and ActionAid in Bangalore were experimenting with local people drawing and analysing their own maps and diagrams. Chambers documented these strands in his 1994 World Development article, which defined participatory rural appraisal as a family of approaches enabling rural people to share, enhance and analyse their own knowledge, plan

TIMELINE, CONTINUED

1985 — International rapid rural appraisal conference at Khon Kaen University, Thailand.

1989 to 1991 — Works in Hyderabad, India, as participatory rural appraisal emerges; teaches a course at the Administrative Staff College of India in 1990.

1994 — Publishes The Origins and Practice of Participatory Rural Appraisal in World Development.

1997 — Publishes Whose Reality Counts? Putting the First Last.

2008 — Publishes the Handbook on Community-Led Total Sanitation with Kamal Kar.

and act. The outsider's role changed from collecting information to facilitating it.

The problem it solved

Top-down planning, brief roadside field visits and slow, extractive questionnaire surveys that ignored local knowledge and arrived too late to be useful.

Impact WHAT CHANGED

PRA spread fast. Chambers's 1994 account records that IIED's Sustainable Agriculture Programme had run about 30 field-based training workshops in 15 countries, that Indian organisations such as MYRADA, ActionAid Bangalore and the Aga Khan Rural Support Programme (India) were early hubs alongside government bodies such as the Karnataka Drylands Development Board, and that international field workshops in India in 1992 and 1993 helped carry the approach to many other countries through South to South sharing. He himself taught a rapid and participatory rural appraisal course at the Administrative Staff College of India in Hyderabad in 1990. IDS credits him with popularising concepts now central to development, including vulnerability, multidimensional poverty, sustainable livelihoods and farmer first.

For evaluators, the legacy is practical. Social maps, seasonal calendars, matrix scoring and wealth ranking appear routinely in NGO baselines and participatory evaluations, and the IDS profile notes that his work has challenged monitoring and evaluation, institutional learning and professional training. In our reading, approaches such as community scorecards and citizen-led monitoring share PRA's premise that communities can generate and judge evidence about their own lives, although a direct line of descent should not be assumed for every method. His involvement in the World Bank's Consultations with the Poor in 1999 and 2000, across 23 countries, also shows the ideas reaching the largest development institutions.

WHERE PEOPLE DISAGREE

Debates

Success brought misuse. Cornwall and Pratt, writing in 2011 and summarised by the IDS Participatory Methods team, documented PRA's rapid spread and frequent mechanical, commodified and over-used applications. Cooke and Kothari's 2001 collection Participation: The New Tyranny? argued that facilitation can become a form of manipulation, and some researchers saw quick PRA exercises as shortcuts that bypass rigorous inquiry. Defenders reply that including local voices is precisely PRA's strength. Chambers himself had warned in 1994 against ritual performance and relabelled extractive work, which suggests the critics and the originator agreed more than the debate sometimes implies.

LIMITS, IN OUR READING

- ☐ Cornwall and Pratt (2011) documented PRA's rapid spread and frequent misuse, including mechanical and commodified applications; Cooke and Kothari (2001) argued that participation can become manipulation through facilitation. Chambers himself reflected on tokenistic uses.
- ☐ Some critics say PRA lacks a systematic, evidence-based methodology; treat this as a general criticism rather than a settled finding.

WHY THIS MATTERS TO YOU

Chambers's first question works anywhere: whom are you not seeing? Before you visit a school, a ward or a village, check your plan against his biases. Are you only meeting the people who are easy to reach, at the time that is easy for you? Then try his deeper lesson. Hand over the pen, or the chalk, and let people draw, map and rank their own lives. You will find evidence that no form can reach. If you work in India, remember that much of this practice was shaped by Indian organisations such as MYRADA, ActionAid and the Aga Khan Rural Support Programme. It is a heritage worth building on.

Did you know?

He was at first turned down for a professorship at IDS because he had too few academic journal publications, according to a 2026 reflection on the IDS Participatory Methods site.

14 printings

English edition of Rural Development: Putting the Last First (1983)

International Institute of Social Studies, Robert Chambers Honorary Fellow page

30 workshops in 15 countries

Field-based PRA training workshops run by IIED's Sustainable Agriculture Programme by the time of his 1994 account

Chambers, The Origins and Practice of Participatory Rural Appraisal, World Development 22(7), 1994

16 books

Books written by Chambers by the time of his 2013 ISS honorary doctorate, the most recent then being Provocations for Development (2012)

International Institute of Social Studies, Robert Chambers Honorary Fellow page

23 countries

Countries covered by the World Bank's Consultations with the Poor (1999 to 2000), in which he took part, with over 2,000 focus groups

International Institute of Social Studies, Robert Chambers Honorary Fellow page

Key concepts

Rural development tourism

Chambers's term for brief, hurried visits by outsiders to rural areas, which tend to see the better off and miss the poorest.

Rapid rural appraisal (RRA)

An approach that emerged in the late 1970s for quicker, cheaper learning about rural life; mainly extractive, with outsiders collecting and analysing information.

Participatory rural appraisal (PRA)

A family of approaches and methods enabling rural people to share, enhance and analyse their own knowledge, and to plan and act, with outsiders as facilitators.

Participatory learning and action (PLA)

A later umbrella term for participatory approaches that grew out of RRA and PRA.

Transect walk

A walk through an area with local people, observing and discussing land use, resources and problems along the way.

Wealth or well-being ranking

A method in which local people sort households by their own criteria of wealth or well-being, revealing local definitions of poverty.

Sources for this chapter: p. 582

THE BIG IDEAS OF THIS CHAPTER

1. **Putting the last first: offsetting the biases of rural development tourism** p. 304

   What outsiders know about poverty is shaped by where, when and whom they visit, so evaluators must name those biases and deliberately reverse them to reach the people usually seen last.

2. **Participatory rural appraisal (PRA)** p. 307

   PRA is a family of visual, group-based methods through which local people produce, analyse and act on their own knowledge, while outsiders act as facilitators.

3. **Behaviour and attitudes before methods** p. 310

   The quality of participatory evidence depends more on how outsiders behave, especially their willingness to hand over control and admit error, than on the correct performance of any tool.

GIANTS CONNECTED TO THIS CHAPTER



Rick Davies and Jess Dart

Most Significant Change brings participants' own accounts into monitoring.

p. 362



Sonal Zaveri

Calls for a more nuanced view of community and participation in South Asia.

p. 536



Bagele M. Chilisa

Indigenous and relational methodologies ask whose knowledge counts at a deeper level.

p. 498

KEY WORKS TO READ IN THE ORIGINAL

- **Rural Development: Putting the Last First** 1983
- **Whose Reality Counts? Putting the First Last** 1997
- **Handbook on Community-Led Total Sanitation** 2008, With Kamal Kar

1

BIG IDEA 1 OF 3 · CHAMBERS

Putting the last first: offsetting the biases of rural development tourism

What outsiders know about poverty is shaped by where, when and whom they visit, so evaluators must name those biases and deliberately reverse them to reach the people usually seen last.

PICTURE IT

Think of a tourist who only visits a city's famous sights, in good weather, guided by the tourist office. They go home sure they have seen the city. Chambers showed that development experts often travel the same way, and that the poorest people live off the tourist route.

The problem

In the late 1970s officials, researchers and evaluators learned about rural poverty mainly through short visits and long questionnaire surveys. Short visits clustered near roads, towns and showcase projects, happened in the comfortable season and met the better off, men and service users. Surveys were slow and often reported after decisions were made. The result was a systematic blind spot: the poorest households, remote hamlets and the hungry months were barely seen, so programmes were designed and judged on a flattering, partial picture.

What it is

Putting the last first is Chambers's argument, set out in Rural Development: Putting the Last First (1983), that development knowledge is distorted by the way it is gathered, and that the distortion can be named and corrected. He grouped the distortions of rural development tourism into six biases: spatial (staying near tarmac, towns and roadsides), project (visiting places where something is being done, often showcase sites), person (meeting elites, men, users of services and the active and healthy), seasonal (visiting in dry, cool, accessible months and missing the wet season of hunger and sickness), diplomatic (not seeking out the poor for fear of embarrassing hosts) and professional (seeing only through one's own discipline). His remedy was a set of reversals: going further, staying longer, sitting, asking and listening, respecting indigenous technical knowledge and learning from the poorest. The underlying

logic is that an evaluator's itinerary is itself a sampling design, and a biased itinerary produces biased evidence however careful the later analysis.

How it works

1. **Write down the plan as it stands**
   List the villages, dates, times and people on the proposed field schedule, and who chose them.
2. **Check it against the six biases**
   For each of the biases Chambers named in 1983, spatial, project, person, seasonal, diplomatic and professional, note how the schedule is exposed.
3. **Redesign to offset each bias**
   Add places off the road and outside the project, visits in the difficult season, and times when working people are free.
4. **Ask who is not here**
   In every meeting, ask who did not come and why, then go and find some of those households.
5. **Sit, ask and listen**
   Spend time learning local people's own explanations and technical knowledge before testing outsiders' assumptions, as Chambers urged in his reversals.
6. **Record what the old plan would have missed**
   Note findings that came only from the added places, seasons and people, and report them alongside the main results.

WORKED EXAMPLE · AN ILLUSTRATION

A team evaluating a watershed programme in a drought-prone district of Karnataka plans five village visits in January, all within a short drive of the block office, with meetings arranged by the implementing NGO. Checking the plan against Chambers's six biases, the team sees spatial, project, seasonal and person bias at once. It adds two hamlets beyond the end of the bus route, one village where the programme never worked, and a second round of visits in the lean months before the monsoon. At each meeting it asks which households are absent and visits several of them. The additional visits show that landless families, who own no fields to treat, have gained little, a finding the original schedule would almost certainly have hidden.

CONCEPT HIGHLIGHT

Rural development tourism

Chambers's term for brief, hurried visits by outsiders to rural areas, which tend to see the better off and miss the poorest.

USE IT WHEN

- Planning field visits, monitoring trips or qualitative fieldwork in rural or dispersed settings.
- Reviewing why a programme's monitoring data looks better than people's lived experience.
- Designing samples where the hardest-to-reach households matter most to the evaluation question.

SKIP IT WHEN

- The evaluation already uses a probability sample from a complete and current listing, and field visits play no part in the evidence.
- Time and budget allow only a single contact with a known population, such as a census of all participants by phone.

COMMON MISREADINGS

- [ ] Treating it as a call to abandon surveys altogether, when Chambers criticised their cost and slowness, not the idea of systematic data.
- [ ] Adding one remote village as a token while keeping every other bias in the schedule unchanged.

WHY IT WAS GROUNDBREAKING

It turned the ordinary field visit, which most professionals treated as neutral, into an object of scrutiny with nameable and correctable biases. In our reading it made the evaluator's own position part of the evidence problem, decades before reflexivity became a common word in evaluation.

Origin

The ideas were developed at Sussex workshops on rural development tourism and rapid rural appraisal between 1977 and 1979, and set out in full in Rural Development: Putting the Last First (Longman, 1983), whose English edition went through 14 printings according to the International Institute of Social Studies. Chambers returned to them in Whose Reality Counts? Putting the First Last (1997).

RELATED IDEAS

Participatory rural appraisal (PRA)

- SZ Zaveri 536
- BC Chilisa 498
- D&D Davies and Dart 362

CONCEPT HIGHLIGHT

Rapid rural appraisal (RRA)

An approach that emerged in the late 1970s for quicker, cheaper learning about rural life; mainly extractive, with outsiders collecting and analysing information.

2

BIG IDEA 2 OF 3 · CHAMBERS

Participatory rural appraisal (PRA)

PRA is a family of visual, group-based methods through which local people produce, analyse and act on their own knowledge, while outsiders act as facilitators.

PICTURE IT

Think of the difference between a doctor writing notes about you and you drawing a map of where it hurts. PRA hands the pen to the people whose lives are being studied, so that they produce and analyse the evidence themselves.

The problem

Rapid rural appraisal had made outsiders' learning faster and cheaper, but it was still extractive: visiting teams collected information and took it away to analyse elsewhere. Communities rarely saw the results, could not correct mistakes and had no say in what was concluded about their own lives. For evaluators this meant findings that missed local logic and priorities, and programmes that treated poor people as respondents rather than as capable analysts of their own conditions.

What it is

In his 1994 World Development account, Chambers defined participatory rural appraisal as a growing family of approaches and methods that enable local people to share, enhance and analyse their knowledge of life and conditions, and to plan and act. He traced it to five streams: activist participatory research influenced by Paulo Freire; agroecosystem analysis, developed in Thailand from 1978 by Gordon Conway and colleagues; applied anthropology; field research on farming systems; and rapid rural appraisal. Its methods include social and resource mapping and modelling on the ground, transect walks, timelines and trend analysis, seasonal calendars, daily time use, Venn diagrams, matrix scoring and ranking, wealth and well-being ranking, and estimation with local materials such as seeds and stones. The logic is a shift along a continuum: information that rapid appraisal owned and used for outsiders becomes, in PRA, shared and owned by local people, who carry the analysis through to planning and action.

How it works

1. **Agree a question worth answering**
   Frame one question, such as access to water or seasonal hunger, that both the community and the programme want answered.
2. **Convene a mixed group and explain the purpose**
   Chambers's 1994 account describes the outsider as convener, catalyst and facilitator, so the team introduces itself and the purpose and then steps back.
3. **Begin with a map or model on the ground**
   Invite people to draw a social or resource map with chalk, sticks or coloured powder, adding homes, wells, fields and services as they see them.
4. **Sequence further methods**
   Follow with a transect walk, seasonal calendar, timeline, wealth ranking or matrix scoring, each building on what the earlier exercise revealed; the 1994 article lists sequences of methods as a method in itself.
5. **Let local people present and check findings**
   Ask the group to explain what its diagrams show and to correct each other, so that analysis happens in the village before anything is recorded.
6. **Compare across groups**
   Repeat key exercises with women, poorer households or other groups and discuss the differences rather than averaging them away.
7. **Move to planning and monitoring**
   Use the outputs for participatory planning and later monitoring, leaving the original charts with the community.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO preparing the baseline for a drinking water project in a tribal block of Odisha replaces part of its household questionnaire with a PRA session. A mixed group draws the village on the ground with chalk and coloured powder, marking handpumps, a seasonal stream and every household. Women then mark which homes run out of water first in summer, and the facilitator walks the route with them on a transect. A second session with older women produces a seasonal calendar showing when girls miss school to fetch water. The group itself identifies three hamlets as the priority and presents its analysis to the panchayat. The NGO photographs the map for its records, but the map stays in the village and becomes the reference point for later monitoring visits.

USE IT WHEN

- Baselines and needs assessments where local definitions of poverty, risk or access matter.
- Participatory monitoring where communities will keep using the information themselves.
- Settings with low literacy, where visual methods let more people take part than a written survey.

SKIP IT WHEN

- The evaluation needs statistically representative estimates for a large population and has no time to pair PRA with a survey.
- Social divisions or risks make open group discussion unsafe for the people whose views matter most.

COMMON MISREADINGS

- Treating PRA as a set of tools for faster data extraction, which Chambers explicitly distinguished from PRA.
- Assuming that a group exercise is participatory simply because many people attended, whatever happened to the analysis afterwards.

WHY IT WAS GROUNDBREAKING

It changed who produces and analyses development data, treating poor and often non-literate people as skilled analysts. It also gave practitioners a portable toolkit that spread through training and South to South exchange rather than through journals.

Origin

Chambers dates the decisive shift to 1988 to 1991: a Kenyan village resource management plan at Mbusanyi, Machakos District, in September 1988, and work with the Aga Khan Rural Support Programme (India) in Gujarat in the same year, followed by MYRADA and ActionAid in Bangalore. He documented the approach in IDS notes in 1990 and in World Development in 1994; participatory learning and action later became an umbrella name.

RELATED IDEAS

Putting the last first

Behaviour and attitudes before methods

D&D Davies and Dart 362 · SZ Zaveri 536 · BC Chilisa 498

CONCEPT HIGHLIGHT

Participatory learning and action (PLA)

A later umbrella term for participatory approaches that grew out of RRA and PRA.

3

BIG IDEA 3 OF 3 · CHAMBERS

Behaviour and attitudes before methods

The quality of participatory evidence depends more on how outsiders behave, especially their willingness to hand over control and admit error, than on the correct performance of any tool.

PICTURE IT

Think of a good host. Fine furniture and an expensive menu matter less than whether guests feel listened to and welcome. Chambers argued that participatory methods work the same way: how the facilitator behaves matters more than the tools.

The problem

As PRA spread in the early 1990s, many organisations adopted its tools while keeping old habits. Facilitators lectured, held the chalk, let village leaders dominate and wrote up conclusions they had brought with them. Manuals encouraged ritual performance of exercises, and some teams relabelled extractive surveys as PRA. For communities this meant being consulted without being heard; for evaluators it meant participatory-looking evidence that quietly reproduced the same blind spots.

What it is

In his 1994 account Chambers described the behaviour and attitudes of outsiders as more important than the methods themselves or their correct performance. The idea has three linked parts. First, personal conduct: outsiders listen, show respect, do not rush, and hand over the stick or chalk so that local people lead. Second, judgement over rules: the only rule PRA proposed was that practitioners use their own best judgement at all times, which is why Chambers warned against manuals and ritual. Third, self-critical awareness: practitioners report their own problems and errors openly, and criticise work that is extractive however it is labelled. The underlying logic, developed further in Challenging the Professions (1993) and Whose Reality Counts? (1997), is that power shapes evidence. Who speaks, who is silent and who holds the pen determine whose reality ends up on the chart, so methods without changed behaviour simply move old biases into a new format.

How it works

1. **Agree team norms before fieldwork**
   Chambers's 1994 list of PRA practices includes team contracts, so the team agrees in advance how it will behave, for example not interrupting and not correcting people's diagrams.
2. **Hand over the chalk**
   Once an exercise begins, the facilitator gives the materials to participants and steps back to observe.
3. **Watch who dominates**
   One team member notes who speaks, who draws and who stays at the edge, and the team creates separate space for those left out.
4. **Use judgement, not the manual**
   Adapt or drop a method when it is not working for the group, following PRA's one stated rule of using your own best judgement.
5. **Debrief self-critically**
   After each day, the team discusses what it got wrong and whose views it may have missed.
6. **Report errors honestly**
   Include problems and mistakes in the report, and do not call an exercise PRA if local people did not analyse the results.

WORKED EXAMPLE · AN ILLUSTRATION

Two teams run wealth ranking in neighbouring villages for the same livelihoods evaluation in Jharkhand. In the first, the sarpanch holds the chalk, the facilitator summarises the ranking and the report calls it a participatory exercise. In the second, the team has agreed beforehand not to touch the materials. One member notices that women and Dalit households sit at the edge, so the team runs a separate ranking with them and then brings both groups together to discuss where the rankings differ. Its daily debrief records that the first session excluded two hamlets, and the report says so. On paper both teams used the same tool; in our reading of Chambers, only the second practised PRA, and only its evidence can be trusted to reflect the poorest households.

CONCEPT HIGHLIGHT

Participatory rural appraisal (PRA)

A family of approaches and methods enabling rural people to share, enhance and analyse their own knowledge, and to plan and act, with outsiders as facilitators.

USE IT WHEN

- Training field teams before any participatory or qualitative fieldwork.
- Reviewing why participatory exercises in a programme produce suspiciously uniform results.
- Commissioning participatory evaluations, as a basis for quality criteria.

SKIP IT WHEN

- Never skip the principle itself; it applies wherever outsiders gather evidence from people.
- A full self-critical debrief cycle may be impractical for a single short consultation, where a brief check on who was excluded is the minimum.

COMMON MISREADINGS

- [ ] Reading it as a rejection of method and rigour, when Chambers argued that good behaviour is what makes the methods rigorous.
- [ ] Treating friendliness as enough, without actually transferring control of the analysis to participants.

WHY IT WAS GROUNDBREAKING

It located the rigour of participatory evidence in the conduct of the facilitator, not only in the tool. It also named the risk of tokenistic participation from inside the movement, years before critics made it a central debate.

Origin

The emphasis appears in Chambers's 1994 World Development article on PRA and in Challenging the Professions (1993). He extended it in Whose Reality Counts? Putting the First Last (1997), and later critics such as Cooke and Kothari's Participation: The New Tyranny? (2001) and Cornwall and Pratt (2011) took up the same concern about tokenistic and commodified participation.

RELATED IDEAS

Participatory rural appraisal (PRA)

G&L Guba and Lincoln 318
DM Mertens 384
SZ Zaveri 536

CONCEPT HIGHLIGHT

Transect walk

A walk through an area with local people, observing and discussing land use, resources and problems along the way.

RELATED GIANTS



Sonal Zaveri

The Indian evaluator who carried participatory, feminist practice from grassroots work into regional networks and real-time evaluation.

p. 536



Bagele M. Chilisa

The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

p. 498



Rick Davies and Jess Dart

Two practitioners who turned the stories people tell about change into a disciplined monitoring method.

p. 362



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318

CONCEPT HIGHLIGHT

Wealth or well-being ranking

A method in which local people sort households by their own criteria of wealth or well-being, revealing local definitions of poverty.

THE BIG IDEAS OF THIS CHAPTER

1. **Putting the last first: offsetting the biases of rural development tourism** p. 304

   What outsiders know about poverty is shaped by where, when and whom they visit, so evaluators must name those biases and deliberately reverse them to reach the people usually seen last.

2. **Participatory rural appraisal (PRA)** p. 307

   PRA is a family of visual, group-based methods through which local people produce, analyse and act on their own knowledge, while outsiders act as facilitators.

3. **Behaviour and attitudes before methods** p. 310

   The quality of participatory evidence depends more on how outsiders behave, especially their willingness to hand over control and admit error, than on the correct performance of any tool.

Put it to work

USE IT ON MONDAY · PARTICIPATORY RURAL APPRAISAL: PUTTING THE LAST FIRST

1. Replace one section of your baseline questionnaire with a participatory session, such as a social map or a wealth ranking, and let the group analyse what they have drawn before you do.

2. Plan field visits that go beyond the road, the office and the project site: the far hamlet, the difficult season, the time of day when working people are free.

3. In every participatory exercise, ask who holds the pen, then hand it over. Note whose reality ends up on the chart and whose does not.

AI prompt **Copy, fill the brackets, then check every line**

Help me plan a participatory rural appraisal session, in the spirit of Robert Chambers, for **[programme]** in **[village or district]**. The question we need to answer is **[question]**. The participants are likely to be **[describe groups]**. Local constraints are **[time, season, language, literacy, social divisions]**. (1) Suggest two or three PRA tools that fit the question and constraints, with step-by-step facilitation notes. (2) List who might be excluded and how to reach them. (3) Suggest prompts that let the group analyse its own output. (4) Suggest how to return findings to the community. Do not predict what the community will say.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a tourist who only visits a city's famous sights, in good weather, guided by the tourist office. They go home sure they have seen the city. Chambers showed that development experts often travel the same way, and that the poorest people live off the tourist route.

YOUR MOVES Three moves from Chambers's path

1 Earn the right to challenge the system

Chambers spent years as an administrator, trainer and researcher before questioning how professionals work, so his critique came from inside.

TRY THIS
Before proposing a change to a system, spend time doing the job you want to change.

2 Credit the people who taught you

He credited practitioners in Kenya and India for key breakthroughs in participatory methods, and documented their work.

TRY THIS
In your next report, name the field staff and community members whose insight shaped the findings.

3 Measure impact by usefulness

Turned down for a professorship over too few journal articles, he still became one of development's most influential writers, and practitioners around the world used his methods.

TRY THIS
Ask three practitioners what would make your work more useful to them, and change one thing.

THE BIG IDEAS OF THIS CHAPTER

1 **Putting the last first: offsetting the biases of rural development tourism** p. 304

What outsiders know about poverty is shaped by where, when and whom they visit, so evaluators must name those biases and deliberately reverse them to reach the people usually seen last.

2 **Participatory rural appraisal (PRA)** p. 307

PRA is a family of visual, group-based methods through which local people produce, analyse and act on their own knowledge, while outsiders act as facilitators.

3 **Behaviour and attitudes before methods** p. 310

The quality of participatory evidence depends more on how outsiders behave, especially their willingness to hand over control and admit error, than on the correct performance of any tool.

CONCEPT HIGHLIGHT

Wealth or well-being ranking

A method in which local people sort households by their own criteria of wealth or well-being, revealing local definitions of poverty.

GIANTS CONNECTED TO THIS CHAPTER



Rick Davies and Jess Dart

Most Significant Change brings participants' own accounts into monitoring.

p. 362



Sonal Zaveri

Calls for a more nuanced view of community and participation in South Asia.

p. 536



Bagele M. Chilisa

Indigenous and relational methodologies ask whose knowledge counts at a deeper level.

p. 498

KEY WORKS TO READ IN THE ORIGINAL

- **Rural Development: Putting the Last First** 1983
- **Whose Reality Counts? Putting the First Last** 1997
- **Handbook on Community-Led Total Sanitation** 2008, With Kamal Kar

CONCEPT HIGHLIGHT

Rural development tourism

Chambers's term for brief, hurried visits by outsiders to rural areas, which tend to see the better off and miss the poorest.

WORKSHEET · TRY THIS TEMPLATE

PRA session plan

Print this page, or copy the headings into your notes.

1 Question we need answered

One question, in plain words, that the community also finds worth answering.

2 Tool

Social map, transect walk, wealth ranking, seasonal calendar, timeline or matrix scoring.

3 Who takes part, who might be missing

Think about gender, caste, age, disability, distance and working hours.

4 Who holds the pen

Plan how the facilitator steps back once the exercise starts.

5 Group analysis

How the group will interpret its own output, and who keeps the original.

6 Return of findings

When and how the results come back to the community.

REFLECT

On your last field visit, whom did you not meet, and why?

CONNECTS TO

D&D **Rick Davies and Jess Dart** Most Significant Change brings participants' own accounts into monitoring. p. 362

SZ **Sonal Zaveri** Calls for a more nuanced view of community and participation in South Asia. p. 536

BC **Bagele M. Chilisa** Indigenous and relational methodologies ask whose knowledge counts at a deeper level. p. 498

14



Emblem: an original HubForge design for this figure, not a likeness.

Egon G. Guba and Yvonna S. Lincoln

Guba 1924 to 2008; Lincoln born 1944 Active 1980s

United States; worldwide influence through qualitative research

Methods, rigour and causality | Participation, equity and voice

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

IN ONE LINE

They showed that listening carefully to how people understand their own lives can be as rigorous as measuring them.

In the Second World War, a young American soldier served with the 381st Combat Engineers Battalion, building bridges, destroying bunkers and, after D-Day, disarming booby traps. Egon Guba came home and became a statistician, earning a Chicago doctorate in statistics and measurement and teaching both. That makes what happened next all the more striking. It was this insider of the measurement tradition who, with Yvonna Lincoln, mounted one of the most influential challenges to it. Their 1985 book Naturalistic Inquiry argued that studies built on interviews and observation should be judged by their own standards of trustworthiness. Four years later, Fourth Generation Evaluation put the people affected by a programme, including those it might harm, at the centre of judging it. Today their four words, credibility, transferability, dependability and confirmability, appear in the methods sections of qualitative studies around the world.

Life CHAPTER 14

Egon Gotthold Guba was born in Chicago in 1924 to Oswald and Roszina Schell Guba, grew up a native German speaker. He studied physics and engineering at Valparaiso University for two years before serving in the US Army in the Second World War with the 381st Combat Engineers Battalion, building bridges, destroying enemy bunkers, disarming booby traps after D-Day and at times working as an interpreter. After the war he completed his bachelor's degree, took a master's degree in education at the University of Kansas and a doctorate at the University of Chicago in educational inquiry, with a focus on statistics and measurement. Among his influences there were Ralph Tyler, J. W. Getzels and Edgar Dale.

Guba taught statistics and measurement at Chicago and

BIG IDEAS

1. **Trustworthiness and authenticity**
2. **Fourth generation evaluation**
3. **Naturalistic and constructivist inquiry**

KEY WORKS

Naturalistic Inquiry
1985 · Lincoln and Guba

But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation
1986 · Lincoln and Guba

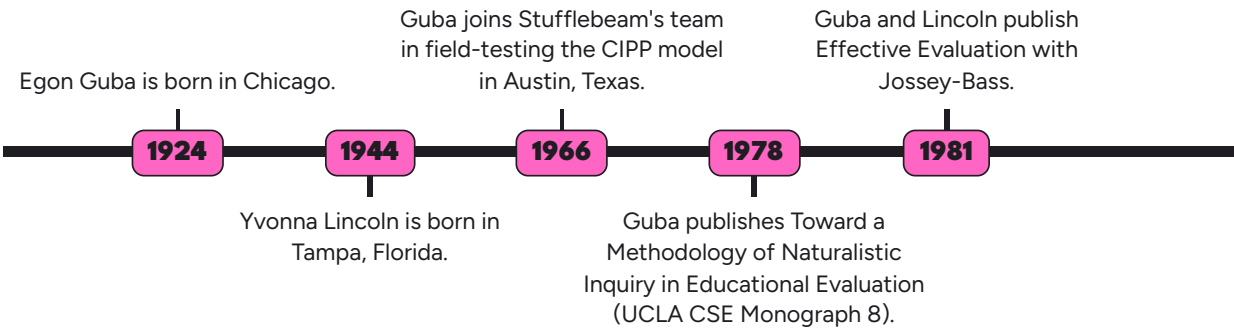
Fourth Generation Evaluation
1989 · Guba and Lincoln

at the University of Kansas City, then spent about ten years at Ohio State University, where he directed the Bureau of Educational Research and Service, the unit that housed Daniel Stufflebeam's new Evaluation Center, and joined Stufflebeam in field-testing the CIPP model in Austin. The final 25 years of his career were at Indiana University Bloomington, where he served as an associate dean of the School of Education. His 1978 monograph for the Center for the Study of Evaluation at UCLA, Toward a Methodology of Naturalistic Inquiry in Educational Evaluation, made the case for naturalistic inquiry as distinct from experimental inquiry. Jennifer Greene later described him as having taken up and championed Robert Stake's responsive and case study ideas.

Yvonna Sessions Lincoln was born in Tampa, Florida, in 1944. She studied history and sociology at Michigan State University, took a master's degree in medieval and Eastern European history at the University of Illinois, and earned a doctorate in higher education at Indiana University, Guba's academic home. She taught at Stephens College, the University of Kansas and Vanderbilt University before joining Texas A&M University in 1991, where she held the Ruth Harrington Chair of Educational Leadership and was a Distinguished Professor of Higher Education; her publisher lists her as Professor Emerita there. She served as president of the American Evaluation Association and of the Association for the Study of Higher Education, and as vice president for postsecondary education of the American Educational Research Association.

The two worked together for decades, producing Effective Evaluation (1981), Naturalistic Inquiry (1985) and Fourth Generation Evaluation (1989), and they shared the American Evaluation Association's Paul F. Lazarsfeld Evaluation Theory Award in 1987. They were also married: Guba's obituary names Lincoln as his wife. Lincoln went on to co-edit, with Norman Denzin, six editions of the Handbook of Qualitative Research, as well as the SAGE Handbook of Critical and Indigenous Methodologies and the journal Qualitative Inquiry. Guba died in College Station, Texas, on 26 March 2008, aged 84.

TIMELINE · LIFE AND WORK



The spark

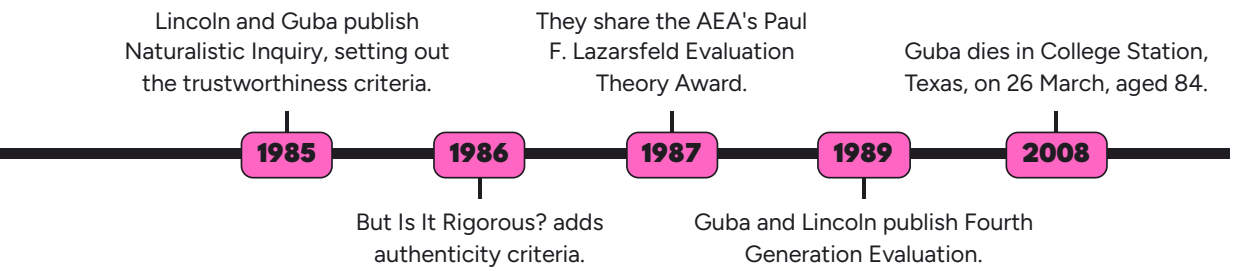
For Guba, the spark was a growing doubt. A statistician who had worked inside the measurement tradition, he came to question whether conventional methods could answer evaluation's real questions. According to the Evaluation Center, one study became the turning point: work he did at Ohio State with the doctoral student Thomas Owens, evaluating the effects of media and audiovisual equipment in four schools for Encyclopaedia Britannica, which became the platform for his naturalistic approach. Lincoln's doctorate in higher education at Indiana brought her to Guba's academic home, and to a partnership that lasted decades.

The breakthrough CHAPTER 14

Imagine presenting a study based on months of interviews and observation in one community, and being told it is soft, because it has no control group, no large sample and no statistical tests. By the 1970s, that was the charge facing qualitative evaluation in the United States. Judged by the standards built for experiments, internal and external validity, reliability and objectivity, work rooted in one setting looked weak. Evaluators were also meeting problems that technical rigour alone could not solve: studies tangled in politics and ethical dilemmas, and conclusions that went unused. Guba, a trained statistician, came to doubt that conventional approaches could answer those questions.

The decisive statement came in Naturalistic Inquiry, published by Sage in 1985. Building on Guba's 1978 monograph and their joint 1981 book Effective Evaluation, Lincoln and Guba argued that the naturalistic paradigm rests on different assumptions from conventional science, and that work within it should be judged by its own criteria of trustworthiness, each designed as a parallel to a conventional standard. In 1986, in But Is It Rigorous? Trustworthiness and Authenticity in Naturalistic Evaluation, published in New Directions for Program Evaluation, they

TIMELINE, CONTINUED



added criteria of authenticity. In Fourth Generation Evaluation (1989) they then turned these ideas into a complete approach to evaluation, organised around negotiation among stakeholders.

The problem it solved

Qualitative evaluators had no defensible answer to questions of rigour, and evaluations that ignored stakeholders ran into problems of politics, ethics and unused findings.

Impact WHAT CHANGED

Few ideas in qualitative research have travelled as far as Lincoln and Guba's trustworthiness criteria. Credibility, transferability, dependability and confirmability appear in university research guidance and are routinely invoked in the methods sections of qualitative studies and evaluations. For MEL practitioners they supply a common language when a funder asks how interview or focus group findings can be trusted, so member checks, thick description and audit trails have become ordinary items in an inception report. Jennifer Greene judged their framing of evaluation as a values-committed, emancipatory practice an even more significant contribution than their constructivism or their methods.

In development and Global South practice their influence is mostly indirect but visible. In our interpretation, participatory and stakeholder-led approaches that treat communities as holders of knowledge share the premise of fourth generation evaluation that values and multiple perspectives belong inside evaluation, although such approaches have several roots. Lincoln's later editorial work, including the Handbook of Qualitative Research and the SAGE Handbook of Critical and Indigenous Methodologies, extended that conversation to critical and indigenous research traditions. Their own practical guide, Guidelines and Checklist for Constructivist Evaluation (2001), distilled the 1989 book into eleven steps, from organising the evaluation and identifying stakeholders, including potential victims, to recycling unresolved issues, making the approach usable well beyond the academy.

WHERE PEOPLE DISAGREE

Debates

Their work sat at the centre of the paradigm debates of the 1980s and 1990s. In general terms, critics from the experimental tradition questioned whether a relativist position could yield findings that decision-makers could rely on, and whether parallel criteria could carry the weight of validity. Their advice against mixing positivist and constructivist propositions in one study ran against the later rise of mixed-methods evaluation, whose advocates argued that paradigms can be combined. In our view, full fourth generation evaluation is slow and costly, and negotiation among stakeholders with unequal power can reproduce that inequality unless it is skilfully facilitated. Lincoln and Guba, for their part, insisted that relativism did not mean that anything goes.

LIMITS, IN OUR READING

- ☐ Critics from the positivist tradition have questioned whether a relativist stance can support claims that decision-makers can rely on.
- ☐ In our view, negotiation among stakeholders can be slow and shaped by unequal power, so it needs skilled facilitation.

WHY THIS MATTERS TO YOU

Lincoln and Guba give anyone who listens to people for a living a way to be taken seriously. When someone asks how a handful of interviews can be trusted, you can answer: I checked my findings with the people I interviewed, described the setting richly, kept a clear record of my decisions and compared several sources. When a programme looks like a success to its managers and a failure to the families it serves, they show you how to bring both views into the room and work out what happens next. Most of all, they remind you that the people you study are not just sources of data. They hold knowledge, and they deserve a say.

6

editions of the Handbook of Qualitative Research co-edited by Yvonna Lincoln with Norman Denzin

Yvonna S. Lincoln, author page, SAGE Publishing

11

steps in their 2001 practitioner guide to constructivist, or fourth generation, evaluation

Guidelines and Checklist for Constructivist (a.k.a. Fourth Generation) Evaluation, Guba and Lincoln (2001)

25

years Guba spent at Indiana University Bloomington, the final stretch of his career

Founders: Ohio State, The Evaluation Center, Western Michigan University; Egon Guba obituary, Hillier Funeral Home (2008)

Key concepts

Trustworthiness
Lincoln and Guba's umbrella term for the quality of naturalistic inquiry, made up of credibility, transferability, dependability and confirmability, which parallel internal validity, external validity, reliability and objectivity.

Member check
Returning data or interpretations to the people who provided them so they can confirm or correct them; a key technique for establishing credibility.

Thick description
A detailed account of the setting and context of a study that allows readers to judge whether findings may transfer to their own situation.

Audit trail
A record of data, decisions and analysis that lets an independent auditor trace conclusions back to their sources, supporting dependability and confirmability.

Authenticity criteria
Fairness and ontological, educative, catalytic and tactical authenticity: standards asking whether an inquiry represented all views and helped stakeholders understand and act.

Claims, concerns and issues
The items stakeholders raise about a programme, which organise a fourth generation evaluation; agreed items go to the report and contested ones to negotiation.

Sources for this chapter: p. 582

Did you know?

Before he became a theorist of qualitative inquiry, Guba served in the Second World War with a combat engineer battalion that disarmed booby traps after D-Day, and he helped blow up the hardened submarine pens at Bremerhaven.

THE BIG IDEAS OF THIS CHAPTER

1. **Trustworthiness and authenticity** p. 326

   Naturalistic and qualitative inquiry should be judged by its own criteria of credibility, transferability, dependability and confirmability, together with authenticity criteria that ask whether it was fair and helped people act.

2. **Fourth generation evaluation** p. 329

   Organise evaluation around the claims, concerns and issues of all stakeholders, and make negotiation among them the heart of the process.

3. **Naturalistic and constructivist inquiry** p. 332

   Social realities are constructed by the people who live them, so inquiry should take place in natural settings, with the human inquirer as its main instrument, and seek to understand multiple constructions.

GIANTS CONNECTED TO THIS CHAPTER



Robert E. Stake

Responsive evaluation also organises the work around stakeholders' concerns.

p. 82



Daniel L. Stufflebeam

Guba's early collaborator, whose decision-oriented model offers a contrasting starting point.

p. 134



Jennifer C. Greene

Carried the engagement with values and multiple perspectives into mixed methods.

p. 340

KEY WORKS TO READ IN THE ORIGINAL

- **Naturalistic Inquiry** 1985, Lincoln and Guba
- **But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation** 1986, Lincoln and Guba
- **Fourth Generation Evaluation** 1989, Guba and Lincoln

1

BIG IDEA 1 OF 3 · GUBA AND LINCOLN

Trustworthiness and authenticity

Naturalistic and qualitative inquiry should be judged by its own criteria of credibility, transferability, dependability and confirmability, together with authenticity criteria that ask whether it was fair and helped people act.

PICTURE IT

Think of a courtroom. A jury does not ask whether a witness's story can be repeated in a laboratory. It asks whether the witness is credible, whether the story hangs together and whether it can be checked. Trustworthiness criteria judge qualitative evidence the way a good jury judges testimony.

The problem

By the late 1970s, evaluators who relied on interviews, observation and case studies faced a damaging charge. Judged by the conventional standards of internal validity, external validity, reliability and objectivity, a study of one setting built on people's own accounts looked soft. Reviewers and funders could dismiss such work as anecdote, and qualitative evaluators lacked a shared, defensible language for showing why their findings deserved trust. As a result, the experiences that qualitative work captured were easy to set aside when decisions about programmes were made.

What it is

In Naturalistic Inquiry (1985), Lincoln and Guba proposed four criteria of trustworthiness, each parallel to a conventional standard. Credibility, the counterpart of internal validity, concerns the truth value of findings; it is supported by prolonged engagement, persistent observation, triangulation, peer debriefing, negative case analysis and member checks. Transferability, the counterpart of external validity, is supported by thick description, so that readers can judge whether findings fit their own settings. Dependability, paralleling reliability, and confirmability, paralleling objectivity, are supported by an inquiry audit, in which an outsider examines the process and traces conclusions through an audit trail back to raw data; a reflexive journal also helps. In 1986 they added authenticity criteria with no conventional parallel: fairness, the balanced representation of all stakeholders' views; ontological authenticity, enlarging people's understanding of their

own situation; educative authenticity, improving their understanding of others; catalytic authenticity, stimulating action; and tactical authenticity, empowering people to act.

How it works

1. **Plan for credibility**
   Budget for prolonged engagement and persistent observation in the setting, and for triangulation across sources, methods and investigators.
2. **Debrief and look for disconfirmation**
   Arrange peer debriefing with a colleague outside the study, and search for negative cases that contradict emerging patterns.
3. **Member check**
   Return data, interpretations and draft conclusions to the people who provided them and record their corrections, a technique Lincoln and Guba treated as central to credibility.
4. **Write thick description**
   Describe settings, people and context richly enough that readers elsewhere can judge transferability for themselves.
5. **Keep an audit trail and reflexive journal**
   Log raw data, coding decisions, analysis notes and changes of design, and record the inquirer's own reflections as the work proceeds.
6. **Commission an inquiry audit**
   Invite an independent auditor to examine the process for dependability and the trail from data to conclusions for confirmability, as described in their 1988 paper on criteria for naturalistic reports.
7. **Check authenticity**
   Ask whether every stakeholder group was fairly represented, and whether participants came to understand themselves and others better and were helped to act.

WORKED EXAMPLE · AN ILLUSTRATION

An evaluator studies how women's self-help groups in three districts of Karnataka use a new digital bookkeeping app. She spends several weeks in each district, attending meetings rather than visiting once, and compares what members say in interviews with what she sees at meetings and in the app's records. A colleague from another organisation questions her emerging themes every fortnight. She takes draft findings back to two groups in each district, who correct her belief that older members avoid the app: in fact, they delegate data entry to their daughters. Her report describes each village's connectivity, literacy and bank access in detail, and a dated log of coding decisions lets a reviewer trace each conclusion to its source. Each step maps onto one of Lincoln and Guba's techniques.

USE IT WHEN

- Designing the interview, focus group or case study components of an evaluation.
- A funder questions whether qualitative findings can be trusted.
- You need to show that every stakeholder group was fairly represented.

SKIP IT WHEN

- The study is a standard survey or experiment, where conventional validity criteria apply directly.
- Time allows only token member checks; in that case state the limits honestly rather than claiming the labels.

COMMON MISREADINGS

- Naming the four criteria in a methods section without actually doing the techniques behind them.
- Treating member checks as asking participants to approve the report, rather than testing and correcting interpretations.

WHY IT WAS GROUNDBREAKING

It gave qualitative evaluators a structured, defensible answer to the charge that their work lacked rigour. The authenticity criteria went further, proposing that inquiry be judged partly by whether it was fair and helped people understand and act, an idea that, in our reading, anticipated later participatory and empowerment approaches.

Origin

Guba laid the groundwork in his 1978 UCLA monograph and, with Lincoln, in Effective Evaluation (1981). Naturalistic Inquiry (Sage, 1985) set out the four trustworthiness criteria and their techniques. But Is It Rigorous? (New Directions for Program Evaluation, 1986) added the authenticity criteria, and their 1988 paper Criteria for Assessing Naturalistic Inquiries as Reports, followed by Fourth Generation Evaluation (1989), applied the criteria to evaluation reports.

RELATED IDEAS

RS Stake 82 · JG Greene 340 · DM Mertens 384 · BC Chilisa 498

CONCEPT HIGHLIGHT

Trustworthiness

Lincoln and Guba's umbrella term for the quality of naturalistic inquiry, made up of credibility, transferability, dependability and confirmability, which parallel internal validity, external validity, reliability and objectivity.

2

BIG IDEA 2 OF 3 · GUBA AND LINCOLN

Fourth generation evaluation

Organise evaluation around the claims, concerns and issues of all stakeholders, and make negotiation among them the heart of the process.

PICTURE IT

Think of planning a neighbourhood park by talking only to the council's engineers, and then imagine planning it with the parents, the skateboarders, the elderly residents and the shopkeepers too. Fourth generation evaluation puts everyone with a stake around the table, and makes their negotiation the heart of the work.

The problem

Guba and Lincoln argued that evaluation's first three generations, focused on measurement, description and judgement, shared three flaws: a tendency towards managerialism, serving whoever commissioned and controlled the programme; a failure to accommodate the differing values of different groups; and overcommitment to the scientific paradigm. People affected by programmes, including those who might be harmed, had little say in what was evaluated or how, and reports often answered questions that mattered mainly to funders.

What it is

Fourth generation evaluation, also called responsive constructivist evaluation, takes the claims, concerns and issues of stakeholders as its organising focus. Claims are favourable assertions about the programme, concerns are unfavourable ones, and issues are matters on which reasonable people may disagree. Stakeholders include agents who run or fund the programme, beneficiaries who gain from it and victims who may be harmed by it. For each stakeholder group, the evaluator forms a hermeneutic dialectic circle, in which each respondent's view is shared with and critiqued by the next, so that a joint construction emerges. Constructions are then compared across groups and enriched with new information. Items on which there is consensus are set aside for the report; unresolved items are prioritised, informed by further data and taken to a negotiation forum of representatives. Results are reported mainly as case reports, and the cycle repeats for whatever remains unresolved, so in the authors' view such an evaluation is never finished, only paused.

How it works

1. **Contract and organise**
   Agree with the client on the evaluand, the purpose and a commitment to a constructivist process, then assemble a team and gain entry to the setting.
2. **Identify stakeholders**
   List agents, beneficiaries and potential victims, and keep searching for groups not yet heard; Guba and Lincoln's 2001 guidelines stress including those who might be harmed.
3. **Build constructions within each group**
   Run a hermeneutic dialectic circle for each group, in which each respondent reacts to the joint construction emerging from those interviewed before.
4. **Enlarge the constructions**
   Introduce new information, from documents, observation, literature and other groups' views, so that each construction becomes better informed.
5. **Sort resolved from unresolved items**
   Set aside claims, concerns and issues on which there is consensus, for the report.
6. **Prioritise and inform the rest**
   Rank the unresolved items with stakeholders and gather information that might help resolve them.
7. **Negotiate**
   Convene a forum of stakeholder representatives with an agenda of unresolved items, and seek agreement on what should happen.
8. **Report and recycle**
   Produce case reports for each audience, and carry items that remain unresolved into a new cycle.

WORKED EXAMPLE · AN ILLUSTRATION

A district water project in Rajasthan installs village tanks and pipelines. A fourth generation evaluator identifies agents (the engineering department, the NGO partner and panchayat members), beneficiaries (households with connections) and potential victims (families at the tail end of pipelines, and households that had sold water from private wells). In separate circles, engineers claim reliable supply, women raise concerns about irregular timings, and tail-end families say pressure is too low. Data on supply hours by hamlet are gathered to inform the disputed items. The consensus that the tanks are well built goes straight into the report. At a forum with representatives of each group, the department and panchayat agree a supply timetable and pressure checks for tail-end hamlets, with a review after six months that begins the next cycle.

USE IT WHEN

- Stakeholders hold sharply different values or views of a programme.
- Community ownership of the findings matters as much as the findings themselves.
- The programme is long-running and can accommodate repeated cycles of inquiry.

SKIP IT WHEN

- A funder needs a quick, independent verdict on effectiveness.
- Power imbalances are so severe that negotiation would simply ratify the strongest voice, unless skilled facilitation can counter them.
- Budget and time cannot support several rounds of circles and negotiation.

COMMON MISREADINGS

- Treating it as a stakeholder satisfaction survey, rather than an iterative process that deepens and negotiates people's constructions.
- Assuming that it means anything goes, when the authors insisted that relativism did not mean that.

WHY IT WAS GROUNDBREAKING

It recast the evaluator from a judge into a facilitator of negotiation among people with a stake. By naming potential victims as stakeholders, it made the voices of those who might be harmed a formal part of evaluation design.

Origin

Guba and Lincoln set out the approach in Fourth Generation Evaluation (Sage, 1989), building on Robert Stake's responsive evaluation and on their own Naturalistic Inquiry (1985). In 2001 they condensed it for practitioners as Guidelines and Checklist for Constructivist (a.k.a. Fourth Generation) Evaluation, which sets out eleven steps from organising the evaluation to recycling unresolved issues.

RELATED IDEAS

RS Stake 82 · RC Chambers 296 · D&D Davies and Dart 362 · JG Greene 340 · DM Mertens 384

CONCEPT HIGHLIGHT

Member check

Returning data or interpretations to the people who provided them so they can confirm or correct them; a key technique for establishing credibility.

3

BIG IDEA 3 OF 3 · GUBA AND LINCOLN

Naturalistic and constructivist inquiry

Social realities are constructed by the people who live them, so inquiry should take place in natural settings, with the human inquirer as its main instrument, and seek to understand multiple constructions.

PICTURE IT

Think of how a family remembers a wedding. Each person tells it differently, and every version is real to them. Naturalistic inquiry studies life in its natural setting and takes those many versions seriously, rather than searching for a single correct account.

The problem

Mid-century evaluation borrowed its philosophy from the physical sciences: a single reality, to be measured objectively with instruments designed in advance and, ideally, controlled experiments. That fitted poorly with programmes whose meaning differed from one participant to the next and whose effects depended on context. Qualitative studies were treated as weaker substitutes for experiments, useful mainly for generating hypotheses. Without a philosophical foundation of their own, interpretive evaluators struggled to explain why their designs were emergent and context-bound, and to defend them as serious inquiry.

What it is

Lincoln and Guba set out naturalistic inquiry as a distinct paradigm, later calling it constructivist. In their 2001 guidelines they summarised three basic assumptions. Its ontology is relativist: there are multiple, socially constructed realities, which they stressed does not mean that anything goes. Its epistemology is subjectivist: claims about truth depend on the meanings and sophistication of those involved, including the inquirer. Its methodology is hermeneutic and dialectic: constructions are elicited, compared and refined. In Naturalistic Inquiry they described the characteristic features of such work: it takes place in natural settings; the human inquirer is the primary instrument, able to respond to the unexpected; tacit as well as explicit knowledge is used; methods are mainly qualitative; sampling is purposive rather than random; analysis is inductive and theory is grounded in the data; the design emerges as the study proceeds; outcomes are negotiated with respondents; and reporting takes the form of case

studies. They noted that the methodology does not dictate quantitative or qualitative methods, but discouraged mixing positivist and constructivist propositions in one study.

How it works

1. **Enter the natural setting**
   Study the programme where it happens, rather than only through instruments designed far away.
2. **Start open and sample purposively**
   Begin with broad questions and choose respondents for the range of perspectives they offer rather than at random, as Naturalistic Inquiry recommends.
3. **Let the design emerge**
   Use early findings to decide whom to talk to next and what to ask, documenting each change.
4. **Analyse inductively**
   Build categories and explanations from the data, grounding any theory in what respondents report and do.
5. **Negotiate meanings**
   Share interpretations with respondents and refine them together, since in this paradigm findings are co-constructed.
6. **Report as case studies**
   Present findings in rich case reports that convey context and multiple perspectives, and judge them by trustworthiness and authenticity criteria.

WORKED EXAMPLE · AN ILLUSTRATION

A state education department wants to understand why attendance at new evening learning centres for adolescent girls in tribal hamlets of Jharkhand is uneven. Instead of designing a survey in the capital, the evaluator spends time in four hamlets. She begins with open conversations with girls, mothers, the centre facilitator and the village headman, and then chooses further respondents to widen the range of views, including girls who never enrolled. Early conversations reveal that session times clash with forest produce collection in some seasons, so she adds interviews with families who gather produce. Her analysis builds categories from what she hears and sees. She shares draft interpretations with a group of girls and the facilitator, who refine them. The final report presents four hamlet case studies, each showing how different people understood the centre's purpose.

CONCEPT HIGHLIGHT

Thick description

A detailed account of the setting and context of a study that allows readers to judge whether findings may transfer to their own situation.

USE IT WHEN

- Meanings and experiences differ widely among participants.
- The programme is new, context-bound or poorly understood.
- You need to discover the right questions, not only answer pre-set ones.

SKIP IT WHEN

- Decision-makers need an estimate of average effect across many sites.
- Questions and measures are well established and comparability across sites matters most.
- The team lacks the time or field skills to act as a sensitive human instrument.

COMMON MISREADINGS

- [ ] Assuming relativism means that every account is equally valid and no judgement is possible, which the authors rejected.
- [ ] Equating it simply with using qualitative methods, when it is a set of assumptions about reality and knowledge that also shapes design, analysis and reporting.

WHY IT WAS GROUNDBREAKING

It gave qualitative evaluation a philosophical foundation of its own instead of treating it as a weaker version of experiments. It helped open the paradigm debates that made room for interpretive, participatory and, later, indigenous approaches.

Origin

Guba's 1978 monograph Toward a Methodology of Naturalistic Inquiry in Educational Evaluation, for UCLA's Center for the Study of Evaluation, first made the case. Lincoln and Guba developed it in Effective Evaluation (1981) and gave the full statement in Naturalistic Inquiry (1985). By Fourth Generation Evaluation (1989) they called the paradigm constructivist, and their 2001 guidelines summarised its assumptions.

RELATED IDEAS

RS Stake 82 · JG Greene 340 · BC Chilisa 498 · DM Mertens 384

CONCEPT HIGHLIGHT

Audit trail

A record of data, decisions and analysis that lets an independent auditor trace conclusions back to their sources, supporting dependability and confirmability.

RELATED GIANTS



Robert E. Stake

The testing specialist who argued that evaluation should listen to what people around a programme actually care about.

p. 82



Jennifer C. Greene

The methodologist who argued that evaluators are never neutral, and that mixing methods should mean mixing ways of seeing.

p. 340



Donna M. Mertens

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

p. 384



Bagele M. Chilisa

The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

p. 498

CONCEPT HIGHLIGHT

Authenticity criteria

Fairness and ontological, educative, catalytic and tactical authenticity: standards asking whether an inquiry represented all views and helped stakeholders understand and act.

CONCEPT HIGHLIGHT

Claims, concerns and issues

The items stakeholders raise about a programme, which organise a fourth generation evaluation; agreed items go to the report and contested ones to negotiation.

Did you know?

The leading champion of naturalistic inquiry was trained as a statistician: Guba's doctorate at Chicago focused on statistics and measurement, which he went on to teach.

Put it to work

USE IT ON MONDAY · NATURALISTIC INQUIRY, TRUSTWORTHINESS AND FOURTH GENERATION EVALUATION

1. In your methods section, set out how you will show that qualitative findings can be trusted, for example by checking interpretations with participants, keeping a record of analytical decisions and describing the context richly enough for readers to judge transfer.

2. Run a claims, concerns and issues round with each stakeholder group early on, then bring the groups together to discuss where they differ.

3. Keep a short reflexive note on your own position and assumptions, and update it as the evaluation proceeds.

AI prompt **Copy, fill the brackets, then check every line**

Help me organise a fourth generation evaluation, in the tradition of Egon Guba and Yvonna Lincoln, for **[programme]**. Here are interview and meeting notes from different stakeholder groups: **[paste notes, labelled by group]**. (1) Sort the material into claims, concerns and issues for each group, citing which note each item comes from. (2) Identify issues where groups disagree and suggest what evidence could inform a negotiated discussion. (3) Draft a short plan showing how I will demonstrate the trustworthiness of my qualitative findings to a sceptical funder. Do not paraphrase anyone in a way that changes their meaning.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

CONCEPT HIGHLIGHT

Claims, concerns and issues

The items stakeholders raise about a programme, which organise a fourth generation evaluation; agreed items go to the report and contested ones to negotiation.

YOUR MOVES Three moves from Guba and Lincoln's path

1 Set the quality bar for your approach

Guba and Lincoln answered critics of qualitative work by defining what trustworthiness looks like, on the critics' own ground.

TRY THIS
Write down the three quality criteria your qualitative work must meet, and share them with your client up front.

2 Be brave enough to rethink foundations

Trained in statistics, Guba came to doubt the methods he had taught and helped build an alternative.

TRY THIS
Pick one method you use by habit and ask whether it fits the question. Try an alternative once.

3 Find a lasting collaborator

Their partnership, sustained across decades and several landmark books, achieved more than either career might have alone.

TRY THIS
Identify one peer whose thinking complements yours and propose a long-term shared project.

Picture it

Think of a courtroom. A jury does not ask whether a witness's story can be repeated in a laboratory. It asks whether the witness is credible, whether the story hangs together and whether it can be checked. Trustworthiness criteria judge qualitative evidence the way a good jury judges testimony.

THE BIG IDEAS OF THIS CHAPTER

1. **Trustworthiness and authenticity** p. 326
   Naturalistic and qualitative inquiry should be judged by its own criteria of credibility, transferability, dependability and confirmability, together with authenticity criteria that ask whether it was fair and helped people act.
2. **Fourth generation evaluation** p. 329
   Organise evaluation around the claims, concerns and issues of all stakeholders, and make negotiation among them the heart of the process.
3. **Naturalistic and constructivist inquiry** p. 332
   Social realities are constructed by the people who live them, so inquiry should take place in natural settings, with the human inquirer as its main instrument, and seek to understand multiple constructions.

GIANTS CONNECTED TO THIS CHAPTER



Robert E. Stake

Responsive evaluation also organises the work around stakeholders' concerns.

p. 82



Daniel L. Stufflebeam

Guba's early collaborator, whose decision-oriented model offers a contrasting starting point.

p. 134



Jennifer C. Greene

Carried the engagement with values and multiple perspectives into mixed methods.

p. 340

KEY WORKS TO READ IN THE ORIGINAL

- **Naturalistic Inquiry** 1985, Lincoln and Guba
- **But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation** 1986, Lincoln and Guba
- **Fourth Generation Evaluation** 1989, Guba and Lincoln

WORKSHEET · TRY THIS TEMPLATE

Claims, concerns and issues log

Print this page, or copy the headings into your notes.

1 Stakeholder group

Who holds this view.

2 Claims

What they say is good or working about the programme.

3 Concerns

What worries them.

4 Issues

Points where reasonable people disagree.

5 Evidence that would help

What information could move the discussion forward.

6 Status after negotiation

Agreed, still contested or set aside, and why.

REFLECT

What would convince a sceptical funder that your interview findings can be trusted?

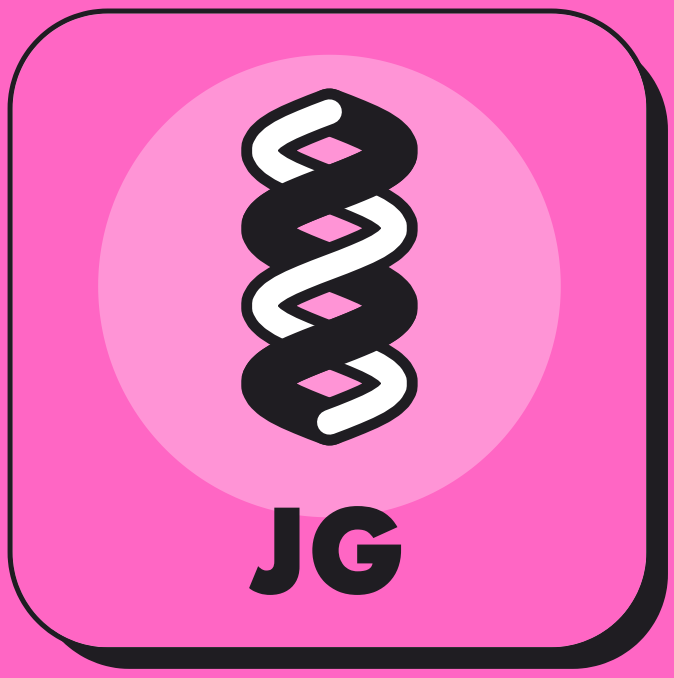
CONNECTS TO

Robert E. Stake (RS)
Responsive evaluation also organises the work around stakeholders' concerns. p. 82

Daniel L. Stufflebeam (DS)
Guba's early collaborator, whose decision-oriented model offers a contrasting starting point. p. 134

Jennifer C. Greene (JG)
Carried the engagement with values and multiple perspectives into mixed methods. p. 340

15



Emblem: an original HubForge design for this figure, not a likeness.

Jennifer C. Greene

Active 2000s to 2010s

United States; international audiences through keynotes and editorial leadership

Value and judgement | Methods, rigour and causality

The methodologist who argued that evaluators are never neutral, and that mixing methods should mean mixing ways of seeing.

IN ONE LINE

Jennifer Greene taught evaluators to listen across differences: between methods, between disciplines, and between the powerful and the unheard.

In the 1980s, social researchers were fighting what came to be called the paradigm wars: numbers on one side, stories on the other, each camp suspicious of the other's evidence. Meanwhile, in the field, evaluators were quietly using both. In 1989 Jennifer Greene, Valerie Caracelli and Wendy Graham did something simple and revealing. They gathered 57 real evaluations that combined qualitative and quantitative methods and examined how they actually did it. The answer was sobering. Studies that claimed to cross-check their findings often had no design capable of doing so, and few brought their methods together when analysing results. From that audit came a framework of five reasons for mixing methods, which the World Bank's Independent Evaluation Group still called seminal in 2017. Greene went further still: mixing methods, she argued, means mixing ways of seeing the world, and no evaluator is ever neutral.

Life CHAPTER 15

Jennifer Caroline Greene took a BA in psychology at Wellesley College and a PhD in educational psychology at Stanford University. She held faculty posts at the University of Rhode Island and then Cornell University, where she was an associate professor when, with Valerie Caracelli, she edited a 1997 issue of New Directions for Evaluation on advances in mixed-method evaluation. By 2010 Arizona State University was describing her as an evaluation scholar and practitioner of more than 30 years' standing.

She moved to the University of Illinois at Urbana-Champaign as Professor of Educational Psychology; her publisher in 2007 described her role there as professor of quantitative and evaluative research methodologies. In 2006 she co-edited the SAGE

BIG IDEAS

1. **Five purposes for mixing methods**
2. **Mixed methods as a dialogue of mental models**
3. **Values-engaged, educative evaluation**

KEY WORKS

Mixed Methods in Social Inquiry · Sole author

Disrupting Program Evaluation and Mixed Methods Research for a More Just Society: The Contributions of Jennifer C. Greene **2023** · Tribute volume about her work, edited by Jori Hall, Ayesha Boyce and Rodney Hopson

Handbook of Evaluation with Ian Shaw and Melvin Mark, a 25-chapter volume with international contributors. In the same year she and colleagues set out a values-engaged, educative approach to evaluation, and in October 2007 Jossey-Bass published her book Mixed Methods in Social Inquiry. The American Evaluation Association had already given her its Paul F. Lazarsfeld Evaluation Theory Award in 2003.

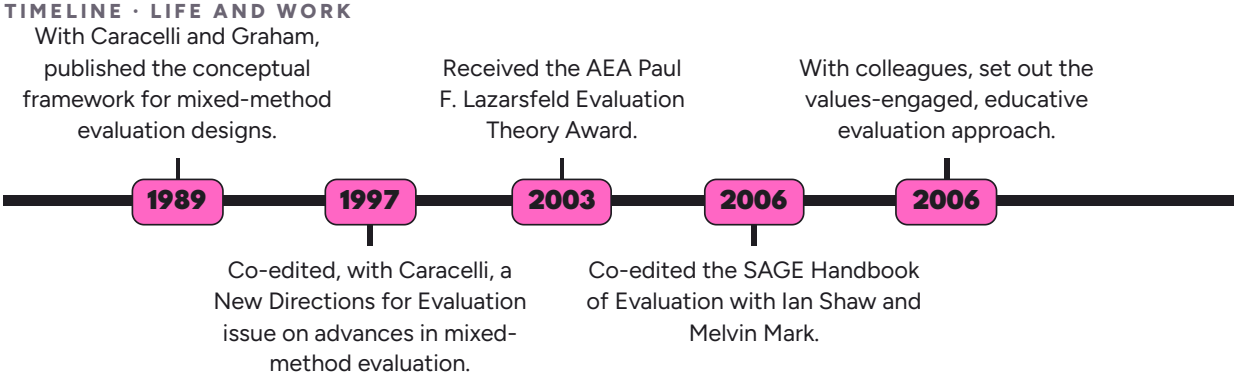
Her leadership roles spanned both of her professional communities. She was named a Fellow of the American Educational Research Association in 2010 and served as president of the American Evaluation Association in 2011, when its conference in Anaheim took Values and Valuing as its theme. She spent six years as co-editor-in-chief of New Directions for Evaluation, served as an associate editor of the Journal of Mixed Methods Research and co-edited the Evaluation and Society book series. She gave a keynote at the 2009 conference of DeGEval, the German-language evaluation society, in Münster, and in 2010 was the featured speaker at the first Delbert and Jewell Lewis Literacy Research Lecture Series at Arizona State University.

Teaching was central to her work. She taught social science methodology to doctoral students, aiming to help them see that method shapes the knowledge a study produces rather than being a neutral technique. She supervised doctoral work that applied her ideas, including a 2014 Illinois dissertation by Ayesha Tillman on using the values-engaged, educative approach in a multi-site science and engineering education evaluation. In 2023 Jori Hall, Ayesha Boyce and Rodney Hopson edited a tribute volume on her contributions, published by Information Age Publishing, which credits her with democratising social science, valuing diverse scholars and perspectives, and opening space for non-traditional methods. Illinois now lists her as Professor Emerita.

The spark

Trained in educational psychology at Stanford, she worked in a field where quantitative and qualitative researchers often talked past each other. Evaluators were combining methods without a guiding theory, and values were treated as something to be kept out of the work.

TIMELINE · LIFE AND WORK



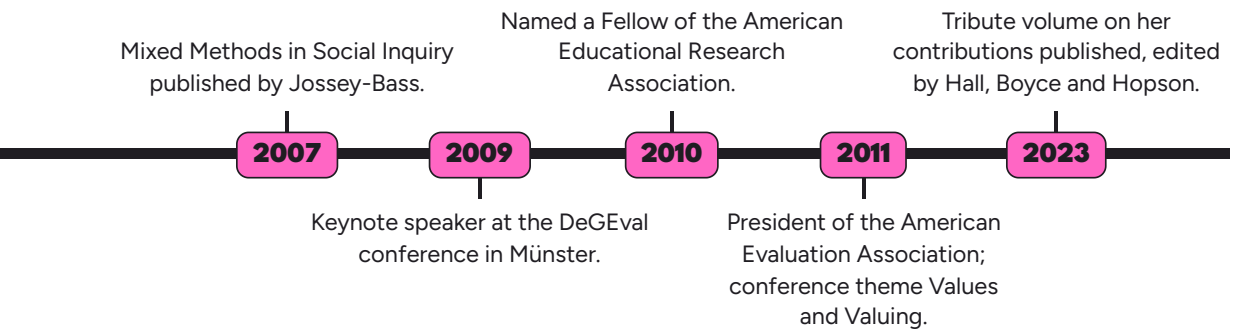
Her university profile describes her work as sitting where social science methods meet social policy, aiming to be both innovative in method and responsible to society.

The breakthrough CHAPTER 15

Imagine a survey shows that a school programme raised attendance, and interviews suggest children feel less welcome. Which do you believe? Should the two methods agree? Were they even meant to answer the same question? Through the 1980s, evaluators of education and social programmes increasingly combined methods, but without a theory to guide why and how. In 1989 Greene, Valerie Caracelli and Wendy Graham published Toward a Conceptual Framework for Mixed-Method Evaluation Designs. They built it from the theoretical literature and refined it by reviewing 57 real mixed-method evaluations. It named five purposes for mixing methods: triangulation, complementarity, development, initiation and expansion. The review also found that studies claiming triangulation often lacked a suitable design, and that few integrated the two kinds of method during analysis.

In the 1997 volume she edited with Caracelli, Greene argued that mixing should be planned at three levels: political, concerning purpose and use; paradigm, concerning assumptions about knowledge; and technical, concerning methods. She described purist, pragmatic and dialectic stances on mixing paradigms, the last treating tensions between paradigms as a source of fuller understanding. Her 2007 book developed this into a view of mixed methods as bringing different mental models into respectful dialogue. Running alongside was a second line of argument: that evaluation is never value-free and should engage the values at stake openly.

TIMELINE, CONTINUED



The problem it solved

The idea of the evaluator as a neutral technician, and a split between qualitative and quantitative camps that left combined studies without a thoughtful rationale.

Impact WHAT CHANGED

Greene's 1989 framework travelled far beyond American education. In a 2017 blog for the World Bank's Independent Evaluation Group, its chief evaluation officer Jos Vaessen called the article seminal and drew on its five purposes to discuss how development evaluations should justify mixed designs, while warning that the mixed methods label is sometimes used to cover weak or cheap designs. The article is also catalogued in ALNAP's library of humanitarian evaluation resources. In our reading, the habit in many development terms of reference of requesting mixed methods, and of asking what each method adds, reflects the framework's logic, although that wider lineage is our interpretation.

Her editorial work shaped what the field read. As co-editor-in-chief of New Directions for Evaluation, associate editor of the Journal of Mixed Methods Research and co-editor of the international SAGE Handbook of Evaluation, she helped set agendas across countries. The 2023 tribute volume credits her with supporting scholars from many disciplines and cultures, including those outside the United States. Her values-engaged approach has been used in evaluations aimed at broadening participation in science education. In our assessment, her insistence that evaluators state their values and engage stakeholders' values speaks directly to current equity-focused and participatory practice in NGOs and philanthropy, though the depth of engagement she describes requires time and skilled facilitation.

KEY WORKS TO READ IN THE ORIGINAL

- **Mixed Methods in Social Inquiry** Sole author
- **Disrupting Program Evaluation and Mixed Methods Research for a More Just Society: The Contributions of Jennifer C. Greene** 2023, Tribute volume about her work, edited by Jori Hall, Ayesha Boyce and Rodney Hopson

WHERE PEOPLE DISAGREE

Debates

Greene's position on values has drawn debate. A 2010 conversation with Michele Tarsilla records her view that advocacy in evaluation is inevitable and that evaluators should therefore take a socially enfranchising role, a stance evaluators who prize neutrality may resist; the same source notes that her thinking developed to hold that evaluators should never take one party's side within an assignment. On mixed methods, she described the pragmatic stance, which sees paradigms as offering little practical guidance, and argued against settling paradigm differences through any single paradigm such as pragmatism, a disagreement that continues in the literature. Vaessen's 2017 critique of loose mixed methods practice is a reminder that her framework is often cited more than applied.

LIMITS, IN OUR READING

- [ ] Her view that advocacy in evaluation is unavoidable, and should be handled in ways that enfranchise people, is contested by those who prefer a more neutral evaluator role.
- [ ] In our view, values-engaged practice needs time and skilled facilitation that short NGO evaluations may not allow.

WHY THIS MATTERS TO YOU

Greene's ideas help anyone who must make sense of evidence that does not agree. Before you combine numbers and stories, ask why: to check one against the other, to fill gaps, to build one from the other, to find surprises or to widen the picture. Each purpose needs a different design. When people around you see the same problem differently, treat their differences as a source of insight rather than an argument to win. And be open about values: ask, out loud, whose idea of success is guiding the work, and make room for the people least often heard. Careful method and fairness, she shows, make each other stronger.

57

empirical mixed-method evaluations reviewed to build the 1989 framework

ALNAP library record of Toward a Conceptual Framework for Mixed-Method Evaluation Designs (1989)

25

chapters in the SAGE Handbook of Evaluation she co-edited with Ian Shaw and Melvin Mark

SAGE Publishing, The SAGE Handbook of Evaluation (2006)

6

years as co-editor-in-chief of New Directions for Evaluation

University of Illinois Experts profile

Key concepts

Triangulation
Using different methods to study the same question in order to test whether their findings converge.

Complementarity
Using one method to elaborate, illustrate or clarify findings from another, so that each covers different facets of the same phenomenon.

Initiation
Mixing methods deliberately to surface paradox or contradiction, which can lead to new questions and perspectives.

Mental model
In Greene's usage, the combination of assumptions, theories, training, values and experience through which an inquirer frames and interprets a study.

Dialectic stance
A position on mixing paradigms that treats them as meaningful and sees the tensions between them as a source of more complete understanding.

Values-engaged, educative evaluation
An approach that openly engages stakeholders' values, centres the interests of underserved groups and aims to help stakeholders learn about their own programme.

Sources for this chapter: p. 583

Did you know?

The 1989 mixed-method framework was built partly by reviewing 57 real evaluations, and that review found that studies claiming triangulation often did not use a design capable of achieving it.

THE BIG IDEAS OF THIS CHAPTER

1. **Five purposes for mixing methods** p. 348
   Decide why you are mixing methods before deciding how, and choose a design that can actually deliver that purpose.
2. **Mixed methods as a dialogue of mental models** p. 351
   Mixing methods means bringing different ways of knowing into respectful dialogue, so that differences between them yield deeper understanding.
3. **Values-engaged, educative evaluation** p. 354
   Because evaluation is never value-free, engage stakeholders' values openly, centre those least well served, and help everyone learn about their own programme.

GIANTS CONNECTED TO THIS CHAPTER



Donna M. Mertens

The transformative paradigm makes a social justice commitment explicit in design.

p. 384



Egon G. Guba and Yvonna S. Lincoln

An earlier case that multiple perspectives and values belong inside evaluation.

p. 318



Michael Scriven

A contrasting route to judgement, through the logic of criteria and evidence.

p. 60

In one line

Jennifer Greene taught evaluators to listen across differences: between methods, between disciplines, and between the powerful and the unheard.

1

BIG IDEA 1 OF 3 · GREENE

Five purposes for mixing methods

Decide why you are mixing methods before deciding how, and choose a design that can actually deliver that purpose.

PICTURE IT

Think of a doctor using an X-ray and a conversation with the patient. Sometimes she uses both to confirm a diagnosis, sometimes because each shows what the other cannot, sometimes because one tells her where to look next. Greene's five purposes ask you to know which of these you are doing before you start.

The problem

By the late 1980s evaluators of educational and social programmes were combining surveys, tests, interviews and observation, but social research was still split by a paradigm debate between quantitative and qualitative camps. Mixing happened by instinct, without a theory of why or how. Studies claimed triangulation without designs capable of it, and analysis rarely brought the strands together. Commissioners paid for two kinds of data and often received two separate stories, missing the fuller understanding that mixing promised.

What it is

The 1989 framework by Jennifer Greene, Valerie Caracelli and Wendy Graham, published in Educational Evaluation and Policy Analysis, identified five purposes for mixed-method evaluation. Triangulation seeks convergence and corroboration of results from different methods studying the same phenomenon, ideally methods whose biases offset each other. Complementarity uses one method to elaborate, illustrate or clarify the results of another, so that each covers different facets of the same phenomenon. Development uses the results of one method to help design or inform another, for example in sampling or building instruments. Initiation deliberately looks for paradox and contradiction, which can prompt new questions and perspectives. Expansion uses different methods for different components of an inquiry to widen its breadth and range. For each purpose the authors recommended a design, described by seven characteristics covering the methods, the phenomena studied, the paradigmatic framework and implementation criteria. The underlying logic is that purpose should drive design, and that each purpose places its own demands on how methods are combined.

How it works

1. **Name the purpose**
   For each part of the study, write down which of the five purposes mixing is meant to serve, following Greene, Caracelli and Graham (1989).
2. **Match the design**
   Choose design features that can deliver that purpose, such as independent methods aimed at the same question for triangulation, or a sequence of methods for development.
3. **Decide component or integrated**
   Greene's 1997 work distinguishes component designs, which keep methods separate until interpretation, from integrated designs, such as iterative or nested ones, which mix during the inquiry.
4. **Plan integration in analysis**
   Build in a step where results from different methods are brought together, since the 1989 review found few studies did this.
5. **Treat divergence as data**
   When methods disagree, investigate rather than average, which is the heart of the initiation purpose.
6. **Report the added value**
   State what each method contributed and what the combination revealed that neither showed alone.

WORKED EXAMPLE · AN ILLUSTRATION

An evaluation of a girls' scholarship scheme in rural Rajasthan uses three methods for three purposes. First, focus groups with parents in six villages identify which barriers matter locally, such as distance, safety and household work. These results are used to design a household survey of 1,200 families, a development purpose. The survey measures how widespread each barrier is and whether scholarship holders attend school more often. Separately, the team observes classrooms and checks school registers to test whether attendance gains reported by households match school records, a triangulation purpose. When the survey shows higher attendance but the registers do not, the team returns to the villages and learns that many girls attend mainly on disbursement days. That contradiction, pursued deliberately, becomes a key finding, an initiation purpose.

CONCEPT HIGHLIGHT

Triangulation

Using different methods to study the same question in order to test whether their findings converge.

USE IT WHEN

- You are designing or reviewing a mixed-methods evaluation, including terms of reference that request mixed methods.
- A team must explain to a commissioner why each method is included.
- A proposal uses the word triangulation loosely and needs testing.

SKIP IT WHEN

- ○ A single, focused question can be answered well by one method.
- ○ The budget allows only token use of a second method, where honesty about limits is better than a mixed label.
- ○ Adding methods would overburden respondents in a small community for little gain.

COMMON MISREADINGS

- ☐ Using triangulation as a catch-all word for any use of several methods, which the 1989 review found was common.
- ☐ Collecting both kinds of data but never integrating them in analysis, so the study is mixed in name only.

WHY IT WAS GROUNDBREAKING

It gave a field that was mixing methods by instinct a shared vocabulary and a design logic. Its empirical review exposed common weaknesses, such as claimed triangulation without a suitable design, and nearly three decades later development evaluators were still using its five purposes to justify mixed designs.

Origin

The framework appeared in Educational Evaluation and Policy Analysis in 1989, built from theoretical literature and refined through a review of 57 empirical mixed-method evaluations. Greene and Caracelli extended it in a 1997 New Directions for Evaluation issue on advances in mixed-method evaluation, adding political, paradigm and technical levels of mixing and the distinction between component and integrated designs. In 2017 the World Bank's Independent Evaluation Group described the 1989 article as seminal.

RELATED IDEAS

Mixed methods as a dialogue of mental models

DM Mertens 384 · LC Cronbach 38 · G&L Guba and Lincoln 318

CONCEPT HIGHLIGHT

Complementarity

Using one method to elaborate, illustrate or clarify findings from another, so that each covers different facets of the same phenomenon.

2

BIG IDEA 2 OF 3 · GREENE

Mixed methods as a dialogue of mental models

Mixing methods means bringing different ways of knowing into respectful dialogue, so that differences between them yield deeper understanding.

PICTURE IT

Think of a group of friends describing the same city: an architect, a chef and a historian. Each sees something real that the others miss. Greene's idea of mixed methods brings those views into conversation, so that their differences reveal more than any single view.

The problem

As mixed methods became popular, many treated it as a technical recipe: add some interviews to a survey and call the result mixed. The deeper differences between traditions, about what counts as knowledge, causality or meaning, generality or particularity, were either ignored or treated as impossible to bridge. Purists said paradigms could not be mixed; pragmatists said paradigms offered little practical guidance. Neither position helped a team whose economists and anthropologists reached different conclusions about the same programme.

What it is

Greene built a theory of mixing in two stages. In 1997, with Valerie Caracelli, she argued that mixing happens at three levels: political, concerning the purpose and use of the evaluation; paradigm, concerning assumptions about the social world, what can be known and the evaluator's role; and technical, concerning methods. She identified the paradigm level as the most controversial and described three stances. The purist stance holds that paradigms cannot sensibly be mixed. The pragmatic stance treats paradigms as useful concepts that offer little practical guidance. The dialectic stance treats paradigms as important guides and sees the tensions between them as a source of more complete and insightful understanding. In Mixed Methods in Social Inquiry (2007) she developed this through the idea of mental models: the blend of philosophical assumptions, theories, disciplinary training, values and experience through which each inquirer frames and interprets a study. Mixing

methods, in this view, brings different mental models into one inquiry for respectful dialogue and mutual learning, with purpose coming before design typologies.

How it works

1. **Surface mental models**
   At the start, ask each team member to describe the assumptions, training and values they bring, and what they regard as good evidence.
2. **Decide at three levels**
   Agree the political purpose, the paradigm stance and the technical methods, following the levels set out by Greene and Caracelli (1997).
3. **Choose a stance consciously**
   Decide whether the team will work dialectically, deliberately keeping different perspectives in tension rather than letting one dominate.
4. **Create spaces for dialogue**
   Schedule meetings during analysis, not only at the end, where findings from different traditions are set side by side.
5. **Engage difference**
   When findings diverge, explore the reasons with both perspectives, treating disagreement as a lead to follow.
6. **Write the synthesis together**
   Report what the dialogue revealed, including tensions that remain, so that readers can weigh them.

WORKED EXAMPLE · AN ILLUSTRATION

A team evaluating a women's self-help group livelihoods programme in Andhra Pradesh includes an economist and an anthropologist. The economist's household panel shows incomes up by a fifth. The anthropologist's life histories describe women working longer hours and borrowing from several groups at once. Instead of writing separate chapters, the team leader schedules three dialogue sessions in which each member explains the assumptions behind their method. Together they design a short follow-up module on debt and time use, and revisit interview transcripts for evidence of who controls the new income. The final report presents a more complex picture: higher incomes, real gains in women's mobility and rising debt stress among the poorest members. The disagreement, engaged rather than smoothed over, produced the most useful findings.

In one line

Decide why you are mixing methods before deciding how, and choose a design that can actually deliver that purpose.

USE IT WHEN

- A multidisciplinary team is evaluating a complex social programme.
- Different stakeholder groups trust different kinds of evidence.
- Results from different methods are likely to diverge and need to be interpreted together.

SKIP IT WHEN

- A small, quick assessment where a single method is adequate.
- The team has no time or budget for joint analysis sessions.
- The commissioner needs a narrowly technical answer, such as a unit cost.

COMMON MISREADINGS

- Reading dialogue as a push for consensus, when the dialectic stance values tensions that may remain unresolved.
- Assuming the dialectic stance relaxes standards, when it asks each tradition to bring its own standards of quality fully into the conversation.

WHY IT WAS GROUNDBREAKING

It moved the mixed methods conversation from techniques to ways of knowing, and made disagreement between methods a resource rather than an embarrassment. It offered an alternative to treating pragmatism as the only basis for mixing.

Origin

Greene and Caracelli set out the three levels of mixing and the purist, pragmatic and dialectic stances in their 1997 New Directions for Evaluation issue on advances in mixed-method evaluation, edited while Greene was at Cornell University. Jossey-Bass published Mixed Methods in Social Inquiry in October 2007, developing the idea of mental models in dialogue. Greene later served as an associate editor of the Journal of Mixed Methods Research.

RELATED IDEAS

Five purposes for mixing methods

G&L Guba and Lincoln 318
DM Mertens 384
RS Stake 82

CONCEPT HIGHLIGHT

Initiation

Mixing methods deliberately to surface paradox or contradiction, which can lead to new questions and perspectives.

3

BIG IDEA 3 OF 3 · GREENE

Values-engaged, educative evaluation

Because evaluation is never value-free, engage stakeholders' values openly, centre those least well served, and help everyone learn about their own programme.

PICTURE IT

Think of a community deciding what makes a good school. Exam results matter, but so do safety, belonging and respect. A values-engaged evaluator brings these questions into the open and makes sure the families with least power are heard.

The problem

Evaluators were long trained to see themselves as neutral technicians who measure and report. Yet decisions about which questions to ask, whose definition of success to use and which groups to sample always carry values. When those choices stay hidden, evaluations tend to reflect the views of funders and managers, while the people least well served by a programme go unheard. In education programmes meant to broaden participation, that blind spot could defeat the programme's own purpose.

What it is

The values-engaged, educative (VEE) approach was set out by Greene with colleagues DeStefano, Burgon and Hall in 2006, with its foundation in evaluating science, technology, engineering and mathematics (STEM) education programmes. It rests on Greene's argument that evaluation cannot be value-free and should engage openly with the values at stake, serving as a venue for democratic dialogue about social and educational issues and as a public good. It has two linked commitments. Being values-engaged means including the interests, concerns, perspectives and values of people who are traditionally unheard or least well served, and attending to the culture of the institution. Being educative means the evaluation aims to help stakeholders learn about their own programme, so that its process and findings support understanding. In STEM settings the approach defines programme quality at the intersection of scientific content, pedagogy and equity, which linked it to the National Science Foundation's agenda on broadening participation. Its logic is that stating and engaging values openly makes evaluation both more honest and more useful.

How it works

1. **Map the values at stake**
   Identify whose values shape the programme and the evaluation, including those of funders, managers, staff and participants.
2. **Centre the least well served**
   Seek out the perspectives of groups that are traditionally unheard, and include their definitions of success in the evaluation criteria.
3. **Read the institutional culture**
   Understand the norms and power relations of the setting; Ayesha Tillman's 2014 study found this was among the features most responsible for a VEE evaluation's success.
4. **Judge quality on several dimensions**
   In STEM programmes, assess content, pedagogy and equity together, as the approach proposes.
5. **Educate as you go**
   Share emerging findings in forums where stakeholders discuss them and learn about their own programme.
6. **State your stance**
   Make your own values and judgements explicit in reports, so that readers can weigh them.

WORKED EXAMPLE · AN ILLUSTRATION

A state government in India funds a residential bridge programme that prepares students from tribal communities for undergraduate science degrees. The evaluator begins by asking the funder, faculty, students and parents what success would mean. Faculty speak of examination results; students speak of confidence and belonging; parents speak of safety and eventual jobs. The evaluation framework includes all three, and judges quality on content, teaching and equity. Interviews reveal that students who speak Gondi at home struggle in English-medium lectures and rarely ask questions. The evaluator presents this at a joint session with faculty and student representatives, who agree to add bilingual tutorials. The report states openly that the evaluation gave weight to students' experience of equity alongside academic results.

CONCEPT HIGHLIGHT

Mental model

In Greene's usage, the combination of assumptions, theories, training, values and experience through which an inquirer frames and interprets a study.

USE IT WHEN

- A programme aims at equity or at broadening participation.
- Stakeholder groups hold sharply different views of what success means.
- The commissioner wants the evaluation process itself to build learning.

SKIP IT WHEN

- The commissioner requires a narrow technical audit and will not accept engagement with values.
- There is no time or facilitation skill for structured dialogue among stakeholders.
- Stakeholders could be put at risk by having their views aired in shared forums, unless safe alternatives are designed.

COMMON MISREADINGS

- Treating values engagement as taking sides with one group, when Greene's thinking, as recorded in a 2010 conversation, came to hold that evaluators should not take one party's side within an assignment.
- Thinking it means abandoning rigour, when it adds values to the criteria of judgement rather than removing standards of evidence.

WHY IT WAS GROUNDBREAKING

It challenged the image of the evaluator as a neutral technician and gave evaluation a deliberate role in public conversation. It tied methodological choices to equity and voice, while keeping the evaluation's purpose centred on learning.

Origin

The VEE approach was published by Greene, DeStefano, Burgon and Hall in 2006 and applied in evaluations of STEM education programmes linked to the National Science Foundation's broadening participation agenda. A 2014 University of Illinois dissertation by Ayesha Tillman examined its use in a longitudinal, multi-site programme. In 2011, during Greene's presidency, the American Evaluation Association conference took Values and Valuing as its theme.

RELATED IDEAS

Mixed methods as a dialogue of mental models

DM Mertens 384 · RS Stake 82 · RC Chambers 296 · BC Chilisa 498

RELATED GIANTS



Donna M. Mertens

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

p. 384



Lee J. Cronbach

The school maths teacher whose reliability statistic sits in almost every survey report, and who later argued that evaluation should serve local learning.

p. 38



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318



Robert E. Stake

The testing specialist who argued that evaluation should listen to what people around a programme actually care about.

p. 82

CONCEPT HIGHLIGHT

Dialectic stance

A position on mixing paradigms that treats them as meaningful and sees the tensions between them as a source of more complete understanding.

CONCEPT HIGHLIGHT

Values-engaged, educative evaluation

An approach that openly engages stakeholders' values, centres the interests of underserved groups and aims to help stakeholders learn about their own programme.

Did you know?

More than a quarter of a century after publication, the 1989 article was singled out by the World Bank's Independent Evaluation Group as a seminal guide to mixing methods in evaluation.

Put it to work

USE IT ON MONDAY · VALUES-ENGAGED EVALUATION AND MIXED METHODS AS DIALOGUE

1. At inception, map the values at stake: ask each stakeholder group what a good programme would look like to them, and record where those pictures differ.

2. For a mixed methods design, write down what each method can see that the others cannot, and plan a moment when results are set side by side. Treat disagreement between methods as a finding, not an error.

3. State your own stance and values in the inception report, briefly and plainly, so the client can weigh your judgements.

AI prompt **Copy, fill the brackets, then check every line**

Help me design a values-engaged mixed methods evaluation, drawing on Jennifer Greene's work, for **[programme]** in **[location]**. Stakeholder groups are **[list]**. Here are notes on what each group has said about success: **[paste notes]**. Our candidate methods are **[list]**. (1) Summarise the values and definitions of success each group expressed, using only my notes. (2) Show where these conflict. (3) For each method, explain what it is likely to see and miss in this setting. (4) Suggest how to bring the methods' results into dialogue during analysis, and how to report divergence honestly.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a doctor using an X-ray and a conversation with the patient. Sometimes she uses both to confirm a diagnosis, sometimes because each shows what the other cannot, sometimes because one tells her where to look next. Greene's five purposes ask you to know which of these you are doing before you start.

YOUR MOVES

Three moves from Greene's path

1 Build theory from real practice

Greene's 1989 framework for mixed methods came from carefully reviewing 57 actual evaluations.

TRY THIS

Collect ten evaluations from your sector and look for the patterns in how they combine methods.

2 Be open about your values

She was explicit about her values and method choices, and showed that openness builds trust rather than undermining it.

TRY THIS

Add a short positionality note to your next evaluation design, saying what you value and why.

3 Build bridges and lift others

A Fellow of AERA and president of AEA, she edited journals and supervised many doctoral students; a 2023 tribute volume credits her with opening space for diverse scholars.

TRY THIS

Invite one emerging evaluator to co-present or co-write with you this year.

THE BIG IDEAS OF THIS CHAPTER

1 **Five purposes for mixing methods** p. 348

Decide why you are mixing methods before deciding how, and choose a design that can actually deliver that purpose.

2 **Mixed methods as a dialogue of mental models** p. 351

Mixing methods means bringing different ways of knowing into respectful dialogue, so that differences between them yield deeper understanding.

3 **Values-engaged, educative evaluation** p. 354

Because evaluation is never value-free, engage stakeholders' values openly, centre those least well served, and help everyone learn about their own programme.

CONCEPT HIGHLIGHT

Values-engaged, educative evaluation

An approach that openly engages stakeholders' values, centres the interests of underserved groups and aims to help stakeholders learn about their own programme.

GIANTS CONNECTED TO THIS CHAPTER



Donna M. Mertens

The transformative paradigm makes a social justice commitment explicit in design.

p. 384



Egon G. Guba and Yvonna S. Lincoln

An earlier case that multiple perspectives and values belong inside evaluation.

p. 318



Michael Scriven

A contrasting route to judgement, through the logic of criteria and evidence.

p. 60

KEY WORKS TO READ IN THE ORIGINAL

- **Mixed Methods in Social Inquiry** Sole author
- **Disrupting Program Evaluation and Mixed Methods Research for a More Just Society: The Contributions of Jennifer C. Greene** 2023, Tribute volume about her work, edited by Jori Hall, Ayesha Boyce and Rodney Hopson

CONCEPT HIGHLIGHT

Triangulation

Using different methods to study the same question in order to test whether their findings converge.

CONCEPT HIGHLIGHT

Complementarity

Using one method to elaborate, illustrate or clarify findings from another, so that each covers different facets of the same phenomenon.

WORKSHEET · TRY THIS TEMPLATE

Values and methods map

Print this page, or copy the headings into your notes.

1 Stakeholder group

Include those with least power in the programme.

2 What success means to them

In their terms.

3 Method that captures it best

Survey, interviews, observation, records or participatory tools.

4 Where findings may diverge

Expected tensions between groups or between methods.

5 How divergence will be discussed

A meeting, a joint analysis session or a section of the report.

6 Our stance

The values the evaluation team brings, stated openly.

REFLECT

What values are you bringing into your current evaluation, and have you told your client?

CONNECTS TO

DM **Donna M. Mertens**
The transformative paradigm makes a social justice commitment explicit in design. p. 384

G&L **Egon G. Guba and Yvonna S. Lincoln**
An earlier case that multiple perspectives and values belong inside evaluation. p. 318

MS **Michael Scriven**
A contrasting route to judgement, through the logic of criteria and evidence. p. 60

16



Emblem: an original HubForge design for this figure, not a likeness.

Rick Davies and Jess Dart

Active 1990s to 2000s

Davies: United Kingdom, with long experience in Africa and Asia, including Bangladesh; Dart: Australia

Participation, equity and voice · Made in the Global South

Two practitioners who turned the stories people tell about change into a disciplined monitoring method.

IN ONE LINE

Rick Davies and Jess Dart showed that a well chosen story, with its reasons written down, can carry a village's judgement all the way to the boardroom.

Rajshahi, north-west Bangladesh, 1994. A development NGO tries something new. The Christian Commission for Development in Bangladesh had more than 500 staff and, by 1996, worked with over 46,000 people in 785 villages, most of them women. Its programmes were far too varied for a single list of indicators. So, working with a doctoral researcher named Rick Davies, it began asking one simple question: what is the most significant change you have seen? Then it did something unusual. It made committees choose between the answers, and explain why. Eleven years later, Davies and Jess Dart, who had trialled the method on dairy farms in Australia, published a free 104 page guide. Today a bibliography on Davies's website lists about 360 documents on the method, and the guide has been translated into many languages, including Hindi and Tamil.

Life CHAPTER 16

Rick Davies is an independent monitoring and evaluation consultant based in Cambridge in the United Kingdom, with about three decades of work for international development organisations in Africa and Asia. Between 1992 and 1995 he carried out fieldwork in Bangladesh for a doctorate on organisational learning in non-government aid organisations, working with the Christian Commission for Development in Bangladesh (CCDB). By 1996 CCDB had more than 500 staff and worked with over 46,000 people in 785 villages, most of its direct participants being women. His thesis, Order and Diversity: Representing and Assisting Organisational Learning in Non-Government Aid Organisations, was completed in 1998 and

BIG IDEAS

1. **Most Significant Change (MSC)**
2. **Monitoring without predefined indicators**
3. **Selection as a way of making values explicit**

KEY WORKS

A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique
2003 · Dart and Davies, American Journal of Evaluation

The Most Significant Change (MSC) Technique: A Guide to Its Use
2005 · Davies and Dart

drew on evolutionary epistemology, the idea that knowledge grows through variation, selection and retention.

Since 1997 Davies has run Monitoring and Evaluation NEWS, a website on monitoring and evaluation methods for social development programmes. Beyond Most Significant Change he has developed a series of other tools: hierarchical card sorting for ethnographic inquiry, the Basic Necessities Survey as a democratic approach to measuring poverty, ParEvo for participatory scenario planning, and EvalC3, an Excel-based tool that combines predictive modelling with a qualitative comparative analysis view of causality. His recent work focuses on quality assurance of evaluations and technical support for evaluation teams.

Jess Dart works in Australia. As part of her doctoral research she tested the technique Davies had devised, running a 12 month trial in the Target 10 dairy extension project in Victoria from May 1998 to May 1999; her 2000 article with colleagues in PLA Notes reported that the approach had apparently not been tried in Australia before. She went on to found Clear Horizon, an Australian specialist evaluation company, which she leads, and has more than 25 years of experience in strategy, design and evaluation, with a particular focus on complex and emergent programmes and systems change. In 2018 the Australian Evaluation Society gave her its Outstanding Contribution to Evaluation Award.

The partnership joined an originator in the United Kingdom with a practitioner who had tested and adapted the method in a high income country. In 2003 Dart and Davies published the first journal article on the technique in the American Journal of Evaluation, and in 2005 they released The Most Significant Change (MSC) Technique: A Guide to Its Use, a 104 page manual made freely available under a Creative Commons licence. MandE NEWS reports translations into languages including French, Spanish, Arabic, Japanese and, in 2026, Chinese, and lists more than 360 documents in its MSC bibliography.

The spark

For Davies, the spark was a researcher's question meeting a real organisation's problem. His doctorate asked how aid organisations learn, drawing on evolutionary epistemology, the idea that

TIMELINE · LIFE AND WORK

1992 to 1995 — Davies carries out PhD fieldwork in Bangladesh with the Christian Commission for Development in Bangladesh.

1994 — CCDB first uses the technique in Rajshahi, Bangladesh.

1996 — Davies publishes the first public description online, An evolutionary approach to facilitating organisational learning.

1997 — Davies begins running the Monitoring and Evaluation NEWS website.

1998 — Davies completes his PhD thesis, Order and Diversity.

knowledge grows the way living things evolve, through variation, selection and retention. CCDB needed a way to summarise progress across very different projects without a fixed list of indicators. The technique joined the two. Dart came to it through her own doctoral research in Australia, testing whether a method born in Bangladesh could work in a dairy extension project in Victoria, where it had apparently not been tried before.

The breakthrough CHAPTER 16

Picture trying to report on hundreds of village groups, each pursuing different goals: one is saving money, another building a school, another challenging a local injustice. Any single list of indicators will miss much of what is changing, and will certainly miss the surprises. That was CCDB's problem, and many NGOs still recognise it. Conventional monitoring also sent information upwards without helping the organisation decide what it valued. Davies, studying how NGOs learn, wanted a process that would first produce many different accounts of change, and then help the organisation choose among them, in the way evolution works through variation and selection.

The result was first used by CCDB in Rajshahi, Bangladesh, in 1994. Field staff gathered accounts of the most significant change they had observed, and committees at successive levels chose the most significant and explained why. Davies published the first public description online in 1996, under the title An evolutionary approach to facilitating organisational learning. The method went under several names, including monitoring without indicators, the story approach and impact monitoring, before the name Most Significant Change was settled in 2000. Dart's 1998 to 1999 Australian trial, the 2003 article and the 2005 guide turned a single NGO's solution into a documented ten step method.

TIMELINE, CONTINUED

1998 to 1999 — Dart runs a 12 month trial in the Target 10 dairy extension project in Victoria, Australia.

2000 — The name Most Significant Change is settled; Dart and colleagues report the Australian trial in PLA Notes.

2003 — Dart and Davies publish the first MSC journal article in the American Journal of Evaluation, 24(2).

2005 — Davies and Dart publish The Most Significant Change (MSC) Technique: A Guide to Its Use.

2018 — Dart receives the Australian Evaluation Society's Outstanding Contribution to Evaluation Award.

The problem it solved

Diverse, hard-to-predict outcomes in participatory programmes that indicator systems captured poorly, including unexpected and negative changes.

Impact WHAT CHANGED

By 2004, according to the 2005 guide, MSC had been used by NGOs and governments in Africa, Asia, Latin America, Europe and Australasia, and the guide became the standard reference. MandE NEWS lists more than 360 documents in its MSC bibliography. In South Asia the method has been used well beyond its Bangladeshi origin: for example, the India HIV/AIDS Alliance used MSC as the participatory method in evaluating a DFID-supported programme on community driven approaches to the feminisation of HIV/AIDS in India, a report Davies listed on MandE NEWS in 2007.

The deeper influence, in our interpretation, is that MSC gave practitioners permission to treat qualitative accounts as a legitimate, structured part of monitoring rather than a decoration for reports. It is now a common choice in participatory, advocacy and community programmes where outcomes are diverse and hard to predict, often alongside indicator-based systems as both authors recommend. Davies's wider body of tools on MandE NEWS, from card sorting to EvalC3, extends the same interest in practical methods for complex programmes, and Dart's later work on evaluating systems change carries the dialogical emphasis into new territory.

KEY WORKS TO READ IN THE ORIGINAL

- **A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique** 2003, Dart and Davies, American Journal of Evaluation
- **The Most Significant Change (MSC) Technique: A Guide to Its Use** 2005, Davies and Dart

WHERE PEOPLE DISAGREE

Debates

MSC's strengths carry costs. Dart's Target 10 trial reported that the process took time, raised questions about informants' confidentiality, and that its competitive element left some people disillusioned when their stories were not chosen. Others have noted that the process can favour well written stories over the most significant ones, that individual stories are hard to compare, and that implementation is often incomplete, with stories collected but never properly selected or fed back. The guide itself warns that MSC is poor at describing average experience and unsuited to accountability reporting or quick, cheap evaluation. Supporters respond that these are reasons to combine MSC with other methods, which is what its authors always advised.

LIMITS, IN OUR READING

- ☐ The selection dialogue is time-consuming and needs skilled facilitation, and participants may distrust its subjectivity or dislike its competitive element.
- ☐ A 2010 WorldFish note warned that the process may favour well-written stories over the most significant ones and that individual stories are hard to compare, so MSC should sit alongside indicator-based systems.

WHY THIS MATTERS TO YOU

Most Significant Change works anywhere people want to understand what really matters. Ask a simple question, what was the most significant change for you?, and listen to the stories. Then do the part most people skip: sit together and choose the most significant, and say why. That conversation reveals what a team, an organisation or even a family truly values. For anyone in monitoring and evaluation, it is a low-cost, disciplined way to bring people's voices into the evidence, alongside the numbers. Run one selection meeting with your team and your funder, and you may learn more about what your programme is for than a year of dashboards could tell you.

Did you know?

Before 2000 the method had several other names, including monitoring without indicators, the story approach, impact monitoring and the evolutionary approach to organisational learning.

About 360 documents

Entries in the MSC bibliography maintained on Monitoring and Evaluation NEWS

Monitoring and Evaluation NEWS, Most Significant Change page

8 languages

Completed translations of the MSC guide listed: French, Spanish, Arabic, Hindi, Sinhala, Tamil, Japanese and Chinese

Monitoring and Evaluation NEWS, Most Significant Change page

5 world regions by 2004

MSC used by NGOs and governments in Africa, Asia, Latin America, Europe and Australasia

Davies and Dart, The Most Significant Change (MSC) Technique: A Guide to Its Use, 2005

785 villages

Reach of CCDB, the Bangladeshi NGO where MSC was first used, by 1996, with over 46,000 people

Monitoring and Evaluation NEWS, Most Significant Change page

Key concepts

Significant change story
A short account, given by a participant or staff member, of a change they consider important during a set period, with their explanation of why it matters.

Domain of change
A broad, deliberately loose category used to group stories, such as changes in livelihoods; guides often suggest three to five, including an open domain.

Selection process
Structured discussion at successive levels of an organisation in which groups choose the most significant stories and record their reasons.

Feedback
Returning the selected stories and the reasons for selection to the people who told and collected them.

Verification
Checking the accuracy of selected stories and gathering more detail, especially for those used beyond the organisation.

Meta-monitoring
Analysing the MSC process itself, for example who tells stories and which kinds are selected, as part of secondary analysis.

Sources for this chapter: p. 583

In one line

Rick Davies and Jess Dart showed that a well chosen story, with its reasons written down, can carry a village's judgement all the way to the boardroom.

THE BIG IDEAS OF THIS CHAPTER

1. **Most Significant Change (MSC)** p. 370

   MSC is a participatory monitoring and evaluation method in which people collect stories of significant change and groups at successive levels select the most significant, recording why, then feed their choices back.

2. **Monitoring without predefined indicators** p. 373

   By asking about significant change within deliberately loose domains rather than counting predefined indicators, MSC lets the people involved decide what counts and surfaces the unexpected, including bad news.

3. **Selection as a way of making values explicit** p. 376

   In MSC the choosing matters more than the stories, because requiring selectors to explain why one change matters more than another makes an organisation's values visible and open to debate.

GIANTS CONNECTED TO THIS CHAPTER



Robert Chambers

Both put participants' own knowledge and analysis at the centre of monitoring.

p. 296



Michael Scriven

Like goal-free evaluation, MSC looks for change without starting from pre-set goals.

p. 60



Egon G. Guba and Yvonna S. Lincoln

The selection dialogue is a practical form of negotiating values among stakeholders.

p. 318

1

BIG IDEA 1 OF 3 · DAVIES AND DART

Most Significant Change (MSC)

MSC is a participatory monitoring and evaluation method in which people collect stories of significant change and groups at successive levels select the most significant, recording why, then feed their choices back.

PICTURE IT

Think of a family looking back at the end of a year. Instead of counting things, each person shares the moment that mattered most to them, and together they choose which story best captures the year. Most Significant Change does this for a programme, carefully and at every level.

The problem

Large participatory programmes such as CCDB's in Bangladesh worked with many groups pursuing different goals. A single set of predefined indicators could not capture what was changing across 785 villages, and risked missing the unexpected entirely. Conventional reporting also flowed upwards without helping the organisation decide what it valued. Staff produced numbers, managers received them, and little learning or dialogue followed, so the people whose lives were changing had almost no voice in how success was judged.

What it is

MSC is a form of participatory monitoring and evaluation built on stories and structured dialogue. The 2005 guide by Davies and Dart sets out ten steps: raising interest; defining domains of change; defining the reporting period; collecting significant change stories; selecting the most significant stories; feeding back the results; verifying stories; quantification; secondary analysis and meta-monitoring; and revising the system. Collection, selection and feedback are the core. At collection, participants or staff are asked to describe the most significant change they have seen in a domain during a recent period, and why it matters to them. At selection, groups at successive levels of the organisation read the stories and choose the most significant, always recording their reasons. At feedback, the choices and reasons return to the storytellers. The logic is cumulative: each round tells everyone what the organisation values, shapes what people look for next time, and keeps a record of change that indicators may miss.

How it works

1. **Raise interest and find champions**
   An enthusiastic person or small group introduces MSC to key people, often with examples from other programmes, and secures senior commitment for several cycles, which the guide names as an enabler.
2. **Define domains and the reporting period**
   Agree three to five broad domains, usually including an open one, and decide how often to collect; the guide reports three-monthly cycles as the most common.
3. **Collect stories**
   Ask participants and field staff, in their own language, for the most significant change in a domain over the period and why it is significant to them, recording consent and context.
4. **Select at successive levels**
   Panels at field, district and head office levels read the stories, choose the most significant using voting, scoring or discussion, and write down their reasons.
5. **Feed back the choices**
   Tell storytellers and collectors which stories were chosen and why, by meeting, newsletter or report.
6. **Verify selected stories**
   Check the facts of chosen stories, usually through a follow-up visit, especially those used outside the organisation.
7. **Quantify and analyse**
   Count recurring changes, analyse all stories including unselected ones, and monitor the process itself, such as who tells stories and whose are chosen.
8. **Revise the system**
   Adjust domains, frequency and panels in light of what the cycles reveal.

WORKED EXAMPLE · AN ILLUSTRATION

A women's self-help group federation in Maharashtra adds MSC to its quarterly monitoring. Field facilitators collect 40 stories in Marathi across three domains: changes in income, changes in decision-making at home, and any other change. Block teams read their stories and choose eight, writing a short paragraph on each choice. The district team chooses two, and the funder's panel reads both with the reasons attached. Feedback meetings in the villages explain which stories travelled furthest and why. In the second cycle, facilitators notice that stories about women speaking at gram sabhas are being selected at every level, and members begin to look for such changes deliberately. Secondary analysis of all 40 stories also reveals a cluster of reports about loan repayment stress, which the federation follows up through its regular indicators.

USE IT WHEN

- Programmes with diverse, emergent outcomes that are hard to predict in advance.
- Organisations that want monitoring to prompt dialogue about values, not only reporting.
- Participatory, advocacy and community programmes where participants' own definitions of change matter.

SKIP IT WHEN

- Simple programmes with easily defined outcomes, such as vaccination coverage, where the guide says quantitative monitoring is enough.
- Accountability reports, quick and cheap evaluations, or questions about average experience, which the guide says other methods handle better.

COMMON MISREADINGS

- [ ] Collecting stories but never running selection and feedback, which removes the core of the method.
- [ ] Treating selected stories as typical, when they are chosen as most significant, not as representative.

WHY IT WAS GROUNDBREAKING

It made participants' own accounts of change a systematic part of monitoring rather than decoration for reports. It also created a repeatable route by which a story told in a village could reach, and be debated by, managers and funders.

Origin

Rick Davies devised the technique during PhD fieldwork with CCDB, which first used it in Rajshahi, Bangladesh, in 1994; he published the first public description online in 1996. Jess Dart trialled it in Victoria, Australia, from 1998 to 1999. The name was settled in 2000, the first journal article appeared in the American Journal of Evaluation in 2003, and the 2005 guide was later translated into languages including Hindi and Tamil.

RELATED IDEAS

Monitoring without predefined indicators

Selection as a way of making values explicit

RC Chambers 296 · MS Scriven 60 · G&L Guba and Lincoln 318

CONCEPT HIGHLIGHT

Significant change story

A short account, given by a participant or staff member, of a change they consider important during a set period, with their explanation of why it matters.

2

BIG IDEA 2 OF 3 · DAVIES AND DART

Monitoring without predefined indicators

By asking about significant change within deliberately loose domains rather than counting predefined indicators, MSC lets the people involved decide what counts and surfaces the unexpected, including bad news.

PICTURE IT

Think of fishing with a net instead of a hook. A hook catches only the fish you were aiming for. A net brings up everything in the water, including the ones you never expected. Asking about significant change, rather than counting set indicators, is the net.

The problem

Indicator systems measure only what designers expected when they wrote the logframe. In complex programmes, the most important changes are often ones nobody planned: a shift in relationships, a new local leader, or an unintended harm. Because indicators fix the question in advance, these changes go unrecorded and unreported, and programmes can look successful while missing what matters to participants. Practitioners needed a disciplined way to look beyond the indicators without abandoning systematic monitoring.

What it is

Before the name Most Significant Change was settled in 2000, the method was known as monitoring without indicators, among other names. The idea is that MSC does not begin with a list of measures. Instead it uses domains of change: broad, deliberately loose categories such as changes in people's lives, often with one fully open domain for any other change. The guide suggests three to five domains as usually manageable and notes that domains are optional. Within each, respondents define significance themselves. This makes the method good at revealing unexpected outcomes, including negative ones; in Dart's Target 10 trial, a lessons learned domain was added about eight months in, after feedback that bad news stories were valuable. The guide is clear that MSC is a complement, not a replacement: it is weak at capturing expected change or average experience, so it is meant to sit beside an indicator system and inform it.

How it works

1. **Map what the indicators already cover**
   List the outcomes the existing monitoring framework measures, so the team can see where its blind spots are likely to be.
2. **Define loose domains**
   Choose three to five broad domains that point to those blind spots, and include one open domain for any other change.
3. **Include a space for bad news**
   Add a domain on lessons or problems, following the Target 10 experience, so that negative change has a legitimate place.
4. **Ask the open question each period**
   Collect stories without suggesting which changes are expected, and record why each teller sees the change as significant.
5. **Analyse all stories for surprises**
   In secondary analysis, look across every story, selected or not, for changes the logframe did not anticipate.
6. **Feed surprises back into the indicator system**
   Where an unexpected change recurs, decide whether to add an indicator, commission a study or change the programme.

WORKED EXAMPLE · AN ILLUSTRATION

A foundational literacy programme in government primary schools in Uttar Pradesh tracks attendance, teacher training and reading scores through indicators. Its MEL team adds MSC with three domains: changes in classroom practice, changes in relationships between teachers and parents, and any other change, plus a lessons learned domain for problems. Cluster resource persons collect stories each term. In the open domain, several stories describe older siblings now staying home on market days to mind younger children so that mothers can attend parent meetings, an unintended cost nobody had planned to measure. The lessons domain surfaces teachers' complaints that new reading materials arrived after the term began. Neither change appeared in the indicator reports. The team adds a question on sibling absence to its next household survey and raises the delivery delay with the district office.

CONCEPT HIGHLIGHT

Domain of change

A broad, deliberately loose category used to group stories, such as changes in livelihoods; guides often suggest three to five, including an open domain.

USE IT WHEN

- The logframe is known to be incomplete, or outcomes are expected to emerge during implementation.
- Programmes where unintended or negative effects are a real risk and need a regular channel.
- Early phases of a programme, to discover which outcomes deserve indicators later.

SKIP IT WHEN

- Outcomes are simple, agreed and well measured, so open domains add cost without new information.
- There is no capacity to analyse stories and act on surprises, so they would be collected and ignored.

COMMON MISREADINGS

- [ ] Reading monitoring without indicators as anti-indicator, when the authors always advised combining MSC with indicator-based systems.
- [ ] Defining domains so tightly that they become indicators in disguise and lose their power to surprise.

WHY IT WAS GROUNDBREAKING

It challenged the assumption that monitoring must start from predefined measures. It offered a disciplined, repeatable way to look for changes that nobody planned, including harms.

Origin

The approach began with CCDB in Rajshahi, Bangladesh, in 1994, under names including monitoring without indicators and the story approach; the name Most Significant Change was settled in 2000. Dart and colleagues reported the Target 10 trial, including the added lessons domain, in PLA Notes in 2000, and the 2005 guide set out the advice on domains and on combining MSC with other methods.

RELATED IDEAS

Most Significant Change (MSC)

MS Scriven 60 · MP Patton 204 · PR Rogers 268

CONCEPT HIGHLIGHT

Selection process

Structured discussion at successive levels of an organisation in which groups choose the most significant stories and record their reasons.

3

BIG IDEA 3 OF 3 · DAVIES AND DART

Selection as a way of making values explicit

In MSC the choosing matters more than the stories, because requiring selectors to explain why one change matters more than another makes an organisation's values visible and open to debate.

PICTURE IT

Think of a film festival jury. The films matter, but what reveals the jury's values is the discussion about why one film wins. In MSC, the stories matter, but the choosing, and the reasons given, show what an organisation truly cares about.

The problem

Every monitoring system embodies judgements about what counts as success, but these are usually hidden inside indicator choices made by planners and donors. Field staff, participants and funders may value quite different changes without ever discussing it. When values stay hidden, disagreements surface only as frustration or as quietly ignored reports, and organisations lose the chance to learn what they are actually trying to achieve.

What it is

The selection step is the heart of MSC. Panels at each level must choose among stories and record their reasons, so the process generates a running record of what different groups consider significant. Davies's doctoral thesis, Order and Diversity (1998), framed this in terms of evolutionary epistemology: stories supply variation, selection panels apply judgement, and retained choices, fed back, shape what people notice in the next round. The 2005 guide describes several ways to select, including majority vote, iterative voting, scoring, pre-scoring followed by discussion, and secret ballot where power imbalances exist, and stresses that reasons must be recorded whatever the method. Recording reasons makes the subjectivity of judgement transparent and open to scrutiny, rather than pretending it away. Dart's Target 10 trial found that the selection meetings were where most discussion of values took place.

How it works

1. **Form panels at each level**
   Set up selection groups for field, district and senior levels, and include participants' representatives where possible, which the guide suggests reduces the risk of imposing programme values.
2. **Read stories within a domain**
   Panels read all stories in one domain at a sitting, keeping sessions to a manageable length.
3. **Choose a selection method**
   Use majority vote, iterative voting, scoring, pre-scoring or secret ballot, as the 2005 guide describes, matching the method to the group's dynamics.
4. **Record the reasons**
   Write down why the chosen story was selected and, where useful, why close contenders were not.
5. **Compare reasons across levels**
   Set the reasons given by different panels side by side and discuss where values diverge.
6. **Feed back and discuss**
   Return the choices and reasons to storytellers and collectors, inviting their response in the next cycle.

WORKED EXAMPLE · AN ILLUSTRATION

A girls' education programme in Rajasthan runs MSC selection at three levels. The block panel of teachers and community mobilisers chooses a story about a girl returning to school after marriage was postponed, citing courage and family change. The district panel chooses a story about a mother elected to the school management committee, citing lasting institutional change. The funder's panel, reading both, prefers the second. Instead of treating the disagreement as noise, the MEL lead circulates all three sets of reasons and holds a short meeting on what the programme is really trying to change: individual girls' trajectories, or the institutions around them. The discussion leads the programme to add a domain on community institutions and to explain its theory of change more clearly to the funder.

CONCEPT HIGHLIGHT

Feedback

Returning the selected stories and the reasons for selection to the people who told and collected them.

USE IT WHEN

- Organisations whose staff, participants and funders may hold different views of success.
- Moments of strategy review, when an organisation needs to clarify what it values.
- Multi-level programmes where field insight rarely reaches senior decision-makers.

SKIP IT WHEN

- Hierarchies are so steep that panels cannot express genuine preferences, unless secret ballot and careful facilitation are possible.
- There is no willingness at senior level to read and respond to the reasons.

COMMON MISREADINGS

- [ ] Treating selection as a contest to find the best written or most moving story for publicity.
- [ ] Skipping the recording of reasons, which removes the transparency that makes the subjectivity defensible.

WHY IT WAS GROUNDBREAKING

It treated disagreement about what matters as data to be surfaced, not noise to be averaged away. In our reading it linked monitoring directly to organisational learning about purpose, which few methods had done.

Origin

The idea is rooted in Davies's thesis Order and Diversity (1998) and his 1996 online paper, An evolutionary approach to facilitating organisational learning. Dart and Davies's 2003 American Journal of Evaluation article presented MSC as a dialogical, story-based tool, and the 2005 guide set out selection methods and the requirement to record reasons.

RELATED IDEAS

Most Significant Change (MSC)

G&L Guba and Lincoln 318
HP Preskill 226
JG Greene 340

CONCEPT HIGHLIGHT

Verification

Checking the accuracy of selected stories and gathering more detail, especially for those used beyond the organisation.

RELATED GIANTS



Robert Chambers

The former colonial administrator who spent a career asking development professionals to learn from the people they usually see last.

p. 296



Michael Scriven

The philosopher who asked evaluators to judge what a programme actually does for the people it serves, not only whether it hit its goals.

p. 60



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204

CONCEPT HIGHLIGHT

Meta-monitoring

Analysing the MSC process itself, for example who tells stories and which kinds are selected, as part of secondary analysis.

THE BIG IDEAS OF THIS CHAPTER

1 **Most Significant Change (MSC)** p. 370

MSC is a participatory monitoring and evaluation method in which people collect stories of significant change and groups at successive levels select the most significant, recording why, then feed their choices back.

2 **Monitoring without predefined indicators** p. 373

By asking about significant change within deliberately loose domains rather than counting predefined indicators, MSC lets the people involved decide what counts and surfaces the unexpected, including bad news.

3 **Selection as a way of making values explicit** p. 376

In MSC the choosing matters more than the stories, because requiring selectors to explain why one change matters more than another makes an organisation's values visible and open to debate.

Put it to work

USE IT ON MONDAY · MOST SIGNIFICANT CHANGE (MSC)

1. Agree two to four broad domains of change with the team, for example changes in livelihoods or in relationships, and leave one domain open for any other change.

2. Ask participants, in their own language, to describe the most significant change they have seen in a domain during a set period, and why it matters to them. Record consent and context with each story.

3. Hold a selection meeting at each level, record the reasons for each choice, feed the choices back to storytellers and verify a sample of stories.

AI prompt **Copy, fill the brackets, then check every line**

Help me analyse Most Significant Change stories, following the approach of Rick Davies and Jess Dart, from **[programme]** in **[location]**. Our domains of change are **[list domains]**. Here are the stories with the reasons storytellers gave: **[paste stories]**. Here are the selection panel's notes: **[paste notes]**. (1) Summarise the themes across stories within each domain, citing story numbers. (2) Summarise what the panel's reasons reveal about what this group values. (3) Note any negative or unexpected changes. (4) Suggest which stories most need verification. Do not rewrite, embellish or merge stories, and do not create example stories.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a family looking back at the end of a year. Instead of counting things, each person shares the moment that mattered most to them, and together they choose which story best captures the year. Most Significant Change does this for a programme, carefully and at every level.

YOUR MOVES Three moves from Davies and Dart's path

1 Solve one real problem brilliantly

Davies built the Most Significant Change technique from one NGO's concrete problem during his doctoral fieldwork in Bangladesh.

TRY THIS
Take one stubborn monitoring problem in your programme and design a simple method to solve it.

2 Test your idea somewhere different

Dart trialled the method in a very different setting in Australia before co-writing the guide, joining invention with independent trial.

TRY THIS
Ask a peer in another context to pilot your tool and tell you what breaks.

3 Give it away so it lasts

They released the guide free under a Creative Commons licence and kept a public bibliography, which helped the method outlast better funded ones.

TRY THIS
Publish your best template under an open licence with a clear how-to.

THE BIG IDEAS OF THIS CHAPTER

1 Most Significant Change (MSC) p. 370

MSC is a participatory monitoring and evaluation method in which people collect stories of significant change and groups at successive levels select the most significant, recording why, then feed their choices back.

2 Monitoring without predefined indicators p. 373

By asking about significant change within deliberately loose domains rather than counting predefined indicators, MSC lets the people involved decide what counts and surfaces the unexpected, including bad news.

3 Selection as a way of making values explicit p. 376

In MSC the choosing matters more than the stories, because requiring selectors to explain why one change matters more than another makes an organisation's values visible and open to debate.

CONCEPT HIGHLIGHT

Meta-monitoring

Analysing the MSC process itself, for example who tells stories and which kinds are selected, as part of secondary analysis.

GIANTS CONNECTED TO THIS CHAPTER



Robert Chambers

Both put participants' own knowledge and analysis at the centre of monitoring.

p. 296



Michael Scriven

Like goal-free evaluation, MSC looks for change without starting from pre-set goals.

p. 60



Egon G. Guba and Yvonna S. Lincoln

The selection dialogue is a practical form of negotiating values among stakeholders.

p. 318

KEY WORKS TO READ IN THE ORIGINAL

- **A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique** 2003, Dart and Davies, American Journal of Evaluation
- **The Most Significant Change (MSC) Technique: A Guide to Its Use** 2005, Davies and Dart

CONCEPT HIGHLIGHT

Significant change story

A short account, given by a participant or staff member, of a change they consider important during a set period, with their explanation of why it matters.

CONCEPT HIGHLIGHT

Domain of change

A broad, deliberately loose category used to group stories, such as changes in livelihoods; guides often suggest three to five, including an open domain.

WORKSHEET · TRY THIS TEMPLATE

MSC story and selection form

Print this page, or copy the headings into your notes.

1 Domain of change

Which agreed domain the story belongs to, or the open domain.

2 Storyteller and context

Role, location, date and consent. Protect identities where needed.

3 The story

What happened, what it was like before and after, who was involved.

4 Why it is significant to the teller

In their terms, kept separate from the evaluator's view.

5 Selection decision and reasons

Which panel chose it, and the reasons they gave.

6 Verification and feedback

How the story was checked, and when the result went back to the teller.

REFLECT

When your team last chose a story for a report, what did that choice reveal about what you value?

CONNECTS TO

RC **Robert Chambers**
Both put participants' own knowledge and analysis at the centre of monitoring. p. 296

MS **Michael Scriven**
Like goal-free evaluation, MSC looks for change without starting from pre-set goals. p. 60

G&L **Egon G. Guba and Yvonna S. Lincoln**
The selection dialogue is a practical form of negotiating values among stakeholders. p. 318

17



Emblem: an original HubForge design for this figure, not a likeness.

Donna M. Mertens

Born 1951 Active 1990s to 2020s

United States; collaborations with scholars in New Zealand, Botswana, South Africa and beyond

Participation, equity and voice · Value and judgement

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

IN ONE LINE

Donna Mertens turned evaluation towards justice, asking every evaluator to begin with the people a system leaves out.

Who gets heard when a programme is judged? For Donna Mertens, that question was never abstract. For more than 30 years she taught research methods and evaluation to deaf and hearing graduate students at Gallaudet University, the university for deaf and hard of hearing students, and her research reached into deaf education, the inclusion of students with disabilities and access to the courts for deaf people. In 1998 she led the American Evaluation Association, whose conference that year in Chicago took the theme Transforming Society Through Evaluation. From her ground she built the transformative paradigm, which places people facing discrimination at the centre of evaluation. In 2025, when The Evaluators' Institute presented its first Social Betterment Award, she was the recipient, nominated by scholars in seven countries, from Botswana and South Africa to New Zealand and Spain.

BIG IDEAS

1. **The transformative paradigm**
2. **Transformative mixed methods**

KEY WORKS

Transformative Research and Evaluation
2008

Indigenous Pathways to Social Research
2013 · Edited with Fiona Cram and Bagele Chilisa

Mixed Methods Design in Evaluation
2018

Life CHAPTER 17

Donna M. Mertens was born in 1951 in Spokane, Washington. She took a BA in psychology at Thomas More College and an MS and a PhD in educational psychology at the University of Kentucky. She then spent her career at Gallaudet University, the university for deaf and hard of hearing students, in the Department of Educational Foundations and Research. For more than 30 years she taught research methods and programme evaluation to deaf and hearing master's and doctoral students.

Her research grew from that setting. SAGE lists among her interests special education services in international settings, the inclusion of students with disabilities in neighbourhood schools, the prevention of sexual abuse in residential schools for deaf students, court access for deaf and hard of hearing people, and teacher

preparation using instructional technology. She edited Creative Ideas for Teaching Evaluation in 1989, co-wrote Parents and Their Deaf Children with Kay Meadow-Orlans and Marilyn Sass-Lehrer, and wrote the textbook Research and Evaluation in Education and Psychology, which reached its sixth edition in 2024.

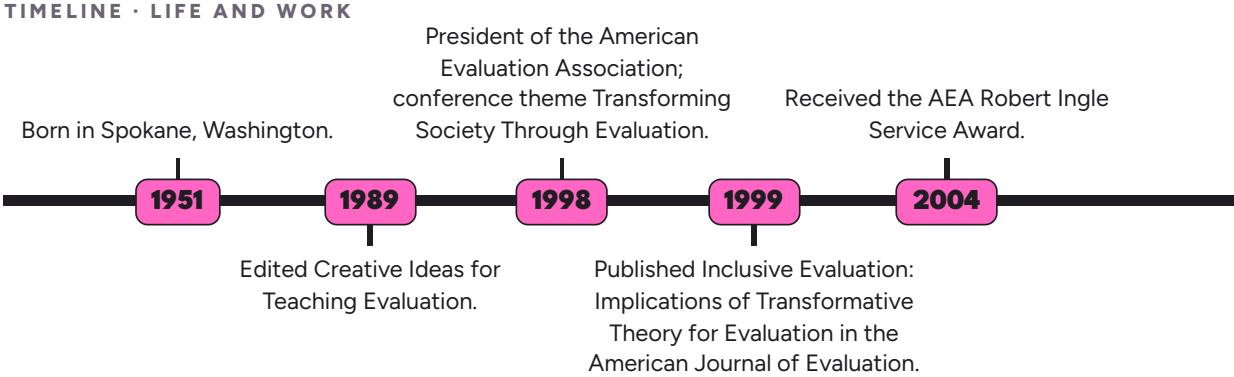
In 1998 she was president of the American Evaluation Association, whose conference in Chicago that year took Transforming Society Through Evaluation as its theme. The following year she published Inclusive Evaluation: Implications of Transformative Theory for Evaluation in the American Journal of Evaluation. She sat on the board of trustees of the International Organization for Cooperation in Evaluation in 2002 and 2003, received AEA's Robert Ingle Service Award in 2004, and set out the transformative paradigm for mixed methods in the first volume of the Journal of Mixed Methods Research in 2007. Guilford published Transformative Research and Evaluation in October 2008, and in 2009 she co-edited the Handbook of Social Research Ethics with Pauline Ginsberg and received AEA's Paul F. Lazarsfeld Evaluation Theory Award. She edited the Journal of Mixed Methods Research from 2009 to 2014.

Her later work reached increasingly across borders. In 2013 she co-edited Indigenous Pathways into Social Research: Voices of a New Generation with Fiona Cram and Bagele Chilisa, a collection of life stories from more than 30 Indigenous researchers on six continents. Mixed Methods Design in Evaluation followed in 2018, and the third edition of Program Evaluation Theory and Practice, with Jori Hall and Amy Wilson, in 2025. On 1 May 2025 The Evaluators' Institute presented her with its inaugural Social Betterment Award, carrying a 10,000 dollar prize, after nominations from scholars in the UK, South Africa, New Zealand, Botswana, the Bahamas, the United States and Spain. In May 2026 she gave a keynote on transformative evaluation in Brasília. She is Professor Emerita at Gallaudet.

The spark

Her spark was her setting. At Gallaudet, Mertens worked with communities that are often left out of decisions about the programmes meant to serve them, and her research addressed deaf education,

TIMELINE · LIFE AND WORK



the inclusion of students with disabilities and access to justice. In her 2007 article she described a study at a state residential school for deaf students, in which surveys answered administrators' questions while qualitative methods were needed to address the harm itself. In our reading, the question of who gets heard became the problem that shaped her work.

The breakthrough CHAPTER 17

Imagine an evaluation of a school programme in which the questions are written by officials, the surveys are in a language some families cannot easily use and the results are reported as averages. The students most affected may never appear in the findings at all. By the late 1990s, in Mertens's framing, evaluators faced a growing call to include the groups usually left out. Her own textbooks later set the transformative paradigm beside two familiar ways of doing research, one built around measurement and one around interpretation. What made hers distinct was its explicit starting point: power, discrimination and social justice.

In her 2007 article in the Journal of Mixed Methods Research she set out the transformative paradigm using the four philosophical assumptions identified by Egon Guba and Yvonna Lincoln: axiology, ontology, epistemology and methodology. Its ethics build on respect, beneficence and justice, interpreted through community norms and human rights. Reality is seen as socially constructed, with privilege shaping which version counts. Knowing requires an interactive, culturally respectful relationship between researchers and participants. Methods may be quantitative, qualitative or mixed, but must address power and discrimination explicitly. Her 2008 book turned this into a framework for methodological decisions.

TIMELINE, CONTINUED

2007 — Set out the transformative paradigm for mixed methods in the Journal of Mixed Methods Research.

2008 — Transformative Research and Evaluation published by Guilford.

2009 — Received the AEA Paul F. Lazarsfeld Evaluation Theory Award and began editing the Journal of Mixed Methods Research.

2013 — Co-edited Indigenous Pathways into Social Research with Fiona Cram and Bagele Chilisa.

2025 — Received the inaugural Social Betterment Award from The Evaluators' Institute.

The problem it solved

Evaluations that excluded the people most affected by decisions, and method choices made by habit rather than by ethical and social aims.

Impact WHAT CHANGED

Mertens's influence is visible in the geography of her recognition. The nominators for her 2025 Social Betterment Award came from the UK, South Africa, New Zealand, Botswana, the Bahamas, the United States and Spain, and a selection committee member described her influence as global in both theory and practice. Her co-edited Indigenous Pathways into Social Research brought together more than 30 Indigenous researchers from six continents, and her collaborations with Chilisa and Cram connected her framework to Indigenous scholarship in Africa and Aotearoa New Zealand. Her co-author Amy Wilson led Gallaudet's international development programmes and later directed deaf education specialists working with parents and educators in economically poor countries, which, in our reading, links her evaluation textbooks to disability-inclusive development practice.

Her ideas reach development evaluation directly. Her 2026 keynote in Brasília, archived by Brazil's National School of Public Administration, used a UN Women study in Pakistan on the economic empowerment of women home-based workers to show how disaggregated findings about women with disabilities changed a funder's approach. In our assessment, her framework gives NGOs and evaluators working on disability, gender and Indigenous rights a vocabulary for insisting that the most excluded help shape evaluation questions, and for asking who a programme leaves out. Her textbooks, now in multiple editions, have carried the transformative paradigm into the methods training of many graduate students.

KEY WORKS TO READ IN THE ORIGINAL

- **Transformative Research and Evaluation** 2008
- **Indigenous Pathways to Social Research** 2013, Edited with Fiona Cram and Bagele Chilisa
- **Mixed Methods Design in Evaluation** 2018

WHERE PEOPLE DISAGREE

Debates

Mertens's explicit commitment to social justice raises questions that she and others have engaged. Michele Tarsilla's 2010 conversation with her was framed around whether the transformative agenda can really alter the status quo; he concluded that it can. A concern often voiced in the field is that an avowed justice stance may weaken an evaluation's perceived independence with some commissioners. Indigenous scholars have articulated their own paradigm, and Mertens's later work treats Indigenous and transformative paradigms as overlapping but distinct. In our view, transformative designs also demand sustained community participation that short project cycles rarely fund.

LIMITS, IN OUR READING

- ☐ Evaluators from more conventional traditions may ask whether an explicit social justice commitment compromises perceived independence.
- ☐ In our view, transformative designs demand sustained community participation, which is hard to fit within short project cycles.

WHY THIS MATTERS TO YOU

Mertens's question belongs in every room where decisions are made about people: who is missing from this conversation? If your work touches disability, gender, caste, tribe or language, she gives you a framework to insist that the people most affected help shape the questions. Make your data collection accessible. Break down results so that excluded groups become visible instead of disappearing into averages. Link findings to action. In her framing this is not a kind extra; it is what makes evaluation ethical and its findings credible. Treat every evaluation as a chance to ask who a programme leaves out, and to help bring them in.

Did you know?

For more than three decades Mertens taught research methods and evaluation to both deaf and hearing graduate students at Gallaudet, the university for deaf and hard of hearing students.

6

editions of Research and Evaluation in Education and Psychology, the sixth published in 2024

SAGE listing via Bokus (2024)

30+

Indigenous researchers from six continents whose life stories appear in Indigenous Pathways into Social Research (2013)

MIT Press Bookstore listing of the Left Coast Press edition

30+

years teaching research methods and evaluation to deaf and hearing graduate students at Gallaudet

SAGE Publishing author page; Guilford Press author details

7

countries from which nominations came for her 2025 Social Betterment Award

Claremont Graduate University news (May 2025)

Key concepts

Transformative paradigm

A framework for research and evaluation that places the experiences of people facing discrimination and oppression at the centre, attends to power and aims to support social justice and human rights.

Axiology

The branch of philosophical assumptions concerned with ethics and values; in the transformative paradigm it is treated as the starting point for other choices.

Inclusive evaluation

Evaluation designed to involve and represent marginalised groups who are often excluded from decisions about programmes and evaluations.

Cultural responsiveness

Designing and conducting evaluation in ways that respect and fit the cultural norms, languages and histories of the communities involved.

Intersectionality

Attention to how identities such as race, disability, gender, sexuality, Deafness, class and migration status combine to shape experiences of privilege and discrimination.

Indigenous paradigm

An approach to research grounded in Indigenous knowledge, centring relationships, interconnectedness, decolonisation and sovereignty, which Mertens treats as distinct from but overlapping with the transformative paradigm.

Sources for this chapter: p. 584

THE BIG IDEAS OF THIS CHAPTER

1. **The transformative paradigm** p. 392
   Start from the reality of injustice, and design research and evaluation to centre people facing discrimination, address power and support social justice and human rights.
2. **Transformative mixed methods** p. 395
   Use mixed methods within a cycle of community participation, so that qualitative and quantitative evidence together expose inequality and help to address it.

GIANTS CONNECTED TO THIS CHAPTER



Bagele M. Chilisa

Co-editor of Indigenous Pathways to Social Research and a leading voice on relational, indigenous methods.

p. 498



Jennifer C. Greene

Shares the view that evaluation cannot be value-free.

p. 340



Marco Segone

Equity-focused evaluation asks who is reached and who is left behind.

p. 516

In one line

Donna Mertens turned evaluation towards justice, asking every evaluator to begin with the people a system leaves out.

1

BIG IDEA 1 OF 2 · MERTENS

The transformative paradigm

Start from the reality of injustice, and design research and evaluation to centre people facing discrimination, address power and support social justice and human rights.

PICTURE IT

Think of a building designed without anyone who uses a wheelchair in the room. It may look perfect on paper and still shut people out at the front door. The transformative paradigm insists that the people most likely to be excluded help design the evaluation from the start.

The problem

Mainstream evaluation often treated communities as subjects to be measured rather than partners. Deaf people and others with disabilities, Indigenous peoples, women and racial minorities were frequently left out of defining problems, sampled as if their groups were uniform, or reached through instruments in languages and formats they could not use. Findings could then misrepresent their lives and reinforce the decisions that excluded them. By the late 1990s, in Mertens's own framing, evaluators faced a call for greater inclusion of marginalised groups.

What it is

The transformative paradigm is a philosophical framework for research and evaluation that begins from the recognition that discrimination and inequality are pervasive, and aims to help reduce them. Mertens defined it through the four sets of assumptions identified by Egon Guba and Yvonna Lincoln. Axiology, treated as the starting point, extends the ethical principles of respect, beneficence and justice: respect is judged by the community's cultural norms, beneficence includes promoting human rights and social justice, and research processes and outcomes are linked to a justice agenda. Ontology holds that realities are socially constructed, that social, political, cultural and economic values shape them, and that power determines whose reality is privileged. Epistemology requires an interactive, trusting link between evaluator and participants, respectful of culture and alert to power. Methodology allows quantitative, qualitative or mixed methods, provided participants help define the problem, methods adapt

to cultural complexity, and power, discrimination and oppression are addressed explicitly. Her textbooks set it beside post-positivist, constructivist, pragmatic and Indigenous paradigms.

How it works

1. **Begin with ethics and power**
   Identify which groups face discrimination in the programme's context, and what respect means by their cultural norms, since axiology is the paradigm's starting point.
2. **Build relationships of trust**
   Invest time in culturally respectful relationships with community members and their organisations before collecting data.
3. **Define the problem with the community**
   Involve representatives of marginalised groups in framing the evaluation questions, as Mertens's 2007 article asks.
4. **Design for diversity and access**
   Avoid assuming groups are uniform; address barriers such as language, transport, childcare and work schedules, and use sign language or local languages where needed.
5. **Collect and disaggregate**
   Use whatever methods serve the justice purpose, and analyse results separately for the groups at risk of exclusion.
6. **Link findings to action**
   Interpret results with the community and connect them to changes that can reduce inequality.

WORKED EXAMPLE · AN ILLUSTRATION

A state rolls out a health insurance scheme for low-income families. A transformative evaluation begins by meeting organisations of people with disabilities, Dalit women's collectives and tribal community leaders, who help frame the questions. They point out that enrolment figures say nothing about whether members can use the benefit. The team designs accessible data collection: interviews in sign language and local languages, survey items read aloud, and venues reachable by wheelchair. Results are disaggregated by disability, caste, gender and location. They show that enrolment among households with a disabled member is high, but that few could use hospital benefits because empanelled hospitals lacked ramps and interpreters. Community representatives present the findings with the evaluators to the state health agency, which adds accessibility to its hospital empanelment criteria.

USE IT WHEN

- A programme serves people who face discrimination, such as people with disabilities, women, minorities or Indigenous peoples.
- Earlier evaluations have excluded or misrepresented the communities concerned.
- The commissioner wants evaluation to support rights and equity goals.

SKIP IT WHEN

- The commissioner will not permit community involvement in framing questions, which would make the approach nominal.
- There is no time or budget to build trust, in which case an honest, narrower scope may serve communities better than a token transformative label.
- The question is purely technical, such as verifying procurement records, though equity concerns may still deserve a separate study.

COMMON MISREADINGS

- [ ] Assuming transformative means qualitative only, when Mertens allows any method, quantitative included, that serves the justice purpose.
- [ ] Equating any participation with transformation, when her 2007 article distinguishes practical participation from transformative participation.

WHY IT WAS GROUNDBREAKING

It made social justice and power an explicit foundation of evaluation design rather than an optional add-on. It gave that commitment a philosophical grounding that placed it alongside, not beneath, the established research paradigms.

Origin

Mertens framed the approach publicly through her 1998 presidency of the American Evaluation Association, whose conference took Transforming Society Through Evaluation as its theme, and her 1999 American Journal of Evaluation article on inclusive evaluation and transformative theory. She set out the paradigm's four assumptions for mixed methods in the Journal of Mixed Methods Research in 2007, and gave it book length in Transformative Research and Evaluation (Guilford, October 2008).

RELATED IDEAS

Transformative mixed methods

BC Chilisa 498 · G&L Guba and Lincoln 318 · JG Greene 340 · RC Chambers 296

2

BIG IDEA 2 OF 2 · MERTENS

Transformative mixed methods

Use mixed methods within a cycle of community participation, so that qualitative and quantitative evidence together expose inequality and help to address it.

PICTURE IT

Think of a lamp and a magnifying glass. Numbers light up the whole room and show the overall pattern; close conversations magnify what is happening to particular people. Transformative mixed methods uses both, with the community, to expose inequality and act on it.

The problem

Mixed methods offered a way to combine depth and breadth, but designs were often chosen by disciplinary habit, with communities consulted, if at all, once at the start. Standardised surveys could miss local realities and the people least able to respond in a dominant language.

Mertens's 2007 article cites a Botswana HIV and AIDS case, reported by Bagele Chilisa, in which a needs assessment relying on literature and a standardised survey targeted English-literate people and produced billboards that offended the community.

What it is

Transformative mixed methods applies the transformative paradigm to the design of mixed-methods studies. In her 2007 article Mertens described a cyclical model in which community participation occurs at the beginning, throughout and at the end of each study, and processes and outcomes are revisited so that the research can be modified to fit community needs and produce sustainable change. Qualitative methods bring community perspectives into every stage, from defining the focus to interpreting results, and can capture perspectives where a language has no written form. Quantitative methods demonstrate outcomes in ways that communities and scholars find credible, and allow results to be disaggregated by group. The choice and sequence of methods follow from the justice purpose and the need for genuine participation, not from disciplinary habit. Mertens distinguishes practical participation from transformative participation, which requires cultural sensitivity and attention to power. Her 2018 book Mixed Methods Design in Evaluation develops these designs for evaluators.

How it works

1. **Engage the community at the start**
   Use focus groups, interviews, surveys or community meetings, with attention to power, to define the evaluation focus, the first point of the cycle in Mertens's 2007 model.
2. **Sample for diversity**
   Avoid assuming that groups are uniform, and plan to reach those facing barriers of language, transport, childcare or work schedules.
3. **Sequence methods for justice**
   Use qualitative work to understand context and quantitative work to show the scale of patterns, in whichever order the purpose requires.
4. **Return to the community midway**
   Share interim findings and adjust instruments or questions in response.
5. **Interpret together**
   Analyse results with community members, especially disaggregated findings about marginalised groups.
6. **Link data to action**
   Close each cycle by connecting findings to concrete changes, then begin the next cycle with what was learned.

WORKED EXAMPLE · AN ILLUSTRATION

A microfinance programme in rural Tamil Nadu wants to know why some women leave their savings groups. The evaluators first hold community dialogues with current and former members, facilitated by women from the area. These reveal reasons such as harassment over repayment, migration for work and husbands' control of loans. The team then designs a survey of 900 members and leavers to measure how common each reason is, with questions read aloud for women who cannot read. Results are disaggregated by caste, disability and household type. Interim findings are taken back to group meetings, where women point out that the survey missed seasonal migration, prompting an added module. In the final cycle, women and programme staff agree changes to repayment rules and meeting times.

CONCEPT HIGHLIGHT

Transformative paradigm

A framework for research and evaluation that places the experiences of people facing discrimination and oppression at the centre, attends to power and aims to support social justice and human rights.

USE IT WHEN

- A mixed-methods evaluation concerns groups whose voices standard surveys tend to miss.
- The evaluation can run over several cycles, allowing findings to return to the community.
- Decision-makers need both credible numbers and an understanding of why patterns occur.

SKIP IT WHEN

- The study is a one-off snapshot with no chance to return to participants.
- Community partners cannot be resourced or compensated for their time.
- Safety risks make shared interpretation sessions inappropriate, unless safer formats can be designed.

COMMON MISREADINGS

- [ ] Adding a few focus groups to a survey and calling the design transformative, without changing who defines the questions.
- [ ] Thinking the cycle is only consultation at the start, when Mertens places participation at the beginning, throughout and at the end.

WHY IT WAS GROUNDBREAKING

It tied method choice to ethical and social aims, giving evaluators a principled reason for mixing. It made community participation part of methodological rigour rather than a separate value.

Origin

Mertens set out the cyclical model in Transformative Paradigm: Mixed Methods and Social Justice, in the first volume of the Journal of Mixed Methods Research in 2007, drawing on cases from deaf education in the United States and from Botswana. She developed it in Transformative Research and Evaluation (2008) and Mixed Methods Design in Evaluation (2018), and edited the Journal of Mixed Methods Research from 2009 to 2014.

RELATED IDEAS

The transformative paradigm

JG Greene 340 · BC Chilisa 498 · SZ Zaveri 536

CONCEPT HIGHLIGHT

Axiology

The branch of philosophical assumptions concerned with ethics and values; in the transformative paradigm it is treated as the starting point for other choices.

Put it to work

USE IT ON MONDAY · THE TRANSFORMATIVE PARADIGM

1. List the groups most likely to be excluded from your programme and your evaluation, such as people with disabilities, minority language speakers or the poorest households, and involve their representatives in setting the evaluation questions.

2. Make data collection accessible: sign language or local language interpreters, accessible venues and formats, and safe settings for sensitive topics. Plan to disaggregate results for these groups.

3. Test every method choice against two questions: does it serve the justice purpose of this evaluation, and does it allow the community to take part meaningfully?

AI prompt **Copy, fill the brackets, then check every line**

Help me apply Donna Mertens's transformative paradigm to an evaluation of **[programme]** in **[location]**. The groups I believe may face discrimination or exclusion are **[list, with what I know]**. The draft evaluation design is **[paste]**. (1) Identify where power differences could shape the evaluation, from question-setting to reporting. (2) Suggest how representatives of excluded groups could shape the questions and tools. (3) Flag accessibility and safety issues in the current design. (4) Review each method against the justice purpose and suggest changes. Do not make assumptions about any group's experiences beyond what I have described.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a building designed without anyone who uses a wheelchair in the room. It may look perfect on paper and still shut people out at the front door. The transformative paradigm insists that the people most likely to be excluded help design the evaluation from the start.

YOUR MOVES

Three moves from Mertens's path

1 Go deep with one community

Decades of teaching and research at Gallaudet, a university for deaf students, grounded a transformative paradigm later recognised by nominators in seven countries.

TRY THIS

Commit to one community long enough to understand it from the inside, and let it shape your questions.

2 Use leadership to open doors

As AEA president in 1998 and editor of the Journal of Mixed Methods Research from 2009 to 2014, she opened space for justice-oriented work.

TRY THIS

Use any role you hold, even a small one, to bring one under-heard voice into the room.

3 Make room for other knowledge

With Fiona Cram and Bagele Chilisa she edited a 2013 volume of life stories from more than 30 Indigenous researchers.

TRY THIS

Ask the communities you evaluate how they judge success, and include their criteria in your framework.

THE BIG IDEAS OF THIS CHAPTER

1 **The transformative paradigm** p. 392

Start from the reality of injustice, and design research and evaluation to centre people facing discrimination, address power and support social justice and human rights.

2 **Transformative mixed methods** p. 395

Use mixed methods within a cycle of community participation, so that qualitative and quantitative evidence together expose inequality and help to address it.

CONCEPT HIGHLIGHT

Indigenous paradigm

An approach to research grounded in Indigenous knowledge, centring relationships, interconnectedness, decolonisation and sovereignty, which Mertens treats as distinct from but overlapping with the transformative paradigm.

GIANTS CONNECTED TO THIS CHAPTER



Bagele M. Chilisa

Co-editor of Indigenous Pathways to Social Research and a leading voice on relational, indigenous methods.

p. 498



Jennifer C. Greene

Shares the view that evaluation cannot be value-free.

p. 340



Marco Segone

Equity-focused evaluation asks who is reached and who is left behind.

p. 516

KEY WORKS TO READ IN THE ORIGINAL

- **Transformative Research and Evaluation** 2008
- **Indigenous Pathways to Social Research** 2013, Edited with Fiona Cram and Bagele Chilisa
- **Mixed Methods Design in Evaluation** 2018

CONCEPT HIGHLIGHT

Transformative paradigm

A framework for research and evaluation that places the experiences of people facing discrimination and oppression at the centre, attends to power and aims to support social justice and human rights.

CONCEPT HIGHLIGHT

Axiology

The branch of philosophical assumptions concerned with ethics and values; in the transformative paradigm it is treated as the starting point for other choices.

WORKSHEET · TRY THIS TEMPLATE

Transformative design check

Print this page, or copy the headings into your notes.

1 Whose experience is centred

Name the groups facing discrimination or exclusion in this context.

2 Power differences

Between funder, implementer, evaluator and community, and within the community.

3 How the community shapes questions

Advisory group, co-design workshop or review of tools.

4 Accessibility and safety

Languages, formats, venues, consent and protection.

5 Method rationale

Why each method serves the purpose and enables participation.

6 Route to change

How findings will support action by and for the community.

REFLECT

Who in the community your programme serves has never been asked to shape an evaluation question?

CONNECTS TO

BC **Bagele M. Chilisa** — Co-editor of Indigenous Pathways to Social Research and a leading voice on relational, indigenous methods. p. 498

JG **Jennifer C. Greene** — Shares the view that evaluation cannot be value-free. p. 340

MS **Marco Segone** — Equity-focused evaluation asks who is reached and who is left behind. p. 516

THE GREAT DEBATES · ACT 4

The Paradigm Wars

THE QUESTION

Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

G&L Guba and Lincoln 318 | JG Greene 340 | DM Mertens 384 | DC Campbell 112

SIDE A

A different paradigm with its own criteria (Guba and Lincoln)

Guba, a statistician who had worked inside the measurement tradition, came to doubt that conventional approaches could answer evaluation questions well. In Naturalistic Inquiry (1985) he and Lincoln argued that the naturalistic paradigm rests on different assumptions from conventional science, so studies within it should be judged by parallel criteria of trustworthiness rather than by internal and external validity. In 1986 they added criteria of authenticity, and in 1989 Fourth Generation Evaluation built an approach around negotiation among stakeholders. They advised against mixing positivist and constructivist propositions in a single study, while insisting that relativism did not mean anything goes. In our reading, their case is that methods carry worldviews.

SIDE B

The experimental tradition replies

Critics from the experimental tradition questioned whether a relativist position could yield findings that decision-makers could rely on, and whether parallel criteria could carry the weight of validity. William Shadish argued in 1995 that few evaluation theorists or practising evaluators were in fact logical positivists, a movement he described as largely dead in philosophy by the late 1940s, and that method choice often does not follow directly from philosophy. He also argued that experiments can include qualitative observation. Campbell himself, in Shadish's account, had written in support of qualitative work. In our reading, their case is that the war was fought partly against a caricature.

WHAT HAPPENED

The field moved towards combination rather than conquest. In 1989 Greene, Caracelli and Graham reviewed 57 mixed-method evaluations and named five purposes for mixing methods. In 1997 Greene and Caracelli described purist, pragmatic and dialectic stances, the last treating tensions between paradigms as a source of fuller understanding. Mertens later built her transformative paradigm on the four philosophical assumptions Guba and Lincoln had identified. Shadish took Denzin and Lincoln's 1994 Handbook of Qualitative Research as a sign that qualitative methods had become established. The argument over paradigms continues in quieter form.

Where you stand

In your current work, do numbers and stories answer different questions, or are they used to confirm each other? Which of Greene's purposes are you really serving?

Is there a finding you trust from interviews that you could not defend to a sceptical statistician? How would you make your case?

Your position

Which side do you find yourself on today, and what evidence would change your mind?

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

THE GIANTS IN THIS DEBATE



Egon G. Guba and Yvonna S. Lincoln

The partnership that answered the hardest question asked of qualitative evaluation: can it be rigorous?

p. 318



Jennifer C. Greene

The methodologist who argued that evaluators are never neutral, and that mixing methods should mean mixing ways of seeing.

p. 340



Donna M. Mertens

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

p. 384



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112

ACT 5 OF 5

5

1990s to 2020s

Evidence for a World in a Hurry



Ray C. Rist

p. 410



Esther Duflo and Abhijit Banerjee

p. 432



Zenda Ofir

p. 454



Howard White

p. 476



Bagele M. Chilisa

p. 498



Marco Segone

p. 516



Sonal Zaveri

p. 536

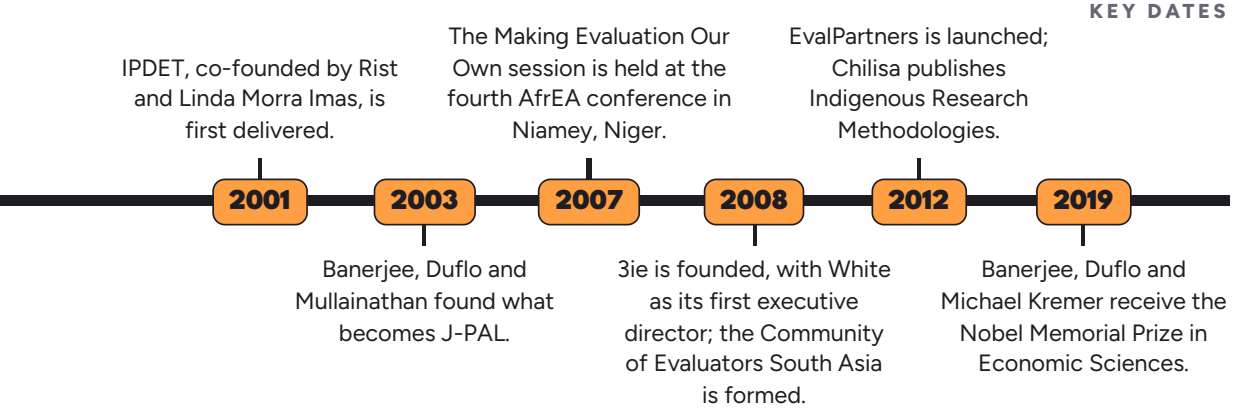
ACT 5 · 1990S TO 2020S

Evidence for a World in a Hurry

At the turn of the millennium, governments in developing countries faced growing pressure to show results, from the Millennium Development Goals adopted in 2000 to poverty reduction strategies and aid conditions. Most public monitoring still checked spending and activities, not outcomes. Ray Rist and Jody Zall Kusek answered in 2004 with Ten Steps to a Results-Based Monitoring and Evaluation System, which began by asking whether anyone in government actually demanded the information; IPDET, which Rist co-founded with Linda Morra Imas, had been training evaluators since 2001. In 2003 Abhijit Banerjee, Esther Duflo and Sendhil Mullainathan founded what became J-PAL, bringing randomised field experiments to questions such as remedial tutoring with Pratham in Vadodara and Mumbai. In 2008 3ie was founded, with Howard White as its first executive director, and in 2009 he argued that impact studies must trace the causal chain, not just report an average.

The rush to rigour provoked a second movement. In January 2007, at the African Evaluation Association conference in Niamey, Zenda Ofir and Sulley Gariba led the session Making Evaluation Our Own, prompted in part by their judgement that imported trial designs did not reflect African ways of knowing. Bagele Chilisa gave the idea theory, from her 2012 Indigenous Research Methodologies to the 2015 synthesis paper on Made in Africa Evaluation. Marco Segone helped launch EvalPartners in 2012 to strengthen national evaluation associations, and in December 2014 the UN General Assembly

KEY DATES



adopted a resolution on country-level evaluation capacity. In South Asia, Sonal Zaveri helped found the Community of Evaluators in 2008 and later led a real-time, gender-focused evaluation during the COVID-19 response.

In 2012 Fred Carden and Marvin Alkin observed that evaluation theory from low and middle income countries seemed to have collective, practice-based origins rather than individual authors. In our reading, this act moves the field's centre of gravity to where programmes happen, and asks that evidence be both rigorous and owned.

THE GREAT DEBATE OF THIS ACT

The Randomistas and Their Critics Are randomised trials the best route to knowing what works against poverty, or one tool among many? p. 554

THE GIANTS OF THIS ACT



Ray C. Rist

The World Bank evaluator who turned results-based M&E into ten steps a government could follow, and helped build a school to teach them.

p. 410



Esther Duflo and Abhijit Banerjee

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

p. 432



Zenda Ofir

The scientist turned evaluator who helped launch the call for evaluation made in Africa, by Africans.

p. 454



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476

Donald T. Campbell, in one line

He gave evaluators a way to be honest about cause when the world will not hold still for an experiment.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340

D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476

BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

18



Emblem: an original HubForge design for this figure, not a likeness.

Ray C. Rist

Active 2000s

United States and the World Bank; evaluation capacity work across low- and middle-income countries

Impact evaluation and evidence systems · Use and learning

The World Bank evaluator who turned results-based M&E into ten steps a government could follow, and helped build a school to teach them.

IN ONE LINE

Ray Rist taught governments to ask who wants the evidence before building the system, and then trained the people to run it.

Within days of the school year starting, a kindergarten teacher had divided her class into reading groups. In 1970 a young sociologist named Ray Rist published what happened next. He had followed one class of children in a low income urban school from kindergarten into second grade, and showed that the groups tracked the children's social class and barely changed. A seating plan made in the first days of school was shaping children's futures. By late 2024 that article had about 3,500 citations. The same careful observer would later direct a division of the United States General Accounting Office, co-write a World Bank handbook on results-based monitoring and, in 2001, co-found a training programme that by 2016 had taught more than 4,000 evaluators from over 80 countries.

Life CHAPTER 18

Ray C. Rist was born in 1944 in Carbondale, Illinois, according to the SAGE Encyclopedia of Evaluation. He studied psychology and sociology at Valparaiso University and took his MA and PhD at Washington University in St Louis. He began as a sociologist of education. His first research paper, published in the Harvard Educational Review in 1970, followed one class of children from low income urban neighbourhoods through kindergarten, first and second grade, and found that the kindergarten teacher placed children in reading groups that tracked their social class, that the groups stayed largely fixed and that differing treatment shaped achievement. By late 2024 it was his most cited article. A book, The Urban School: A Factory for Failure, followed from MIT Press in 1973.

His career then moved between universities and government. He taught at Portland State University and later at Cornell, Johns Hopkins and George

BIG IDEAS

1. **Ten steps to a results-based monitoring and evaluation system**
2. **Readiness assessment and political demand**
3. **Evaluation capacity development through training and networks**

KEY WORKS

Ten Steps to a Results-Based Monitoring and Evaluation System
2004 · With Jody Zall Kusek, World Bank

The Capacity to Evaluate: Why Countries Need It
2006 · With Linda G. Morra Imas

The Road to Results: Designing and Conducting Effective Development Evaluations
2009 · With Linda G. Morra Imas

Washington University, and was a Senior Fulbright Fellow at the Max Planck Institute in Germany in 1976 and 1977, studying German housing, education and employment policy and guestworkers. In 1981 he joined the United States General Accounting Office, the audit and evaluation arm of Congress, where from 1987 to 1993 he was Director of the General Government Division. He then returned to academia as professor and director of the Center for Policy Studies at George Washington University from 1993 to 1997.

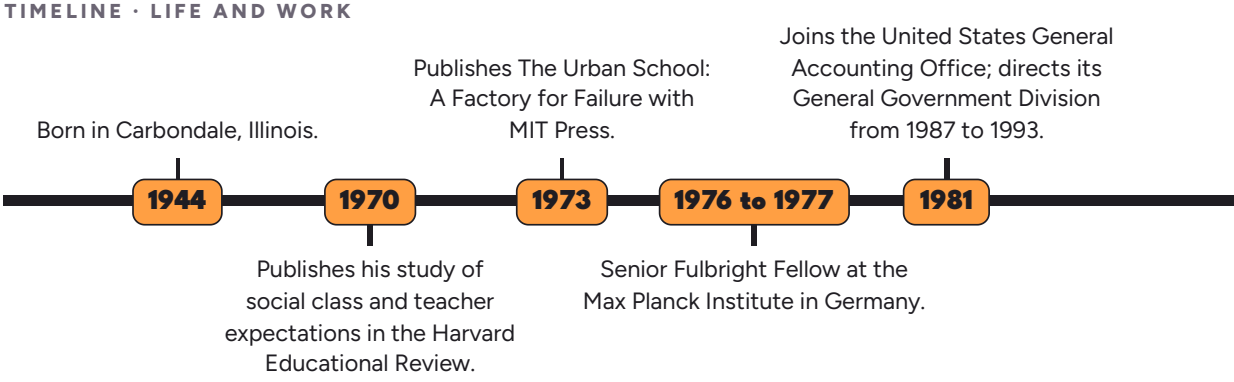
In 1997 Rist moved to the World Bank as an evaluation adviser, and from 2000 to 2006 he was a senior evaluation adviser in what became the Independent Evaluation Group. In 2001, with Linda Morra Imas, he founded the International Program for Development Evaluation Training (IPDET), which he co-directed until 2016. In 2004, with Jody Zall Kusek, he published Ten Steps to a Results-Based Monitoring and Evaluation System, and with Morra Imas he wrote The Road to Results (2009). From 2008 to 2014 he was president of the International Development Evaluation Association (IDEAS).

Running through all of this was a long commitment to comparative scholarship. For almost 40 years he chaired INTEVAL, the International Research Group for Policy and Program Evaluation, and was final editor of its Comparative Policy Evaluation series, which produced nearly 40 books. Among them were Carrots, Sticks and Sermons (1998), on policy instruments, co-edited with Marie-Louise Bemelmans-Videc and Evert Vedung; the International Atlas of Evaluation (2002), co-edited with Jan-Eric Furubo and Rolf Sandahl; and From Studies to Streams (2006). A 2025 special issue of the Journal of MultiDisciplinary Evaluation, devoted to his contributions, counted about 190 publications and nearly 13,000 citations.

The spark

When Rist arrived at the World Bank in 1997, a problem was waiting. Governments everywhere were being asked to show results. Yet their monitoring mostly tracked activities, not outcomes, and systems designed on paper kept collapsing because nobody in government actually asked for the information. A study by the World Bank and the African Development Bank had

TIMELINE · LIFE AND WORK



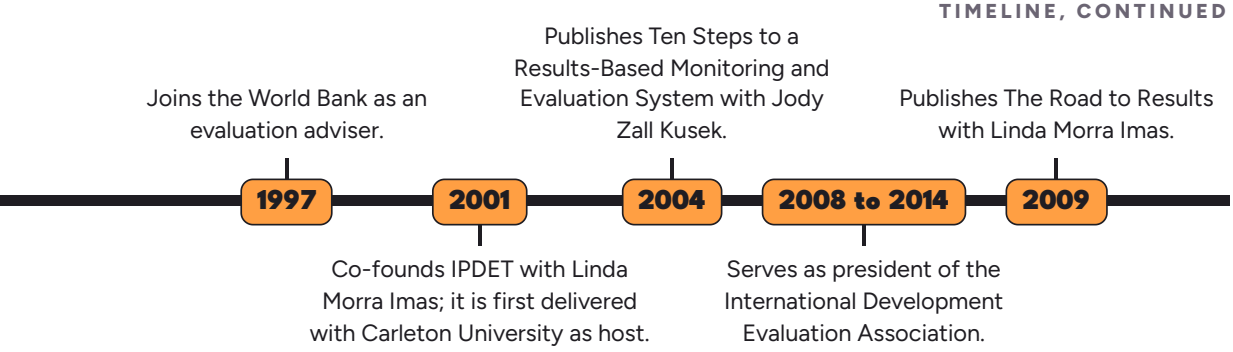
named weak demand as the key constraint. Rist arrived with two decades of experience inside government audit and evaluation, and knew how systems fail.

The breakthrough CHAPTER 18

Picture a ministry that spends years building a monitoring system: forms, indicators, software, training. A year later, nobody uses it. Ministers do not ask for the reports, managers do not read them and the data quietly stops arriving. In the early 2000s this was a common story. Governments faced growing pressure to show results, from the Millennium Development Goals adopted in 2000 to the conditions attached to debt relief and European Union membership. But most public monitoring checked whether money was spent and activities carried out, not whether lives changed. And systems kept failing because, as a study by the World Bank and the African Development Bank found in Sub-Saharan Africa, there was too little demand for the information.

Drawing on World Bank work with countries including Bangladesh, Egypt and Romania, and funded through a Dutch government trust fund, Kusek and Rist published Ten Steps to a Results-Based Monitoring and Evaluation System in 2004. Its central move was to put a readiness assessment first: before choosing a single indicator, ask whether there is demand, a political champion, clear roles and the capacity to sustain a system. Three years earlier, IPDET had begun to tackle the parallel problem of too few trained evaluators in the countries where programmes operate. Together they addressed the system and the people needed to run it.

TIMELINE, CONTINUED



The problem it solved

Governments and agencies tracked inputs and activities but could not show results, and M&E systems built on paper failed because nobody demanded or used the data.

Impact WHAT CHANGED

Ten Steps became one of Rist's two most cited publications, with about 1,300 citations by late 2024, and the handbook remains available free from the World Bank's open repositories, where it is widely used as a starting template for government and NGO monitoring systems. IPDET's numbers give a sense of reach: more than 4,000 participants from over 80 countries by 2016, and roughly 150 to 250 a year at Bern since 2018. Since 2021 IPDET has formed part of the Global Evaluation Initiative, the network the Independent Evaluation Group set up to manage its evaluation capacity development portfolio, which ties the school Rist co-founded to wider support for national evaluation systems.

In our interpretation, Rist's distinctive contribution was to connect three worlds that often stay apart: academic policy research through INTEVAL, government audit and evaluation through the GAO, and development practice through the World Bank. The idea that monitoring systems must begin with demand and champions now appears, in various forms, in national evaluation policies and in the work of evaluation societies across Africa, Asia and Latin America. For Indian practitioners building state-level dashboards or management information systems, the Ten Steps offers a structured check that the system will actually be used.

KEY WORKS TO READ IN THE ORIGINAL

- **Ten Steps to a Results-Based Monitoring and Evaluation System** 2004, With Jody Zall Kusek, World Bank
- **The Capacity to Evaluate: Why Countries Need It** 2006, With Linda G. Morra Imas
- **The Road to Results: Designing and Conducting Effective Development Evaluations** 2009, With Linda G. Morra Imas

WHERE PEOPLE DISAGREE

Debates

Results-based management has strong critics, although their target is the wider results agenda rather than Rist personally. Andrew Natsios, a former head of USAID, argued in a 2010 Center for Global Development paper that a counter-bureaucracy of compliance and measurement privileges what is easy to count, such as immunisations delivered, over institution building and policy reform that are harder to measure and slower to show results. Others warn that indicator systems can crowd out evaluation of why results occur. The Ten Steps itself includes a step on using evaluation information, and Rist's later work on streams of evidence engages with exactly this risk, so the tension is acknowledged inside his own body of work.

LIMITS, IN OUR READING

- ☐ Results-based approaches have been criticised for privileging what is easy to measure; Andrew Natsios (Center for Global Development, 2010) made this case about aid agencies' results agenda generally, not about Rist personally.
- ☐ Rist himself noted that monitoring often crowds out evaluation inside organisations, a practical limit on results-based systems.

WHY THIS MATTERS TO YOU

Rist's first lesson saves time and money in any organisation: before you build a system to collect information, find out who wants it and what they will do with it. If nobody will use it, do not build it yet; find a champion first. His Ten Steps give you a roadmap you can share with any government or NGO partner, and his career shows the value of moving between research, government and practice. His earliest work teaches something just as lasting. Watch closely, because small, early decisions, like where a child sits on the first day of school, can shape everything that follows.

Did you know?

The evaluator best known for a World Bank handbook on monitoring systems first made his name with a classroom observation study of kindergarten reading groups; by late 2024 that 1970 article had about 3,500 citations, more than any of his evaluation books.

About 190 publications and nearly 13,000 citations

Rist's body of work as counted in a 2025 special issue

Journal of MultiDisciplinary Evaluation, Linking Evaluation Theory and Practice: Exploring Ray Rist's Enduring Legacy, 2025

More than 4,000

Professionals trained by IPDET, which he co-founded in 2001, from over 80 countries by 2016

Capacity Development in Evaluation: The Role of IPDET, Journal of MultiDisciplinary Evaluation, 2025; IPDET website

Nearly 40 books

Volumes in INTEVAL's Comparative Policy Evaluation series, for which he was final editor over almost 40 years

Journal of MultiDisciplinary Evaluation, 2025 special issue

About 1,300 citations

Citations of Ten Steps to a Results-Based Monitoring and Evaluation System by late 2024

Journal of MultiDisciplinary Evaluation, 2025 special issue

Key concepts

Results-based monitoring and evaluation

A system that tracks outcomes and impacts, not only inputs, activities and outputs, to show whether policies and programmes achieve their goals.

Implementation-focused monitoring

Traditional monitoring that checks compliance: whether money was spent, activities completed and outputs delivered.

Readiness assessment

The first of the Ten Steps: a diagnosis of demand, incentives, roles, champions and capacity before a monitoring system is designed.

Baseline

The value of an indicator at or before the start of an intervention, against which later change is measured.

Evaluation capacity development

Strengthening the skills, institutions and demand that allow a country or organisation to commission, conduct and use evaluation.

Policy instruments

The tools governments use to influence behaviour, grouped in Carrots, Sticks and Sermons as economic incentives, regulation and information.

Sources for this chapter: p. 585

THE BIG IDEAS OF THIS CHAPTER

1. **Ten steps to a results-based monitoring and evaluation system** p. 418

   The Ten Steps gives governments and organisations a sequenced roadmap from political demand to sustained use, shifting monitoring from whether activities happened to whether results changed.

2. **Readiness assessment and political demand** p. 421

   Before designing any monitoring system, diagnose who wants the information, why, who will champion it and whether there is capacity to sustain it.

3. **Evaluation capacity development through training and networks** p. 424

   Countries need their own trained evaluators and professional networks, not only visiting experts, so building people and institutions is itself core evaluation work.

GIANTS CONNECTED TO THIS CHAPTER



Daniel L. Stufflebeam

An earlier model of evaluation built to serve management decisions.

p. 134



Marco Segone

Took national evaluation capacity into a global partnership and UN resolution.

p. 516



Howard White

Built institutions for impact evaluation and evidence synthesis that complement results monitoring.

p. 476

In one line

Ray Rist taught governments to ask who wants the evidence before building the system, and then trained the people to run it.

1

BIG IDEA 1 OF 3 · RIST

Ten steps to a results-based monitoring and evaluation system

The Ten Steps gives governments and organisations a sequenced roadmap from political demand to sustained use, shifting monitoring from whether activities happened to whether results changed.

PICTURE IT

Think of building a bridge. You do not pour concrete on the first day. You survey the ground, plan the spans, build step by step and then maintain it for years. The Ten Steps lay out a monitoring system the same way, from the first survey of demand to long-term use.

The problem

By 2000, developing country governments faced mounting pressure to show results, from the Millennium Development Goals, debt relief initiatives and Poverty Reduction Strategy Papers. Yet most public monitoring was implementation focused: it checked whether money was spent and activities and outputs delivered, not whether people's lives improved. Many systems were designed on paper and then collapsed because nobody demanded or used the information. Ministries were reporting busily while learning little about outcomes for citizens.

What it is

Ten Steps to a Results-Based Monitoring and Evaluation System (2004), by Jody Zall Kusek and Ray Rist, treats monitoring and evaluation as a public management tool. Its ten steps are: conducting a readiness assessment; agreeing on outcomes to monitor and evaluate; selecting key performance indicators to monitor outcomes; setting baselines and gathering data on indicators; planning for improvement by selecting results targets; monitoring for results; the role of evaluations; reporting findings; using findings; and sustaining the system within the organisation. The central distinction is between implementation-focused monitoring, which tracks inputs, activities and outputs, and results-based monitoring, which asks whether those outputs are leading to outcomes and impacts. The steps are sequenced so that political demand and agreed outcomes come before technical choices about indicators, and evaluation

is built in to explain why results are or are not being achieved. The handbook presents systems as continuous works in progress.

How it works

1. **Assess readiness**
   Diagnose demand, champions, roles and capacity before designing anything, the step the handbook treats as the foundation.
2. **Agree outcomes with stakeholders**
   Settle, through a participatory process with main stakeholders, the outcomes the system will track, before discussing indicators.
3. **Select key indicators**
   Choose performance indicators for each outcome, accepting that good indicators are built iteratively.
4. **Set baselines**
   Establish the starting value of each indicator, identifying data sources and piloting collection methods.
5. **Select targets**
   Set realistic interim targets from baseline levels and the desired improvement.
6. **Monitor for results**
   Track both implementation and results, with attention to data quality.
7. **Evaluate, report and use findings**
   Use evaluations to explain why results are or are not occurring, tailor reports to audiences and feed findings into decisions.
8. **Sustain the system**
   Maintain demand, clear roles, credible information, accountability, capacity and incentives over time.

WORKED EXAMPLE · AN ILLUSTRATION

A state school education department reports each year the number of teachers trained and workbooks distributed. Applying the Ten Steps, its new monitoring cell first runs a readiness assessment and finds that the education secretary wants outcome data for the state budget. Senior officials then agree an outcome: more Class 3 children reading with comprehension. Only after this do they choose indicators, such as the share of children reading a short paragraph on a simple oral assessment, and set a baseline from a sample survey of 300 schools. Targets are set for each of three years. Quarterly monitoring tracks both training delivered and reading levels, and a mid-term evaluation examines why some districts lag. The dashboard comes last, built around decisions officials have already said they will make.

USE IT WHEN

- Designing or overhauling a government or organisation-wide monitoring system.
- A dashboard or MIS project risks starting with indicators before outcomes and users are agreed.
- Reviewing why an existing system produces reports nobody uses.

SKIP IT WHEN

- The task is a single, time-bound evaluation rather than an ongoing system.
- The programme is highly emergent, so fixed outcomes and targets would mislead; adaptive approaches may suit better.

COMMON MISREADINGS

- Treating the steps as a strictly linear checklist, when the handbook presents systems as continuous works in progress.
- Skipping step one and jumping to indicators, which is the failure the handbook was written to prevent.

WHY IT WAS GROUNDBREAKING

It gave governments a single, sequenced roadmap from political demand to sustained use, written for practitioners rather than researchers. Its free availability let it become a common template far beyond World Bank projects.

Origin

Kusek and Rist published the handbook through the World Bank in 2004, supported by a Dutch government trust fund, drawing on readiness assessments in Bangladesh, Egypt and Romania and experience in OECD countries. Rist and Linda Morra Imas extended the approach to designing and conducting evaluations in The Road to Results (2009). The handbook remains freely available from the World Bank.

RELATED IDEAS

Readiness assessment and political demand

- DS Stufflebeam 134
- MS Segone 516
- HW White 476

CONCEPT HIGHLIGHT

Results-based monitoring and evaluation

A system that tracks outcomes and impacts, not only inputs, activities and outputs, to show whether policies and programmes achieve their goals.

2

BIG IDEA 2 OF 3 · RIST

Readiness assessment and political demand

Before designing any monitoring system, diagnose who wants the information, why, who will champion it and whether there is capacity to sustain it.

PICTURE IT

Think of opening a restaurant. Before buying the ovens, you check whether anyone in the neighbourhood wants to eat there. A readiness assessment asks whether anyone in government truly wants the information before a single indicator is chosen.

The problem

Technically sound monitoring systems repeatedly failed in practice. Consultants designed indicators and software, but no one in government needed the data for decisions, managers had no incentive to report bad news, and the system died when the project ended. The Ten Steps handbook cites a World Bank and African Development Bank study identifying lack of demand as the key constraint on monitoring and evaluation capacity in Sub-Saharan Africa. The missing ingredient was not technique but politics.

What it is

Step one of the Ten Steps is a readiness assessment: a structured diagnosis of whether a country or organisation can build and use a results system. The handbook groups it into three parts: the incentives and demands for building the system; the roles, responsibilities and existing structures for assessing performance; and the capacity-building requirements. It poses eight key questions: what pressures are driving the need, and why; who will advocate for the system; what motivates that champion; who will own the system, who will benefit and how much information they really want; how the system will support better resource allocation and the achievement of goals; how the organisation will respond to negative information; where existing capacity lies; and how the system will link project, programme, sector and national goals. The handbook gives champions such as Egypt's Minister of Finance and Zambia's Secretary to the Cabinet as examples, and notes that Bangladesh at the time lacked one.

How it works

1. **Identify the pressures**
   Ask what is driving the call for a results system now, such as budget reform, a donor requirement or a political commitment.
2. **Find the champion and the motive**
   Identify who will advocate for the system and why, since the handbook treats champions as essential for setting up and sustaining systems.
3. **Map owners and users**
   Establish who will own the system, who will benefit and how much information they actually want.
4. **Test the link to decisions**
   Ask how the information would improve resource allocation and progress towards goals.
5. **Probe the response to bad news**
   Ask how leaders and staff will react to negative findings, a strong signal of whether the system will be used.
6. **Map existing capacity and data**
   Find out where skills, systems and reliable data already exist.
7. **Check links across levels**
   See how project, programme, sector and national goals connect, so the system fits the wider architecture.
8. **Decide what to build first**
   If demand is weak, make building demand and a champion the first task, rather than indicators or software.

WORKED EXAMPLE · AN ILLUSTRATION

A district administration in Madhya Pradesh asks a consultant to design a monitoring dashboard for nutrition schemes. Before choosing indicators, the consultant runs a readiness assessment over two weeks. Interviews show that the district collector wants outcome data for monthly review meetings, which gives a champion with a clear motive. Block officials, however, fear that showing high stunting will bring censure, and anganwadi supervisors already enter data in three separate systems. Capacity to analyse data exists only in one office. The consultant's readiness note recommends agreeing with the collector how negative findings will be treated, merging data entry with an existing system, and training two block staff, all before any dashboard is built. The district agrees, and the dashboard design starts with the review meeting's actual decisions.

USE IT WHEN

- Before any new MIS, dashboard or results framework is designed.
- When inheriting a system that produces data no one seems to use.
- When a donor requirement, rather than internal demand, is driving the request.

SKIP IT WHEN

- A small team is setting up simple project monitoring for its own use, where a brief conversation can replace a formal assessment.
- Demand and ownership are already proven and the task is a technical upgrade of a well used system.

COMMON MISREADINGS

- Treating readiness as a yes or no gate, when its purpose is to show what must be built first.
- Running it as a technical capacity audit while ignoring the political questions of demand and response to bad news.

WHY IT WAS GROUNDBREAKING

It shifted attention from data design to institutional politics, explaining why so many technically sound systems failed. It made demand, incentives and champions a formal stage of system design rather than an afterthought.

Origin

The readiness assessment is step one of Kusek and Rist's Ten Steps (World Bank, 2004), which presents assessments from Bangladesh, Egypt and Romania and a capacity assessment survey in its annexes. It built on World Bank evaluation capacity development work from the late 1990s, and Rist and Morra Imas carried the demand argument into The Capacity to Evaluate (2006).

RELATED IDEAS

Ten steps to a results-based monitoring and evaluation system

MP Patton 204 CW Weiss 152 MS Segone 516

CONCEPT HIGHLIGHT

Implementation-focused monitoring

Traditional monitoring that checks compliance: whether money was spent, activities completed and outputs delivered.

3

BIG IDEA 3 OF 3 · RIST

Evaluation capacity development through training and networks

Countries need their own trained evaluators and professional networks, not only visiting experts, so building people and institutions is itself core evaluation work.

PICTURE IT

Think of a country that relies entirely on foreign doctors. When they leave, the hospitals empty. Training local doctors and building medical schools changes that for good. Rist argued that evaluation needs the same: trained evaluators and strong networks in every country.

The problem

In the 1990s most development evaluations were designed and led by consultants from donor countries. Ministries and local organisations in the countries where programmes ran had few trained evaluators and few places to learn. Findings arrived from outside, were often not owned and could not be sustained once a project ended. Without national capacity to commission, conduct and use evaluation, even the best results frameworks depended on external help indefinitely.

What it is

Evaluation capacity development means strengthening the skills, institutions and demand that let a country or organisation commission, conduct and use evaluation itself. Rist pursued it through two linked institutions. The first was the International Program for Development Evaluation Training (IPDET), which he founded in 2001 with Linda Morra Imas under the World Bank's evaluation department, with Carleton University in Ottawa as host. It combined an intensive core course on evaluation fundamentals with a mixed cohort of participants from governments, NGOs and donors. The second was the International Development Evaluation Association (IDEAS), which he led as president from 2008 to 2014, a professional network for development evaluators. The logic is that trained people, a shared language and peer networks are the infrastructure on which national evaluation systems rest. A 2025 tribute in the Journal of MultiDisciplinary Evaluation credits him with championing evaluators from the Global South.

How it works

1. **Identify who must evaluate**
   Map the people in ministries, NGOs and institutions who need to commission, conduct or use evaluations.
2. **Offer intensive foundation training**
   Provide a structured core course in evaluation fundamentals, as IPDET has run annually since 2001.
3. **Mix participants across sectors and countries**
   Bring together government, civil society and donor staff from many countries so that they learn from each other.
4. **Connect trainees to networks**
   Link graduates to alumni networks and professional bodies such as IDEAS for continued learning.
5. **Apply learning to institutions**
   Encourage participants to use their training to design evaluation policies and systems in their own organisations.
6. **Sustain through partnerships**
   Anchor training in durable institutions, as IPDET moved from Carleton University to the University of Bern and CEval and later joined the Global Evaluation Initiative.

WORKED EXAMPLE · AN ILLUSTRATION

A state planning department in India sets up an evaluation directorate and wants it to commission and manage evaluations rather than depend on external consultants. It sends four officers to an intensive international training programme of the kind IPDET offers, and pairs them on their return with an experienced evaluator for their first two commissions. The officers draft a state evaluation policy, set quality standards for terms of reference and join a national evaluation network to share practice with peers in other states. Within three years the directorate manages its own evaluation calendar and reviews consultants' reports critically. The training was the start; the capacity came from linking trained people to real tasks, institutions and a professional community.

USE IT WHEN

- A government or organisation is establishing an evaluation function.
- Evaluations are repeatedly commissioned but poorly used or poorly managed.
- Building a national or regional community of practice for evaluators.

SKIP IT WHEN

- The need is a single urgent evaluation, where training cannot substitute for experienced evaluators.
- Trainees will have no role, mandate or resources to apply what they learn.

COMMON MISREADINGS

- [ ] Equating capacity development with running workshops, without changing roles, institutions or demand.
- [ ] Assuming capacity can be built only by international programmes, when national networks and mentoring are equally part of the idea.

WHY IT WAS GROUNDBREAKING

It treated trained people and professional networks as the infrastructure of evaluation, not a side activity. It helped make evaluation capacity a recognised goal of development cooperation and gave practitioners from the Global South a sustained route into the field.

Origin

IPDET was founded in 2001 by Rist and Linda Morra Imas, first hosted by Carleton University; by 2016 it had trained more than 4,000 participants from over 80 countries. After 2017 the University of Bern and CEval took over delivery, and since 2021 it has been part of the Global Evaluation Initiative. Rist and Morra Imas set out the case in The Capacity to Evaluate (2006).

RELATED IDEAS

Readiness assessment and political demand

MS Segone 516 · HP Preskill 226 · HW White 476

RELATED GIANTS



Daniel L. Stufflebeam

The former school band director who gave evaluation four questions decision-makers could use at every stage: context, input, process and product.

p. 134



Marco Segone

The UN evaluator who helped turn scattered evaluation associations into a global movement for national evaluation capacity and equity.

p. 516



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204

CONCEPT HIGHLIGHT

Readiness assessment

The first of the Ten Steps: a diagnosis of demand, incentives, roles, champions and capacity before a monitoring system is designed.

CONCEPT HIGHLIGHT

Baseline

The value of an indicator at or before the start of an intervention, against which later change is measured.

CONCEPT HIGHLIGHT

Evaluation capacity development

Strengthening the skills, institutions and demand that allow a country or organisation to commission, conduct and use evaluation.

CONCEPT HIGHLIGHT

Policy instruments

The tools governments use to influence behaviour, grouped in Carrots, Sticks and Sermons as economic incentives, regulation and information.

Did you know?

He chaired the INTEVAL research group for almost 40 years; after its original funding ended, members kept the comparative book series going on a voluntary basis.

Put it to work

USE IT ON MONDAY · TEN STEPS TO A RESULTS-BASED M&E SYSTEM, STARTING WITH READINESS

1. Before designing any dashboard or MIS, run a short readiness assessment: who is asking for results information, who will use it, what rewards or penalties follow, and who has the skills to run it.

2. Agree the outcomes with senior decision-makers before you discuss indicators. Indicators chosen first tend to measure what is easy rather than what matters.

3. Protect space for evaluation inside the system: schedule at least one question each year about why results are or are not happening, not only what the numbers show.

AI prompt **Copy, fill the brackets, then check every line**

Help me prepare a readiness assessment for a results-based M&E system, following the first of the Ten Steps by Jody Zall Kusek and Ray Rist, for **[government department or organisation]** in **[country or state]**. Here is what I know about the organisation's mandate, current data systems and stakeholders: **[paste]**. (1) Draft interview questions for senior leaders, managers and data staff covering demand, use, incentives, roles and capacity. (2) List the documents I should request. (3) Suggest warning signs that the system may exist only on paper. (4) Outline a short readiness report structure.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of building a bridge. You do not pour concrete on the first day. You survey the ground, plan the spans, build step by step and then maintain it for years. The Ten Steps lay out a monitoring system the same way, from the first survey of demand to long-term use.

YOUR MOVES

Three moves from Rist's path

1 Observe closely, then scale up

Rist began with a classroom study of children's reading groups and went on to shape national monitoring systems.

TRY THIS

Before designing a system, observe one site for a full day and let what you see shape the indicators.

2 Check readiness before indicators

His ten-step handbook with Jody Kusek starts with a readiness assessment, checking demand and incentives first.

TRY THIS

Before your next M&E system design, ask who wants the data and what they will do with it.

3 Invest in other evaluators

He co-founded IPDET and led IDEAS, multiplying his influence well beyond his own publications.

TRY THIS

Teach one session a year for new evaluators, at a network, university or within your organisation.

THE BIG IDEAS OF THIS CHAPTER

1 Ten steps to a results-based monitoring and evaluation system p. 418

The Ten Steps gives governments and organisations a sequenced roadmap from political demand to sustained use, shifting monitoring from whether activities happened to whether results changed.

2 Readiness assessment and political demand p. 421

Before designing any monitoring system, diagnose who wants the information, why, who will champion it and whether there is capacity to sustain it.

3 Evaluation capacity development through training and networks p. 424

Countries need their own trained evaluators and professional networks, not only visiting experts, so building people and institutions is itself core evaluation work.

CONCEPT HIGHLIGHT

Policy instruments

The tools governments use to influence behaviour, grouped in Carrots, Sticks and Sermons as economic incentives, regulation and information.

GIANTS CONNECTED TO THIS CHAPTER



Daniel L. Stufflebeam

An earlier model of evaluation built to serve management decisions.

p. 134



Marco Segone

Took national evaluation capacity into a global partnership and UN resolution.

p. 516



Howard White

Built institutions for impact evaluation and evidence synthesis that complement results monitoring.

p. 476

KEY WORKS TO READ IN THE ORIGINAL

- **Ten Steps to a Results-Based Monitoring and Evaluation System** 2004, With Jody Zall Kusek, World Bank
- **The Capacity to Evaluate: Why Countries Need It** 2006, With Linda G. Morra Imas
- **The Road to Results: Designing and Conducting Effective Development Evaluations** 2009, With Linda G. Morra Imas

CONCEPT HIGHLIGHT

Results-based monitoring and evaluation

A system that tracks outcomes and impacts, not only inputs, activities and outputs, to show whether policies and programmes achieve their goals.

CONCEPT HIGHLIGHT

Implementation-focused monitoring

Traditional monitoring that checks compliance: whether money was spent, activities completed and outputs delivered.

WORKSHEET · TRY THIS TEMPLATE

M&E readiness check

Print this page, or copy the headings into your notes.

1 Demand

Who is asking for results information, and why now?

2 Use

Which decisions it will feed, and who makes them.

3 Incentives

What happens to a manager whose numbers are good or bad?

4 Capacity

People, skills, systems and budget available to run the system.

5 Existing data and roles

What is already collected, by whom, and how reliable it is.

6 Champion and risks

Who will sustain the system, and what could stop it.

REFLECT

Who in your client organisation would be worse off if the M&E system stopped working tomorrow?

CONNECTS TO

DS **Daniel L. Stufflebeam** — An earlier model of evaluation built to serve management decisions. p. 134

MS **Marco Segone** — Took national evaluation capacity into a global partnership and UN resolution. p. 516

HW **Howard White** — Built institutions for impact evaluation and evidence synthesis that complement results monitoring. p. 476

19



Portrait adapted from a photograph; credit on p. 590.

Esther Duflo and Abhijit Banerjee

Banerjee born 1961; Duflo born 1972 Active 2000s to 2010s

India, France and the United States; research in India, Kenya and other countries

Impact evaluation and evidence systems **Methods, rigour and causality**

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

IN ONE LINE

Esther Duflo and Abhijit Banerjee showed that the biggest questions about poverty can be answered one careful, testable question at a time.

Udaipur District, Rajasthan. In villages offered regular immunisation camps, plus a kilogram of lentils for each vaccination, 38.3 per cent of young children ended up fully immunised. In comparison villages, the figure was 6.2 per cent. A small, practical change had multiplied the result about six times, and the only reason anyone could be sure of it was the way the study was designed. On 14 October 2019 the Royal Swedish Academy of Sciences awarded the economics prize to three researchers for exactly this kind of work: an experimental approach to alleviating global poverty. Two of them, Abhijit Banerjee, born in Mumbai, and Esther Duflo, born in Paris, had spent years testing ideas in Indian villages and classrooms. The Nobel Committee stated that more than five million Indian children had benefited from remedial tutoring programmes as a direct result of one of the laureates' studies.

Life CHAPTER 19

Abhijit Banerjee was born in Mumbai on 21 February 1961 into a family of economists. His Nobel autobiography describes a father who left college in Kolkata, worked as a clerk and then as a manual labourer in London while taking night classes, won a scholarship to the London School of Economics, and later headed the economics department at Presidency College, Kolkata; his mother had also studied at the LSE. Banerjee briefly studied at the Indian Statistical Institute before taking economics at Presidency College and a master's degree at Jawaharlal Nehru University in Delhi. He then completed a PhD at Harvard, taught at Princeton from about 1987, returned to Harvard and moved to MIT in 1993.

Esther Duflo was born in Paris on 25 October 1972, the daughter of a paediatrician and a mathematics

BIG IDEAS

1. **Randomised evaluation of specific programme designs**
2. **Teaching at the Right Level: evidence that builds over many trials**
3. **Economists as plumbers: partnership and implementation detail**

KEY WORKS

Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty
2011 · Banerjee and Duflo

Handbook of Field Experiments, Volumes 1 and 2
2017 · Edited by Duflo and Banerjee

Good Economics for Hard Times
2019 · Banerjee and Duflo

professor, and grew up in the suburb of Asnières. She entered the École Normale Supérieure, majoring in history with economics as a second subject, and wrote her history master's dissertation in Moscow, where in 1993 and 1994 she also worked as a teaching assistant and on advisory work on Russia's economic reforms. Encouraged by Thomas Piketty, she applied to MIT, began her PhD there in 1995, took a development course taught by Banerjee and Michael Kremer, and was supervised by Banerjee and Joshua Angrist. She completed her PhD in 1999 and joined the MIT faculty the same year.

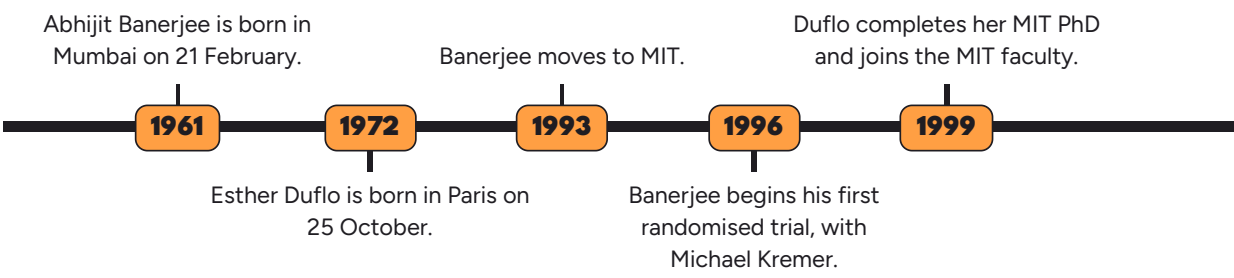
In 2003 Banerjee, Duflo and Sendhil Mullainathan founded the Poverty Action Lab at MIT, pooling funds MIT had offered to keep them there. Rachel Glennerster became its early leader, and an endowment from Mohammed Jameel led to the name Abdul Latif Jameel Poverty Action Lab, or J-PAL. A South Asia office opened in 2007 at the Institute for Financial Management and Research in Chennai. The two economists, who later married, wrote Poor Economics (2011) and Good Economics for Hard Times (2019), and edited the two volume Handbook of Field Experiments (2017).

On 14 October 2019 the Royal Swedish Academy of Sciences awarded Banerjee, Duflo and Kremer the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for their experimental approach to alleviating global poverty. At 46, Duflo was reported to be the youngest winner of the economics prize and the second woman to receive it. Both gave their Nobel lectures in Stockholm on 8 December 2019. In October 2025 the University of Zurich announced that from July 2026 they would co-lead its new Lemann Center for Development, Education and Public Policy, while keeping part-time positions at MIT and continuing to co-lead J-PAL.

The spark

Their paths turned at specific moments. Banerjee was trained as an economic theorist. According to his Nobel autobiography, his turn to experiments came after Michael Kremer described a randomised trial he was running around 1995, and the two began a trial together in 1996. Duflo studied history and economics in Paris and worked on Russia's economic reforms in Moscow in 1993 and 1994, before taking the

TIMELINE · LIFE AND WORK



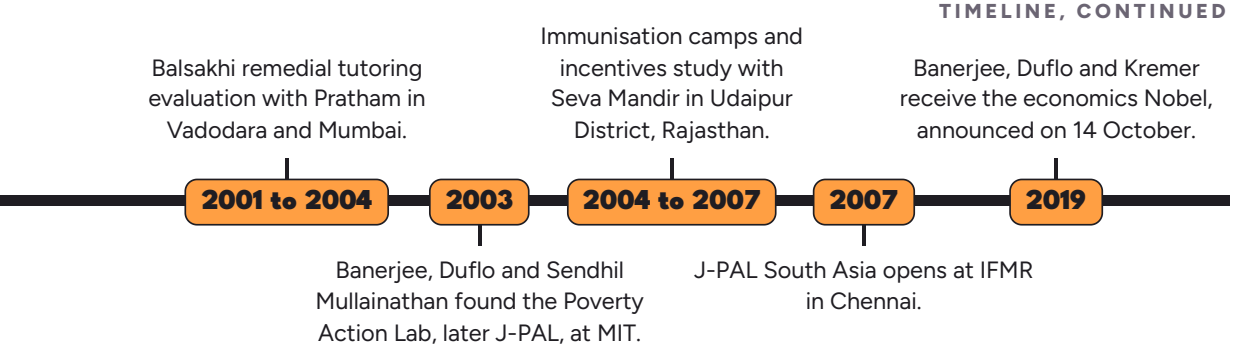
development course taught by Banerjee and Kremer at MIT. Both were drawn to a question the field had not answered well: what actually works, for whom and why?

The breakthrough CHAPTER 19

For decades, debates about poverty were conducted at a very high level. Does foreign aid cause growth? Economists argued over data from whole countries, while governments and NGOs scaled up programmes on the strength of plausible stories and simple before-and-after comparisons. Imagine being a minister choosing between two ways to help children learn to read. You could rarely know which worked better, or whether either worked at all. The Nobel Committee's 2019 announcement described the laureates' answer: break the giant question of global poverty into smaller, sharper ones, such as which approach best improves children's learning or health, and answer each through carefully designed experiments among the people most affected.

The approach grew from field experiments that Kremer and colleagues ran in western Kenya from the mid-1990s on improving school results. Banerjee and Duflo, often with Kremer, carried it into other issues and countries, and India was central. Between 2001 and 2004 they and colleagues worked with the NGO Pratham in Vadodara and Mumbai to randomise which schools received a balsakhi, a young woman from the community paid to tutor children who were falling behind. From 2004 to 2007, with the NGO Seva Mandir in Udaipur District, Rajasthan, they tested ways to raise child immunisation. The founding of J-PAL in 2003 turned these projects into a sustained programme of research and policy partnership.

TIMELINE, CONTINUED



The problem it solved

Programmes were scaled or cut without credible evidence of what actually changed outcomes, and broad debates about whether aid works gave policymakers little practical guidance.

Impact WHAT CHANGED

The Nobel Committee stated that, as a direct result of one of the laureates' studies, more than five million Indian children had benefited from programmes of remedial tutoring in schools, and that heavy subsidies for preventive health care had been introduced in many countries. Duflo's lecture reported that Haryana-style partnerships for Teaching at the Right Level had reached 21.3 million children across India, with versions spreading to Kenya and Ghana. J-PAL reports more than 2,700 randomised evaluations led by a network of over 1,100 researchers, and more than 850 million people reached by programmes informed by its affiliates' evaluations. In India it began a long-term partnership with India's National Academy of Administration in 2011 to train senior Indian Administrative Service officers in evidence use, and Duflo cited Tamil Nadu's memorandum of understanding with J-PAL as an example of a government building its own culture of experimentation.

In the wider field, the Nobel press release observed that experimental methods now dominate development economics. In our interpretation, their larger effect on monitoring and evaluation practice has been to make the counterfactual question routine: commissioners now ask whether a credible comparison exists, and governments increasingly host their own evidence units. This influence is uneven, and many programmes still cannot be randomised, but the expectation of rigour has shifted.

WHERE PEOPLE DISAGREE

Debates

The approach has drawn serious criticism. Angus Deaton, writing in the Journal of Economic Literature in 2010, argued that randomisation has no special epistemic status, that trials estimate average effects that may hide harm to some, that results may not transfer to other settings or to scale, and that knowledge accumulates only when studies reveal mechanisms. A 2012 DFID working paper led by Elliot Stern concluded that experimental designs suit only a minority of programmes. Banerjee's Nobel lecture replied that replication across sites, combined with theory, can generalise, and Duflo stressed evidence built from many studies. The debate continues.

LIMITS, IN OUR READING

- ☐ Angus Deaton (Journal of Economic Literature, 2010) argued that randomisation has no special status over other evidence, that results from one trial may not transfer, and that understanding mechanisms matters more than project verdicts.
- ☐ Stern and colleagues (DFID Working Paper 38, 2012) found that experimental designs suit only a small share of programmes and argued for theory-based, case-based and participatory designs for complex interventions.

WHY THIS MATTERS TO YOU

Banerjee and Duflo changed the question anyone can ask about an idea. Not whether it works, in general and forever, but which version works better, for whom and at what cost. Even if you never run a trial, their thinking sharpens everyday decisions: look for a fair comparison before you believe a success story; pay attention to the small, practical details of how something is delivered; and treat scaling up as a chance to keep learning. In India you are surrounded by their partners, from Pratham to state evaluation units. Their path shows that careful evidence and practical partnership can work together, and that a well framed question can reach millions.

Did you know?

According to Banerjee's Nobel autobiography, he initially turned down Duflo's application to MIT's PhD programme, and Thomas Piketty helped rescue it.

More than 5 million

Indian children who benefited from remedial tutoring programmes as a direct result of one of the laureates' studies

Nobel Prize press release, The Prize in Economic Sciences 2019

21.3 million

Children reached across India by Haryana-style Teaching at the Right Level partnerships

Duflo, Field Experiments and the Practice of Policy, Nobel lecture, 2019

More than 2,700

Randomised evaluations led by J-PAL's network of over 1,100 affiliated researchers

J-PAL, About J-PAL

More than 850 million

People reached by programmes informed by J-PAL affiliates' evaluations

J-PAL, About J-PAL

Key concepts

Randomised controlled trial (RCT)

An evaluation in which units are assigned to treatment or comparison groups by chance, so that differences in outcomes can be attributed to the intervention.

Counterfactual

What would have happened to participants without the programme; in an RCT the comparison group provides an estimate of it.

Cluster randomisation

Randomly assigning whole groups, such as villages or schools, rather than individuals, as in the 134 village Udaipur immunisation study.

Standard deviation effect size

A way of expressing an impact relative to the spread of outcomes, allowing comparison across studies; the Balsakhi programme raised scores by about 0.28 standard deviations in its second year.

External validity

The extent to which a result found in one setting holds in other places, populations or at larger scale.

Teaching at the Right Level

An approach developed with Pratham in which children are assessed, grouped by learning level rather than grade and taught at that level.

Sources for this chapter: p. 585

THE BIG IDEAS OF THIS CHAPTER

1. **Randomised evaluation of specific programme designs** p. 440

   By using a lottery to decide who receives a programme, or which version of it, evaluators create comparable groups whose later differences can be credited to the programme.

2. **Teaching at the Right Level: evidence that builds over many trials** p. 443

   Good policy evidence is built from a sequence of trials and process studies that refine an idea until it works inside real systems, not from one study and an automatic scale-up.

3. **Economists as plumbers: partnership and implementation detail** p. 446

   Much useful evaluation is plumbing: working with implementers over years to find and test fixes in how an existing programme actually runs.

GIANTS CONNECTED TO THIS CHAPTER



Donald T. Campbell

Treating reforms as experiments, and the threats to validity that randomisation is designed to remove.

p. 112



Howard White

Joined counterfactual designs to theory of change and built bodies for evidence synthesis.

p. 476



Sonal Zaveri

Argues that participation should feature in all designs, including randomised trials.

p. 536

KEY WORKS TO READ IN THE ORIGINAL

- **Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty** 2011, Banerjee and Duflo
- **Handbook of Field Experiments, Volumes 1 and 2** 2017, Edited by Duflo and Banerjee
- **Good Economics for Hard Times** 2019, Banerjee and Duflo

1

BIG IDEA 1 OF 3 · DUFLO AND BANERJEE

Randomised evaluation of specific programme designs

By using a lottery to decide who receives a programme, or which version of it, evaluators create comparable groups whose later differences can be credited to the programme.

PICTURE IT

Think of a fair coin toss deciding which half of a class tries a new teaching method first. Because chance, not choice, divided the groups, any later difference between them can be credited to the method. A randomised evaluation is that coin toss done carefully, at scale.

The problem

For decades, debates about poverty were conducted at the level of grand theory and cross-country data, and programmes were scaled up on plausible stories and before-and-after comparisons. People who chose to join a programme often differed from those who did not, so simple comparisons confused the effect of the programme with the effect of who took part. Governments and NGOs could rarely say with confidence whether an intervention caused a change, or which of two designs worked better.

What it is

A randomised evaluation assigns people, schools or villages to receive a programme, or a particular version of it, by chance. Because chance rather than choice decides membership, the groups are alike on average at the start in both observed and unobserved ways, and later differences in outcomes estimate the programme's effect against a credible counterfactual. The approach Banerjee and Duflo championed has several distinctive parts: breaking a large question into a precise one, often a choice between designs; working with an implementing partner who can act on the answer; choosing the level of randomisation, individual or cluster, with attention to spillovers; calculating sample size and statistical power; and analysing carefully when reality departs from the plan, for example when some people assigned to a programme do not take it up. The 2006 toolkit by Duflo, Rachel Glennerster and Michael Kremer also stresses using theory to design trials and interpret results.

How it works

1. **Sharpen the question with a partner**
   Turn a broad question into a specific, testable design choice the implementer genuinely wants answered, as the Nobel Committee described the laureates doing.
2. **Choose the unit of randomisation**
   Decide whether to randomise individuals, households, schools or villages, taking account of spillovers and what is practical for the partner, as the 2006 toolkit discusses.
3. **Calculate power and stratify**
   Work out with a statistician how many units are needed to detect a meaningful effect, and stratify the lottery by key characteristics.
4. **Randomise transparently**
   Run the assignment by public lottery or computer draw, sometimes as a phased roll-out so that everyone eventually receives the programme.
5. **Measure the same way in every group**
   Collect baseline and outcome data with identical instruments and timing in treatment and comparison groups.
6. **Analyse honestly**
   Estimate effects by original assignment, and handle imperfect compliance and spillovers as the toolkit explains.
7. **Interpret with theory and process evidence**
   Use implementation data and theory to explain why the result occurred and whether it may hold elsewhere.

WORKED EXAMPLE · AN ILLUSTRATION

A maternal health NGO in Bihar wants to know whether reminder calls in local dialect increase the number of antenatal check-ups. Rather than asking whether its whole programme works, it frames a specific choice: calls or no calls. With 2,000 registered pregnant women across 100 villages, the team randomises by village, because health workers might otherwise remind neighbours in the comparison group.

A statistician confirms the sample is large enough to detect a change the NGO would regard as worthwhile, and the lottery is stratified by distance from the health centre. Both groups' check-ups are recorded from the same clinic registers. Process data show that calls reached only some women, because phones are often held by husbands, a finding that shapes the next design: calls timed for when men are at home.

In one line

By using a lottery to decide who receives a programme, or which version of it, evaluators create comparable groups whose later differences can be credited to the programme.

USE IT WHEN

- A partner faces a genuine choice between designs and can deliver more than one version.
- Demand exceeds supply, or roll-out is phased, so a lottery is a fair way to allocate places.
- The decision is large enough to justify the cost and time of a trial.

SKIP IT WHEN

- The programme is a single national policy or a complex, adaptive intervention that cannot be assigned by chance.
- Withholding or delaying the programme would be unethical, or numbers are too small for adequate statistical power.

COMMON MISREADINGS

- [ ] Treating an RCT as a verdict on a whole sector, when it answers a specific question in a specific setting.
- [ ] Believing that randomisation alone explains why something works, without process data or theory.

WHY IT WAS GROUNDBREAKING

It brought the logic of clinical trials to everyday questions of programme design in low income settings, with implementers as partners. It showed that small and cheap design details, such as a kilogram of lentils at an immunisation camp, could change outcomes substantially and could be tested credibly.

Origin

Randomised field experiments in development grew from Michael Kremer's school studies in western Kenya from the mid-1990s; Banerjee began his first trial with Kremer in 1996. Real sourced example: the Udaipur immunisation study with Seva Mandir (2004 to 2007) randomised 134 villages, reporting full immunisation of 38.3 per cent with incentives, 16.6 per cent with camps alone and 6.2 per cent in comparison villages (J-PAL). The 2006 Duflo, Glennerster and Kremer toolkit codified practice.

RELATED IDEAS

Teaching at the Right Level · DC Campbell 112 · HW White 476 · SZ Zaveri 536

CONCEPT HIGHLIGHT

Randomised controlled trial (RCT)

An evaluation in which units are assigned to treatment or comparison groups by chance, so that differences in outcomes can be attributed to the intervention.

2

BIG IDEA 2 OF 3 · DUFLO AND BANERJEE

Teaching at the Right Level: evidence that builds over many trials

Good policy evidence is built from a sequence of trials and process studies that refine an idea until it works inside real systems, not from one study and an automatic scale-up.

PICTURE IT

Think of how a great recipe is perfected. The first version is promising, then the cook adjusts the spice, the timing and the heat, testing each change, until it works every time in an ordinary kitchen. Teaching at the Right Level was built the same way, through a sequence of trials.

The problem

A common caricature of experimental evaluation is one small trial, one policy brief and a leap to national scale. In practice, interventions that work when run by a committed NGO often fail when handed to government. Indian classrooms illustrated the stakes: many children in later primary grades could not read simple text, yet teachers were expected to teach the grade curriculum. A programme that helped them had to work not only in a pilot but in thousands of government schools.

What it is

Teaching at the Right Level, developed with the NGO Pratham, assesses children with a simple tool, groups them by actual learning level rather than by grade, and teaches with activities matched to that level for a period each day. In her 2019 Nobel lecture Duflo used it as her main example of how evidence accumulates, comparing it to a pointillist painting built from many dots. The sequence had distinct stages. The Balsakhi study (2001 to 2004) found gains of about 0.28 standard deviations by the second year, concentrated among the weakest children. Trials in Bihar and Uttarakhand found government teachers did not adopt the approach despite training. Analysing why, the researchers redesigned it and tested two models: in Haryana, teachers delivered it in a dedicated daily hour with support from trained resource persons; in Uttar Pradesh, Pratham volunteers ran intensive learning camps. The idea is that each trial answers one question and generates the next.

How it works

1. **Find a promising idea with a partner**
   Start from an implementer's innovation, as with Pratham's approach of teaching children at their actual level.
2. **Test it rigorously in one setting**
   Run a randomised evaluation, as with the Balsakhi study in Vadodara and Mumbai, to see whether it raises learning.
3. **Test delivery models at scale**
   Trial versions run by different people, such as volunteers or government teachers, in different states.
4. **Collect process data**
   Observe whether the approach actually happens in classrooms; in Bihar and Uttarakhand this showed that teachers felt bound to finish the curriculum.
5. **Redesign and re-test**
   Change the design to address the barrier, such as a dedicated daily hour in Haryana, and evaluate again; Banerjee and colleagues reported this in NBER Working Paper 22746 in 2016.
6. **Scale with government and keep learning**
   Expand through state partnerships, continuing to monitor implementation as the model spreads.

WORKED EXAMPLE · AN ILLUSTRATION

A state education department introduces new workbooks for foundational numeracy in Classes 3 to 5, but end-line scores barely move. The MEL team resists concluding that the materials failed. Classroom observation in a sample of schools shows that grouping by level, which the workbooks assume, rarely happens, because the timetable leaves no protected period and teachers prioritise the textbook. Following the logic of the Teaching at the Right Level sequence, the department designs a second trial: half of the schools in two districts, chosen by lottery, receive a protected daily numeracy hour and monthly visits from cluster mentors, while the rest continue as before. Process monitoring tracks whether grouping happens. Only after this second round does the department decide whether to scale, and which version.

CONCEPT HIGHLIGHT

Counterfactual

What would have happened to participants without the programme; in an RCT the comparison group provides an estimate of it.

USE IT WHEN

- An approach has worked in a pilot and a government or large NGO is considering scale.
- Implementation in different systems is likely to vary, so delivery models need testing.
- Partners can commit to several years of joint learning.

SKIP IT WHEN

- The decision must be made immediately and there is no prospect of further rounds of evidence.
- The intervention is so context-specific that sequential testing across settings is not meaningful.

COMMON MISREADINGS

- Citing the first trial's result as proof that any version will work at scale.
- Treating failed scale-up trials as wasted, when in this story they were what revealed the barrier and the fix.

WHY IT WAS GROUNDBREAKING

It showed how a sequence of randomised trials, combined with process evidence, could refine an idea until it worked inside government systems. It reframed scale-up as a learning process rather than a single decision.

Origin

Real sourced case: the Balsakhi evaluation with Pratham ran from 2001 to 2004 (J-PAL). Banerjee, Banerji, Berry, Duflo and colleagues reported the Bihar, Uttarakhand, Haryana and Uttar Pradesh trials in NBER Working Paper 22746 (2016), with effects of about 0.15 and 0.70 standard deviations in Haryana and Uttar Pradesh. Duflo's 2019 Nobel lecture reported partnerships reaching 21.3 million children.

RELATED IDEAS

Randomised evaluation of specific programme designs

Economists as plumbers SS Sridharan 246

PR Rogers 268

CONCEPT HIGHLIGHT

Cluster randomisation

Randomly assigning whole groups, such as villages or schools, rather than individuals, as in the 134 village Udaipur immunisation study.

3

BIG IDEA 3 OF 3 · DUFLO AND BANERJEE

Economists as plumbers: partnership and implementation detail

Much useful evaluation is plumbing: working with implementers over years to find and test fixes in how an existing programme actually runs.

PICTURE IT

Think of a plumber fixing a leak. The building's design may be fine, but water is escaping at one joint. Duflo argued that much useful evaluation is plumbing: finding and testing small fixes in how a real programme runs.

The problem

Evaluations often deliver a verdict on whole programmes, worth funding or not, while the real problems lie in the working parts: a form too complex to complete, an incentive that rewards the wrong behaviour, an audit paid for by the firm being audited. Verdicts of this kind give managers little to act on, and programmes with sound ideas can be cut because of fixable details. Practitioners needed evaluation that engaged with operational design, in partnership with those who run the system.

What it is

In her 2019 Nobel lecture Duflo argued that much policy work resembles plumbing: fixing how existing programmes function rather than deciding whether to fund whole sectors. The idea has three parts. First, long-term partnership with governments and NGOs, so that researchers understand how a system really operates and implementers trust the evidence. Second, attention to detail: identifying specific operational features, such as who pays, who checks and how information flows, that drive outcomes. Third, co-designed experiments that test changes to those features and feed directly into rules. Her lecture's example came from Gujarat, where researchers working with the state Pollution Control Board found that third-party environmental audits, paid for by the audited firms, were often inaccurate. A randomised trial of reforms, with auditors randomly assigned, paid from a central pool and checked through back-checks, produced more accurate reports and lower pollution, and the rules were changed in 2015.

How it works

1. **Build a lasting partnership**
   Work with an implementing agency over years rather than for a single study, so that questions come from its real problems.
2. **Diagnose how the system runs**
   Trace the actual flow of money, information and incentives, as the Gujarat team did with the audit system.
3. **Identify a fixable feature**
   Pinpoint a specific design element, such as who pays an auditor, that may be driving poor outcomes.
4. **Co-design the fix**
   Develop changes with the partner that are feasible within its rules and budget.
5. **Test it rigorously**
   Randomise the change where possible, as with the random assignment of auditors in Gujarat, and measure the outcome that matters.
6. **Help change the rules**
   Support the partner to adopt the fix formally, then monitor whether it holds.

WORKED EXAMPLE · AN ILLUSTRATION

A post-matric scholarship scheme for students from marginalised communities in a southern Indian state has low uptake, and a review recommends cutting it. The state's evaluation unit instead examines how the scheme runs. Application data show that many eligible students start online forms but abandon them at the income certificate upload. Working with the department, the team co-designs two changes: a simplified form that fetches certificates from an existing state database, and SMS reminders before deadlines. Colleges are randomly assigned to receive the new form, reminders, both or neither for one admission cycle. Completed applications rise most where both are used, and the department adopts both statewide. The scheme survives because the evaluation found the leak and helped fix it.

USE IT WHEN

- A programme's core idea is sound but implementation is leaking value.
- There is an established partnership with an agency willing to change its operating rules.
- Operational data exist that can reveal where the process breaks down.

SKIP IT WHEN

- The fundamental question is whether a policy should exist at all, which plumbing cannot answer.
- No partner is willing or able to act on operational findings.

COMMON MISREADINGS

- [ ] Reading plumbing as small or unambitious, when operational fixes can affect millions of users.
- [ ] Assuming the economist designs the fix alone, rather than with the implementer.

WHY IT WAS GROUNDBREAKING

It moved evaluation from verdicts on whole programmes to improving their working parts. It placed long-term partnership with implementers at the centre of rigorous evidence.

Origin

Duflo set out the plumbing metaphor in her Richard T. Ely Lecture, published as The Economist as Plumber in the American Economic Review in 2017, and developed it in her Nobel lecture, Field Experiments and the Practice of Policy, delivered in Stockholm on 8 December 2019, which described the Gujarat environmental audit reforms and the 2015 rule change.

RELATED IDEAS

Teaching at the Right Level

MP Patton 204
HW White 476
RR Rist 410

RELATED GIANTS



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476



Sonal Zaveri

The Indian evaluator who carried participatory, feminist practice from grassroots work into regional networks and real-time evaluation.

p. 536



Sanjeev Sridharan

The evaluator of complex health programmes who argues that designs should fit the programme, and that evidence lives in ecologies, not silos.

p. 246

CONCEPT HIGHLIGHT

External validity

The extent to which a result found in one setting holds in other places, populations or at larger scale.

CONCEPT HIGHLIGHT

Standard deviation effect size

A way of expressing an impact relative to the spread of outcomes, allowing comparison across studies; the Balsakhi programme raised scores by about 0.28 standard deviations in its second year.

CONCEPT HIGHLIGHT

Teaching at the Right Level

An approach developed with Pratham in which children are assessed, grouped by learning level rather than grade and taught at that level.

THE BIG IDEAS OF THIS CHAPTER

1 **Randomised evaluation of specific programme designs** p. 440

By using a lottery to decide who receives a programme, or which version of it, evaluators create comparable groups whose later differences can be credited to the programme.

2 **Teaching at the Right Level: evidence that builds over many trials** p. 443

Good policy evidence is built from a sequence of trials and process studies that refine an idea until it works inside real systems, not from one study and an automatic scale-up.

3 **Economists as plumbers: partnership and implementation detail** p. 446

Much useful evaluation is plumbing: working with implementers over years to find and test fixes in how an existing programme actually runs.

Did you know?

In the Udaipur study, offering a kilogram of lentils per vaccination roughly halved the cost per fully immunised child, from about 56 US dollars with camps alone to about 28 dollars, even after paying for the lentils.

Put it to work

USE IT ON MONDAY · RANDOMISED EVALUATIONS OF SPECIFIC, TESTABLE QUESTIONS

1

When a client asks for an RCT, first turn their question into a specific design choice, for example whether tutoring by trained volunteers beats tutoring by teachers, rather than whether the whole programme works.

2

Check feasibility honestly with a statistician: can units be assigned fairly and ethically by lottery, is the sample big enough, and can the partner deliver both versions?

3

Plan process and qualitative data alongside any trial, so that you can explain why the result came out as it did and whether it might transfer elsewhere.

AI prompt **Copy, fill the brackets, then check every line**

Help me screen whether a randomised evaluation, in the style of J-PAL's work led by Esther Duflo and Abhijit Banerjee, is feasible for **[programme]** run by **[partner]** in **[location]**. The client's question is **[question]**. The budget is **[budget]**, the timeline is **[timeline]**, and the programme reaches about **[number, if known] [units]**. (1) Rewrite the client's question as one or more specific, testable design choices. (2) Suggest possible units of assignment and randomisation approaches, with ethical and practical issues for each. (3) List the information a statistician would need for a power calculation; do not perform one with assumed numbers. (4) If an RCT is not feasible, suggest credible alternatives and what each could and could not tell us.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a fair coin toss deciding which half of a class tries a new teaching method first. Because chance, not choice, divided the groups, any later difference between them can be credited to the method. A randomised evaluation is that coin toss done carefully, at scale.

YOUR MOVES Three moves from Duflo and Banerjee's path

1 Be bold enough to switch paths

Banerjee, trained as an economic theorist, began his first field trial in 1996, and Duflo came to economics from history. Both changes opened the work that made their names.

TRY THIS

Learn one new method this year that you could use to answer a question your current methods cannot.

2 Treat partners as part of the method

Long partnerships with Pratham, Seva Mandir and state governments made their research possible.

TRY THIS

Spend as much effort building trust with your implementing partner as you spend on your sampling plan.

3 Turn an approach into an institution

By founding J-PAL in 2003 with Sendhil Mullainathan, with offices including Chennai from 2007, they made a research approach outlast any single study.

TRY THIS

Package what you do into a repeatable offer: a training, a toolkit or a service others can join.

THE BIG IDEAS OF THIS CHAPTER

1 **Randomised evaluation of specific programme designs** p. 440

By using a lottery to decide who receives a programme, or which version of it, evaluators create comparable groups whose later differences can be credited to the programme.

2 **Teaching at the Right Level: evidence that builds over many trials** p. 443

Good policy evidence is built from a sequence of trials and process studies that refine an idea until it works inside real systems, not from one study and an automatic scale-up.

3 **Economists as plumbers: partnership and implementation detail** p. 446

Much useful evaluation is plumbing: working with implementers over years to find and test fixes in how an existing programme actually runs.

CONCEPT HIGHLIGHT

Teaching at the Right Level

An approach developed with Pratham in which children are assessed, grouped by learning level rather than grade and taught at that level.

GIANTS CONNECTED TO THIS CHAPTER



Donald T. Campbell

Treating reforms as experiments, and the threats to validity that randomisation is designed to remove.

p. 112



Howard White

Joined counterfactual designs to theory of change and built bodies for evidence synthesis.

p. 476



Sonal Zaveri

Argues that participation should feature in all designs, including randomised trials.

p. 536

KEY WORKS TO READ IN THE ORIGINAL

- **Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty** 2011, Banerjee and Duflo
- **Handbook of Field Experiments, Volumes 1 and 2** 2017, Edited by Duflo and Banerjee
- **Good Economics for Hard Times** 2019, Banerjee and Duflo

CONCEPT HIGHLIGHT

Randomised controlled trial (RCT)

An evaluation in which units are assigned to treatment or comparison groups by chance, so that differences in outcomes can be attributed to the intervention.

WORKSHEET · TRY THIS TEMPLATE

RCT feasibility screen

Print this page, or copy the headings into your notes.

1 Specific question

A design choice the partner genuinely wants answered.

2 Unit of assignment

Individuals, households, schools, villages or blocks, and why.

3 Fair and ethical assignment

Is a lottery, phased roll-out or encouragement design acceptable to the community and partner?

4 Sample and power

To be calculated with a statistician from real figures; note what is still unknown.

5 Partner capacity

Can the partner deliver to plan and keep groups separate?

6 Mechanism and context

What process, cost and qualitative data will explain the result?

REFLECT

Is the question your client is asking really a yes-or-no question about impact, or a choice between two ways of doing the work?

CONNECTS TO

Donald T. Campbell (DC)
Treating reforms as experiments, and the threats to validity that randomisation is designed to remove.
p. 112

Howard White (HW)
Joined counterfactual designs to theory of change and built bodies for evidence synthesis.
p. 476

Sonal Zaveri (SZ)
Argues that participation should feature in all designs, including randomised trials.
p. 536

20



Emblem: an original HubForge design for this figure, not a likeness.

Zenda Ofir

Active 2000s to 2020s

South Africa; evaluation work in about 40 countries, mainly in Africa and Asia

Made in the Global South · Impact evaluation and evidence systems

The scientist turned evaluator who helped launch the call for evaluation made in Africa, by Africans.

IN ONE LINE

Zenda Ofir asked who gets to define good evaluation, and helped a continent, and then a profession, begin to answer for itself.

Niamey, Niger, January 2007. More than 100 evaluators fill a room for a day-long session with a pointed title: Making Evaluation Our Own. Its organisers, Zenda Ofir and Sulley Gariba, were worried that evaluation in Africa was largely commissioned, funded and designed somewhere else, and that randomised trials, then spreading fast among development economists, oversimplified communal, complex societies. By the end of the day a small group of African volunteers had drafted a statement calling for evaluation rooted in and led from Africa. Ofir regards that session as the birth of Made in Africa Evaluation. She had arrived in that room by an unusual road: a PhD in ecological chemistry, years managing national science grants, and from 2000 an evaluation career that would take her to about 40 countries.

Life CHAPTER 20

Zenda Ofir trained as a scientist before she became an evaluator. She holds a PhD in ecological chemistry, and her early career was in research management in South Africa. At a national science council she managed research grant programmes covering higher education and industry research in fields such as biotechnology, food production and food security. She then became Director of Research at the University of Pretoria, with responsibility for university research and international affairs.

From 2000 she worked as an independent evaluator and adviser. Her clients and panels have included IUCN, GAVI, CGIAR, IFAD, AGRA, the Rockefeller Foundation, WHO-TDR, UNDP's Independent Evaluation Office and other UN and bilateral bodies, and she reports work in about 40 countries, mainly in Africa and Asia, and more than 100 major evaluations. She

BIG IDEAS

1. **Made in Africa Evaluation**
2. **Evaluation for transformation**
3. **Research Quality Plus (RQ+)**

KEY WORKS

Making Evaluation Our Own (statement from the 4th AfrEA conference, Niger)
2007 · Session led with Sulley Gariba

Post on the International Evaluation Academy and evaluation for transformation (AEA365)
2022 · With Scott Chaplowe

taught a module on aid or development effectiveness at the United Nations University in Tokyo for several years, has been a visiting professor at Hiroshima University and has held an honorary professorship at Stellenbosch University since 2014. In 2018 she was described as based near Geneva.

Alongside consulting she built the profession. Her LinkedIn record lists her as founder of the South African Evaluation Network and a founding member of the South African Monitoring and Evaluation Association, co-convenor of the third AfrEA conference in Cape Town and co-initiator of what became Made in Africa Evaluation at the fourth AfrEA conference in Niamey in January 2007. She served as AfrEA President, as Vice-President of IOCE and of IDEAS, on the first steering committee of the Network of Networks on Impact Evaluation, and on the board of the American Evaluation Association, where Bounce Beyond describes her as the first board member living outside North America.

In the 2020s her focus moved to transformation. By February 2022 she chaired the Council of the new International Evaluation Academy, later described as its Scientific Leader and Council Chair, and from May 2024 she co-led its Three Horizons Evaluation Initiative. She has been a Richard von Weizsaecker Fellow of the Robert Bosch Academy in Berlin, a Resident Senior Impact Advisor to the Mastercard Foundation's Pan-African programmes, and the South African member of the High-Level Evaluation Advisory Committee of the BRICS New Development Bank. She writes a long-running blog, published under the name Evaluation for Development and now titled Tomorrow and Then, on evaluation, transformation and Global South perspectives.

The spark

Ofir was a scientist uneasy with simple answers to complex problems. In her own 2018 account, she and Gariba judged the randomised trials then gaining ground in African development research to be extractive and narrow, treating communities rather like experimental subjects and pulling attention away from problems that needed whole-system solutions. Her earlier career, managing

TIMELINE – LIFE AND WORK

2000 — Begins working internationally as an independent evaluator and adviser, after research management roles in South Africa.

2005 — Invited international expert at a JICA and JBIC workshop in Tokyo, 20 to 21 March.

2007 — Co-organises, with Sulley Gariba, the Making Evaluation Our Own stream at the fourth AfrEA conference in Niamey, Niger, in January.

2013 — The AfrEA Board adopts the name Made in Africa Evaluation and drafts a strategy.

2014 — Appointed Honorary Professor at Stellenbosch University.

research on biotechnology and food security and directing research at the University of Pretoria, had been about judging the quality of science. In our reading, the question that would not leave her alone was this: who decides what counts as good evidence, and for whom?

The breakthrough CHAPTER 20

Imagine your country's programmes are almost always judged by teams flown in from elsewhere, using frameworks designed elsewhere, answering questions chosen elsewhere. That, in Ofir's account, was the situation of evaluation in much of Africa by the mid-2000s. Then randomised trials arrived as the new gold standard. Ofir and Sulley Gariba, a former AfrEA President and the first IOCE President, judged that these designs did not reflect African philosophy or ways of knowing, could be extractive and oversimplified communal, complex societies. Meanwhile African evaluators were few, overstretched and had almost no academic centres of their own.

Their response came at the fourth AfrEA conference in Niamey, Niger, in January 2007. With NORAD funding they organised a special stream whose day-long session, Making Evaluation Our Own, drew more than 100 participants and international figures including Michael Quinn Patton, Robert Picciotto, Donna Mertens and Fred Carden. A small group of African volunteers drafted a statement calling for evaluation that was rooted in and led from Africa, grounded in African values, with African academic centres, mentoring, theory and a journal. Ofir regards this as the origin of Made in Africa Evaluation. Momentum faded, then revived at a 2012 Bellagio forum; in 2013 the AfrEA Board adopted the name Made in Africa Evaluation, and the African Evaluation Journal followed.

TIMELINE, CONTINUED

2016 — Co-authors IDRC's position paper Research Quality Plus: A Holistic Approach to Evaluating Research.

2018 — Gives a CECAN seminar in London on 27 February questioning routine use of the DAC criteria; publishes her account of the origins of Made in Africa Evaluation on 18 August.

2022 — Writes on the International Evaluation Academy and evaluation for transformation as its Council Chairperson, 20 February.

2024 — The International Evaluation Academy's Three Horizons Evaluation Initiative, which she co-leads, begins in May.

The problem it solved

Evaluations in Africa were largely commissioned and designed by external funders using foreign frameworks.

Impact WHAT CHANGED

Ofir's influence runs mainly through institutions and agendas rather than a single method. Made in Africa Evaluation, which she co-initiated, became AfrEA's named approach in 2013, helped give rise to the African Evaluation Journal and has generated a stream of African scholarship, including Chilisa's 2015 synthesis paper, a 2021 American Journal of Evaluation article by Chilisa and Mertens, and special collections in the African Evaluation Journal. Her work on RQ+ shaped how IDRC, a major funder of Southern research, judges the quality of the research it supports, and IDRC still used the approach in 2026.

Our interpretation is that her distinctive contribution is to carry Global South concerns into global rooms. Her positions in AfrEA, IOCE, IDEAS, AEA, NONIE and the International Evaluation Academy gave African and Asian perspectives a voice in bodies that set professional norms, and her blog has offered practitioners a running commentary on evaluation, decolonisation and transformation from a Global South standpoint. For practitioners in India and elsewhere in the South, the practical legacy is a set of questions that can be asked of any assignment: where the framework came from, who leads it, and whether the evaluation looks at system change as well as project results.

WHERE PEOPLE DISAGREE

Debates

Ofir's 2018 blog post records that Made in Africa Evaluation was widely misunderstood and that some evaluators rejected the idea that African evaluation should differ from evaluation elsewhere. Chilisa's 2024 article adds that the approach is still rarely applied in practice, citing funder and methodological colonialism and the reluctance of Western-trained evaluators. On evaluation for transformation, our reading is that the agenda is broad and open-ended, as the Academy itself acknowledges, which makes it harder to translate into criteria that commissioners can contract and test.

LIMITS, IN OUR READING

- [ ] Made in Africa Evaluation has been hard to turn into shared practice: Bagele Chilisa's 2024 article notes that the 2015 synthesis paper stayed in draft and was not widely known, and that scholars debate Ubuntu-based approaches.
- [ ] In our view, a fellowship body modelled on science academies faces an open question about how inclusive it can be, even when it aims not to be elitist.

WHY THIS MATTERS TO YOU

Ofir's questions are worth asking anywhere outside experts arrive with ready-made answers. Where did this framework come from? Who chose the questions? Whose idea of success is being used? She also asks a bigger question that applies to any effort to improve the world: does this project move the whole system towards a better future, or does the system quietly absorb it and carry on as before? And her work on research quality reminds you to judge evidence by whether it is trustworthy, relevant and useful to the people who need it, not only by how often it is cited. For evaluators in India, Kenya or Brazil, these questions open the door to leadership.

40+

countries in which she has worked on the ground as an evaluator and adviser

Robert Bosch Academy fellow profile

100+

major evaluations she has led or contributed to

Robert Bosch Academy fellow profile

100+

participants in the 2007 Making Evaluation Our Own session in Niamey

Zenda Ofir, Made in Africa Evaluation 1. The beginning (2018)

170

projects in the retrospective analysis IDRC used to validate RQ+, the approach she co-authored

Lebel and McLean, Nature 559 (2018), as summarised by Research Impact Canada

Did you know?

Before evaluation she was a research chemist: her PhD is in ecological chemistry, and she first managed national science grants in fields such as biotechnology and food security.

Key concepts

Made in Africa Evaluation
An evolving set of evaluation approaches rooted in African philosophies, cultures and knowledge, led by Africans and promoted by AfrEA since the 2007 Niamey conference.

Making Evaluation Our Own
The 2007 AfrEA special stream and statement, organised by Gariba and Ofir, that called for Africa-rooted and Africa-led evaluation.

DAC criteria
The OECD Development Assistance Committee criteria such as relevance, effectiveness, efficiency, impact and sustainability, widely used to structure development evaluations.

Prague Declaration
A 2019 statement adopted at the IDEAS Global Assembly committing evaluators to support the systemic changes needed for the 2030 Agenda.

Three Horizons (H2+ and H2-)
A futures framework used by the International Evaluation Academy: H1 is the present system, H3 the desired future, and H2 the transition; H2+ efforts transform, H2- efforts are absorbed by the status quo.

RQ+
IDRC's Research Quality Plus approach for assessing research for development using context-sensitive rubrics beyond scientific rigour.

Sources for this chapter: p. 586

THE BIG IDEAS OF THIS CHAPTER

1 **Made in Africa Evaluation** p. 462
Evaluation in Africa should be rooted in African values, knowledge and worldviews and led by African evaluators, rather than imported wholesale.

2 **Evaluation for transformation** p. 465
Evaluate whether interventions help shift whole systems towards sustainable, equitable development, not only whether projects deliver their outputs.

3 **Research Quality Plus (RQ+)** p. 468
Judge research for development on integrity, legitimacy, importance and positioning for use, in light of its context, not on rigour or citations alone.

GIANTS CONNECTED TO THIS CHAPTER



Bagele M. Chilisa

Wrote the synthesis paper that gave Made in Africa Evaluation a theoretical grounding.

p. 498



Marco Segone

Built the global partnership of evaluation associations through which such ideas travel.

p. 516



Sonal Zaveri

A parallel push for regional evaluation leadership in South Asia.

p. 536

KEY WORKS TO READ IN THE ORIGINAL

- **Making Evaluation Our Own (statement from the 4th AfrEA conference, Niger)** 2007, Session led with Sulley Gariba
- **Post on the International Evaluation Academy and evaluation for transformation (AEA365)** 2022, With Scott Chaplowe

In one line

Zenda Ofir asked who gets to define good evaluation, and helped a continent, and then a profession, begin to answer for itself.

1

BIG IDEA 1 OF 3 · OFIR

Made in Africa Evaluation

Evaluation in Africa should be rooted in African values, knowledge and worldviews and led by African evaluators, rather than imported wholesale.

PICTURE IT

Think of a meal cooked for you by a visiting chef who has never tasted your local food. It may be technically perfect and still miss what you value. Made in Africa Evaluation asks that evaluation be cooked in the local kitchen, with local ingredients and local cooks.

The problem

In the mid-2000s most evaluation in Africa was commissioned, funded and designed outside the continent. African evaluators were few and overstretched, with almost no academic centres or journals of their own. Ofir's 2018 account records her and Sulley Gariba's concern that randomised trials, then promoted by development economists, oversimplified communal, complex societies and could be extractive. For African communities this meant programmes judged by criteria they had no hand in setting; for African evaluators it meant a junior role in their own field.

What it is

Made in Africa Evaluation (MAE) is not a single method. In Bagele Chilisa's 2024 definition it is an umbrella for evolving approaches guided by African philosophies, cultures, values, histories, languages, indigenous knowledge and experience, with a decolonising intent and the application of AfrEA principles. It has three linked parts. The first is intellectual leadership: Africans set the questions, develop theory, and lead and author evaluations. The second is grounding: relational and communal ways of knowing shape what counts as value and as evidence. The third is infrastructure: academic centres, mentoring, publishing and a journal, all named in the 2007 Niamey statement. Chilisa's article lists nine ethical principles built on Ubuntu, among them relationality, reciprocity, respectful representation, reflexivity and decolonisation, which complement UN evaluation norms. MAE does not reject other knowledge systems. Its logic is that evaluation becomes more valid and more useful when the people who live with a programme's consequences help set its terms.

How it works

1. **Ask where the framework came from**
   Before design, identify who chose the evaluation approach and criteria, and for which context; Ofir and Gariba framed the issue first as one of who leads and whose theories apply.
2. **Put African evaluators in lead roles**
   Give national or regional evaluators leadership of design, analysis and authorship, in line with the 2007 Niamey call for Africa-rooted and Africa-led evaluation.
3. **Ground criteria in local values**
   Work with communities to define success in their own terms and languages, and carry those definitions into criteria and indicators.
4. **Apply the AfrEA principles**
   Use the principles listed in Chilisa's 2024 article, such as relationality, reciprocity, respectful representation and community ownership of data, as a design and ethics checklist.
5. **Mix knowledge systems**
   Combine indigenous methods, such as the Lekgotla consensus forum that Chilisa's article names as a main qualitative tool, with conventional qualitative and quantitative methods.
6. **Return and publish on the continent**
   Share findings back with communities and publish in African outlets such as the African Evaluation Journal, so the knowledge base grows where it is used.

WORKED EXAMPLE · AN ILLUSTRATION

A foundation funds a smallholder dairy programme in western Kenya and asks for an end-line evaluation. Taking a Made in Africa approach, the commissioning team appoints a Kenyan evaluator as team leader and lead author, with an international adviser in support. Before fixing indicators, the team holds community dialogues with farmer groups, women's cooperatives and elders to ask how they judge whether a household is better off. Alongside milk yields and income, the groups name the ability to meet obligations to extended family and to keep land in good health, and these become criteria. The team combines a household survey with community consensus forums, agrees with villages how data will be stored and shared, returns draft findings for correction, and submits a methods paper to the African Evaluation Journal.

CONCEPT HIGHLIGHT

Making Evaluation Our Own

The 2007 AfrEA special stream and statement, organised by Gariba and Ofir, that called for Africa-rooted and Africa-led evaluation.

USE IT WHEN

- The evaluation takes place in an African or other formerly colonised setting where imported frameworks dominate by default.
- Success depends on community, family, land or culture, so local definitions of value matter.
- Commissioners are willing to share leadership and authorship with national evaluators.

SKIP IT WHEN

- The label would be used without any change in who leads or how criteria are set; it is better not to claim it.
- The team cannot invest time in relationships and community dialogue, in which case state the limits honestly.
- The task is a narrow technical check, such as a financial compliance audit, where cultural grounding changes little.

COMMON MISREADINGS

- Reading it as protectionism or as a rejection of all Western methods; Chilisa's 2024 article lists such readings among the myths that slow its uptake.
- Treating it as one fixed model, or as interchangeable with generic participatory evaluation, when it rests on African philosophies and African intellectual leadership.

WHY IT WAS GROUNDBREAKING

It turned a widespread complaint about donor control into a named, continent-wide agenda with an institutional home in AfrEA. It treated authorship and worldview as part of evaluation quality rather than an afterthought. In our reading, it also gave other Southern regions a model for claiming their own evaluation traditions.

Origin

The idea took shape in the Making Evaluation Our Own stream at the fourth AfrEA conference in Niamey, Niger, in January 2007, organised by Ofir and Gariba with NORAD funding (Ofir, 2018). A 2012 Bellagio thought leaders' forum revived it, and in 2013 the AfrEA Board coined the name Made in Africa Evaluation. Chilisa's synthesis paper (final draft 2015) and her 2024 African Evaluation Journal review developed it further.

RELATED IDEAS

BC Chilisa 498 · African relational evaluation

DM Mertens 384 · SZ Zaveri 536

RC Chambers 296

2

BIG IDEA 2 OF 3 · OFIR

Evaluation for transformation

Evaluate whether interventions help shift whole systems towards sustainable, equitable development, not only whether projects deliver their outputs.

PICTURE IT

Think of fixing one pothole versus redesigning a city's whole drainage system so that potholes stop forming. Both matter, but only one changes the system. Evaluation for transformation asks whether a programme is helping to redesign the system.

The problem

Most development evaluations follow the five DAC criteria or the questions commissioners ask. In a 2018 London seminar Ofir argued that this habit can make evaluators overlook the issues that matter most for sustainable development and the SDGs, and so create false confidence about success. A project can meet its targets while the wider system, from soil and water to markets and power, deteriorates. With climate change, inequality and ecosystem loss accelerating, evaluations that look only at project results risk certifying activity that does not change the trajectory.

What it is

Evaluation for transformation asks evaluators to give as much attention to what they evaluate, and why, as to how. It applies a complex systems lens, which in Ofir's 2018 CECAN seminar meant examining interactions between goals, interventions and activities, the links between impact and sustainability, and ways of joining results-based with adaptive management. The 2019 Prague Declaration committed evaluators to support transformational and systemic change, and the International Evaluation Academy, whose Council Ofir chaired by 2022, made transformation its central aim. Since May 2024 the Academy's Three Horizons Evaluation Initiative, which she co-leads, has used the Three Horizons futures framework. Horizon 1 maps the current system, Horizon 3 an aspirational future set at 2035, and Horizon 2 the transition between them. Within Horizon 2 it separates H2+ initiatives, which are genuinely transformational, from H2- initiatives, which look transformational but are captured by the status quo. The logic is that evaluation should help societies tell real transitions from cosmetic ones.

How it works

1. **Question the default criteria**
   Before adopting the DAC criteria, ask what matters for sustainable development in this setting and what those criteria might hide, as Ofir urged in her 2018 CECAN seminar.
2. **Draw the system**
   Map the wider system around the intervention, including interactions with other goals and interventions and the links between impact and sustainability.
3. **Set out the three horizons**
   Describe the current system (H1), a desired future (H3) and the transition pathway between them (H2), as the Academy's initiative does with systems maps.
4. **Test for capture**
   Judge whether the intervention is H2+, moving the system towards H3, or H2-, absorbed by the status quo; the initiative cites carbon offsets and ESG as possible H2- examples.
5. **Look for leverage points**
   Identify points where change could spread across the system, a priority named by Ofir and Scott Chaplowe in their 2022 AEA365 post.
6. **Join accountability and adaptation**
   Report on agreed results while feeding learning into adaptive management, rather than choosing one over the other.

WORKED EXAMPLE · AN ILLUSTRATION

A state agriculture department in India asks for an evaluation of a crop insurance and input subsidy scheme for rice farmers. A conventional design would check coverage, claim settlement times and yields against targets. A transformation lens adds further questions. The team maps the system around the scheme: groundwater levels, soil health, farm debt, women's land rights and the market power of input dealers. Using Three Horizons, it describes today's system, a 2035 future of resilient, diversified farms, and the transition between them. It then asks whether the scheme moves farmers along that path or locks in water-intensive monocropping, which would make it an H2- effort despite meeting its targets. The final report gives the department both a verdict on delivery and a judgement on direction of travel.

CONCEPT HIGHLIGHT

Made in Africa Evaluation

An evolving set of evaluation approaches rooted in African philosophies, cultures and knowledge, led by Africans and promoted by AfrEA since the 2007 Niamey conference.

USE IT WHEN

- A programme or portfolio claims to contribute to systemic goals such as climate resilience or the SDGs.
- Strategy, sector or portfolio evaluations need to know whether many initiatives together are changing a trajectory.
- Commissioners are open to questions beyond their own results framework.

SKIP IT WHEN

- A small, well-bounded service with an accepted model needs only a standard performance answer.
- The commissioner needs a narrow compliance verdict on a short timeline and has no scope to act on systems findings.

COMMON MISREADINGS

- Treating transformation as a slogan attached to any evaluation, without changing its questions, boundaries or criteria.
- Assuming it abandons accountability for results, when Ofir's 2018 framing linked results-based and adaptive management rather than choosing between them.

WHY IT WAS GROUNDBREAKING

It challenged criteria that most development evaluations still apply by default. It asked the profession to judge the direction of travel of whole systems alongside project delivery, and to do so under deep uncertainty.

Origin

Ofir set out her critique of routine DAC use in a CECAN seminar in London on 27 February 2018. The IDEAS Prague Declaration followed in 2019, and on 20 February 2022 she and Scott Chaplowe described the International Evaluation Academy's transformational agenda on AEA365. The Three Horizons framework comes from futures practice, described by Sharpe and colleagues in Ecology and Society (2016); the Academy began applying it to evaluation in May 2024.

RELATED IDEAS

Made in Africa Evaluation

Research Quality Plus (RQ+)

MP Patton 204

PR Rogers 268

RC Chambers 296

Did you know?

The 2007 Niamey session that she regards as the birth of Made in Africa Evaluation was prompted in part by the rise of randomised controlled trials in African development research.

3

BIG IDEA 3 OF 3 · OFIR

Research Quality Plus (RQ+)

Judge research for development on integrity, legitimacy, importance and positioning for use, in light of its context, not on rigour or citations alone.

PICTURE IT

Think of judging a restaurant only by how many reviews it has. That tells you it is popular, not whether the food is safe, honest or right for the people who eat there. RQ+ judges research on integrity, legitimacy, importance and usefulness, not citations alone.

The problem

Funders of research in the Global South usually judged quality through peer review and citation counts. Writing in Nature in 2018, Jean Lebel and Robert McLean of IDRC argued that such measures undervalue local researchers because they ignore context and different motivations. A study by a team in a fragile state, working with poor data and political risk, could be useful and well executed yet score badly on conventional metrics. That bias shaped who was funded and whose knowledge reached policy, with consequences for Southern research and its users.

What it is

Research Quality Plus (RQ+) is IDRC's approach to evaluating the quality of research for development. It has three components. First, contextual factors: five influences on quality, namely the maturity of the research field, research capacity strengthening, and risk in the data, research and political environments. Second, a multidimensional view of quality with four dimensions. Research integrity covers the technical quality and rigour of design and execution. Research legitimacy covers how well the research accounts for the concerns and insights of relevant stakeholders. Research importance covers the value of the knowledge to key intended users. Positioning for use covers how well the research was managed and its outputs prepared to make use and influence more likely. The first three appear in most quality frameworks; positioning for use is the plus. Third, evaluative rubrics and empirical evidence make each judgement explicit. The logic is that rigour is required but does not by itself make research valuable to society.

How it works

1. **Select the sample**
   Choose a representative sample of projects and their outputs, with data sources such as documents and interviews, the first of IDRC's four published steps.
2. **Characterise the context**
   Rate each of the five contextual factors, in the 2016 version on a low, medium or high scale, using project documents, interviews and reviewers' expertise.
3. **Rate the quality dimensions**
   Score each dimension and sub-dimension against an evaluative rubric; the 2016 version used an eight-point scale.
4. **Synthesise the ratings**
   Combine ratings from parallel reviewers and roll them up from project to programme or portfolio level.
5. **Adapt the rubric to the purpose**
   Customise rubrics for formative learning, partnership building or summative review, uses IDRC lists for the approach.

WORKED EXAMPLE · AN ILLUSTRATION

An Indian foundation funds twelve policy research projects on urban sanitation across five states and wants to know which deserve a second phase. Using an RQ+ style review, two reviewers independently assess each project. They first rate context: one project worked in a city with little reliable data and a hostile municipal administration, so data and political risks are high. They then score integrity, legitimacy, importance and positioning for use on a rubric. A project led by a small local institute scores only moderately on methodological polish but highly on legitimacy, because sanitation workers helped frame its questions, and on positioning for use, because the municipal commissioner adopted its ward-level findings. The foundation gains a fuller picture of quality than a citation count could give it.

USE IT WHEN

- A funder or institute is assessing a portfolio of research for development.
- Calls for proposals or grant reviews need criteria beyond publication record.
- Research is co-produced with policymakers or communities, for which IDRC has developed an adapted framework.

SKIP IT WHEN

- The task is peer review of a single academic paper, where journal norms already apply.
- The question is long-term impact itself, which positioning for use points towards but does not assess.

COMMON MISREADINGS

- [ ] Treating a high score for positioning for use as proof of impact, when it assesses steps that make use more likely, not use itself.
- [ ] Using contextual risk to inflate scores quietly, when the context ratings are meant to inform judgement openly.

WHY IT WAS GROUNDBREAKING

It gave funders a structured alternative to citation counts and narrow rigour, with context and use built into the definition of quality. Validated on a retrospective analysis of 170 projects, it challenged assumptions about Southern-led research and about any trade-off between rigour and usefulness.

Origin

IDRC first designed RQ+ in 2015 for its external programme reviews. The position paper Research Quality Plus: A Holistic Approach to Evaluating Research (2016) was written by Ofir, Thomas Schwandt, Colleen Duggan and Robert McLean. A 2022 IDRC book, Research Quality Plus: Evaluating Research Differently, with Ofir among five authors, extended it, and IDRC revised its assessment instrument in 2026.

RELATED IDEAS

Evaluation for transformation

CW Weiss 152

MP Patton 204

BC Chilisa 498

RELATED GIANTS



Bagele M. Chilisa

The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

p. 498



Donna M. Mertens

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

p. 384



Sonal Zaveri

The Indian evaluator who carried participatory, feminist practice from grassroots work into regional networks and real-time evaluation.

p. 536



Robert Chambers

The former colonial administrator who spent a career asking development professionals to learn from the people they usually see last.

p. 296

CONCEPT HIGHLIGHT

DAC criteria

The OECD Development Assistance Committee criteria such as relevance, effectiveness, efficiency, impact and sustainability, widely used to structure development evaluations.

CONCEPT HIGHLIGHT

Prague Declaration

A 2019 statement adopted at the IDEAS Global Assembly committing evaluators to support the systemic changes needed for the 2030 Agenda.

CONCEPT HIGHLIGHT

Three Horizons (H2+ and H2-)

A futures framework used by the International Evaluation Academy: H1 is the present system, H3 the desired future, and H2 the transition; H2+ efforts transform, H2- efforts are absorbed by the status quo.

CONCEPT HIGHLIGHT

RQ+

IDRC's Research Quality Plus approach for assessing research for development using context-sensitive rubrics beyond scientific rigour.

In one line

Evaluation in Africa should be rooted in African values, knowledge and worldviews and led by African evaluators, rather than imported wholesale.

Put it to work

USE IT ON MONDAY · MADE IN AFRICA EVALUATION

1. Before adopting an evaluation framework, ask local evaluators and community members which ideas of value and success already exist in this setting, and how the framework would need to change to respect them.

2. Put national and local evaluators in lead roles, in design, analysis and authorship, not only in data collection.

3. Add one systems question to your evaluation: does this programme contribute to deeper change, for example on inequality or the environment, or does it only deliver its outputs?

AI prompt **Copy, fill the brackets, then check every line**

Help me review an evaluation design for **[programme]** in **[country or region]** through the lens of Made in Africa Evaluation, associated with Zenda Ofir and Sulley Gariba, and evaluation for transformation. Here is the draft design: **[paste]**. Here is what I know about the local context, languages and evaluation community: **[paste]**. (1) Identify which elements of the design are imported and may not fit, and explain why. (2) Suggest questions to ask local evaluators and communities about values and success. (3) Suggest how leadership of the evaluation could sit more with national or regional evaluators. (4) Propose one systems-change question. Do not describe the beliefs or philosophies of any community unless I have provided them.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a meal cooked for you by a visiting chef who has never tasted your local food. It may be technically perfect and still miss what you value. Made in Africa Evaluation asks that evaluation be cooked in the local kitchen, with local ingredients and local cooks.

YOUR MOVES Three moves from Ofir's path

1 Bring your old field into your new one

Ofir moved from ecological chemistry and science management into evaluation in 2000, bringing a scientist's concern with quality and systems.

TRY THIS
If you are changing careers, list three skills you carry over and lead with them.

2 Build movements with allies

Made in Africa Evaluation grew from a session she co-organised with Sulley Gariba in 2007, was revived at Bellagio in 2012 and was named by the AfrEA Board in 2013.

TRY THIS
Find two allies who share a cause you care about in evaluation, and organise one session together.

3 Serve the profession to shape its norms

As AfrEA President, Vice-President of IOCE and IDEAS and Council Chair of the International Evaluation Academy, she carried Global South concerns into the rooms that set norms.

TRY THIS
Volunteer for one committee or task force in your national or regional evaluation association.

THE BIG IDEAS OF THIS CHAPTER

1. **Made in Africa Evaluation** p. 462
   Evaluation in Africa should be rooted in African values, knowledge and worldviews and led by African evaluators, rather than imported wholesale.
2. **Evaluation for transformation** p. 465
   Evaluate whether interventions help shift whole systems towards sustainable, equitable development, not only whether projects deliver their outputs.
3. **Research Quality Plus (RQ+)** p. 468
   Judge research for development on integrity, legitimacy, importance and positioning for use, in light of its context, not on rigour or citations alone.

CONCEPT HIGHLIGHT

RQ+

IDRC's Research Quality Plus approach for assessing research for development using context-sensitive rubrics beyond scientific rigour.

GIANTS CONNECTED TO THIS CHAPTER



Bagele M. Chilisa

Wrote the synthesis paper that gave Made in Africa Evaluation a theoretical grounding.

p. 498



Marco Segone

Built the global partnership of evaluation associations through which such ideas travel.

p. 516



Sonal Zaveri

A parallel push for regional evaluation leadership in South Asia.

p. 536

KEY WORKS TO READ IN THE ORIGINAL

- **Making Evaluation Our Own (statement from the 4th AfrEA conference, Niger)** 2007, Session led with Sulley Gariba
- **Post on the International Evaluation Academy and evaluation for transformation (AEA365)** 2022, With Scott Chaplowe

CONCEPT HIGHLIGHT

Made in Africa Evaluation

An evolving set of evaluation approaches rooted in African philosophies, cultures and knowledge, led by Africans and promoted by AfrEA since the 2007 Niamey conference.

CONCEPT HIGHLIGHT

Making Evaluation Our Own

The 2007 AfrEA special stream and statement, organised by Gariba and Ofir, that called for Africa-rooted and Africa-led evaluation.

WORKSHEET · TRY THIS TEMPLATE

Contextual grounding check

Print this page, or copy the headings into your notes.

1 Where the framework came from

Who designed it, for which context, and who chose it here.

2 Local ideas of value and success

Concepts, languages and philosophies already in use.

3 Who leads

Nationality and location of the people leading design, analysis and authorship.

4 Knowledge systems respected

Oral, communal or indigenous forms of knowledge and how they are treated.

5 Systems change question

One question about deeper, longer-term transformation.

6 Adaptations we will make

Concrete changes to the framework, tools or team.

REFLECT

On your current assignment, who chose the evaluation framework, and where was it made?

CONNECTS TO

BC **Bagele M. Chilisa** — Wrote the synthesis paper that gave Made in Africa Evaluation a theoretical grounding. p. 498

MS **Marco Segone** — Built the global partnership of evaluation associations through which such ideas travel. p. 516

SZ **Sonal Zaveri** — A parallel push for regional evaluation leadership in South Asia. p. 536

21



Emblem: an original HubForge design for this figure, not a likeness.

Howard White

Active 2000s to 2020s

International; worked with institutions in Europe, the United States and India

Impact evaluation and evidence systems · Theory and programme logic

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

IN ONE LINE

Howard White taught development evaluators to open the black box, and then to look up and see the whole landscape of evidence.

In 2006 a working group convened by the Center for Global Development asked a question that stung: When Will We Ever Learn? Billions of dollars were being spent every year on social programmes in developing countries, the group argued, yet very few had been studied in a way that could show whether they made any difference. Its answer became an institution: the International Initiative for Impact Evaluation, 3ie, whose founding document dates from 2008. Its first executive director was Howard White. By the time he handed over in 2015, 3ie had completed 63 impact evaluations, committed close to US$99 million in grants and made its New Delhi office a full branch. Then he pushed further: from single studies to whole bodies of evidence, and from producing evidence to asking who would actually use it.

Life CHAPTER 21

Howard White's career began in academic research on development economics at the Institute of Social Studies in The Hague and the Institute of Development Studies at the University of Sussex. The World Health Organization lists his research interests as aid and development effectiveness, anti-poverty interventions, macroeconomic modelling and policy, evaluation and evidence synthesis methods, knowledge translation and mixed methods. Several speaker biographies describe his aim as bringing policy relevance to academic work and academic rigour to policy work, and note that outside work he enjoys long-distance running, walking in remote places and reading English history.

From academia he moved to the World Bank's Independent Evaluation Group, where he led the impact evaluation programme. A 2006 paper drawing on that experience presented impact evaluation as a way to judge how far aid

BIG IDEAS

1. **Theory-based impact evaluation**
2. **Evidence and gap maps**
3. **The four waves of the evidence revolution**

KEY WORKS

Theory-based impact evaluation: principles and practice
2009 · Journal of Development Effectiveness 1(3)

Journal of Development Effectiveness · Founding managing editor

3ie evidence gap maps
2015 · First five maps published by 3ie

actually reduces poverty, answering critics who said such impact had rarely been measured. In parallel he edited journals: he was managing editor of the Journal of Development Studies from 2002 to 2013 and founding managing editor of the Journal of Development Effectiveness.

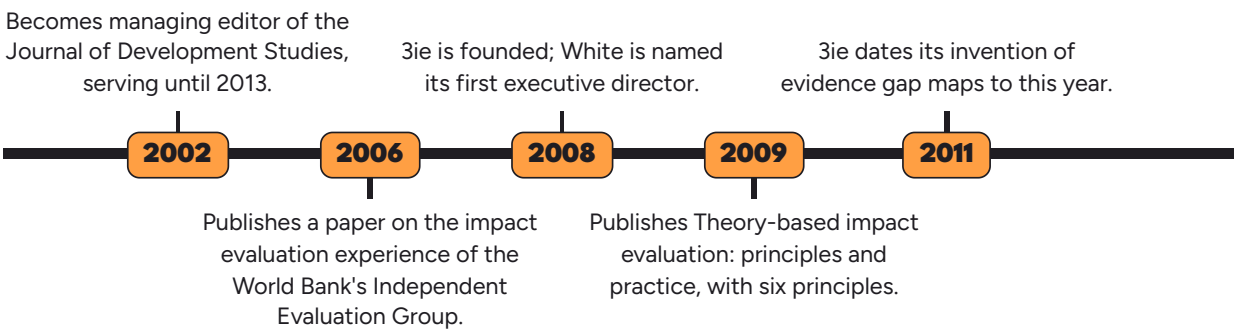
His best-known role came next. In 2006 the Center for Global Development's Evaluation Gap Working Group, convened in 2004, recommended that governments, aid agencies and foundations jointly create an independent body to fund rigorous impact evaluations. That body became the International Initiative for Impact Evaluation, 3ie, whose founding document dates from 2008. The Center for Global Development reported that White, then based in Cairo and experienced in evaluation across Africa and Asia, had been named its first director. By 2015 3ie had offices in New Delhi, Washington DC and London, had completed 63 impact evaluations and had committed close to US$99 million in grants. That year Emmanuel Jimenez succeeded him.

White then served as chief executive of the Campbell Collaboration, the international network producing systematic reviews in social policy, from 2015 to 2022, and sat on the Global Commission on Evidence to Address Societal Challenges, whose report appeared in 2022. Current biographies list him as Director of Evaluation and Evidence Synthesis at the Global Development Network in New Delhi, Professor of Evidence-Based Social Policy at Lanzhou University and Research Director of the Centre of Excellence for Development Impact and Learning. The governments of Benin and Uganda have given him awards for services to evaluation.

The spark

A development economist working on aid effectiveness and anti-poverty programmes, he knew cases like the Bangladesh Integrated Nutrition Project, which his 2009 paper discussed: it had been scaled up on the strength of monitoring data before anyone knew whether it worked. As 3ie's first director, he argued that spending on such a scale should produce evidence of what works, where and when, with voices from developing countries shaping the agenda.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 21

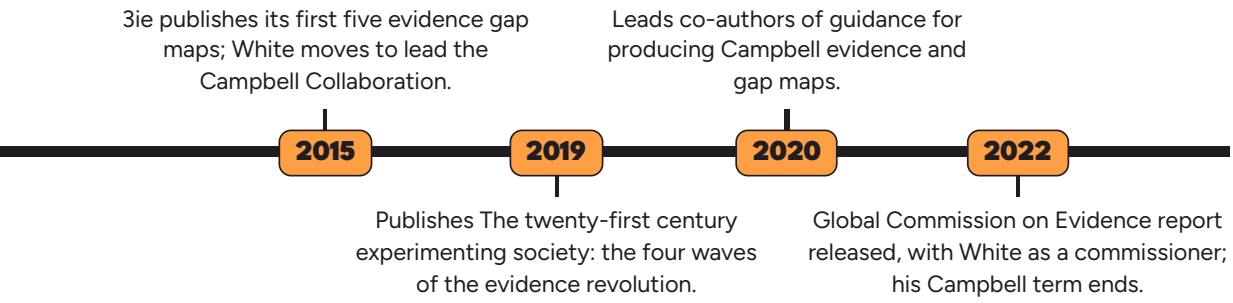
Imagine a nutrition programme that has been rolled out across a country because its monitoring reports looked good. Then a careful study finds no significant overall impact on children's nutrition. What went wrong? A typical impact study of the mid-2000s could not tell you. Many reported a single average effect and stopped there. White called this a black box approach: it could say whether a programme worked on average, but not why, for whom or what to change. And the Center for Global Development's 2006 report had shown how rarely even that much was known: billions were spent each year on social programmes, yet very few had been rigorously tested.

His answer appeared in June 2009 as 3ie Working Paper 3 and, the same year, as Theory-based impact evaluation: principles and practice in the Journal of Development Effectiveness 1(3). The paper observed that almost everyone endorsed theory-based evaluation while its application remained weak, and set out six principles for doing it properly. Its central case was the Bangladesh Integrated Nutrition Project, which had been scaled up on the strength of monitoring data before later analysis found no significant overall impact. Tracing the causal chain showed where it broke: mis-targeting, leakage of supplementary food, a gap between mothers' knowledge and practice, and the influence of mothers-in-law over household decisions.

Did you know?

When the Center for Global Development reported his appointment as 3ie's first director, White was based in Cairo.

TIMELINE, CONTINUED



The problem it solved

Impact studies that reported whether a programme worked without showing why, and decision-makers lost among scattered studies.

Impact WHAT CHANGED

White's most tangible legacy is institutional. 3ie was designed to fund evaluations in low- and middle-income countries and to involve teams based there. In 2015 alone it awarded 28 new impact evaluation grants in 16 such countries, and its New Delhi office, home to three of its five teams, became a full branch office that year. By 2023 the Development Evidence Portal hosted by 3ie listed more than 10,000 impact evaluations and nearly 1,000 systematic reviews. His later base at the Global Development Network in New Delhi, and awards from the governments of Benin and Uganda, point to long engagement with evaluation systems in the Global South.

His ideas travelled with the institutions. In our reading, the six principles of theory-based impact evaluation have become a common reference point in impact evaluation guidance, and evidence and gap maps have become a standard Campbell product with formal guidance since 2020. Our interpretation is that his larger influence lies in the sequence he helped build: run better single studies, pool them, map the gaps, then press governments to use what is known. For MEL practitioners in India and elsewhere, the practical result is that a credible evaluation plan now usually starts with a search for existing reviews and maps and ends with a causal chain, not just an effect size.

KEY WORKS TO READ IN THE ORIGINAL

- **Theory-based impact evaluation: principles and practice** 2009, Journal of Development Effectiveness 1(3)
- **Journal of Development Effectiveness** Founding managing editor
- **3ie evidence gap maps** 2015, First five maps published by 3ie

WHERE PEOPLE DISAGREE

Debates

The wider impact evaluation movement that 3ie belongs to has been contested. A 2012 DFID working paper by Stern and colleagues argued that experimental and statistical designs suit only a small share of development programmes and called for a broader range of designs; this was a critique of the agenda rather than of White personally. White himself defended quantitative methods, including randomised trials, as often the best available, while arguing in a 2009 paper that attribution need not mean an intervention was the sole cause. His own 2009 assessment that theory-based practice remained weak is also a limit worth noting. In our reading, evidence maps can be mistaken for findings, although their own guidance stresses they show where evidence is, not what it says.

LIMITS, IN OUR READING

- ☐ Stern and colleagues (DFID Working Paper 38, 2012) argued that experimental and statistical designs apply to only a small share of programmes; this critique concerned the wider impact evaluation agenda, not White personally.
- ☐ White's own 2009 paper noted that the practice of theory-based impact evaluation remained weak, a limit on the approach as actually applied.

WHY THIS MATTERS TO YOU

White's habits are useful whenever you want to know whether something really works. Before starting anything new, find out what is already known, because someone may have tested your idea already. When you draw a plan, treat each arrow as a claim to check, not a decoration. When something fails, look for the exact link that broke, because that is the finding that lets people fix a good idea instead of abandoning it. And remember that evidence only matters when someone with the power to act asks for it and uses it. That craft is open to anyone.

In one line

Howard White taught development evaluators to open the black box, and then to look up and see the whole landscape of evidence.

63

impact evaluations completed by 3ie by 2015, under his leadership as founding executive director

3ie Annual Report 2015

US$98.7 million

committed by 3ie across all its grant windows by December 2015

3ie Annual Report 2015

47+

low- and middle-income countries where 3ie-funded studies were informing decisions by 2015

3ie Annual Report 2015

10,000+

impact evaluations listed in the Development Evidence Portal hosted by 3ie by 2023, alongside nearly 1,000 systematic reviews

3ie blog, 15 years of 3ie (March 2023)

Key concepts

Counterfactual

An estimate of what would have happened to the same people without the intervention, usually built from a comparison or control group.

Causal chain

The sequence of links from inputs to activities, outputs, outcomes and impact that a programme assumes; theory-based impact evaluation tests each link.

Heterogeneity

Differences in how a programme affects different groups, places or circumstances, which an average effect can hide.

Factual analysis

Rigorous description of what actually happened along the causal chain, such as who was reached and what they did, as distinct from counterfactual estimation.

Systematic review

A study that locates, appraises and synthesises all research meeting explicit criteria on a defined question, to say what the body of evidence shows.

Evidence and gap map

A visual, systematically produced grid showing where studies exist and where they are missing; it maps evidence rather than summarising findings.

Sources for this chapter: p. 587

THE BIG IDEAS OF THIS CHAPTER

1. **Theory-based impact evaluation** p. 484

   Estimate a programme's impact against a credible counterfactual and test, link by link, the causal chain that is supposed to produce it, so the study explains why as well as whether.

2. **Evidence and gap maps** p. 487

   Systematically chart where evidence on a policy area exists and where it is missing, so that decisions and new research start from the whole evidence base.

3. **The four waves of the evidence revolution** p. 490

   See the evidence movement as four building waves, results, impact evaluation, systematic reviews and knowledge brokering, and recognise that supplying evidence is not enough without demand for its use.

GIANTS CONNECTED TO THIS CHAPTER



Huey-tsyh Chen

Theory-driven evaluation supplied the programme theory side of the pairing.

p. 174



Esther Duflo and Abhijit Banerjee

Randomised evaluations supplied much of the counterfactual evidence base.

p. 432



Donald T. Campbell

The threats-to-validity tradition behind credible counterfactuals.

p. 112

1

BIG IDEA 1 OF 3 · WHITE

Theory-based impact evaluation

Estimate a programme's impact against a credible counterfactual and test, link by link, the causal chain that is supposed to produce it, so the study explains why as well as whether.

PICTURE IT

Think of a string of fairy lights that will not switch on. Knowing the string is dead is not enough; you need to find the faulty bulb. Theory-based impact evaluation measures whether the lights came on and then tests each bulb in the chain, to show exactly where it broke.

The problem

By the late 2000s development agencies were commissioning more impact evaluations, yet many reported a single average effect and little else. White called these black box studies. A positive or null result told managers nothing about which link had held or failed, for whom, or what to change. The Bangladesh Integrated Nutrition Project showed the cost: it was scaled up on the strength of monitoring data before rigorous analysis found no significant overall impact, and resources meant for malnourished children went into a model that did not work as designed.

What it is

Theory-based impact evaluation, as set out in White's 3ie Working Paper 3 (June 2009), combines two things often done separately: a rigorous counterfactual estimate of impact, and a systematic test of the programme theory behind it. It rests on six principles. First, map the causal chain from inputs to impact, with the assumptions behind each link. Second, understand the social, political and economic context. Third, anticipate heterogeneity, specifying in advance the subgroups for whom impact may differ. Fourth, estimate impact rigorously against a credible counterfactual. Fifth, carry out rigorous factual analysis, because many links, such as who was targeted or whether food leaked to others, are factual rather than counterfactual questions. Sixth, use mixed methods. The underlying logic is that the average effect remains essential, not least for cost-effectiveness, but only the chain explains it. Where causal channels are too numerous to separate, a reduced-form estimate may be the best available.

How it works

1. **Map the causal chain**
   Review project documents and logframes, check them with managers, field staff and beneficiaries, and draw each link from inputs to impact with its assumptions; White's 2009 paper asks for a dynamic map that also considers competing theories and unintended effects.
2. **Read the context**
   Use project documents and wider literature, including political economy and anthropology, to understand the conditions that shape each link and limit how far results generalise.
3. **Specify subgroups in advance**
   Name the groups for whom impact may differ and size the sample so those comparisons are possible, which also guards against false positives from multiple testing.
4. **Build the counterfactual**
   Choose an experimental or quasi-experimental design that avoids selection bias, ideally with baseline panel data, and watch for spillovers into the comparison group and contamination from other programmes.
5. **Test the factual links**
   Use representative data to examine targeting, participation, leakage and behaviour change along the chain, starting with descriptive tabulations and using regression to explain what drives participation.
6. **Mix methods and interpret**
   Combine time in the field, focus groups and other qualitative inquiry with the quantitative estimates, then report the average effect together with where and why the chain held or broke.

WORKED EXAMPLE · AN ILLUSTRATION

A state government in India pilots a programme in which anganwadi workers counsel mothers of underweight children and provide take-home rations. The evaluators draw the chain: children are weighed, faltering children are correctly identified, rations reach them, mothers learn new feeding practices and put them into practice, and nutritional status improves. They compare pilot blocks with matched comparison blocks to estimate the effect on underweight prevalence, and household surveys test each link. Suppose the estimate shows no overall change. The factual analysis then locates the breaks: many eligible children were never enrolled, rations were shared with siblings, and mothers knew the advice but grandmothers decided what was cooked. The state can now fix targeting and involve whole households, rather than abandoning the programme or scaling it up blindly.

USE IT WHEN

- A programme is being considered for scale-up and decision-makers need to know which components drive results.
- You can build a credible comparison group and also collect data at each link of the causal chain.
- Earlier studies of similar programmes report mixed results that need explaining.

SKIP IT WHEN

- ○ The question concerns implementation alone and no impact estimate is needed.
- ○ Budget, timing or ethics rule out any credible counterfactual; a theory-based design without an impact estimate may then fit better.
- ○ The programme is so emergent that its causal chain cannot yet be specified.

COMMON MISREADINGS

- ☐ Treating the theory of change as decoration and reporting only the effect size, which is exactly the black box study White criticised.
- ☐ Assuming theory-based means qualitative only, when the paper insists on a rigorous counterfactual alongside factual analysis.

WHY IT WAS GROUNDBREAKING

It moved the central question of development impact evaluation from whether a programme worked to why, and for whom. It offered a working bridge between economists who favoured experiments and evaluators who favoured programme theory and qualitative inquiry. In our reading, it also made the broken link, not just the failed programme, the unit of learning.

Origin

White set out the six principles in June 2009 as 3ie Working Paper 3 and, the same year, in the Journal of Development Effectiveness 1(3), drawing on analysis by the World Bank's evaluation department of the Bangladesh Integrated Nutrition Project. A companion 2009 paper, Some reflections on current debates in impact evaluation, defended counterfactual designs while arguing that attribution need not mean an intervention was the sole cause.

RELATED IDEAS

HC Chen 174 · CW Weiss 152 · DC Campbell 112 · D&B Duflo and Banerjee 432 · PR Rogers 268

2

BIG IDEA 2 OF 3 · WHITE

Evidence and gap maps

Systematically chart where evidence on a policy area exists and where it is missing, so that decisions and new research start from the whole evidence base.

PICTURE IT

Think of an old explorer's map, with well-charted coasts and large blank spaces marked unknown. An evidence and gap map shows, at a glance, where research on a topic is plentiful and where it is missing, so new studies fill the blanks.

The problem

Decision-makers faced thousands of scattered studies, and new evaluations were often commissioned without anyone checking what was already known. Systematic reviews helped, but each answered one narrow question and took a long time. Funders could not see at a glance which interventions had been studied many times, which outcomes were neglected or which regions were missing. The result was duplication in some areas, blind spots in others, and programmes for poor people designed on thinner evidence than anyone realised.

What it is

An evidence and gap map (EGM) is a systematic, visual display of the evidence in a broad policy area. It uses the search, screening and coding rules of a systematic review, but codes studies by where they sit rather than by what they find. Its core is a matrix, usually with interventions as rows and outcomes as columns; the subcategories are called labels. Each cell shows the studies that fall there, often as bubbles sized by number. Filters such as region, population or study design let users narrow the view. Standard maps include primary studies and reviews; mega-maps include only reviews and maps; maps of maps include only maps. The 2020 Campbell guidance stresses that maps show what evidence exists, not what it says, so the direction and size of effects are usually not coded. Empty or thin cells reveal evidence gaps, and cells with many studies but no review reveal synthesis gaps.

How it works

1. **Build the framework with users**
   Agree rows, columns and filters with stakeholders and the commissioner, adapting existing typologies and piloting over several rounds; the 2020 Campbell guidance suggests roughly four to six headings.
2. **Write the scope and protocol**
   Define the scope, check for overlapping maps, and record the framework, search strategy, coding form, appraisal tools and analysis plan.
3. **Search widely**
   Search several databases, unpack the included studies of existing reviews, snowball from key studies and check registries for ongoing research.
4. **Screen in duplicate**
   Screen records against the inclusion criteria with two reviewers, using machine learning only as an aid; the guidance notes that maps often screen 15,000 to 30,000 records.
5. **Code each study**
   Pilot the coding form on 20 to 30 studies, then code interventions, outcomes and filters, and clean the data.
6. **Appraise and visualise**
   Critically appraise studies where feasible, build the interactive matrix, often in EPPI-Mapper, and write a descriptive report with a plain-language summary.
7. **Keep it current**
   Create records for included studies and update the map regularly, annually where possible.

WORKED EXAMPLE · AN ILLUSTRATION

A state department of school education and its funders want to decide where to invest in foundational learning. Instead of commissioning a new trial straight away, they commission an evidence and gap map. Rows list interventions such as structured pedagogy, teacher coaching, remedial camps, school meals and community libraries; columns list outcomes such as attendance, reading fluency, numeracy and dropout; filters cover South Asia, girls and study design. The finished map shows dense clusters of studies on structured pedagogy and reading, several reviews on school meals and attendance, and almost nothing on community libraries or on learning among tribal children. The department commissions a rapid review of the dense cluster to inform this year's plan, and directs new evaluation money to the empty cells.

Did you know?

Several speaker biographies list his interests outside work as long-distance running, walking in remote places and reading English history.

USE IT WHEN

- Setting a research or evaluation agenda for a sector or a funder.
- Deciding whether a systematic review is feasible or a new primary study is needed.
- Briefing decision-makers who need an overview of the field, not a verdict.

SKIP IT WHEN

- You need to know whether an intervention works, which calls for a systematic review or a study.
- The question is narrow enough for a single review.
- You lack the time or team for systematic searching and screening, which the guidance puts at six to twelve months.

COMMON MISREADINGS

- Reading a large bubble as proof that an intervention works, when it shows only that it has been studied often.
- Treating an empty cell as evidence of no effect, rather than as an absence of evidence.

WHY IT WAS GROUNDBREAKING

Maps let funders and governments see the shape of an entire evidence base before paying for new studies. They turned evidence gaps from a closing line in a literature review into a planning tool. In our reading, they also made synthesis visible to people who will never read a systematic review.

Origin

3ie dates its invention of evidence gap maps to 2011 and published its first five maps in 2015, the year White handed over its leadership. At the Campbell Collaboration maps became a formal product, and in 2020 White led ten co-authors in publishing Guidance for producing a Campbell evidence and gap map in Campbell Systematic Reviews 16(4). The format is a team achievement, not his alone.

RELATED IDEAS

DC Campbell 112
D&B Duflo and Banerjee 432
SS Sridharan 246
PR Rogers 268

CONCEPT HIGHLIGHT

Counterfactual

An estimate of what would have happened to the same people without the intervention, usually built from a comparison or control group.

3

BIG IDEA 3 OF 3 · WHITE

The four waves of the evidence revolution

See the evidence movement as four building waves, results, impact evaluation, systematic reviews and knowledge brokering, and recognise that supplying evidence is not enough without demand for its use.

PICTURE IT

Think of a library full of books that nobody borrows. Building the library was only half the job. White's four waves end with knowledge brokering, because evidence changes nothing until people who make decisions come looking for it.

The problem

By the late 2010s monitoring units, impact evaluators, systematic reviewers and What Works centres often worked as separate communities, each treating its own tool as the answer. Practitioners struggled to see how results frameworks, trials and reviews fitted together, and governments with impressive data systems still took decisions without evidence. Without a shared picture of the whole system, investment kept flowing into producing more studies, while the harder problem of getting evidence used by the people who decide received far less attention.

What it is

In a 2019 paper in Palgrave Communications 5, written while he led the Campbell Collaboration and offered as a personal perspective drawn mainly from international development, White described an evidence revolution unfolding over about 30 years in four waves. The first was the results agenda of the 1990s, linked to New Public Management, which made performance and outcomes visible. The second was the rise of rigorous impact evaluation, especially randomised controlled trials, from the early 2000s. The third was the growth of systematic reviews, which pool the findings of many studies. The fourth was knowledge brokering: institutionalising the use of evidence, most visibly through the What Works movement in the United States and United Kingdom. Each wave builds on the one before, and together they form the supply side of an evidence architecture. White suggested that artificial intelligence, machine learning and

big data could drive a fifth wave, and proposed demand-side measures such as Evidence Needs Assessments and awards for the use of evidence.

How it works

1. **Locate your system**
   Ask which waves are already present, for example a results framework and a few impact evaluations but no reviews and no brokering unit.
2. **Strengthen results data**
   Make sure monitoring tracks outcomes credibly, since every later wave depends on it.
3. **Add impact studies where they answer priority questions**
   Commission rigorous evaluations with a credible counterfactual on the decisions that matter most.
4. **Synthesise before commissioning**
   Use systematic reviews and evidence maps so that any single study is read in the light of the wider evidence.
5. **Broker knowledge**
   Create or link to a unit that turns evidence into briefings and advice timed to decisions.
6. **Build demand**
   Following White's 2019 proposals, run an Evidence Needs Assessment with decision-makers and recognise good use of evidence, for example through awards.

WORKED EXAMPLE · AN ILLUSTRATION

A state rural livelihoods mission reviews its evidence system. It finds strong first-wave foundations: a management information system tracks self-help group formation, credit and incomes. Second-wave work is patchy: two impact evaluations exist, both commissioned by donors. There is no third-wave synthesis and no fourth-wave brokering unit. The mission's MEL team proposes a staged plan. It begins with an Evidence Needs Assessment, asking district and state officials which decisions they find hardest. It then commissions an evidence gap map on livelihoods interventions in South Asia and sets up a small unit that writes two-page evidence briefs before each annual planning cycle. The officials' answers, not the evaluators' interests, shape the next round of impact studies. A fifth-wave pilot using machine learning to flag struggling groups is left for later.

CONCEPT HIGHLIGHT

Causal chain

The sequence of links from inputs to activities, outputs, outcomes and impact that a programme assumes; theory-based impact evaluation tests each link.

USE IT WHEN

- Diagnosing the evidence system of a ministry, mission or large organisation.
- Making the case for investing in synthesis or knowledge brokering rather than more single studies.
- Helping new MEL staff see how their own work fits a wider system.

SKIP IT WHEN

- You need a method for designing a single evaluation.
- Basic results data are missing, so the first wave must be built before the others make sense.

COMMON MISREADINGS

- Reading the waves as replacing one another, when each builds on the one before.
- Assuming that more supply of evidence will automatically lead to use, which is precisely the gap the demand-side proposals address.

WHY IT WAS GROUNDBREAKING

It gave practitioners a historical map of how monitoring, impact evaluation and synthesis fit into one system. It shifted attention from producing evidence to the harder work of creating demand for it and getting it used.

Origin

White published the framework in 2019 as The twenty-first century experimenting society: the four waves of the evidence revolution, in Palgrave Communications 5 (the journal is now Humanities and Social Sciences Communications), while chief executive of the Campbell Collaboration. In our reading, it distils a career that had passed through three of the waves: impact evaluation at the World Bank and 3ie, synthesis at Campbell, and work on evidence use.

RELATED IDEAS

DC Campbell 112
CW Weiss 152
MP Patton 204
RR Rist 410
D&B Duflo and Banerjee 432

CONCEPT HIGHLIGHT

Heterogeneity

Differences in how a programme affects different groups, places or circumstances, which an average effect can hide.

RELATED GIANTS



Huey-tsyh Chen

The sociologist who refused black-box verdicts and asked evaluators to test how a programme is supposed to work.

p. 174



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Donald T. Campbell

The psychologist who gave field evaluators a disciplined vocabulary for doubt: threats to validity.

p. 112



Esther Duflo and Abhijit Banerjee

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

p. 432

CONCEPT HIGHLIGHT

Factual analysis

Rigorous description of what actually happened along the causal chain, such as who was reached and what they did, as distinct from counterfactual estimation.

CONCEPT HIGHLIGHT

Systematic review

A study that locates, appraises and synthesises all research meeting explicit criteria on a defined question, to say what the body of evidence shows.

CONCEPT HIGHLIGHT

Evidence and gap map

A visual, systematically produced grid showing where studies exist and where they are missing; it maps evidence rather than summarising findings.

Put it to work

USE IT ON MONDAY · THEORY-BASED IMPACT EVALUATION AND EVIDENCE SYNTHESIS

1. Before designing a new study, search for systematic reviews and evidence gap maps on the intervention. Tell the client what is already known and where the real gap is.

2. Draw the causal chain from inputs to outcomes and mark which links are well evidenced elsewhere and which need testing in this setting.

3. Use a counterfactual for the final outcome where feasible, and process and qualitative data for the intermediate links, so the study explains as well as measures.

AI prompt **Copy, fill the brackets, then check every line**

Help me build a theory-based impact evaluation plan, following the principles in Howard White's 2009 paper, for **[programme]** in **[location]**. Here is the programme's theory of change: **[paste]**. Here is a list of existing studies, reviews or evidence maps I have found, with their key findings as I understand them: **[paste list]**. (1) Lay out the causal chain and the assumptions behind each link. (2) Using only the evidence I have listed, mark which links are supported elsewhere and which need testing here. (3) Suggest methods for each untested link and a credible counterfactual for the main outcome. (4) Suggest subgroups to examine. Do not cite studies I have not provided, and say so if my evidence list looks too thin.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a string of fairy lights that will not switch on. Knowing the string is dead is not enough; you need to find the faulty bulb. Theory-based impact evaluation measures whether the lights came on and then tests each bulb in the chain, to show exactly where it broke.

YOUR MOVES

Three moves from White's path

1 Build the infrastructure of evidence

White moved from academic research to leading impact evaluation at the Independent Evaluation Group and then became 3ie's first executive director.

TRY THIS
Contribute to one shared evidence resource: an evidence map, a repository or a review.

2 Step back and see the whole body of work

Leading the Campbell Collaboration from 2015 to 2022, he moved from single studies to synthesis.

TRY THIS
Before your next study, read what existing reviews already say about the question.

3 Build evidence with the countries it serves

From 3ie's New Delhi office to recognition by the governments of Benin and Uganda, his evidence work gained weight in the Global South.

TRY THIS
Involve national or state government counterparts in shaping your evaluation questions from the start.

THE BIG IDEAS OF THIS CHAPTER

1 **Theory-based impact evaluation** p. 484
Estimate a programme's impact against a credible counterfactual and test, link by link, the causal chain that is supposed to produce it, so the study explains why as well as whether.

2 **Evidence and gap maps** p. 487
Systematically chart where evidence on a policy area exists and where it is missing, so that decisions and new research start from the whole evidence base.

3 **The four waves of the evidence revolution** p. 490
See the evidence movement as four building waves, results, impact evaluation, systematic reviews and knowledge brokering, and recognise that supplying evidence is not enough without demand for its use.

CONCEPT HIGHLIGHT

Evidence and gap map

A visual, systematically produced grid showing where studies exist and where they are missing; it maps evidence rather than summarising findings.

GIANTS CONNECTED TO THIS CHAPTER



Huey-tsyh Chen

Theory-driven evaluation supplied the programme theory side of the pairing.

p. 174



Esther Duflo and Abhijit Banerjee

Randomised evaluations supplied much of the counterfactual evidence base.

p. 432



Donald T. Campbell

The threats-to-validity tradition behind credible counterfactuals.

p. 112

KEY WORKS TO READ IN THE ORIGINAL

- **Theory-based impact evaluation: principles and practice** 2009, Journal of Development Effectiveness 1(3)
- **Journal of Development Effectiveness** Founding managing editor
- **3ie evidence gap maps** 2015, First five maps published by 3ie

CONCEPT HIGHLIGHT

Counterfactual

An estimate of what would have happened to the same people without the intervention, usually built from a comparison or control group.

CONCEPT HIGHLIGHT

Causal chain

The sequence of links from inputs to activities, outputs, outcomes and impact that a programme assumes; theory-based impact evaluation tests each link.

WORKSHEET · TRY THIS TEMPLATE

Causal chain test plan

Print this page, or copy the headings into your notes.

1 Link in the chain

For example, from training delivered to practices changed.

2 Assumption

What must hold for this link to work.

3 Existing evidence

What reviews or gap maps already say, with references.

4 Evidence we will collect

Data source, method and timing.

5 Counterfactual needed?

Yes or no, and what comparison is credible here.

6 Subgroups to examine

Groups for whom the link may work differently.

REFLECT

Before you design your next study, what does the existing evidence already say?

CONNECTS TO

HC **Huey-tsyh Chen** Theory-driven evaluation supplied the programme theory side of the pairing. p. 174

D&B **Esther Duflo and Abhijit Banerjee** Randomised evaluations supplied much of the counterfactual evidence base. p. 432

DC **Donald T. Campbell** The threats-to-validity tradition behind credible counterfactuals. p. 112

22



Emblem: an original HubForge design for this figure, not a likeness.

Bagele M. Chilisa

Active 2010s to 2020s Botswana; research and evaluation across Africa

Made in the Global South Participation, equity and voice

The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

IN ONE LINE

Bagele Chilisa gave the communities that research once studied a say in deciding what counts as knowledge.

What if the research methods you were taught carried hidden assumptions about whose knowledge counts? For students in many African universities, that question had no textbook. Then, in 2012, a professor at the University of Botswana wrote one. Bagele Chilisa's Indigenous Research Methodologies named the colonial assumptions inside standard research methods and offered alternatives drawn from African and other indigenous knowledge: proverbs, folk stories, counter-narratives and the idea of the researcher as healer. She knew the mainstream from the inside, having trained in research design, statistics and measurement at the University of Pittsburgh. The second edition, published in November 2019, added chapters on decolonising evaluation. In 2018 she was appointed to UNDP's Evaluation Advisory Panel, and she became principal investigator of a Mastercard Foundation funded consortium to develop evaluation tools rooted in African traditions.

BIG IDEAS

1. **Postcolonial indigenous research paradigm**
2. **African relational evaluation**

KEY WORKS

Indigenous Research Methodologies
2012 · Second edition 2019

Decolonising and Indigenising Evaluation Practice in Africa: Toward African Relational Evaluation Approaches
2016

Made in Africa Evaluation, African Evaluation Journal 12(2)
2024 · Commissioned and funded by EvalPartners

Life CHAPTER 22

Bagele Chilisa grew up in Botswana, where she went to primary and secondary school. Her professional story begins with a decision to master the mainstream research methods she would later question.

She studied at the University of Pittsburgh in the United States, earning an MA in research methodology, covering research design, statistics and measurement, and a PhD in policy, planning and evaluation. She then joined the University of Botswana, where she rose to full professor in the Post Graduate Research and Evaluation Programme, teaching research methods, educational evaluation, questionnaire and survey design, project design and

evaluation, and policy planning. She has served on the boards of the university's Centre for Scientific Research, Indigenous Knowledge and Innovations and of its Research Ethics Committee.

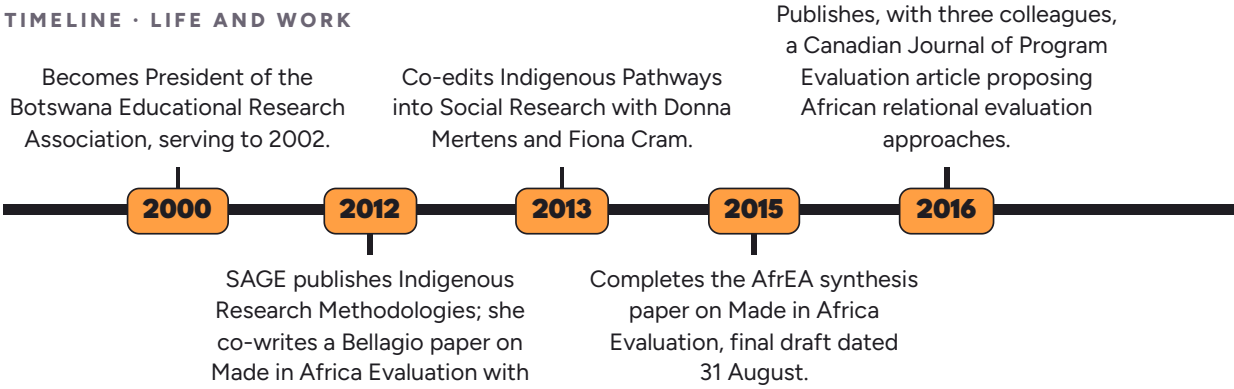
Her professional service began in educational research. She was a founding member of the Botswana, Lesotho, Swaziland and Namibia Research Association, President of the Botswana Educational Research Association from 2000 to 2002 and editor of the Botswana Educational Research Journal. As an evaluator she has worked for, among others, UNICEF, WHO, the UN Economic Commission for Africa, the World Bank, ILO, UNESCO, UNDP, DFID and the Forum for African Women Educationalists, and took part in the African Development Bank's evaluation of its assistance to Botswana (2004 to 2013) and the Netherlands WOTRO joint MFS II evaluations (2012 to 2015).

From 2012 her writing turned conference debates into theory: her textbook Indigenous Research Methodologies, the 2015 AfrEA synthesis paper on Made in Africa Evaluation, and a series of articles on African relational evaluation. In 2018 she was appointed to the UNDP Independent Evaluation Office's Evaluation Advisory Panel. She has been principal investigator of the Pan African Indigenous Evaluation Consortium, funded by the Mastercard Foundation with about P10 million, and in May 2024 she opened the London School of Hygiene and Tropical Medicine's lecture series on decolonial global health evaluation.

The spark

Chilisa's early research included work on gender-fair assessment in 2000 and a 2005 critique, from a postcolonial standpoint, of HIV and AIDS research in Botswana. In our reading, that work exposed the problem she would pursue for two decades: research methods built elsewhere could misread the very communities they were meant to serve, and African scholars had few written alternatives to teach.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 22

Imagine studying your own community with tools designed to study strangers. The survey assumes that each person answers alone, when decisions are made by families. The consent form assumes individual ownership of a story that belongs to a village. The analysis treats local knowledge as opinion and outside theory as fact. Reviewers of Chilisa's textbook summarise the problem it addressed: Euro-Western methods could disconnect researchers from communities, treat formerly colonised and indigenous peoples as objects, and leave African researchers trained in the West reproducing colonial knowledge systems without noticing. In evaluation, the 2007 Niamey call for Africa-rooted practice had produced resolutions but little written theory.

Her key idea appeared in Indigenous Research Methodologies, published by SAGE in 2012, which set out a postcolonial indigenous research paradigm in which knowledge is relational and holistic, and offered methods such as the use of proverbs, folk stories and counter-narratives, and the researcher as healer. A second edition in November 2019 added chapters on decolonising evaluation and on indigenous mixed methods in programme evaluation. She carried the idea into evaluation in a 2012 Bellagio paper with Chiku Malunga, the 2015 AfrEA synthesis paper on Made in Africa Evaluation, and a 2016 Canadian Journal of Program Evaluation article proposing an African relational evaluation paradigm.

In one line

Bagele Chilisa gave the communities that research once studied a say in deciding what counts as knowledge.

TIMELINE, CONTINUED



The problem it solved

Research and evaluation designed in the Global North that ignored or misrepresented indigenous communities and missed real effects on their lives and livelihoods.

Impact WHAT CHANGED

Chilisa's textbook became a widely adopted graduate text: it is in its second edition, has been reviewed in the African Evaluation Journal and the International Journal of Critical Indigenous Studies, and appears on reading lists for qualitative and indigenous methods courses outside Africa. Her 2013 co-edited volume Indigenous Pathways into Social Research, with Donna Mertens and Fiona Cram, and her 2021 American Journal of Evaluation article with Mertens on indigenous Made in Africa frameworks carried African perspectives into mainstream evaluation journals. The University of Botswana reports that evaluators in South Africa's Community Development Workers Programme proposed using her publications to discuss Africanising development evaluation.

Our interpretation is that her main contribution to the Global South is to make decolonial evaluation teachable and fundable. The Pan African Indigenous Evaluation Consortium, linking the University of Botswana, Africa Nazarene University and the University of Ghana with consultancies and AfrEA, was set up to develop tools and protocols rooted in African traditions, and her seat on UNDP's Evaluation Advisory Panel places these concerns inside a major UN evaluation office. For evaluators in India and other postcolonial settings, her framework offers a vocabulary for asking whose knowledge defines a good outcome.

KEY WORKS TO READ IN THE ORIGINAL

- **Indigenous Research Methodologies** 2012, Second edition 2019
- **Decolonising and Indigenising Evaluation Practice in Africa: Toward African Relational Evaluation Approaches** 2016
- **Made in Africa Evaluation, African Evaluation Journal 12(2)** 2024, Commissioned and funded by EvalPartners

WHERE PEOPLE DISAGREE

Debates

Chilisa engages her critics directly. Her 2024 article discusses Uwizeyimana (2020), who argued that Ubuntu is not shared by all groups living in Africa, so an Ubuntu-based approach may not suit everyone, and that indigenous causal beliefs cannot be scientifically proven. She responds that Ubuntu reflects a widely shared relational outlook and that dismissing community beliefs is culturally insensitive. A 2012 review of her textbook judged that it covers too much, leaving topics such as validity thin. She herself records that Made in Africa Evaluation remains rarely applied in practice.

LIMITS, IN OUR READING

- ☐ Her own 2024 article records that the 2015 Made in Africa synthesis paper remained a draft years later and was not widely known, so uptake has been uneven.
- ☐ The same article discusses a critique of Ubuntu-based evaluation (Uwizeyimana, 2020), showing that African philosophical foundations for evaluation are contested, not settled.

WHY THIS MATTERS TO YOU

Chilisa's ideas help anyone who works with communities whose ways of knowing differ from their own training. Start by examining the assumptions built into your own methods. Agree relationships and protocols with a community before you collect anything. Let people help define what success means for them. And remember that the effects of a programme travel through families and communities, not only through individuals. For evaluators working with Adivasi communities, rural women's collectives or any group too often studied rather than heard, her work opens the door to partnerships in which communities co-own the knowledge.

2

editions of Indigenous Research Methodologies, the second published in November 2019

SAGE Publishing; African Evaluation Journal review (2013)

3

new chapters on evaluation and mixed methods added to the 14-chapter second edition

SAGE Publishing

P10 million

approximate Mastercard Foundation funding for the Pan African Indigenous Evaluation Consortium she leads as principal investigator

University of Botswana

Key concepts

Indigenous research methodologies
Approaches to inquiry that draw on the knowledge systems, values and ways of knowing of indigenous and formerly colonised peoples.

Relational ontology, epistemology and axiology
The view that reality, knowledge and ethics are all grounded in relationships among people, communities and the environment.

Methodological colonialism
The imposition of research and evaluation methods developed in the Global North without regard to local worldviews; a theme of her 2024 LSHTM lecture.

Ubuntu
A southern African philosophy holding that a person's humanity is realised through relationships with others; one of several African worldviews her work draws on.

Evaluation tree
A metaphor, adapted from Marvin Alkin, that arranges evaluation theorists on branches; Chilisa proposed African relational and context and needs branches.

Epistemic violence
Harm done when one culture's way of knowing is used to judge and override another's, a concept used in her 2021 article with Donna Mertens.

Sources for this chapter: p. 587

Did you know?

Her academic training in the United States was quantitative at its core: her Pittsburgh MA covered research design, statistics and measurement, and the second edition of her indigenous methods textbook expanded its material on indigenous quantitative research.

THE BIG IDEAS OF THIS CHAPTER

1 **Postcolonial indigenous research paradigm** p. 506

Research and evaluation should rest on relational ways of knowing and on accountability to communities, decolonising the Euro-Western assumptions built into standard methods.

2 **African relational evaluation** p. 509

Ground evaluation theory in African relational worldviews, and give African theorists their own branches on the map of evaluation thinking.

GIANTS CONNECTED TO THIS CHAPTER



Zenda Ofir

Co-originator of the Made in Africa Evaluation idea whose synthesis Chilisa wrote.

p. 454



Donna M. Mertens

Co-editor with Chilisa of Indigenous Pathways to Social Research (2013).

p. 384



Robert Chambers

An earlier challenge to whose reality counts in development knowledge.

p. 296

1

BIG IDEA 1 OF 2 · CHILISA

Postcolonial indigenous research paradigm

Research and evaluation should rest on relational ways of knowing and on accountability to communities, decolonising the Euro-Western assumptions built into standard methods.

PICTURE IT

Think of a map of a village drawn by an outsider from a satellite photo, and another drawn by the elders who know every path, sacred tree and boundary. Both are maps, but only one shows what matters to the people who live there. Chilisa's paradigm asks research to start from the second kind of map.

The problem

Social research and evaluation taught in African universities were built on Euro-Western paradigms. Reviewers of Chilisa's textbook describe the consequences: researchers disconnected from communities, formerly colonised peoples treated as objects or others, and Western-trained African scholars reproducing colonial knowledge systems without noticing. Chilisa's own 2005 critique of HIV and AIDS research in Botswana raised such concerns in a field of urgent public health importance. Students had no single text that named these assumptions and offered alternatives grounded in their own societies.

What it is

Chilisa's postcolonial indigenous research paradigm, set out in Indigenous Research Methodologies, starts from the claim that every method carries assumptions about reality, knowledge and ethics, and that dominant assumptions were shaped by colonial history. Its parts follow the classic elements of a paradigm. A relational ontology holds that people exist in relationships with one another, with the living and the non-living, and with the environment. A relational epistemology holds that knowledge is produced through those relationships and is historically located. A relational axiology makes ethics a matter of accountability to relationships, expressed in four Rs: relational accountability; respectful representation; reciprocal appropriation, meaning benefits must flow back to communities; and rights and regulations that give communities ownership of research and knowledge. Methodologically, the paradigm decolonises standard tools such as the interview and makes room for proverbs,

folk stories, counter-narratives and the researcher as healer, alongside participatory, feminist and mixed methods. Its logic is that research which honours relationships is both more ethical and more valid.

How it works

1. **Examine your paradigm**
   Identify the assumptions about reality, knowledge and values in your chosen design, as the textbook's opening chapters on knowledge systems and research paradigms ask.
2. **Read the literature responsibly**
   Ask whose knowledge existing studies represent and where communities have been othered, the theme of the chapter on discovery and recovery.
3. **Agree relationships and protocols**
   Settle with community gatekeepers how the work will proceed, who owns the knowledge and how benefits return, applying the four Rs.
4. **Choose decolonised methods**
   Adapt tools such as interviews and add indigenous sources such as proverbs, stories and counter-narratives, alongside surveys where they fit.
5. **Integrate knowledge systems**
   Combine indigenous and Western methods deliberately, as the second edition's chapters on decolonising mixed methods and on indigenous mixed methods in programme evaluation describe.
6. **Return findings and reflect**
   Take interpretations back to the community for correction, and examine your own position in power structures before publishing.

WORKED EXAMPLE · AN ILLUSTRATION

A university team in Telangana evaluates a tribal health programme that trains community health volunteers in Adivasi villages. Following a relational paradigm, the team first meets village councils and elders to agree protocols: which places and topics are sensitive, who owns recordings and how findings will come back. Alongside a household survey on clinic visits, researchers collect local stories and sayings about illness and healing, which suggest that volunteers are trusted when they work with traditional healers rather than against them. The team revises its interview guide to drop questions that framed traditional practice as superstition. Draft findings are discussed in village meetings, and the community's corrections are recorded in the report. Benefits return as a plain-language summary in the local language and a joint presentation to the district health office.

USE IT WHEN

- Research or evaluation involves indigenous, tribal or formerly colonised communities.
- Standard instruments carry outsider assumptions about family, health, land or success.
- You are teaching research methods and want students to examine the assumptions behind them.

SKIP IT WHEN

- The team cannot commit to protocols, shared ownership and returning findings, which the paradigm treats as essential rather than optional.
- The aim is to apply one community's worldview to diverse groups without their agreement; critics such as Uwizeyimana (2020) warn against assuming any single philosophy is shared by all.

COMMON MISREADINGS

- [ ] Treating it as a toolkit of proverbs and stories that adds colour while power relationships and ownership of knowledge stay unchanged.
- [ ] Assuming it rejects quantitative or Western methods, when the second edition expands indigenous quantitative and mixed methods.

WHY IT WAS GROUNDBREAKING

It gave students in formerly colonised societies a single, teachable text that named the colonial assumptions in standard methods and offered workable alternatives. It placed African experience alongside other indigenous scholarship as a source of theory, not merely a case study.

Origin

The paradigm appeared in Indigenous Research Methodologies (SAGE), dated 2011 in Chilisa's own reference lists and 2012 in some catalogues and reviews. The second edition, published in November 2019 with 392 pages and 14 chapters, added chapters on decolonising evaluation, decolonising mixed methods and indigenous mixed methods in programme evaluation, and expanded coverage of indigenous quantitative research.

RELATED IDEAS

African relational evaluation

- DM Mertens 384
- RC Chambers 296
- ZO Ofir 454
- SZ Zaveri 536

CONCEPT HIGHLIGHT

Indigenous research methodologies

Approaches to inquiry that draw on the knowledge systems, values and ways of knowing of indigenous and formerly colonised peoples.

2

BIG IDEA 2 OF 2 · CHILISA

African relational evaluation

Ground evaluation theory in African relational worldviews, and give African theorists their own branches on the map of evaluation thinking.

PICTURE IT

Think of the tree diagrams used to show how evaluation theories grew from their roots. For years, almost every branch was North American or European. African relational evaluation adds new branches grown from African worldviews, where people are understood through their relationships.

The problem

By the 2010s culturally responsive evaluation had become a global theme, yet, as Chilisa and colleagues noted in 2016, African voices were largely absent from it. The best-known map of evaluation theory, Marvin Alkin's evaluation tree, featured mainly North American and European names. The 2007 Niamey call for Africa-rooted evaluation had produced resolutions but little written theory. Without theory, African evaluators had little to cite in terms of reference or to teach in classrooms, and commissioners had little reason to change practice.

What it is

African relational evaluation grounds evaluation in African worldviews in which a person exists through relationships with others, the community, the environment and the spiritual world. In a 2016 Canadian Journal of Program Evaluation article, Chilisa, Thenjiwe Major, Michael Gaotlhobogwe and Hildah Mokgolodi set out this paradigm, illustrated it with relational evaluation studies and proposed an African evaluation tree drawing on methods developed in Africa by African theorists. The tree adapts the familiar branches of methods, values and use, to which Mertens had added social justice; Chilisa and Mertens (2021) proposed a further needs and context branch for approaches rooted in the cultures and histories of formerly colonised peoples. Chilisa's 2024 article summarises the paradigm: realities that link the living and the non-living; relational, historically located knowledge; values of social, epistemic and ecological justice; and a participatory methodology mixing indigenous and Western

methods, including the Lekgotla consensus forum. Its logic is that relationships supply evaluative criteria, not merely a setting for data collection.

How it works

1. **Map the relational web**
   Identify the relationships the programme touches, among individuals, families, communities and the environment, since the paradigm treats them as part of what is evaluated.
2. **Ask whose programme it is**
   Apply the reciprocity question in Chilisa's 2024 list of AfrEA principles: who initiated the programme, who owns it and who benefits.
3. **Set relational criteria**
   With communities, define success in relational terms, such as belonging, interdependence and harmony, alongside individual outcomes.
4. **Use consensus methods**
   Gather and interpret evidence through forums such as the Lekgotla, which Chilisa's article names as a main tool for building relationships and consensus, mixed with surveys and other methods.
5. **Respect beliefs in interpretation**
   Seek to understand participants' own explanations of cause, including spiritual ones, rather than dismissing them, as Chilisa argues in reply to her critics.
6. **Place and document the design**
   Locate the evaluation among theories on the tree, including the needs and context branch, and publish it so that African approaches build a citable record.

WORKED EXAMPLE · AN ILLUSTRATION

An NGO in Botswana runs a youth enterprise programme, and its funder asks for a report on individual income gains. Applying an African relational approach, the evaluator first asks village leaders and young people whose programme it is and who benefits from it. In a series of community consensus forums modelled on the Lekgotla, participants say that a successful young entrepreneur also supports parents, contributes to community events and helps younger siblings. The team adds these relational outcomes to the income survey, interviews families as well as participants, and records how the businesses affect relations between generations. The final report shows modest income gains but strong effects on family support, a result an individual income measure would have missed, and explains to the funder why both matter.

USE IT WHEN

- Programmes operate in communal settings where outcomes travel through families and communities.
- A funder's individual-level indicators miss what participants and communities value.
- Evaluators want to build and document theory grounded in their own region.

SKIP IT WHEN

- ○ The people involved do not share the relational worldview in question; Uwizeyimana (2020) warned that Ubuntu is not held by every group in Africa.
- ○ The commissioner will not accept relational criteria and they cannot be negotiated, in which case say plainly that the design is conventional.

COMMON MISREADINGS

- ☐ Equating it with Ubuntu alone, when Chilisa presents Ubuntu as one of several relational worldviews found across Africa under different names.
- ☐ Treating it as just another participatory approach; Chilisa lists that view among the myths that slow the uptake of Made in Africa Evaluation.

WHY IT WAS GROUNDBREAKING

It inserted African theorists into a map of evaluation theory that had featured almost only North American and European names. It argued that relationships are a source of evaluative criteria, not just a context for collecting data.

Origin

Chilisa and Chiku Malunga proposed a relational branch in a 2012 Bellagio paper on Made in Africa Evaluation. The 2016 Canadian Journal of Program Evaluation article, volume 30, issue 3, set out the African relational evaluation paradigm and an African evaluation tree. Chilisa and Mertens proposed a needs and context branch in the American Journal of Evaluation in 2021, and Chilisa's 2024 African Evaluation Journal article mapped a regional Made in Africa tree.

RELATED IDEAS

Postcolonial indigenous research paradigm

Made in Africa Evaluation

ZO Ofir 454

DM Mertens 384

RC Chambers 296

CONCEPT HIGHLIGHT

Relational ontology, epistemology and axiology

The view that reality, knowledge and ethics are all grounded in relationships among people, communities and the environment.

Put it to work

USE IT ON MONDAY · INDIGENOUS AND AFRICAN RELATIONAL EVALUATION

1. Put relationships first in your work plan: time for introductions, community protocols and agreement on how the evaluation will proceed, before any data collection.

2. Ask the community who defines a good outcome and how they would recognise it, and bring those definitions into your indicators and judgement criteria.

3. Agree upfront who owns the knowledge produced, how it will be returned, and how the community can challenge your interpretation.

AI prompt **Copy, fill the brackets, then check every line**

Help me design a relational evaluation approach for **[programme]** in **[community and country]**, drawing on Bagele Chilisa's work on indigenous research methodologies and African relational evaluation. Here is what the community and local partners have told me about their protocols, values and priorities: **[paste notes]**. Here is the draft design: **[paste]**. (1) Identify where the design may conflict with what the community has shared. (2) Suggest how community definitions of a good outcome could shape criteria and indicators. (3) Propose arrangements for knowledge ownership, consent and return of findings. (4) Suggest questions to ask local partners where my notes are silent. Do not describe or generalise about any community's beliefs beyond what my notes contain.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a map of a village drawn by an outsider from a satellite photo, and another drawn by the elders who know every path, sacred tree and boundary. Both are maps, but only one shows what matters to the people who live there. Chilisa's paradigm asks research to start from the second kind of map.

YOUR MOVES Three moves from Chilisa's path

1 Master the mainstream, then challenge it

Trained in research design, statistics and measurement at Pittsburgh, Chilisa went on to write a leading text on Indigenous research methodologies.

TRY THIS
Learn the dominant method in your field thoroughly, then ask what it leaves out in your context.

2 Write the idea down

By turning conference resolutions into a textbook, a synthesis paper and journal articles, she gave others something to cite, test and teach.

TRY THIS
Turn your best idea from a recent meeting into a two-page paper and share it.

3 Carry ideas through partnerships

Her Pan African Indigenous Evaluation Consortium linked universities, consultancies and AfrEA, taking her ideas beyond one campus.

TRY THIS
Connect one university, one practitioner group and one network around a shared evaluation question.

THE BIG IDEAS OF THIS CHAPTER

1. **Postcolonial indigenous research paradigm** p. 506
   Research and evaluation should rest on relational ways of knowing and on accountability to communities, decolonising the Euro-Western assumptions built into standard methods.

2. **African relational evaluation** p. 509
   Ground evaluation theory in African relational worldviews, and give African theorists their own branches on the map of evaluation thinking.

CONCEPT HIGHLIGHT

Epistemic violence

Harm done when one culture's way of knowing is used to judge and override another's, a concept used in her 2021 article with Donna Mertens.

GIANTS CONNECTED TO THIS CHAPTER



Zenda Ofir

Co-originator of the Made in Africa Evaluation idea whose synthesis Chilisa wrote.

p. 454



Donna M. Mertens

Co-editor with Chilisa of Indigenous Pathways to Social Research (2013).

p. 384



Robert Chambers

An earlier challenge to whose reality counts in development knowledge.

p. 296

KEY WORKS TO READ IN THE ORIGINAL

- **Indigenous Research Methodologies** 2012, Second edition 2019
- **Decolonising and Indigenising Evaluation Practice in Africa: Toward African Relational Evaluation Approaches** 2016
- **Made in Africa Evaluation, African Evaluation Journal 12(2)** 2024, Commissioned and funded by EvalPartners

CONCEPT HIGHLIGHT

Indigenous research methodologies

Approaches to inquiry that draw on the knowledge systems, values and ways of knowing of indigenous and formerly colonised peoples.

CONCEPT HIGHLIGHT

Relational ontology, epistemology and axiology

The view that reality, knowledge and ethics are all grounded in relationships among people, communities and the environment.

WORKSHEET · TRY THIS TEMPLATE

Relational evaluation plan

Print this page, or copy the headings into your notes.

1 Relationships and protocols

Whose permission, which customs and what time are needed before starting.

2 Local knowledge and worldviews

Ways of knowing to respect and learn from, as described by the community.

3 Who defines a good outcome

How community definitions will shape criteria and indicators.

4 Language and concepts

Key terms that do not translate directly, and how you will handle them.

5 Ownership and return of knowledge

Data ownership, storage, consent and how findings go back.

6 Accountability to the community

How the community can question or correct the evaluation.

REFLECT

In your evaluation, whose way of knowing decides what counts as evidence?

CONNECTS TO

Zenda Ofir (ZO)
Co-originator of the Made in Africa Evaluation idea whose synthesis Chilisa wrote. p. 454

Donna M. Mertens (DM)
Co-editor with Chilisa of Indigenous Pathways to Social Research (2013). p. 384

Robert Chambers (RC)
An earlier challenge to whose reality counts in development knowledge. p. 296

23



Emblem: an original HubForge design for this figure, not a likeness.

Marco Segone

Active 2010s to 2020s

Worked in Niger, Brazil, Latin America and the Caribbean, Europe and Central Asia, and with NGOs in Asia, Africa and Europe

Impact evaluation and evidence systems | Participation, equity and voice

The UN evaluator who helped turn scattered evaluation associations into a global movement for national evaluation capacity and equity.

IN ONE LINE

Marco Segone helped turn scattered evaluation associations into a movement, and asked every evaluation to look first at those left behind.

Accra, April 2012. At a board meeting of the International Organization for Cooperation in Evaluation, board members and UNICEF conceive a partnership called EvalPartners. Marco Segone, then at UNICEF's Evaluation Office, becomes a co-founder and co-chair. The idea was simple and ambitious. Associations of evaluators had multiplied around the world, from 15 in 1999 to more than 188 by the end of 2013, yet most worked in isolation from the governments whose policies they evaluated. What if they acted together? Within three years, the movement had declared 2015 the International Year of Evaluation, the UN General Assembly had adopted resolution 69/237 on evaluation capacity in December 2014, and 92 events had been organised worldwide to mark the year. Segone went on to lead UNFPA's Independent Evaluation Office until 2025.

Life CHAPTER 23

Marco Segone holds a BA and an MSc in political science. He first worked on integrated development projects with international NGOs in Bangladesh, Pakistan, Thailand, Uganda and Albania.

He joined UNICEF's Regional Office for Latin America and the Caribbean in 1996 and served as monitoring and evaluation officer for UNICEF Niger from 1999 to 2001, where his work included a participatory evaluation of child poverty. He went on to roles in the Brazil country office and as Regional Chief of Monitoring and Evaluation for Latin America and the Caribbean and then for Europe and Central Asia. In UNICEF's Evaluation Office he was responsible for the decentralised evaluation function and the national evaluation capacity development portfolio. In this period he edited a series of UNICEF volumes on country-led monitoring and evaluation systems with IDEAS, IOCE and other

BIG IDEAS

1. **Country-led monitoring and evaluation systems**
2. **Equity-focused evaluation**
3. **A global movement for evaluation**

KEY WORKS

How to Design and Manage Equity-Focused Evaluations

Evaluation for Equitable Development Results

Evaluating the SDGs with a no one left behind lens

DID YOU KNOW?

To mark the International Year of Evaluation, an Evaluation Torch was lit at the UN Secretariat building in December 2014 and passed through evaluation events during 2015; Segone handed it on in Kathmandu that November.



partners (2008 to 2010), and co-wrote, with Michael Bamberger, the 2011 guide How to Design and Manage Equity-Focused Evaluations. He also served as a Vice-President of IOCE.

In April 2012, at an IOCE board meeting in Accra, IOCE board members and UNICEF conceived EvalPartners, a global partnership for national evaluation capacities, which Segone co-founded and co-chaired. He received the American Evaluation Association's Alva and Gunnar Myrdal Evaluation Practice Award in 2012. He then became Director of the Evaluation Office at UN Women and, later, Director of UNFPA's Evaluation Office, later known as its Independent Evaluation Office. He chaired the United Nations Evaluation Group, the network of evaluation offices of about 46 UN entities, and co-founded and co-chaired EvalGender+.

At UNFPA his office partnered with EvalYouth and the Global Parliamentarians Forum for Evaluation to launch the Eval4Action campaign in 2020. In January 2025 he presented the independent evaluation of the UNFPA Strategic Plan 2022 to 2025 to the Executive Board. A UNFPA note to member states records that he gave notice of resignation as Director of the Independent Evaluation Office, effective 31 August 2025.

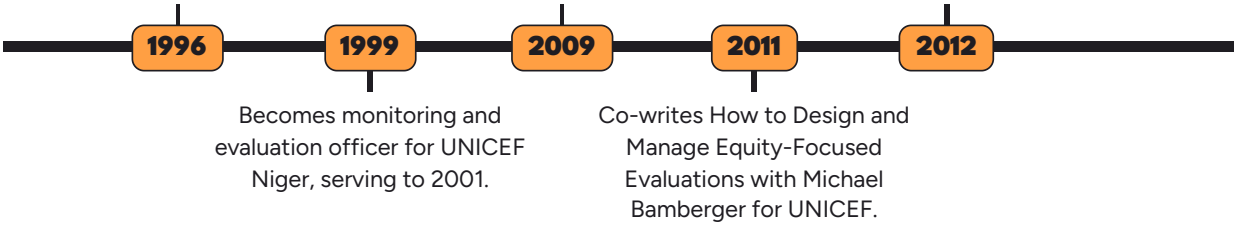
The spark

After working with international NGOs in Bangladesh, Pakistan, Thailand, Uganda and Albania, he served as UNICEF's monitoring and evaluation officer in Niger from 1999 to 2001, where his work included a participatory evaluation of child poverty. His 2010 volume names the problem that ran through his later career: countries adopted good policies that faltered in practice, while donors' parallel systems weakened the national evidence needed to fix them.

The breakthrough CHAPTER 23

Picture a country that has passed an excellent law to protect children, yet sees little change in children's lives. The policy is good; the problem is putting it into practice. And the evidence needed to fix it is weak, because each donor runs its own monitoring system alongside the government's, instead of strengthening

TIMELINE · LIFE AND WORK



one national system. That was the bottleneck Segone described in his 2010 UNICEF volume, From Policies to Results. Meanwhile, building evaluation skills was fragmented, often driven by donors, and largely ignored the national associations of evaluators appearing across the Global South.

His answer developed in two steps. In UNICEF's 2008 to 2010 series, edited by Segone with partners including IDEAS, IOCE, the World Bank and UNECE, he set out the case for country-led monitoring and evaluation systems owned by governments and supported, not replaced, by donors, with capacity built on both the supply and demand sides of evidence. In April 2012 the idea took organisational form in EvalPartners, born from a joint vision of IOCE board members and UNICEF; its first Global Evaluation Forum in Chiang Mai in December 2012 brought associations, governments, UN agencies and donors into one movement. A parallel strand, from 2011, applied an equity lens to evaluation itself.

The problem it solved

Averages that hid whether the poorest and most marginalised benefited, and fragmented, donor-driven capacity building that left local evaluation communities weak.

Impact WHAT CHANGED

Segone's impact is institutional and measured in networks and policies. EvalPartners grew from an Accra board meeting into a partnership whose history includes the 2015 International Year of Evaluation, the first General Assembly resolution on evaluation capacity, the Global Evaluation Agenda 2016 to 2020

TIMELINE, CONTINUED

2013 EvalPartners declares 2015 the International Year of Evaluation at the São Paulo conference on national evaluation capacities.

2014 The UN General Assembly adopts resolution 69/237 on evaluation capacity at country level, 19 December.

2015 The Kathmandu Global Evaluation Forum, 23 to 25 November, launches the Global Evaluation Agenda 2016 to 2020 and EvalGender+.

2020 UNFPA's evaluation office, EvalYouth and parliamentarians launch Eval4Action; its first Twitter chat is held on 21 May.

2025 Presents the evaluation of the UNFPA Strategic Plan 2022 to 2025 in January; his resignation as IEO Director takes effect on 31 August.

and five global networks, among them EvalYouth and EvalGender+. His EvalPartners tenure, by its own account, included 92 events to mark 2015 and the first global evaluation forum held in a national parliament, in Nepal. By 2020 Eval4Action reported more than 130 partners and six regional consultations on action plans.

Our interpretation is that the Global South gained most from the shift in where evaluation capacity was thought to live: in national associations, parliaments and governments rather than donor projects. His supply and demand framework remains a practical tool for governments in South Asia and Africa, and the equity guide put disaggregation and the worst-off at the centre of UNICEF evaluations long before the SDGs. Within UNFPA, the office he led reported in 2020 that evaluation investment had risen from 0.45 per cent of programme expenditure in 2014 to 0.83 per cent and that coverage reached about 97 per cent of country offices.

WHERE PEOPLE DISAGREE

Debates

In our reading, one limit is that his approach works through UN agencies and global partnerships, and a fair question is how far agendas set in New York or Geneva shape the priorities of national associations that the movement aims to empower. Campaigns such as EvalYear and Eval4Action measure success partly by commitments and events, which are easier to count than changes in whether governments actually use evaluations. UNFPA's own 2020 report noted that implementation of centralised evaluation recommendations was below target.

LIMITS, IN OUR READING

- ☐ In our view, his influence is institutional and UN-centred, and critics of top-down capacity building might ask how far Global North agencies shape the agenda of national evaluation associations.

WHY THIS MATTERS TO YOU

Segone's lessons apply to any cause you care about. First, look beyond averages: ask what a programme achieves for the people who are worst off, and whether it narrows the gaps. Second, remember that evidence needs two things: good studies, and people who want to use them. If either is missing, nothing changes. Third, do not work alone. A coalition of associations, governments, parliaments and international agencies can achieve what no single organisation can. If you belong to an evaluation association, a youth network or a gender network, his work shows that your collective voice can reach ministries, parliaments and the UN.

92

events worldwide to celebrate 2015 as the International Year of Evaluation, which he lists among his EvalPartners achievements

EvalPartners, Previous co-chairs

28

countries visited by the EvalYear Evaluation Torch, lit at the UN in December 2014 and handed on by Segone in Kathmandu in 2015

EvalPartners history; EvalPartners, Previous co-chairs

5

global EvalNetworks established in 2015, among them EvalYouth and EvalGender+

EvalPartners history

40

authors in From Policies to Results (2010), the UNICEF volume he edited

UNICEF, From Policies to Results

Key concepts

National evaluation capacity

A country's ability to commission, produce and use evaluations through its own institutions, people and policies.

Supply and demand for evidence

Segone's distinction between the ability to produce credible evaluation and the willingness of decision-makers to ask for and use it.

Voluntary organisation for professional evaluation (VOPE)

A national or regional evaluation association or network; EvalPartners was built around strengthening them.

Equity-focused evaluation

Evaluation designed to judge whether interventions reach and benefit the worst-off groups and reduce inequity.

EvalYear

The International Year of Evaluation, 2015, declared by EvalPartners in 2013 and marked by events worldwide.

Eval4Action

The Decade of Evaluation for Action campaign launched in 2020 by UNFPA's evaluation office, EvalYouth and the Global Parliamentarians Forum for Evaluation.

Act 5

Sources for this chapter: p. 588

KEY WORKS TO READ IN THE ORIGINAL

- **How to Design and Manage Equity-Focused Evaluations**
- **Evaluation for Equitable Development Results**
- **Evaluating the SDGs with a no one left behind lens**

1

BIG IDEA 1 OF 3 · SEGONE

Country-led monitoring and evaluation systems

Build evaluation capacity that countries own, by strengthening both the supply of credible evidence and the demand to use it.

PICTURE IT

Think of the difference between a country that rents generators from foreign companies and one that builds its own power grid. The generators help for a while; the grid lasts. Country-led evaluation systems are the national grid for evidence, with supply and demand both built at home.

The problem

By the late 2000s many countries had adopted ambitious social policy reforms that were not delivering for children and women. Segone's 2010 volume argued that the bottleneck was implementation, and that the evidence needed to fix it was weak because donors often ran their own parallel monitoring and evaluation systems. Capacity building was fragmented and donor driven: it trained individuals while ignoring whether anyone in government wanted or used the results. Reforms stalled, and the people they were meant to help waited.

What it is

A country-led monitoring and evaluation system is one in which the country decides what is evaluated, which questions are asked, which methods and analysis are used, and how findings are communicated and used. Donors strengthen such systems instead of replacing them, applying the ownership principle of the Paris Declaration and Accra Agenda to evidence. Segone's framework has two axes. On the supply side, professionals must produce sound evidence; on the demand side, policymakers must understand and use it, which depends on incentives.

Combining the two gives four situations: a vicious circle, where evidence is weak and little used; evidence supply constrained; evidence demand constrained; and the virtuous circle, the goal, where robust evidence is increasingly used. Capacity is then built at three levels: individuals, such as managers, champions and evaluators; institutions, with evaluation policies, units, independent funding and quality assurance; and the enabling environment of legislation, transparency and national evaluation associations.

How it works

1. **Diagnose the existing function**
   Assess the current evaluation function and its bottlenecks before designing a policy or system, as Segone's lead article in the 2010 volume recommends.
2. **Place the country on the matrix**
   Judge whether the situation is a vicious circle, supply constrained, demand constrained or a virtuous circle.
3. **Choose the matching strategy**
   Raise both supply and demand in a vicious circle, prioritise evidence quality where supply is weak, and build incentives and dialogue where demand is weak.
4. **Work at three levels**
   Plan for individuals, such as champions and managers; institutions, such as policies and units; and the enabling environment, such as legislation and associations.
5. **Build on ownership and existing capacity**
   Ground the strategy in national plans and existing capacities, and tailor it to each country and institution.
6. **Learn by doing**
   Use demand-driven technical assistance and joint country-led evaluations, and track reform milestones against annual targets agreed with stakeholders.

WORKED EXAMPLE · AN ILLUSTRATION

A state government in India has a capable directorate of economics and statistics that runs regular surveys, but ministers rarely ask for evaluations and findings seldom reach budget decisions. Using Segone's matrix, a support team classes the state as evidence demand constrained. Rather than training more enumerators, it works on demand. The finance department agrees that new spending proposals above a set size must cite existing evaluation evidence. Two departmental secretaries who have shown interest become champions and co-commission a joint evaluation of a nutrition scheme with the state's evaluation office. The team helps draft a state evaluation policy and links the office with the national evaluation community. After two budget cycles it reviews milestones: proposals citing evidence, evaluations commissioned and recommendations acted upon.

CONCEPT HIGHLIGHT

National evaluation capacity

A country's ability to commission, produce and use evaluations through its own institutions, people and policies.

USE IT WHEN

- You are designing an evaluation capacity programme with a national or state government.
- Good evaluations are being produced but are not used in decisions.
- A government is drafting an evaluation policy or setting up an evaluation unit.

SKIP IT WHEN

- The task is a quick independent verdict on one programme, and system building is outside the mandate.
- Political conditions would let government ownership be used to suppress findings; then independent and civil society evaluation may need to lead.

COMMON MISREADINGS

- Treating capacity development as training individuals while ignoring institutions and the enabling environment.
- Assuming that more and better data will automatically be used, when the framework shows that demand is often the binding constraint.

WHY IT WAS GROUNDBREAKING

It moved evaluation capacity from a donor training exercise to a question of national ownership and political demand. Its simple matrix gave governments and advisers a diagnostic that still applies.

Origin

Segone edited UNICEF's Evaluation Working Papers on the theme: Bridging the Gap (2008); Country-led Monitoring and Evaluation Systems (2009), published with partners including the World Bank, IDEAS, IOCE and UNECE; and From Policies to Results (2010), with 40 authors. The supply and demand framing and the four situations appear in his lead article in the 2010 volume. From 2012 EvalPartners carried the idea into a global partnership.

RELATED IDEAS

Equity-focused evaluation

A global movement for evaluation

RR Rist 410 CW Weiss 152 MP Patton 204

CONCEPT HIGHLIGHT

Supply and demand for evidence

Segone's distinction between the ability to produce credible evaluation and the willingness of decision-makers to ask for and use it.

2

BIG IDEA 2 OF 3 · SEGONE

Equity-focused evaluation

Judge programmes by what they achieve for the worst-off groups and whether they narrow gaps, not by averages alone.

PICTURE IT

Think of a class where the average mark rose, while the students who were struggling most fell further behind. The average hides them. Equity-focused evaluation looks at the gap, and at what happened to those at the back.

The problem

Averages can hide what matters most for equity. A programme can raise mean enrolment or immunisation while the poorest, girls, children with disabilities or remote communities fall further behind. Many evaluations were not designed to see such gaps: data were rarely disaggregated, samples were too small for minority groups, and theories of change seldom named the worst-off. Without evidence on who was being left behind, managers could not redirect resources to them, and the children with the most to gain stayed invisible.

What it is

The 2011 UNICEF guide by Michael Bamberger and Marco Segone defines equity as all children having the opportunity to survive, develop and reach their full potential without discrimination. Equity differs from equality: the aim is to remove unfair and avoidable circumstances, not to make everyone the same. An equity-focused evaluation is a rigorous, systematic and objective judgement of the relevance, effectiveness, efficiency, impact and sustainability of interventions aimed at equitable results, adding coverage, connectedness and coherence in humanitarian settings. Each criterion carries an equity question. Is the intervention relevant to the rights and needs of the worst-off? Are equity objectives achieved, and what constraints stand in the way? How do the costs of reaching the worst-off compare with average costs? Have gaps between best-off and worst-off groups narrowed? Will benefits for the worst-off last? The logic is that results must be examined for the worst-off and for the gaps between groups, not only for the average.

How it works

1. **Assess evaluability**
   Check whether equity was considered in the intervention's design, implementation and monitoring, and what approach is realistic, the first task in Part 2 of the 2011 guide.
2. **Map stakeholders and users**
   Decide how and when worst-off groups and other stakeholders will take part, and form a steering committee that includes the intended users.
3. **Anticipate challenges**
   Plan early for political resistance to disaggregated results, missing data, extra cost and governance obstacles.
4. **Frame equity questions and terms of reference**
   Build questions from the theory of change and select a technically strong, culturally sensitive team.
5. **Design for the gaps**
   Choose a theory-based or bottleneck analysis framework, mixed methods and samples large enough to report on worst-off groups.
6. **Work within real-world constraints**
   Reconstruct missing baselines or theories of change where needed, and adapt to budget, time and political limits.
7. **Use the findings**
   Report in formats suited to each audience, disseminate widely and prepare a management response.

WORKED EXAMPLE · AN ILLUSTRATION

A district in Rajasthan reports that a school health and nutrition programme has raised average attendance. An equity-focused evaluation begins with an evaluability check and finds that monitoring data are not split by caste, disability or type of settlement. The team reconstructs the theory of change with district officials and a steering committee that includes a disabled people's organisation and a women's collective. It oversamples children from remote hamlets and children with disabilities, and uses a bottleneck analysis to trace supply and demand constraints. Results show strong gains in roadside villages but almost none in remote hamlets, where meals arrive late, or among children with mobility impairments, who cannot reach the school. The cost of closing those gaps is compared with average costs, and the district prepares a management response.

In one line

Build evaluation capacity that countries own, by strengthening both the supply of credible evidence and the demand to use it.

USE IT WHEN

- A programme has explicit equity or targeting goals.
- A universal service reports good averages that may hide gaps between groups.
- SDG or national reporting requires results disaggregated by group.

SKIP IT WHEN

- The budget cannot support samples large enough to say anything about worst-off groups; then state clearly that equity questions cannot be answered rather than imply they were.
- Everyone in the programme already belongs to the same worst-off group and the question is purely one of overall cost-effectiveness.

COMMON MISREADINGS

- Reducing it to disaggregated tables added at the end, rather than building equity into questions, sampling and participation from the start.
- Confusing equity with equality, when the guide aims to remove unfair and avoidable disadvantage, which may require unequal effort for different groups.

WHY IT WAS GROUNDBREAKING

It made the distribution of results, not only the average, a standard question for a major UN agency's evaluations. It anticipated the leave no one behind principle that later became central to the SDGs.

Origin

Bamberger and Segone published How to Design and Manage Equity-Focused Evaluations with UNICEF in 2011, a 124-page guide with eight UNICEF-supported case studies and a further case on social assistance in Albania. Segone edited the companion volume Evaluation for Equitable Development Results in 2012. Later work with Bamberger and Florencia Tateossian linked the approach to evaluating the SDGs with a no one left behind lens.

RELATED IDEAS

Country-led monitoring and evaluation systems

DM Mertens 384 · SZ Zaveri 536 · RC Chambers 296

In one line

Judge programmes by what they achieve for the worst-off groups and whether they narrow gaps, not by averages alone.

3

BIG IDEA 3 OF 3 · SEGONE

A global movement for evaluation

Evaluation gains influence when a coalition of national associations, governments, parliaments, UN agencies and donors advocates for it together.

PICTURE IT

Think of how one voice in a crowd is easily ignored, while a well organised choir fills the hall. EvalPartners brought national associations, governments, parliaments, UN agencies and donors together, so that the case for evaluation could be heard.

The problem

Voluntary organisations for professional evaluation multiplied from 15 in 1999 to more than 188 by the end of 2013, but most were small, volunteer-run and weakly connected to governments. Capacity building ran through separate donor projects, and evaluation had little political profile: few parliaments asked for it, and no UN General Assembly resolution had addressed evaluation capacity. National associations had energy and expertise but little leverage over whether evaluations were commissioned, funded or used in national decisions.

What it is

Segone's institutional idea was that no single agency could make evaluation matter, but a partnership might. EvalPartners, conceived in April 2012 at an IOCE board meeting in Accra as a joint vision of IOCE board members and UNICEF, brought voluntary organisations for professional evaluation (VOPEs) together with governments, UN agencies and donors. The movement had several parts. Strengthening VOPEs gave national evaluators a collective voice. Global forums, in Chiang Mai in 2012 and Kathmandu in 2015, set shared agendas. Advocacy moments raised the profile of evaluation: EvalPartners declared 2015 the International Year of Evaluation at the São Paulo conference in 2013, and the UN General Assembly followed with resolution 69/237 on evaluation capacity in December 2014. Thematic networks, EvalYouth, EvalGender+, EvalIndigenous, EvalSDGs and a parliamentarians' forum, took the work to new constituencies. The logic is that evaluation use rises when demand is organised politically, not only when supply improves.

How it works

1. **Convene a coalition**
   Bring associations, governments, UN agencies and donors into one partnership with a shared goal, as IOCE and UNICEF did in Accra in April 2012.
2. **Strengthen national associations**
   Invest in VOPEs so that national evaluators can advocate, learn from peers and engage government.
3. **Agree a shared agenda**
   Use global and regional forums to set common priorities, as the Kathmandu forum did with the Global Evaluation Agenda 2016 to 2020.
4. **Create advocacy moments**
   Use symbolic milestones, such as EvalYear 2015 and its Evaluation Torch, which travelled to 28 countries, to win political attention.
5. **Reach new constituencies**
   Build networks for young evaluators, gender, Indigenous peoples and parliamentarians, so that demand comes from many directions.
6. **Seek formal commitments**
   Turn momentum into policy, such as UN resolution 69/237 and, through Eval4Action from 2020, pledges to strengthen national evaluation systems.

WORKED EXAMPLE · AN ILLUSTRATION

A national evaluation association in South Asia has 120 volunteer members and no contact with parliament. Drawing on the EvalPartners model, it joins a regional network, seeks peer support from an association in another region, and uses an international evaluation campaign as the hook for a national event. It invites members of a parliamentary committee, who form an informal caucus on evaluation, while a youth chapter organises training for young evaluators. Over two years the association helps the planning ministry draft a national evaluation policy, and parliament asks for evaluation evidence when reviewing two flagship schemes. The association's influence comes less from its own resources than from the coalition it belongs to and the political moments it can join.

USE IT WHEN

- You want to build demand for evaluation at national level.
- A national or regional association needs allies and visibility.
- Advocates are seeking a policy or parliamentary commitment to evaluation.

SKIP IT WHEN

- The task is a single evaluation assignment, where movement building is outside the mandate.
- Events and pledges are being counted as progress without checking whether governments actually commission and use evaluations.

COMMON MISREADINGS

- [ ] Measuring success by events and declarations alone, when the aim is changed use of evaluation in national decisions.
- [ ] Treating it as a UN programme rather than a partnership in which national associations are meant to help set priorities.

WHY IT WAS GROUNDBREAKING

It linked grassroots evaluation associations to the UN General Assembly, giving evaluation a political profile it had lacked. It built durable networks for young evaluators, gender and parliamentarians that outlasted any single post.

Origin

EvalPartners was conceived in April 2012 in Accra and held its first Global Evaluation Forum in Chiang Mai in December 2012. It declared 2015 the International Year of Evaluation in São Paulo in 2013, and UN resolution 69/237 followed on 19 December 2014. The Kathmandu forum of November 2015 launched the Global Evaluation Agenda 2016 to 2020 and five networks, and Eval4Action extended the approach from 2020.

RELATED IDEAS

Country-led monitoring and evaluation systems

SZ Zaveri 536 · ZO Ofir 454 · RR Rist 410

RELATED GIANTS



Ray C. Rist

The World Bank evaluator who turned results-based M&E into ten steps a government could follow, and helped build a school to teach them.

p. 410



Carol H. Weiss

The sociologist who asked what a programme assumes, and showed that evidence often changes minds slowly rather than decisions overnight.

p. 152



Michael Quinn Patton

The sociologist who built a career on one stubborn question: will anyone actually use this evaluation?

p. 204



Donna M. Mertens

The Gallaudet professor who put people facing discrimination at the centre of evaluation design.

p. 384

CONCEPT HIGHLIGHT

Voluntary organisation for professional evaluation (VOPE)

A national or regional evaluation association or network; EvalPartners was built around strengthening them.

CONCEPT HIGHLIGHT

Equity-focused evaluation

Evaluation designed to judge whether interventions reach and benefit the worst-off groups and reduce inequity.

CONCEPT HIGHLIGHT

EvalYear

The International Year of Evaluation, 2015, declared by EvalPartners in 2013 and marked by events worldwide.

CONCEPT HIGHLIGHT

Eval4Action

The Decade of Evaluation for Action campaign launched in 2020 by UNFPA's evaluation office, EvalYouth and the Global Parliamentarians Forum for Evaluation.

In one line

Evaluation gains influence when a coalition of national associations, governments, parliaments, UN agencies and donors advocates for it together.

Put it to work

USE IT ON MONDAY · EQUITY-FOCUSED EVALUATION AND NATIONAL EVALUATION CAPACITY

1. Plan disaggregation from the start: decide which groups at risk of being left behind your sample must include, and make sure it is large enough to say something about them.

2. Ask directly about barriers: for each group, what stops them from accessing or benefiting from the programme?

3. Involve your national or regional evaluation association, or local evaluators, in the team, and align your indicators with national systems and SDG reporting where you can.

AI prompt **Copy, fill the brackets, then check every line**

Help me make the evaluation of **[programme]** in **[country or state]** equity-focused, drawing on the equity-focused evaluation approach associated with Marco Segone and UNICEF. Here is the draft design and sampling plan: **[paste]**. Here is what I know about groups at risk of being left behind in this context: **[paste]**. (1) Identify where the design relies on averages that could hide differences between groups. (2) Suggest a disaggregation plan and questions about barriers to access and benefit. (3) Flag whether the sample is likely to include enough people from each group, and what information a statistician would need to check this. (4) Suggest how these groups could have a voice in the evaluation.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of the difference between a country that rents generators from foreign companies and one that builds its own power grid. The generators help for a while; the grid lasts. Country-led evaluation systems are the national grid for evidence, with supply and demand both built at home.

YOUR MOVES

Three moves from Segone's path

1 Let field work ground your system thinking

Segone moved from NGO projects in five countries and a country office post in Niger to regional and global roles, carrying lessons about national ownership.

TRY THIS

Keep one foot in field work, even when your role becomes strategic.

2 Build networks that outlast your job

EvalPartners, EvalGender+ and Eval4Action, which he helped found or launch, reached far beyond any single post he held.

TRY THIS

Start one network or initiative that could continue without you, and give others real ownership.

3 Pair tools with advocacy

His 2011 equity guide gave evaluators tools, while the movement work built the political demand to use them.

TRY THIS

For every tool you release, plan one conversation with a decision-maker about why it matters.

THE BIG IDEAS OF THIS CHAPTER

1 **Country-led monitoring and evaluation systems** p. 522

Build evaluation capacity that countries own, by strengthening both the supply of credible evidence and the demand to use it.

2 **Equity-focused evaluation** p. 525

Judge programmes by what they achieve for the worst-off groups and whether they narrow gaps, not by averages alone.

3 **A global movement for evaluation** p. 528

Evaluation gains influence when a coalition of national associations, governments, parliaments, UN agencies and donors advocates for it together.

CONCEPT HIGHLIGHT

Eval4Action

The Decade of Evaluation for Action campaign launched in 2020 by UNFPA's evaluation office, EvalYouth and the Global Parliamentarians Forum for Evaluation.

GIANTS CONNECTED TO THIS CHAPTER



Ray C. Rist

Evaluation capacity development and results-based systems for governments.

p. 410



Sonal Zaveri

Both co-founded or co-chaired parts of the EvalPartners family, including EvalGender+.

p. 536



Donna M. Mertens

The transformative paradigm takes equity into the heart of evaluation design.

p. 384

KEY WORKS TO READ IN THE ORIGINAL

- **How to Design and Manage Equity-Focused Evaluations**
- **Evaluation for Equitable Development Results**
- **Evaluating the SDGs with a no one left behind lens**

CONCEPT HIGHLIGHT

National evaluation capacity

A country's ability to commission, produce and use evaluations through its own institutions, people and policies.

CONCEPT HIGHLIGHT

Supply and demand for evidence

Segone's distinction between the ability to produce credible evaluation and the willingness of decision-makers to ask for and use it.

WORKSHEET · TRY THIS TEMPLATE

Equity lens checklist

Print this page, or copy the headings into your notes.

1 Groups at risk of being left behind

Defined for this context, for example by gender, disability, caste, location or income.

2 Disaggregation plan

Which results will be broken down, and by what.

3 Sampling reach

Will the sample include enough people from each group?

4 Barriers to access and benefit

Questions and methods to uncover them.

5 Voice in the evaluation

How these groups shape questions and interpret findings.

6 Link to national systems

National indicators, SDG targets and local evaluation partners.

REFLECT

Which group would disappear from view if your programme's results were reported only as averages?

CONNECTS TO

RR Ray C. Rist
Evaluation capacity development and results-based systems for governments. p. 410

SZ Sonal Zaveri
Both co-founded or co-chaired parts of the EvalPartners family, including EvalGender+. p. 536

DM Donna M. Mertens
The transformative paradigm takes equity into the heart of evaluation design. p. 384

24



Emblem: an original HubForge design for this figure, not a likeness.

Sonal Zaveri

Active 2000s to 2020s

India; work in more than 25 countries across Asia, East and West Africa, the Middle East and Eastern Europe

Made in the Global South · Participation, equity and voice

The Indian evaluator who carried participatory, feminist practice from grassroots work into regional networks and real-time evaluation.

IN ONE LINE

Sonal Zaveri showed that participation means seeking out the voices a community keeps quiet, and that evaluation can listen to them even in a crisis.

September 2020. COVID-19 is reshaping health, education and protection services across South Asia, and decisions are being made week by week. An independent evaluator from India sets out to do something unusual: evaluate whether UNICEF's response is meeting the needs of women and girls while the response is still unfolding. Over the following year, Sonal Zaveri's real-time evaluation covered eight countries, from Afghanistan to Sri Lanka, paired feminist criteria with the standard OECD DAC criteria and produced four staged reports, feeding findings back to country programmes as they emerged. In 2022 it received an honourable mention in the IDEAS Evaluation as Transformation category. Her path to that work ran through some three decades that began in grassroots and community settings, and a founding role in the Community of Evaluators South Asia in 2008.

Life CHAPTER 24

Sonal Zaveri is an independent evaluator based in India. She holds a PhD in social work and was a Fleishman Fellow at Duke University in the United States. Her career, by her profiles' accounts, spans about three decades and began in grassroots and community work before moving into strategic planning, programme design, capacity building, mentoring and evaluation at sub-national, national, regional and international levels. She has worked in more than 25 countries across South and Central Asia, the Asia-Pacific, East and West Africa, the Middle East and Eastern Europe, with a particular commitment to young people and gender.

Her practice combines participatory, feminist, outcome-based and utilisation-focused approaches. Between 2009 and 2012 she was one of

BIG IDEAS

1. **Participation in context**
2. **Real-time feminist evaluation**

KEY WORKS

The Importance of Context in Participatory Evaluations: Reflections from South Asia
2014 · Chapter in Making Evaluation Matter: Writings from South Asia, edited by Katherine Eve Hay and Shubh Kumar-Range

Real-Time Evaluation of Gender Integration in the UNICEF COVID-19 Response in South Asia
2020 to 2021 · Evaluation led with UNICEF gender and M&E staff

DID YOU KNOW?

GENSA was built with no budget: by Zaveri's account no Northern institution was involved and no external funding was sought, yet it gained 150 members in its first six months.

three regional mentors in IDRC's Developing Evaluation Capacity in ICTD project, coaching research teams in Asia, including a Mongolian programme and an internet innovation fund, through utilisation-focused evaluation, preparing training modules and presenting on the method at evaluation meetings in Sri Lanka in 2011 and Ghana in 2012.

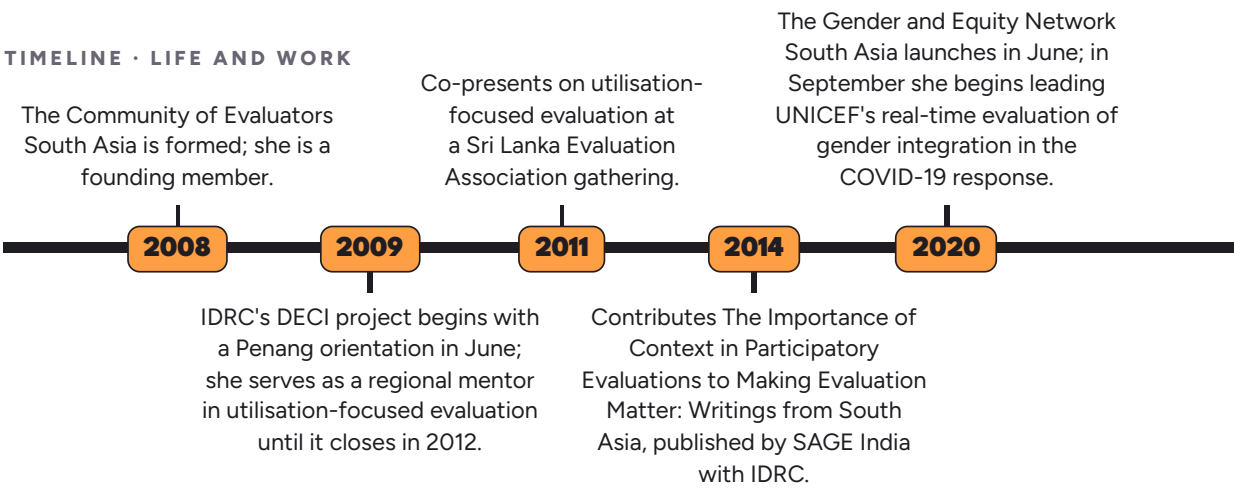
She is a founding member of the Community of Evaluators South Asia, a regional network formed in 2008, and has served as its Vice-President and as a board member. In 2014 she contributed a chapter on context in participatory evaluation to the network's IDRC-supported book Making Evaluation Matter: Writings from South Asia. Out of discussions in that community grew the Gender and Equity Network South Asia, launched in June 2020 as the network's first special interest group, which she has coordinated or convened. She co-chairs EvalGender+, sits on the EvalPartners coordination committee, is a core member of South to South Evaluation and has served on the boards of IDEAS and the International Evaluation Academy and on CGIAR's Evaluation Reference Group.

From September 2020 to September 2021 she led the real-time evaluation of gender integration in UNICEF's COVID-19 response across eight South Asian countries, which received an honourable mention in the IDEAS 2022 Evaluation as Transformation category.

The spark

Her career began in grassroots and community work, and her 2014 chapter, as its editors summarise it, addressed a gap in participatory evaluation in South Asia: imported ideas of community overlooked the hierarchies of gender, age and social position that decide who may speak. In our reading, the question running through her work is how to make participation real for those least likely to be heard.

TIMELINE · LIFE AND WORK



The breakthrough CHAPTER 24

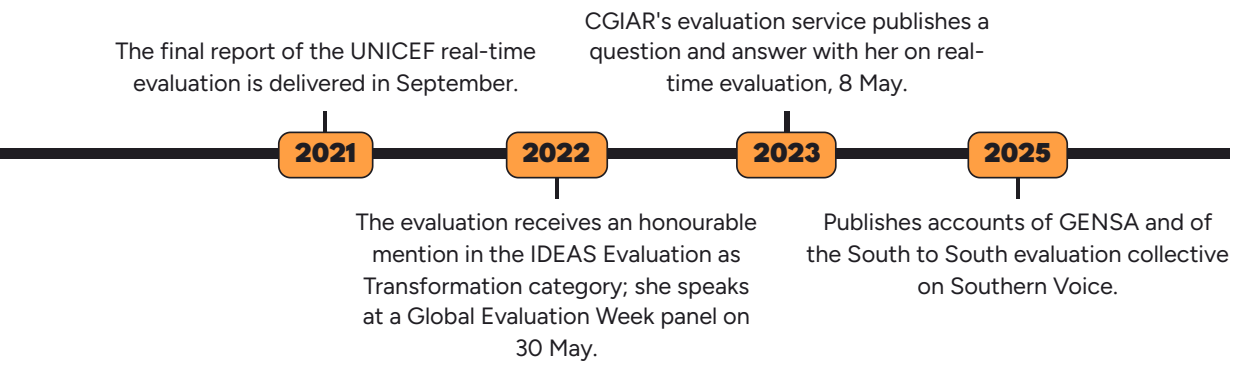
Picture a village meeting called to evaluate a programme. The headman speaks first, then the older men. Women sit at the back, young people say little and the poorest families may not be there at all. Everyone has been invited, but not everyone can speak. Participatory evaluation reached South Asia largely through imported toolkits that treated a community as a single, willing group. Zaveri's 2014 chapter, as the editors of Making Evaluation Matter summarise it, pointed out that deep-seated beliefs about children, women and older people, together with social stratification, create unspoken but powerful norms that decide who can speak.

Her key contribution took two forms. In the 2014 chapter she argued that evaluators must work with communities within communities, and that participation should be built into all designs, including randomised trials, because participatory processes help address power inequities. During the pandemic she applied these commitments under pressure: the UNICEF real-time evaluation, conceptualised in May 2020 and run from September 2020, combined feminist criteria with OECD DAC criteria, shared evidence with country programmes as it emerged and co-created tools with gender specialists, showing that gender-transformative evaluation is possible even in an emergency.

Did you know?

GENSA was built with no budget: by Zaveri's account no Northern institution was involved and no external funding was sought, yet it gained 150 members in its first six months.

TIMELINE, CONTINUED



The problem it solved

Imported designs that treated communities as data sources rather than partners, overlooked gender inequalities and local culture, and delivered reports after decisions had already been made.

Impact WHAT CHANGED

Zaveri's impact is concentrated in South Asia and in the global networks that link it to other regions. As a founder of the Community of Evaluators South Asia she helped create one of the region's main professional homes for evaluators, whose 2014 volume was the first of its kind on evaluation from South Asia. GENSA has added a regional focus on gender and equity, building alliances with EvalGender+ and with networks such as AGDEN and ReLAC, and contributing to what she describes as Asia's only course on gender and evaluation. In 2022 she spoke on a Global Evaluation Week panel on practical evaluation capacity development in South Asia.

Our interpretation is that her UNICEF real-time evaluation is her most transferable contribution: it offers a documented model for evaluating gender integration while an emergency is still unfolding, with six published lessons on demystifying feminist concepts, timeliness, co-creation and the evaluator's role in creating safe spaces. For Indian practitioners working with government and large agencies, her career shows how feminist and participatory values can be expressed through criteria and products that commissioners already recognise, and how regional networks can extend an independent consultant's reach.

WHERE PEOPLE DISAGREE

Debates

In our reading of her published account, her approach has some limits. In the UNICEF real-time evaluation a second gender expert left midway through, and time and resources were limited, which suggests that real-time designs strain small teams. Efficiency and impact were excluded from the evaluation's scope, so it could not answer questions about cost or long-term change.

LIMITS, IN OUR READING

- ☐ In the UNICEF real-time evaluation, a second gender expert left partway through, showing how small teams strain under real-time demands.
- ☐ The criteria used in that evaluation excluded efficiency and impact, which limited what it could say on those questions.

WHY THIS MATTERS TO YOU

Zaveri's lesson holds in any meeting, office or community: the loudest voices are rarely the whole story. Look for the communities within a community, and build ways for quieter people to be heard into every plan, even the most technical one. Her UNICEF evaluation shows that careful attention to gender and equity can keep pace with a crisis, using criteria decision-makers already recognise. And her networks show that you do not need a big budget to build a professional community: a few committed people and a shared purpose can start one. She opens a door for South Asian evaluators to lead, write and shape the field from where they stand.

25+

countries in which she has worked

CGIAR IAES Evaluation Reference Group profile

30+

years of experience in planning, programme design, capacity building, mentoring and evaluation

CGIAR IAES Evaluation Reference Group profile

8

South Asian countries covered by the UNICEF real-time evaluation of gender integration she led

European Evaluation Society TWG1 Newsletter (2022)

150

members gained by GENSA in its first six months, by her account

Southern Voice (2025)

THE BIG IDEAS OF THIS CHAPTER

1 **Participation in context** p. 544

Participation cannot be assumed: evaluators must recognise the hierarchies inside a community and build real participation into every design, randomised trials included.

2 **Real-time feminist evaluation** p. 547

In a crisis, evaluate gender integration while the response is still unfolding, combining feminist and standard criteria and feeding findings back fast enough to change decisions.

Key concepts

Participatory evaluation
Evaluation in which stakeholders, including programme participants, help shape questions, methods, interpretation and use.

Communities within communities
Zaveri's point that a community contains groups divided by gender, age and social position, whose voices must be sought separately.

Real-time evaluation
Evaluation conducted during an intervention, feeding findings back quickly so that decisions can change while action is under way.

Feminist evaluation criteria
Criteria such as empowerment, reflection, social justice, participation and inclusion, used alongside conventional criteria in her UNICEF evaluation.

Utilisation-focused evaluation
An approach, associated with Michael Quinn Patton, that designs every step around the intended use by intended users; she mentored teams in it through IDRC's DECI project.

Gender scale
A simple tool used in the UNICEF evaluation to frame discussion of how far responses moved from gender blind towards gender transformative.

Sources for this chapter: p. 589

GIANTS CONNECTED TO THIS CHAPTER



Robert Chambers
Participatory methods that she argues need a more nuanced view of community in South Asia.
p. 296



Marco Segone
EvalPartners and EvalGender+ form part of the global network in which she works.
p. 516



Michael Quinn Patton
Developmental evaluation shares the commitment to feedback as events unfold.
p. 204

KEY WORKS TO READ IN THE ORIGINAL

- **The Importance of Context in Participatory Evaluations: Reflections from South Asia** 2014, Chapter in Making Evaluation Matter: Writings from South Asia, edited by Katherine Eve Hay and Shubh Kumar-Range
- **Real-Time Evaluation of Gender Integration in the UNICEF COVID-19 Response in South Asia** 2020 to 2021, Evaluation led with UNICEF gender and M&E staff

In one line

Sonal Zaveri showed that participation means seeking out the voices a community keeps quiet, and that evaluation can listen to them even in a crisis.

1

BIG IDEA 1 OF 2 · ZAVERI

Participation in context

Participation cannot be assumed: evaluators must recognise the hierarchies inside a community and build real participation into every design, randomised trials included.

PICTURE IT

Think of a family photograph in which a few people stand at the front and others are half hidden behind them. Everyone is in the picture, but not everyone can be seen. Participation in context means rearranging the photograph so that those at the back come into view.

The problem

Participatory evaluation reached South Asia largely through imported toolkits that treated a village or settlement as one willing community. In practice, as the editors of Making Evaluation Matter summarise Zaveri's 2014 chapter, deep-seated beliefs about children, women and older people, together with social stratification, create unspoken but powerful norms about who may speak. A meeting dominated by older men or dominant groups can then be recorded as community voice, while the people a programme most needs to hear stay silent.

What it is

Zaveri's argument, in The Importance of Context in Participatory Evaluations: Reflections from South Asia, chapter 4 of Making Evaluation Matter (2014), has three linked parts, as the book's editors summarise it. First, participatory evaluation must be understood in the specific context of the region, where complex social hierarchies make it hard first to recognise, and then to work with, the communities within a community, divided by gender, age and social position. Second, participation is needed for empowerment but is harder to achieve where inequities run deep, so evaluators must think carefully about how they use participatory methods and ground them in real conditions rather than ideals. Third, participation should feature in all evaluation designs, including randomised controlled trials, because it can help address power inequities between evaluators and those evaluated. The logic is that who participates shapes what is learned, so attention to context is a matter of validity as well as ethics.

How it works

1. **Map the communities within the community**
   Before fieldwork, identify groups divided by gender, age, caste, class and location, following Zaveri's call, as her editors summarise it, to recognise communities within communities.
2. **Read the unspoken norms**
   Ask local partners which norms decide who attends, who speaks and whose views count, and how these could shape meetings and interviews.
3. **Create separate spaces**
   Use separate discussions or methods for groups unlikely to speak in mixed settings, such as adolescent girls, older women or landless households.
4. **Agree what participation means here**
   Decide with these groups at which stages they will shape questions, measures, interpretation and use, rather than assuming one model fits.
5. **Build participation into any design**
   In our reading of her argument, even a randomised trial can involve participants in defining outcomes and interpreting results while leaving random assignment intact.
6. **Report on power openly**
   Record whose voices shaped the findings and whose did not, and state these limits in the report.

WORKED EXAMPLE · AN ILLUSTRATION

A research organisation is running a randomised trial of a sanitation subsidy in 80 villages in Bihar. Design meetings in the villages are attended mostly by older men from landowning families, who say the main benefit of a toilet is convenience. Applying Zaveri's argument, the team holds separate discussions with women, adolescent girls, landless households and older people. Women and girls name safety and dignity after dark, and landless families explain that they lack space to build. These groups help choose additional outcome measures, such as reported safety, and later help interpret why uptake is low in some hamlets. The random assignment of villages is unchanged, but the trial now measures what matters to those who were silent in the first meetings, and its report says whose views shaped it.

CONCEPT HIGHLIGHT

Participatory evaluation

Evaluation in which stakeholders, including programme participants, help shape questions, methods, interpretation and use.

USE IT WHEN

- The evaluation works in socially stratified communities where some groups rarely speak in public.
- You are planning community consultations or participatory exercises for an evaluation.
- A trial's outcomes have been defined by outsiders without input from participants.

SKIP IT WHEN

- Speaking out could expose vulnerable people to harm, unless safe and confidential methods are available.
- There is no time or budget to engage marginal groups properly; then report the gap rather than label a dominant-group meeting as community participation.

COMMON MISREADINGS

- [ ] Counting attendance at community meetings as participation without asking who was absent or silent.
- [ ] Assuming participation is incompatible with randomised trials, when Zaveri argues it belongs in all designs.

WHY IT WAS GROUNDBREAKING

It challenged the habit of treating participation and experimental rigour as opposites. It gave South Asian practitioners a written, regional argument grounded in their own social realities rather than in imported toolkits.

Origin

Zaveri set out the argument in chapter 4 of Making Evaluation Matter: Writings from South Asia, edited by Katherine Eve Hay and Shubh Kumar-Range and published by SAGE India with IDRC in 2014, a volume linked to the Community of Evaluators South Asia formed in 2008. The concern with voice carried into her later gender and equity work, including the Gender and Equity Network South Asia, launched in June 2020.

RELATED IDEAS

Real-time feminist evaluation

- RC Chambers 296
- BC Chilisa 498
- D&B Duflo and Banerjee 432
- DM Mertens 384

CONCEPT HIGHLIGHT

Communities within communities

Zaveri's point that a community contains groups divided by gender, age and social position, whose voices must be sought separately.

2

BIG IDEA 2 OF 2 · ZAVERI

Real-time feminist evaluation

In a crisis, evaluate gender integration while the response is still unfolding, combining feminist and standard criteria and feeding findings back fast enough to change decisions.

PICTURE IT

Think of a ship's navigator checking the course during a storm, rather than writing a report after the ship reaches port. Real-time feminist evaluation steers while the crisis is under way, so that women and girls are not left behind.

The problem

When COVID-19 spread across South Asia in 2020, governments and agencies redesigned services within weeks. A conventional evaluation would report a year or more later, after the key decisions had been taken. Gender evaluation also carried a reputation for abstract language: Zaveri's account lists demystifying gender and feminist concepts as its first lesson. Agencies needed evidence on whether their responses addressed gender inequalities in time to adjust them, without giving up rigour or independence.

What it is

Real-time feminist evaluation, as Zaveri practised it for UNICEF's Regional Office for South Asia, evaluates an intervention while it is under way and uses findings to adapt it. Her 2022 account describes a hybrid design that was developmental, learning focused and use focused, and participatory, with UNICEF gender and M&E staff closely engaged. It paired five feminist criteria, empowerment, reflection, social justice, participation and inclusion, with four OECD DAC criteria, relevance, effectiveness, coherence and connectedness, and sustainability, leaving efficiency and impact out of scope. Methods included perception surveys, interviews, iterative peer reviews of findings, polls, a desk review of COVID-19 studies, virtual observation of meetings and learning case studies with country offices. A gender scale running from discriminatory and blind through sensitive and responsive to transformative helped demystify concepts and frame discussion. Products were staged so that each could inform the next phase of the response.

How it works

1. **Conceptualise early**
   Agree the purpose with the commissioner while the crisis unfolds; UNICEF's regional office conceptualised this evaluation in May 2020, and the work began in September.
2. **Combine criteria**
   Pair feminist criteria with the DAC criteria the agency already uses, and agree openly which criteria are out of scope.
3. **Demystify concepts**
   Introduce a simple tool such as the gender scale, so that staff share a language for judging how far responses move towards gender transformative.
4. **Stage the products**
   Tie short products to decisions: here an inception report (October 2020), an operational review (November 2020), a gender integration and effectiveness review (June 2021) and a final report (September 2021).
5. **Feed back and co-create**
   Share emergent findings with country programmes as they arise, run peer reviews and co-create tools so gender specialists own the results.
6. **Support self-assessment and redesign**
   Help stakeholders assess gender and equity in their own interventions through case studies, and then redesign them.
7. **Consolidate the lessons**
   Bring the staged products together in a final report; Zaveri's account records six lessons, including timeliness, co-creation and the evaluator's listening role.

WORKED EXAMPLE · AN ILLUSTRATION

After severe floods in Assam, a state disaster authority and its UN partners launch a relief and recovery programme and commission a six-month real-time evaluation of gender integration. In the first month the evaluator agrees criteria with the client: relevance, effectiveness and coherence, together with participation, inclusion and empowerment. An operational review at week six, based on phone surveys and virtual observation of coordination meetings, finds that relief camps lack separate toilets and that women-headed households are missing from cash transfer lists. Findings go straight to the coordination group, which adjusts camp layouts and list verification. Gender focal points use a simple gender scale to rate their own activities and co-design fixes. A final report at month six consolidates the reviews and documents what changed and why.

USE IT WHEN

- A crisis or fast-moving response is making decisions month by month.
- Gender integration is new to staff, so co-creation can build understanding and ownership.
- The commissioner is committed to adapting the response on the basis of findings.

SKIP IT WHEN

- ○ The main questions concern efficiency or long-term impact, which this design left out of scope.
- ○ The team is too small to sustain continuous feedback; in Zaveri's case a second gender expert left midway, which shows the strain.
- ○ The commissioner cannot or will not act on findings during the response.

COMMON MISREADINGS

- ☐ Equating real time with rushed, when the design depended on sustained dialogue and about a year of immersion.
- ☐ Assuming feminist criteria replace standard ones, when the evaluation paired them with DAC criteria the agency already used.

WHY IT WAS GROUNDBREAKING

It showed that a feminist, participatory evaluation could keep pace with a humanitarian emergency instead of reporting after the fact. Its blend of feminist and DAC criteria gave a large agency a familiar frame and a transformative lens at once.

Origin

UNICEF's Regional Office for South Asia conceptualised the evaluation in May 2020, and Zaveri led it from September 2020 to September 2021 across Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka. She described its design and six lessons in the European Evaluation Society's TWG1 newsletter in 2022. It received an honourable mention in the IDEAS 2022 Evaluation as Transformation category, and CGIAR interviewed her on real-time evaluation in May 2023.

RELATED IDEAS

Participation in context · MP Patton 204 · DM Mertens 384 · MS Segone 516

CONCEPT HIGHLIGHT

Real-time evaluation

Evaluation conducted during an intervention, feeding findings back quickly so that decisions can change while action is under way.

Put it to work

USE IT ON MONDAY · PARTICIPATORY EVALUATION IN CONTEXT, IN REAL TIME

1. Map who in the community participates and who does not, including differences of gender, age, caste and location, before you decide what participation will mean in your design.

2. Agree a rapid feedback rhythm with the client, for example a short review every few weeks timed to real decision points, and record each adaptation in your methods.

3. Pair standard criteria such as relevance and effectiveness with explicit gender and equity questions, so the client gets both a familiar frame and a feminist lens.

AI prompt **Copy, fill the brackets, then check every line**

Help me plan a real-time, participatory evaluation with a gender and equity lens, drawing on Sonal Zaveri's work, for **[programme or response]** in **[location]**. Upcoming decision points are **[list with dates]**. Our team is **[size and roles]**, and stakeholders available to co-create tools are **[list]**. (1) Propose a feedback rhythm linked to the decision points. (2) Suggest rapid questions and light methods for each cycle. (3) Suggest how to combine standard criteria with gender and equity questions. (4) Draft an adaptation log format. (5) Flag risks to a small team under real-time pressure and how to manage them.

GUARDRAILS Rules: use only the information I have given you. Do not invent data, results, quotations, statistics, sources or stakeholder views. Where something is missing, write NEEDS INPUT and tell me the question I should ask the client or the team. Clearly label anything that is your suggestion rather than evidence.

Picture it

Think of a family photograph in which a few people stand at the front and others are half hidden behind them. Everyone is in the picture, but not everyone can be seen. Participation in context means rearranging the photograph so that those at the back come into view.

YOUR MOVES

Three moves from Zaveri's path

1 Carry grassroots experience to every level

Zaveri's career moved from community work to regional and international evaluation in more than 25 countries.

TRY THIS

Keep a record of what you learn in community settings, and bring it into every senior meeting.

2 Build the community you need

She was a founding member of the Community of Evaluators South Asia in 2008 and helped launch GENSA in 2020 with no budget.

TRY THIS

If the network you need does not exist, invite five people and start it this month.

3 Translate values into frames clients know

Her UNICEF evaluation paired feminist criteria with the DAC criteria and earned international recognition in 2022.

TRY THIS

Map your values onto the criteria your client already uses, so they can act on what you find.

THE BIG IDEAS OF THIS CHAPTER

1 Participation in context p. 544

Participation cannot be assumed: evaluators must recognise the hierarchies inside a community and build real participation into every design, randomised trials included.

2 Real-time feminist evaluation p. 547

In a crisis, evaluate gender integration while the response is still unfolding, combining feminist and standard criteria and feeding findings back fast enough to change decisions.

CONCEPT HIGHLIGHT

Gender scale

A simple tool used in the UNICEF evaluation to frame discussion of how far responses moved from gender blind towards gender transformative.

GIANTS CONNECTED TO THIS CHAPTER



Robert Chambers

Participatory methods that she argues need a more nuanced view of community in South Asia.

p. 296



Marco Segone

EvalPartners and EvalGender+ form part of the global network in which she works.

p. 516



Michael Quinn Patton

Developmental evaluation shares the commitment to feedback as events unfold.

p. 204

KEY WORKS TO READ IN THE ORIGINAL

- **The Importance of Context in Participatory Evaluations: Reflections from South Asia** 2014, Chapter in Making Evaluation Matter: Writings from South Asia, edited by Katherine Eve Hay and Shubh Kumar-Range
- **Real-Time Evaluation of Gender Integration in the UNICEF COVID-19 Response in South Asia** 2020 to 2021, Evaluation led with UNICEF gender and M&E staff

CONCEPT HIGHLIGHT

Participatory evaluation

Evaluation in which stakeholders, including programme participants, help shape questions, methods, interpretation and use.

CONCEPT HIGHLIGHT

Communities within communities

Zaveri's point that a community contains groups divided by gender, age and social position, whose voices must be sought separately.

WORKSHEET · TRY THIS TEMPLATE

Real-time evaluation rhythm

Print this page, or copy the headings into your notes.

1 Decision moments

Upcoming decisions the evaluation should inform, with dates.

2 Rapid questions

What decision-makers need to know before each moment.

3 Quick evidence

Light methods that can deliver in time: short calls, reviews of records, rapid surveys.

4 Co-creation

Which stakeholders help design tools and interpret results.

5 Feedback product

Operational review, brief or meeting, and who receives it.

6 Adaptation log

What changed in the evaluation design, when and why.

REFLECT

When did your last evaluation's findings arrive: before or after the decisions they were meant to inform?

CONNECTS TO

RC **Robert Chambers** Participatory methods that she argues need a more nuanced view of community in South Asia. p. 296

MS **Marco Segone** EvalPartners and EvalGender+ form part of the global network in which she works. p. 516

MP **Michael Quinn Patton** Developmental evaluation shares the commitment to feedback as events unfold. p. 204

THE GREAT DEBATES · ACT 5

The Randomistas and Their Critics

THE QUESTION

Are randomised trials the best route to knowing what works against poverty, or one tool among many?

D&B Duflo and Banerjee 432 | HW White 476 | ZO Ofir 454 | BC Chilisa 498

SIDE A

Break big questions into testable ones (Banerjee and Duflo)

The Nobel committee's 2019 account summarised the approach: break the large question of global poverty into smaller, precise questions, such as which interventions best improve schooling or child health, and answer them through carefully designed experiments among the people most affected. In their 2009 review of the experimental approach, Banerjee and Duflo argued that its main advantage is close collaboration between researchers and implementers, which allows questions to be tested that could not otherwise be evaluated. They took the criticisms seriously but concluded that many were not specific to experiments. In his Nobel lecture Banerjee argued that replication across sites, combined with theory, can support generalisation.

SIDE B

No special access to truth (Deaton, Stern, Ofir)

Angus Deaton argued in the Journal of Economic Literature in 2010 that randomised trials and quasi-experiments give no special access to credible knowledge, that they often fail to handle heterogeneous effects, and that evaluating projects without understanding mechanisms is unlikely to advance knowledge of development. With Nancy Cartwright he later argued that results strictly describe the trial sample and that trials should be combined with other methods and theory. A 2012 DFID working paper led by Elliot Stern concluded that experimental designs suit only a minority of programmes. In Africa, Zenda Ofir and Sulley Gariba judged that such designs did not reflect African ways of knowing and could oversimplify complex communal societies.

WHAT HAPPENED

Both camps reshaped practice. Banerjee, Duflo and Michael Kremer received the 2019 Nobel prize, and the committee credited one of their Indian studies with leading to remedial tutoring that has benefited more than five million children. Deaton himself welcomed a shift in development experiments from project evaluation towards testing mechanisms. Howard White argued in 2009 that impact evaluations should trace the full causal chain, and the Stern paper pressed for a broader range of designs. Ofir's concern helped spark Made in Africa Evaluation, which Chilisa developed into theory. The debate continues.

Where you stand

For the programme you know best, what is the one question a randomised trial could answer well, and the one it could not?

If a trial found an average effect of zero, what would you want to know about who gained and who lost before you advised ending the programme?

Your position

Which side do you find yourself on today, and what evidence would change your mind?

THE FIVE ACTS AT A GLANCE

ACT 1 **What Did They Learn? Measuring What Matters** 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 **The Experimenting Society** 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 **Will Anyone Use It? The Use Revolution** 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 **Many Voices, Many Truths** 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340 · D&D Davies and Dart 362 · DM Mertens 384

ACT 5 **Evidence for a World in a Hurry** 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476 · BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

THE GIANTS IN THIS DEBATE



Esther Duflo and Abhijit Banerjee

Two economists who broke the big question of whether aid works into small questions that a lottery could answer.

p. 432



Howard White

The institution builder who insisted that impact evaluations explain why, and that decision-makers see the whole body of evidence.

p. 476



Zenda Ofir

The scientist turned evaluator who helped launch the call for evaluation made in Africa, by Africans.

p. 454



Bagele M. Chilisa

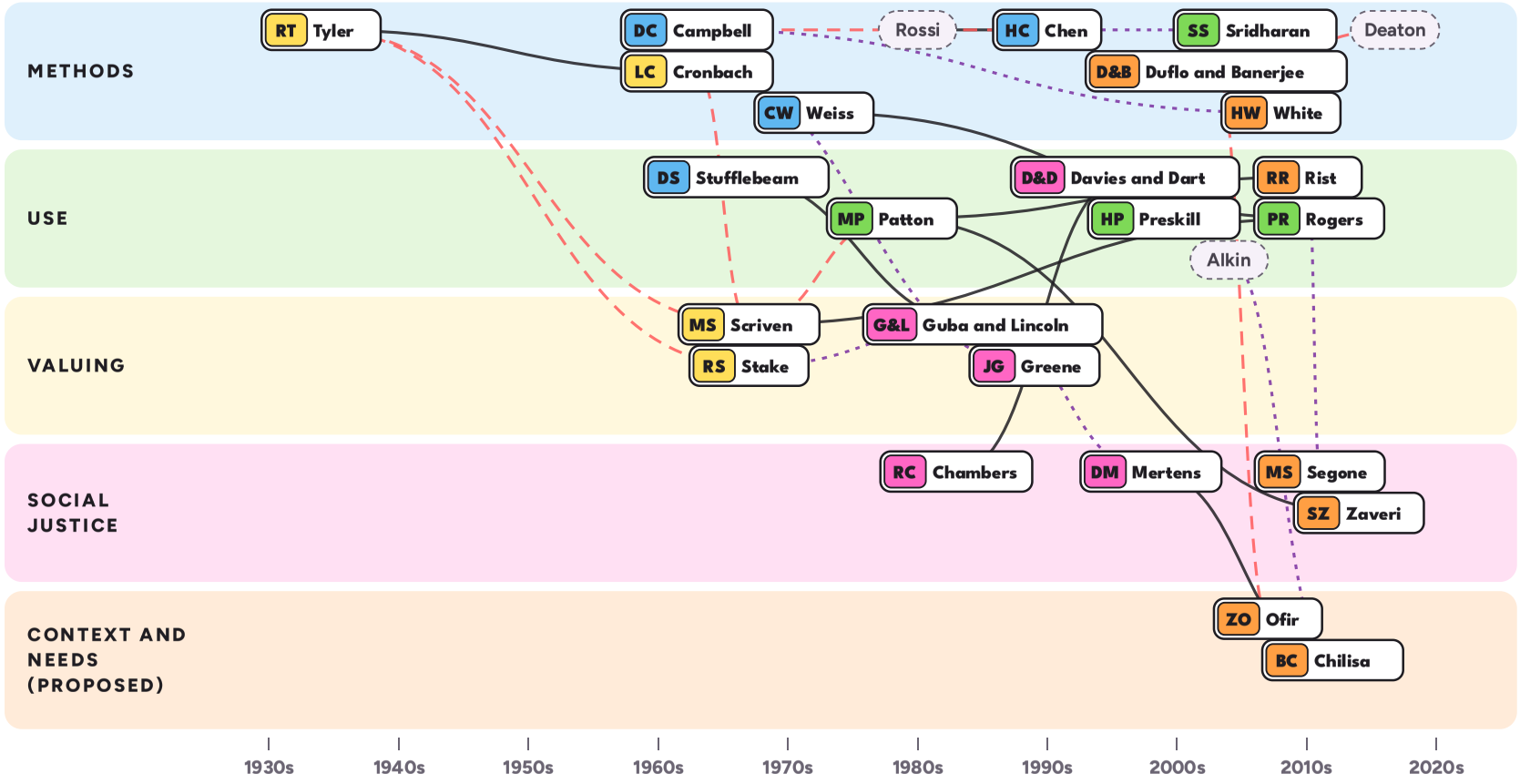
The Botswana scholar who gave decolonial and indigenous evaluation a written theory that others could cite, test and teach.

p. 498

THE FAMILY TREE

Who shaped whom

ADAPTED FROM THE EVALUATION THEORY TREE OF ALKIN AND CHRISTIE



Influenced — Challenged — Extended — Other key figures, no chapter of their own

TURN THE BOOK TO READ THE TREE

How to read it. Each edge reads as a sentence: from, type, to. For example, cronbach challenged campbell; tyler influenced cronbach; chilisa extended Alkin. Lanes are branches; positions run from the 1930s on the left to the 2020s on the right.

The five branches of the family tree

Methods

Theories that emphasise research design, causal inference and generalisation. One of the three original branches of Alkin and Christie's evaluation theory tree; Christie (2024) describes Campbell's work as foundational to it. A position on a branch shows relative emphasis, not exclusive focus (Mark 2024).

RT Tyler 20 · DC Campbell 112 · LC Cronbach 38 · CW Weiss 152 · HC Chen 174 · D&B Duflo and Banerjee 432 · SS Sridharan 246 · HW White 476

Use

Theories that emphasise how evaluation serves decisions, intended users and organisational learning. An original branch of the tree; in the 2004 book its part includes Stufflebeam, Wholey, Patton, Alkin, Fetterman, Cousins, King, Preskill and Owen.

DS Stufflebeam 134 · MP Patton 204 · D&D Davies and Dart 362 · HP Preskill 226 · RR Rist 410 · PR Rogers 268

Valuing

Theories that emphasise how evaluators reach judgements of merit and worth. An original branch; in the second edition it was split into an objectivist side, rooted in Scriven, and a subjectivist side, rooted in Stake (Christie 2024).

MS Scriven 60 · RS Stake 82 · G&L Guba and Lincoln 318 · JG Greene 340

Social justice

A fourth branch added by Donna Mertens, aligned with the transformative paradigm (Mertens and Wilson 2012). Mark (2024) notes that Alkin and Christie's most recent version instead places more of these theories on the valuing branch. We use it here for evaluators who put power, equity and participation first.

RC Chambers 296 · DM Mertens 384 · MS Segone 516 · SZ Zaveri 536

Context and needs (proposed fifth branch)

A proposed branch, not part of Alkin and Christie's tree. Mertens (2023) proposed a fifth branch named context and needs, fitting an Indigenous paradigm. Chilisa (2024) describes the case for it, made in her own work and with Mertens, as drawing on the culture, philosophy, history and experience of formerly colonised peoples in Africa and Indigenous peoples elsewhere. Earlier, Chilisa and colleagues (2016) proposed an African relational evaluation paradigm and an African evaluation tree.

ZO Ofir 454 · BC Chilisa 498

Method note. This family tree is our adaptation of a published framework, the evaluation theory tree of Marvin Alkin and Christina Christie, introduced in Evaluation Roots (2004) and revised in later editions (a second edition, then a third in 2023, which classifies theories rather than theorists). Placements marked as Alkin and Christie's are inferred from the part of Evaluation Roots in which a theorist's chapter appears (Methods, Valuing or Use), as listed by SAGE for the 2004 book, or from the 2023 Guilford contents for Greene, Chen and Preskill; the tree's own diagram was not consulted directly, so leaf positions within a branch are not reproduced. The social justice branch is Mertens's revision, not Alkin and Christie's. The context and needs branch is a proposal by Mertens and Chilisa, not an accepted part of the tree. All placements of Chambers, Davies and Dart, Rogers, Sridharan, Rist, Duflo and Banerjee, White, Ofir, Chilisa, Segone, Zaveri and Deaton are ours, as are era years, which mark a defining publication or event in each figure's profile. Edges summarise sourced relationships; where a note says our reading, the connection is interpretive. Patton himself has suggested a river metaphor and found the tree too constrained (Christie 2024), and Mertens (2024) notes that a tree does not show interrelationships between paradigms: read this as one map among several.

EPILOGUE

Your Turn to Ask

Look back across these lives and a pattern appears. Tyler was asked to show what around thirty schools had achieved. Chambers noticed that visitors only ever saw the villages beside the tarmac road. Davies needed a way for one Bangladeshi NGO to learn from changes no indicator had predicted. Ofir and Gariba decided that evaluation in Africa should be led from Africa. None of them set out to found a field. Each met a practical problem, refused an easy answer, and wrote down what they learned so that others could use it. In our reading, that is the whole secret of the field, and it is open to anyone.

You may be a field coordinator filling in a monitoring form, a programme manager staring at a results framework, or a student wondering whether evaluation is a career at all. The questions in this book are your questions. Did it work? For whom? Why? By whose standards? Who will use what we found? Every time you ask one of them with care, in a district office or a village meeting, you are doing the work these people did. The field still has open problems: evidence that travels, methods that respect local knowledge, systems that governments actually use, and evaluation that keeps pace with crises.

So treat this book as an invitation rather than a monument. Read one of the original works. Try one method on a real programme. Argue with a theorist you disagree with, as they argued with each other. Join an evaluation association or a community of practice near you. The people in these pages built a field by asking how we would know. The next chapter will be written by practitioners who keep asking, and it could begin with you.

What comes next: AI and evaluation

OPEN QUESTIONS

The story does not stop in 2026. Artificial intelligence is arriving in evaluation work: drafting tools, coding interviews, summarising evidence. We do not know yet how it will reshape the field. The people in this book left us questions worth asking of it.

Threats to validity (Campbell)
when an AI tool summarises interviews or cleans data, what new rival explanations and biases does it introduce, and how would we detect them?

Whose reality counts (Chambers, Chilisa, Ofir)
if AI tools learn mostly from text produced in the Global North, whose knowledge do they amplify and whose do they miss?

Intended use (Patton)
who are the intended users of an AI-generated finding, and do they understand how it was produced?

Equity (Segone, Mertens)
who gains access to AI-enabled evaluation, and who is left behind?

Trustworthiness (Guba and Lincoln)
what would it take to show that AI-assisted qualitative analysis can be trusted, and who should judge?

Evidence ecologies (White, Sridharan)
can AI help decision-makers see whole bodies of evidence, or will it flood them with plausible text that has no evidence behind it?

These are open questions, not predictions. Every AI prompt in this book is written to stop the tool from inventing evidence; treat that as a minimum standard, not a guarantee.

Patricia J. Rogers, in one line

Patricia Rogers showed evaluators that every model needs a purpose, and then gave the whole world a free map of the work.

Find the problem on your desk

PROBLEM CROSSWALK

Start from the problem you face this week. Read the figures it points to, then borrow their idea and their template.

01

The client wants to know if it worked.

Name the rival explanations first (Campbell), build a credible counterfactual and test each link in the causal chain (White), and use randomisation where it is feasible and fair (Duflo and Banerjee).

DC Campbell 112 · HW White 476 · D&B Duflo and Banerjee 432

02

Nobody uses our reports.

Design around named intended users and uses (Patton), plan for slow, conceptual use over time (Weiss) and build learning sessions into the team's routine (Preskill).

MP Patton 204 · CW Weiss 152 · HP Preskill 226

03

The programme keeps changing.

Use developmental evaluation for innovations in motion (Patton), fit the design to complexity (Sridharan) and run real-time feedback cycles with an adaptation log (Zaveri).

MP Patton 204 · SS Sridharan 246 · SZ Zaveri 536

04

Communities feel extracted from.

Hand over the pen with participatory methods (Chambers), put relationships and knowledge ownership first (Chilisa) and centre those facing discrimination (Mertens).

RC Chambers 296 · BC Chilisa 498 · DM Mertens 384

05

We need a theory of change.

Surface and test assumptions (Weiss), separate the change model from the action model (Chen) and decide what the diagram is for before drawing it (Rogers).

CW Weiss 152 HC Chen 174 PR Rogers 268

06

The ToR asks for an RCT on a tiny budget.

Shrink the question to one testable design choice (Duflo and Banerjee), consider a quasi-experimental design with its threats stated openly (Campbell), and check existing reviews and evidence maps before collecting new data (White).

D&B Duflo and Banerjee 432 DC Campbell 112 HW White 476

07

We cannot predict the outcomes, so indicators feel forced.

Collect and select stories of significant change (Davies and Dart) and look for effects without starting from goals (Scriven).

D&D Davies and Dart 362 MS Scriven 60

08

The donor's framework does not fit our context.

Ground evaluation in local values and leadership (Ofir), let communities define a good outcome (Chilisa) and combine local, feminist lenses with criteria the donor recognises (Zaveri).

ZO Ofir 454 BC Chilisa 498 SZ Zaveri 536

09

We have to judge whether it was good, not just describe what happened.

Make criteria of worth explicit and reason from evidence to judgement (Scriven), and openly engage the competing values at stake (Greene).

MS Scriven 60 JG Greene 340

10

Our qualitative findings are dismissed as anecdotes.

Show trustworthiness with explicit quality criteria (Guba and Lincoln), use well-chosen in-depth cases (Stake) and use a structured story-selection process with verification (Davies and Dart).

G&L Guba and Lincoln 318 RS Stake 82 D&D Davies and Dart 362

11

A government department wants an M&E system.

Start with a readiness assessment (Rist), tie evaluation to real management decisions (Stufflebeam) and connect to national evaluation capacity and associations (Segone).

RR Rist 410 DS Stufflebeam 134 MS Segone 516

12

Average results hide who is left behind.

Plan disaggregation and barrier analysis from the start (Segone) and centre excluded groups in setting the questions (Mertens).

MS Segone 516 DM Mertens 384

13

We are not sure our survey measures what we think it measures.

Report reliability honestly, know what alpha cannot show and find the biggest source of error (Cronbach), and check validity across different methods (Campbell).

LC Cronbach 38 DC Campbell 112

14

Stakeholders disagree about what success looks like.

Map values openly (Greene), log claims, concerns and issues and negotiate (Guba and Lincoln), and organise the study around stakeholders' issues (Stake).

JG Greene 340 G&L Guba and Lincoln 318 RS Stake 82

15

Results disappointed: was it the idea or the delivery?

Check the action model before blaming the change model (Chen), and look at process alongside product (Stufflebeam).

HC Chen 174 DS Stufflebeam 134

16

There are too many studies and nobody knows what is really known.

Use systematic reviews and evidence gap maps (White), and think of evidence as an ecology of many forms rather than single studies (Sridharan).

HW White 476 SS Sridharan 246

Write your own story

WORKSHEET · THE 25TH STORY

Every person in this book started with a problem they could not leave alone. Use these prompts to write the first page of your own evaluation story. Come back to it in a year.

1 How did you come to evaluation? Write the honest version: the job, the accident, the person or the frustration that brought you here.

2 What is the one question in your work that you cannot stop asking? Write it as a single sentence.

3 Which two people in this book feel closest to how you work, and which one challenges you most? What would you borrow from the one who challenges you?

4 Describe a turning point in your practice: a moment when a method failed you, or a community, client or colleague changed how you see evaluation.

5 What would you like to have changed in your corner of the field in ten years? Name one tool, network or habit you could build towards it.

6 Who will you teach, mentor or collaborate with so that your ideas travel further than your own projects?

Glossary 20 TERMS

Counterfactual
What would have happened without the programme. Impact evaluations estimate it, often with a comparison or control group.

Randomised evaluation
An impact design in which chance, such as a lottery, decides who receives a programme, so differences in outcomes can be credited to it.

Quasi-experimental design
A design that compares groups or time periods without random assignment, while ruling out rival explanations as far as possible.

Threats to validity
Rival explanations or weaknesses that could make a causal claim, or its generalisation, wrong.

Coefficient alpha
A statistic estimating how consistently a set of survey or test items measures something. It is a lower bound on reliability and does not prove a scale measures a single thing.

Formative evaluation
Evaluation done to improve something while it is being developed or delivered.

Summative evaluation
Evaluation that judges the overall worth of something, usually to inform a decision about it.

Goal-free evaluation
Judging what a programme actually does against the needs of those it serves, without being steered by its stated goals.

Responsive evaluation
Evaluation organised around the concerns and issues of people connected with a programme, often using case study, and reported in forms they can use.

CIPP model
Context, Input, Process, Product: four linked evaluation questions designed to serve decisions throughout a programme.

Theory of change
An explicit account of how and why a programme's activities are expected to lead to its intended outcomes, including the assumptions behind each step.

Change model and action model
In theory-driven evaluation, the change model describes how an intervention produces outcomes; the action model describes how the programme must be delivered to make that happen.

Utilization-focused evaluation
An approach that judges evaluations by whether intended users actually use them, and designs everything around those users and uses.

Developmental evaluation
Evaluation that works alongside innovators in complex, changing situations, giving real-time feedback to support adaptation.

Participatory rural appraisal (PRA)
A family of visual, group-based methods, such as mapping, transect walks and wealth ranking, through which local people produce and analyse their own information.

Trustworthiness
In naturalistic inquiry, the set of criteria used to judge the quality of qualitative work in place of conventional validity and reliability.

Most Significant Change (MSC)

A participatory monitoring technique in which stories of change are collected and then discussed and selected at different levels, revealing what stakeholders value.

Transformative paradigm

A framework that centres people facing discrimination or oppression, attends to power and designs research and evaluation to support social justice.

Made in Africa Evaluation

The call for evaluation in Africa to be grounded in African philosophies, contexts and values and led by African evaluators.

Evidence gap map

A visual summary of where evidence on a set of interventions and outcomes exists and where it is missing.

THE FIVE ACTS AT A GLANCE

ACT 1 What Did They Learn? Measuring What Matters — 1930s to 1960s

RT Tyler 20 · LC Cronbach 38 · MS Scriven 60 · RS Stake 82

ACT 2 The Experimenting Society — 1960s to 1990s

DC Campbell 112 · DS Stufflebeam 134 · CW Weiss 152 · HC Chen 174

ACT 3 Will Anyone Use It? The Use Revolution — 1970s to 2010s

MP Patton 204 · HP Preskill 226 · SS Sridharan 246 · PR Rogers 268

ACT 4 Many Voices, Many Truths — 1970s to 2010s

RC Chambers 296 · G&L Guba and Lincoln 318 · JG Greene 340 · D&D Davies and Dart 362 · DM Mertens 384

ACT 5 Evidence for a World in a Hurry — 1990s to 2020s

RR Rist 410 · D&B Duflo and Banerjee 432 · ZO Ofir 454 · HW White 476 · BC Chilisa 498 · MS Segone 516 · SZ Zaveri 536

Why this book exists

Every tool in your MEL kit was once somebody's answer to a hard question.

Notes and sources

TOOLS AT THE BACK

The public sources behind each part of the book, numbered by section and by figure, with key works to read in the original. Click any address to open it. Facts were checked as of October 2026.

Prologue, acts and epilogue

1 **Evaluation Roots: Tracing Theorists' Views and Influences (Alkin, ed.), first edition, table of contents grouped into Methods, Valuing and Use.** SAGE Publishing, 2004. in.sagepub.com/en-in/sas/evaluation-ro...

2 **Evaluation Roots, library catalogue record with chapter list.** Auckland University of Technology Library, 2004. librarysearch.aut.ac.nz/vufind/Record...

3 **Evaluation Roots: Theory Influencing Practice, third edition (Alkin and Christie, eds.).** Guilford Press, 2023. guilford.com/books/Evaluation-Roots/Al...

4 **A Tree: Planted and Growing (Christie).** Journal of MultiDisciplinary Evaluation 20(48), 2024. jmde.journals.publicknowledgeproject.o...

5 **Visualizing Evaluation Theory: Current Reflections and Future Projections (Mark).** Journal of MultiDisciplinary Evaluation 20(48), 2024. jmde.journals.publicknowledgeproject.o...

6 **Visualizing Evaluation Theory: Tree, Forest, or Ocean Currents? (Mertens).** Journal of MultiDisciplinary Evaluation 20(48), 2024. jmde.journals.publicknowledgeproject.o...

7 **The Pursuit of Social, Economic, and Environmental Justice through Evaluation: Learning from Indigenous Scholars and the Fifth Branch of the Evaluation Theory Tree (Mertens).** Journal of MultiDisciplinary Evaluation 19(44), 2023. jmde.journals.publicknowledgeproject.o...

8 **Program Evaluation Theory and Practice: A Comprehensive Guide (Mertens and Wilson), first edition, publisher description and contents.** Guilford Press via Lavoisier, 2012. lavoisier.fr/livre/autre/program-evalu...

9 **Embedding evaluation theory on African philosophies: An asset to evaluation transformation (Chilisa).** African Evaluation Journal 12(2), 2024. aejonline.org/index.php/aej/article/vi...

10 **Decolonizing and Indigenizing Evaluation Practice in Africa: Toward African Relational Evaluation Approaches (Chilisa, Major, Gaotlhobogwe and Mokgolodi).** Canadian Journal of Program Evaluation 30(3), 2016. ces.journals.uvic.ca/index.php/cjpe/ar...

11 **Indigenous Made in Africa Evaluation Frameworks: Addressing Epistemic Violence and Contributing to Social Transformation (Chilisa and Mertens).** American Journal of Evaluation, via ALNAP, 2021. alnap.org/help-library/resources/indig...

12 **Evaluation Roots: An International Perspective (Carden and Alkin).** Journal of MultiDisciplinary Evaluation 8(17), 2012. jmde.journals.publicknowledgeproject.o...

13 **Foundations of Program Evaluation: Theories of Practice (Shadish, Cook and Leviton), catalogue record and contents.** Technical University of Munich Library, 1991. search.ub.tum.de/vufind/Record/DE-604...

14 **Foundations of Program Evaluation (publisher page).** SAGE Publishing, 1991. sagepub.com/shop/buy-a-book/foundation...

15 **Cronbach, Lee J. (entry).** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

16 **Designing Evaluations of Educational and Social Programs (Cronbach, with Shapiro), catalogue record.** Jossey-Bass, via NIER Library, 1982. nierlib.nier.go.jp/opac/opac_link/bibi...

17 **Elementary and Secondary Education Act (entry).** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

18 **1986 Articles of Consolidation (Evaluation Research Society and Evaluation Network).** American Evaluation Association public library, 1986. comm.eval.org/upgradedcopy/viewdocumen...

19 **Qualitative-Quantitative Debate in Evaluation (entry).** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

20 **The Paradigm Wars and Their Aftermath (Gage), Educational Researcher 18.** STARS, University of Central Florida (CIRS index), 1989. stars.library.ucf.edu/cirs/1820

21 **Philosophy of Science and the Quantitative-Qualitative Debates: Thirteen Common Errors (Shadish).** Evaluation and Program Planning 18(1), 1995. people.ucsc.edu/~ktellez/stadish.pdf

22 **Evaluation Practice 17(2), issue contents: Patton, A World Larger than Formative and Summative; Scriven, Types of Evaluation and Types of Evaluator.** SAGE Journals, 1996. journals.sagepub.com/toc/ajeb/17/2

23 **Instruments, Randomization, and Learning about Development (Deaton), Journal of Economic Literature 48(2).** IDEAS/RePEc record, 2010. ideas.repec.org/a/aea/jeclit/v48y2010i...

24 **Understanding and Misunderstanding Randomized Controlled Trials (Deaton and Cartwright), NBER Working Paper 22595, published in Social Science and Medicine.** NBER, 2016 (working paper); 2018 (journal). nber.org/papers/w22595

25 **The Experimental Approach to Development Economics (Banerjee and Duflo), NBER Working Paper 14467; Annual Review of Economics 1.** NBER, 2008 (working paper); 2009 (journal). nber.org/papers/w14467

26 **The Prize in Economic Sciences 2019, press release.** NobelPrize.org, 2019. nobelprize.org/prizes/economic-science...

27 **Field Experiments and the Practice of Economics (Banerjee, Nobel lecture).** NobelPrize.org, 2019. nobelprize.org/uploads/2019/10/banerje...

28 **Broadening the range of designs and methods for impact evaluations (Stern and colleagues, DFID Working Paper 38).** GSDRC, 2012. gsdrc.org/document-library/broadening...

29 **Made in Africa Evaluation 1. The beginning (Ofir).** Zenda Ofir, Tomorrow and Then blog, 2018. zendaofir.com/made-africa-evaluation-p...

30 **Theory-based impact evaluation: principles and practice (White, 3ie Working Paper 3).** 3ie, 2009. cnxus.org/wp-content/uploads/2022/04/W...

31 **Scriven's Goal-Free Evaluation (Youker).** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...

32 **Experimental and Quasi-Experimental Designs for Research (Campbell and Stanley).** Rand McNally / Houghton Mifflin, 1963. jwilson.coe.uga.edu/emat7050/articles...

33 **Reforms as Experiments (Campbell).** American Psychologist, 1969. sfu.ca/~palys/Campbell-1969-ReformsAsE...

34 **Course Improvement through Evaluation (Cronbach).** Teachers College Record 64(8), 1963. journals.sagepub.com/doi/10.1177/01614...

35 **Lee J. Cronbach 1916 to 2001, Biographical Memoir (Shavelson).** National Academy of Sciences, 2009. nasonline.org/wp-content/uploads/2024...

36 **Eight-Year Study.** Education Encyclopedia, StateUniversity.com. education.stateuniversity.com/pages/19...

37 **Toward a Conceptual Framework for Mixed-Method Evaluation Designs (Greene, Caracelli and Graham).** Educational Evaluation and Policy Analysis, via ALNAP, 1989. library.alnap.org/node/57956

38 **Toward a Methodology of Naturalistic Inquiry in Educational Evaluation (Guba).** CRESST, UCLA, 1978. cresst.org/publication/toward-a-method...

39 **Fourth Generation Evaluation (publisher page).** SAGE Publishing, 1989. collegepublishing.sagepub.com/products...

40 **The origins and practice of participatory rural appraisal (Chambers), World Development 22(7).** University of Hohenheim reading copy, 1994. entwicklungspolitik.uni-hohenheim.de/u...

41 **Current Empirical Research on Evaluation Utilization (Cousins and Leithwood).** Review of Educational Research, via Western Michigan University, 1986. files.wmich.edu/s3fs-public/attachment...

42 **Developmental Evaluation: The Experience and Reflections of Early Adopters (Cabaj).** University of Waterloo, 2011. dspacemainprd01.lib.uwaterloo.ca/serve...

43 **Ten Steps to a Results-Based Monitoring and Evaluation System (Kusek and Rist).** World Bank, 2004. documents1.worldbank.org/curated/en/63...

44 **EvalPartners history.** EvalPartners. evalpartners.org/about/evalpartners-hi...

45 **Building capacity for the evaluation of development activities at the country level, A/RES/69/237.** UN Digital Library, 2014. digitallibrary.un.org/record/789573

46 **About J-PAL.** J-PAL. povertyactionlab.org/about-j-pal

47 **Balsakhi remedial tutoring in Vadodara and Mumbai, India.** J-PAL. povertyactionlab.org/evaluation/balsak...

48 **Making Evaluation Matter: Writings from South Asia.** IDRC and SAGE India, 2014. idrc-crdi.ca/sites/default/files/opene...

49 **Utilization focused evaluation: a primer for evaluators (Ramirez and Brodhead).** IDRC Digital Library, 2013. idl-bnc-idrc.dspacedirect.org/items/80...

50 **Theorists' Theories of Evaluation: A Conversation with Jennifer Greene (Tarsilla).** Journal of MultiDisciplinary Evaluation, 2010. jmde.journals.publicknowledgeproject.o...

Great debate: Proof Here or Promise Elsewhere?

1 **Experimental and Quasi-Experimental Designs for Research (Campbell and Stanley).** Rand McNally / Houghton Mifflin, 1963. jwilson.coe.uga.edu/emat7050/articles...

2 **Reforms as Experiments (Campbell).** American Psychologist, 1969. sfu.ca/~palys/Campbell-1969-ReformsAsE...

3 **Cronbach, Lee J. (entry).** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

4 **Designing Evaluations of Educational and Social Programs (Cronbach, with Shapiro), catalogue record.** Jossey-Bass, via NIER Library, 1982. nierlib.nier.go.jp/opac/opac_link/bibi...

5 **Foundations of Program Evaluation: Theories of Practice (Shadish, Cook and Leviton), catalogue record and contents.** Technical University of Munich Library, 1991. search.ub.tum.de/vufind/Record/DE-604...

6 **Evaluation Roots, library catalogue record with chapter list.** Auckland University of Technology Library, 2004. librarysearch.aut.ac.nz/vufind/Record...

Great debate: Goals or No Goals?

1. **Eight-Year Study.** Education Encyclopedia, StateUniversity.com. education.stateuniversity.com/pages/19...
2. **Scriven's Goal-Free Evaluation (Youker).** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...
3. **A Tree: Planted and Growing (Christie).** Journal of MultiDisciplinary Evaluation 20(48), 2024. jmde.journals.publicknowledgeproject.o...

Great debate: Use or Judgement?

1. **Evaluation Practice 17(2), issue contents: Patton, A World Larger than Formative and Summative; Scriven, Types of Evaluation and Types of Evaluator.** SAGE Journals, 1996. journals.sagepub.com/toc/ajeb/17/2
2. **Current Empirical Research on Evaluation Utilization (Cousins and Leithwood).** Review of Educational Research, via Western Michigan University, 1986. files.wmich.edu/s3fs-public/attachment...
3. **A Tree: Planted and Growing (Christie).** Journal of MultiDisciplinary Evaluation 20(48), 2024. jmde.journals.publicknowledgeproject.o...
4. **Evaluation Roots: Theory Influencing Practice, third edition (Alkin and Christie, eds.).** Guilford Press, 2023. guilford.com/books/Evaluation-Roots/Al...
5. **Developmental Evaluation: The Experience and Reflections of Early Adopters (Cabaj).** University of Waterloo, 2011. dspacemainprd01.lib.uwaterloo.ca/serve...

Great debate: The Paradigm Wars

1. **Qualitative-Quantitative Debate in Evaluation (entry).** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...
2. **The Paradigm Wars and Their Aftermath (Gage), Educational Researcher 18.** STARS, University of Central Florida (CIRS index), 1989. stars.library.ucf.edu/cirs/1820
3. **Philosophy of Science and the Quantitative-Qualitative Debates: Thirteen Common Errors (Shadish).** Evaluation and Program Planning 18(1), 1995. people.ucsc.edu/~ktellez/stadish.pdf
4. **Toward a Conceptual Framework for Mixed-Method Evaluation Designs (Greene, Caracelli and Graham).** Educational Evaluation and Policy Analysis, via ALNAP, 1989. library.alnap.org/node/57956
5. **Toward a Methodology of Naturalistic Inquiry in Educational Evaluation (Guba).** CRESST, UCLA, 1978. cresst.org/publication/toward-a-method...
6. **Fourth Generation Evaluation (publisher page).** SAGE Publishing, 1989. collegepublishing.sagepub.com/products...

Great debate: The Randomistas and Their Critics

1. **The Prize in Economic Sciences 2019, press release.** NobelPrize.org, 2019. nobelprize.org/prizes/economic-science...
2. **The Experimental Approach to Development Economics (Banerjee and Duflo), NBER Working Paper 14467; Annual Review of Economics 1.** NBER, 2008 (working paper); 2009 (journal). nber.org/papers/w14467
3. **Field Experiments and the Practice of Economics (Banerjee, Nobel lecture).** NobelPrize.org, 2019. nobelprize.org/uploads/2019/10/banerje...

4 **Instruments, Randomization, and Learning about Development (Deaton), Journal of Economic Literature 48(2).** IDEAS/RePEc record, 2010.
ideas.repec.org/a/aea/jeclit/v48y2010i...

5 **Understanding and Misunderstanding Randomized Controlled Trials (Deaton and Cartwright), NBER Working Paper 22595, published in Social Science and Medicine.** NBER, 2016 (working paper); 2018 (journal).
nber.org/papers/w22595

6 **Broadening the range of designs and methods for impact evaluations (Stern and colleagues, DFID Working Paper 38).** GSDRC, 2012.
gsdrc.org/document-library/broadening...

7 **Made in Africa Evaluation 1. The beginning (Ofir).** Zenda Ofir, Tomorrow and Then blog, 2018.
zendaofir.com/made-africa-evaluation-p...

8 **Theory-based impact evaluation: principles and practice (White, 3ie Working Paper 3).** 3ie, 2009.
cnxus.org/wp-content/uploads/2022/04/W...

The family tree

1 **Evaluation Roots: Tracing Theorists' Views and Influences (Alkin, ed.), first edition, table of contents grouped into Methods, Valuing and Use.** SAGE Publishing, 2004.
in.sagepub.com/en-in/sas/evaluation-ro...

2 **Evaluation Roots, library catalogue record with chapter list.** Auckland University of Technology Library, 2004.
librarysearch.aut.ac.nz/vufind/Record...

3 **Evaluation Roots: Theory Influencing Practice, third edition (Alkin and Christie, eds.).** Guilford Press, 2023.
guilford.com/books/Evaluation-Roots/Al...

4 **A Tree: Planted and Growing (Christie).** Journal of MultiDisciplinary Evaluation 20(48), 2024.
jmde.journals.publicknowledgeproject.o...

5 **Visualizing Evaluation Theory: Current Reflections and Future Projections (Mark).** Journal of MultiDisciplinary Evaluation 20(48), 2024.
jmde.journals.publicknowledgeproject.o...

6 **Visualizing Evaluation Theory: Tree, Forest, or Ocean Currents? (Mertens).** Journal of MultiDisciplinary Evaluation 20(48), 2024.
jmde.journals.publicknowledgeproject.o...

7 **The Pursuit of Social, Economic, and Environmental Justice through Evaluation: Learning from Indigenous Scholars and the Fifth Branch of the Evaluation Theory Tree (Mertens).** Journal of MultiDisciplinary Evaluation 19(44), 2023.
jmde.journals.publicknowledgeproject.o...

8 **Program Evaluation Theory and Practice: A Comprehensive Guide (Mertens and Wilson), first edition, publisher description and contents.** Guilford Press via Lavoisier, 2012.
lavoisier.fr/livre/autre/program-evalu...

9 **Embedding evaluation theory on African philosophies: An asset to evaluation transformation (Chilisa).** African Evaluation Journal 12(2), 2024.
aejonline.org/index.php/aej/article/vi...

10 **Decolonizing and Indigenizing Evaluation Practice in Africa: Toward African Relational Evaluation Approaches (Chilisa, Major, Gaotlhobogwe and Mokgolodi).** Canadian Journal of Program Evaluation 30(3), 2016.
ces.journals.uvic.ca/index.php/cjpe/ar...

11 **Indigenous Made in Africa Evaluation Frameworks: Addressing Epistemic Violence and Contributing to Social Transformation (Chilisa and Mertens).** American Journal of Evaluation, via ALNAP, 2021.
alnap.org/help-library/resources/indig...

12 **Evaluation Roots: An International Perspective (Carden and Alkin).** Journal of MultiDisciplinary Evaluation 8(17), 2012.
jmde.journals.publicknowledgeproject.o...

13 **Foundations of Program Evaluation: Theories of Practice (Shadish, Cook and Leviton), catalogue record and contents.** Technical University of Munich Library, 1991. search.ub.tum.de/vufind/Record/DE-604...

RT 01 Ralph W. Tyler

Chapter on p. 20

KEY WORKS Basic Principles of Curriculum and Instruction (1949); Evaluation volume of the Eight-Year Study (1942), With Eugene R. Smith.

1 **Obituary of Ralph W. Tyler.** Stanford University News Service, 1994. news.stanford.edu/pr/94/940228Arc4425...

2 **Tyler, Ralph W. (1902 to 1994).** Encyclopedia of Education, via Encyclopedia.com. encyclopedia.com/people/history/histor...

3 **Ralph W. Tyler Papers, finding aid and biographical note.** University of Chicago Library, Special Collections Research Center. lib.uchicago.edu/ead/rlg/ICU.SPCL.TYLE...

4 **Eight-Year Study.** Education Encyclopedia, StateUniversity.com. education.stateuniversity.com/pages/19...

5 **History of NAEP, The Nation's Report Card.** National Center for Education Statistics. nces.ed.gov/nationsreportcard/about/ne...

6 **Books of the Century (Craig Kridel).** Educational Leadership, ASCD, 2000. ascd.org/el/articles/books-of-the-cent...

7 **Basic Principles of Curriculum and Instruction.** Encyclopedia of Curriculum Studies, SAGE, 2010. sk.sagepub.com/ency/edvol/embed/curric...

8 **Lee J. Cronbach 1916 to 2001, Biographical Memoir (Richard J. Shavelson).** National Academy of Sciences, 2009. nasonline.org/wp-content/uploads/2024...

9 **Objectives-Based Evaluation.** Encyclopedia of Evaluation, SAGE (ed. Sandra Mathison), 2005. sk.sagepub.com/ency/edvol/embed/evalua...

10 **Tyler, Ralph W.** Encyclopedia of Evaluation, SAGE (ed. Sandra Mathison), 2005. sk.sagepub.com/ency/edvol/embed/evalua...

11 **Understanding the Tyler rationale: Basic Principles of Curriculum and Instruction in historical context (William G. Wraga).** Espacio, Tiempo y Educacion 4(2), 2017. dialnet.unirioja.es/descarga/articulo...

12 **Basic Principles of Curriculum and Instruction, with a foreword by Peter S. Hlebowitsh (publisher listing).** University of Chicago Press, 2013. chicagoreference.com/ucp/books/book/ch...

13 **An Interview with Ralph Tyler (Jeri Ridings Nowakowski, interview of November 1981).** Journal of MultiDisciplinary Evaluation 7(15), 2011. jmde.journals.publicknowledgeproject.o...

14 **Educational Objectives: The Blueprint for Instruction and Evaluation (George F. Madaus).** The Irish Journal of Education 1(1), via Educational Research Centre, 1967. erc.ie/wp-content/uploads/2025/08/vol0...

LC 02 Lee J. Cronbach

Chapter on p. 38

KEY WORKS Coefficient Alpha and the Internal Structure of Tests (1951); Construct Validity in Psychological Tests (1955), With Paul E. Meehl; Toward Reform of Program Evaluation (1980), With Stanford Evaluation Consortium colleagues.

1 **Lee J. Cronbach 1916 to 2001, Biographical Memoir (Richard J. Shavelson).** National Academy of Sciences, 2009. nasonline.org/wp-content/uploads/2024...

2 **Lee J. Cronbach (1916 to 2001), obituary (Richard J. Shavelson and Goldine Gleser).** American Psychologist, 57(5), 2002. Www.Gwern.net/doc/psychology/2002-shav...

3 **Cronbach, Lee J(oseph) 1916 to 2001.** Contemporary Authors, via Encyclopedia.com. encyclopedia.com/arts/educational-maga...

4 **Cronbach, Lee J.** Encyclopedia of Evaluation, SAGE (ed. Sandra Mathison), 2005. sk.sagepub.com/ency/edvol/embed/evalua...

5 **Course Improvement through Evaluation.** Teachers College Record, 64(8), 1963. journals.sagepub.com/doi/10.1177/01614...

6 **Center for Instructional Research and Curriculum Evaluation.** Encyclopedia of Evaluation, SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

7 **Construct Validity in Psychological Tests (Cronbach and Meehl), Psychological Bulletin 52, 281 to 302.** Classics in the History of Psychology, York University, 1955. psychclassics.yorku.ca/Cronbach/constr...

8 **Invited discussion of Cronbach (1951) Coefficient alpha and the internal structure of tests (Klaas Sijtsma).** Psychometrika, via Tilburg University Research Portal, 2016. research.tilburguniversity.edu/en/publ...

9 **Coefficients and Indices in Generalizability Theory (Robert L. Brennan), CASMA Research Report 1.** University of Iowa, 2003. education.uiowa.edu/sites/education.ui...

MS 03 Michael Scriven

Chapter on p. 60

KEY WORKS Evaluation Thesaurus (1977), Later edition 1991; The Logic of Evaluation (1980); Key Evaluation Checklist (2007), Evaluation Checklist Project.

1 **Even So, Too Soon: Michael John Scriven (28 March 1928 to 28 August 2023), by Gene V. Glass.** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...

2 **Scriven's Goal-Free Evaluation, by Brandon W. Youker.** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...

3 **Revisiting and Expanding Scriven's Fallacies About Formative and Summative Evaluation, by Janet Clinton and John Hattie.** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...

4 **Michael Scriven and The Missing Half of Quantitative Evaluation, by Brian Yates and Nadini Persaud.** Journal of MultiDisciplinary Evaluation 20(47), 2024. jmde.journals.publicknowledgeproject.o...

5 **In memoriam: AERA Past President Michael Scriven.** American Educational Research Association, 2023. aera.net/tabid/20327/mid/52573/default...

6 **Remembering Michael Scriven.** Claremont Graduate University, 2023. cgu.edu/?p=33132

7 **Honoring the legacy of Michael Scriven.** International Evaluation Academy, 2023. ieac.global/honoring-the-legacy-of-mic...

8 **Michael Scriven (1928 to 2023).** European Evaluation Society, 2023. europeanevaluation.org/2023/09/05/mich...

9 **In memoriam: Michael Scriven (1928 to 2023).** Leiter Reports, 2023. leiterreports.com/2023/09/18/in-memori...

10 **Obituary report on Michael Scriven.** Point Reyes Light, 2023. ptreyeslight.com/?p=49070

11 **Goal-free evaluation, by Brandon W. Youker.** SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, 2018. methods.sagepub.com/ency/edvol/embed/s...

12 **Past award winners.** American Evaluation Association. eval.org/About/Awards/Past-Award-Winners

13 **The Logic and Methodology of Checklists, by Michael Scriven (revised December 2007).** The Evaluation Center, Western Michigan University, via MandE NEWS, 2007. mande.co.uk/wp-content/uploads/2007/20...

14 **The Unruly Logic of Evaluation, by Michael Scriven.** Studies in Critical Thinking, Oklahoma State University Open Textbooks. open.library.okstate.edu/kathye/?p=520

15 **Goal-Free Evaluation in Theory and Practice (Evaluation Cafe slides), by Brandon W. Youker.** The Evaluation Center, Western Michigan University, 2017.
files.wmich.edu/s3fs-public/attachment...

16 **Formative, Preformative, and Proformative Evaluation, by Michael Scriven.** Journal of MultiDisciplinary Evaluation 8(18), 2012.
jmde.journals.publicknowledgeproject.o...

RS 04 Robert E. Stake

Chapter on p. 82

KEY WORKS The Countenance of Educational Evaluation (1967), Teachers College Record; The Art of Case Study Research (1995); Qualitative Research: Studying How Things Work (2010).

1 **Stake, Robert E.** Encyclopedia of Evaluation, SAGE (ed. Sandra Mathison), 2005.
sk.sagepub.com/ency/edvol/embed/evalua...

2 **Countenance Model of Evaluation.** Encyclopedia of Evaluation, SAGE, 2005.
sk.sagepub.com/ency/edvol/embed/evalua...

3 **Responsive Evaluation (Tineke A. Abma).** The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, 2018.
methods.sagepub.com/ency/edvol/embed/s...

4 **Center for Instructional Research and Curriculum Evaluation.** Encyclopedia of Evaluation, SAGE, 2005.
sk.sagepub.com/ency/edvol/embed/evalua...

5 **Intrinsic Case Study.** Encyclopedia of Case Study Research, SAGE, 2010.
methods.sagepub.com/ency/edvol/embed/e...

6 **Standards-Based and Responsive Evaluation (publisher listing and author biography).** SAGE Publications, via Woodslane, 2004.
woodslane.com.au/products/9780761926658

7 **Honors item on Robert Stake's honorary doctorate from the University of Valladolid.** Inside Illinois, University of Illinois News Bureau, 2009.
news.illinois.edu/view/6367/210551

8 **Program Evaluation, Particularly Responsive Evaluation (in Evaluation Models, eds Stufflebeam, Madaus and Kellaghan).** Kluwer Academic Publishers, via Illinois Experts, 2002.
experts.illinois.edu/en/publications/p...

9 **A Guide to Responsive Evaluation.** Canadian Institute for Substance Use Research, University of Victoria.
uvic.ca/research/centres/cisur/assets...

10 **Program Evaluation, Particularly Responsive Evaluation (Robert E. Stake, 1973 address; CIRCE Occasional Paper 5).** Journal of MultiDisciplinary Evaluation 7(15), 2011.
jmde.journals.publicknowledgeproject.o...

11 **The Art of Case Study Research (publisher listing).** Sage Publications, via Bookshop.org, 1995.
bookshop.org/p/books/the-art-of-case-s...

12 **Multiple Case Study Analysis (publisher listing).** Guilford Press, 2005.
guilford.com/books/Multiple-Case-Study...

13 **Instrumental Case Study.** Encyclopedia of Case Study Research, SAGE, 2010.
methods.sagepub.com/ency/edvol/embed/e...

DC 05 Donald T. Campbell

Chapter on p. 112

KEY WORKS Experimental and Quasi-Experimental Designs for Research (1963), With Julian C. Stanley; Reforms as Experiments (1969); Quasi-Experimentation: Design and Analysis Issues for Field Settings (1979), With Thomas D. Cook.

1 **Donald T. Campbell: Social Psychologist and Scholar, 1916 to 1996.** APS Observer, Association for Psychological Science, 1996.
psychologicalscience.org/observer/dona...

2 **Donald T. Campbell memorial.** Lehigh University Libraries.
jsaw.lib.lehigh.edu/campbell/memorial...

3 **Experimental and Quasi-Experimental Designs for Research (Campbell and Stanley).** Houghton Mifflin; originally in N. L. Gage (ed.), Handbook of Research on Teaching, Rand McNally, 1963. jwilson.coe.uga.edu/emat7050/articles...

4 **Reforms as Experiments.** American Psychologist, 1969. sfu.ca/~palys/Campbell-1969-ReformsAsE...

5 **Methods for the Experimenting Society.** Prepared for 1971 presentations; revised version in E. S. Overman (ed.), Methodology and Epistemology for Social Science, University of Chicago Press, 1988. sfu.ca/~palys/Campbell-1991-MethodsFor...

6 **Assessing the Impact of Planned Social Change.** Public Affairs Center, Dartmouth College, Occasional Paper 8 (reprinted in Journal of MultiDisciplinary Evaluation, 2011), 1976. jmde.journals.publicknowledgeproject.o...

7 **Modern Regression Discontinuity Analysis (Howard S. Bloom).** MDRC, 2009. mdrc.org/sites/default/files/Regressio...

8 **Collaboration links real life to research on what works.** Penn Today, University of Pennsylvania, 2000. penntoday.upenn.edu/node/148817

9 **International Development Coordinating Group.** 3ie. 3ieimpact.org/resources/International...

10 **Cronbach, Lee J.** Encyclopedia of Evaluation, SAGE (ed. Sandra Mathison), 2005. sk.sagepub.com/ency/edvol/embed/evalua...

DS 06 Daniel L. Stufflebeam

Chapter on p. 134

KEY WORKS Educational Evaluation and Decision Making, With the Phi Delta Kappa National Study Committee on Evaluation; The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability (2017), With Guili Zhang.

1 **Remembering Daniel L. Stufflebeam.** The Evaluation Center, Western Michigan University, 2017. wmich.edu/evaluation/stufflebeam

2 **The parent of evaluation: Daniel Stufflebeam, 1936 to 2017.** Social Science Space, 2017. socialsciencespace.com/2017/08/parent...

3 **Daniel Stufflebeam obituary.** Langeland Family Funeral Homes, 2017. langelands.com/obituaries/daniel-stuff...

4 **Founders: Ohio State.** The Evaluation Center, Western Michigan University. wmich.edu/evaluation/about-0/founders...

5 **Founders.** The Evaluation Center, Western Michigan University. wmich.edu/evaluation/about-0/founders

6 **Standards of Excellence from a Leader of Excellence, by Michael Quinn Patton.** Journal of MultiDisciplinary Evaluation 13(29), 2017. jmde.journals.publicknowledgeproject.o...

7 **Program evaluation standards.** SAGE Encyclopedia of Evaluation, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

8 **Educational Evaluation and Decision Making (catalogue record).** Texas A&M University Libraries, 1971. libcat.library.tamu.edu/Record/in00000...

9 **Evaluation Models: New Directions for Evaluation 89.** AbeBooks listing of the Jossey-Bass volume, 2001. abebooks.fr/9780787957551/Evaluation-M...

10 **About the evaluation checklists.** The Evaluation Center, Western Michigan University. wmich.edu/evaluation/checklists/checkl...

11 **The Application of the CIPP Evaluation Model in Educational Programs in Indonesia.** Scaffolding: Jurnal Pendidikan Islam dan Multikulturalisme, 2025. ejournal.insuriponorogo.ac.id/index.ph...

12 **Programme Evaluation of B.Ed. English at a Remote Higher Education Institution in Nepal Using CIPP Model.** KMC Journal, 2026. nepjol.info/index.php/kmcj/article/vie...

13 **CIPP Evaluation Model Checklist, by Daniel L. Stufflebeam.** BetterEvaluation (listing of the Evaluation Center checklist), 2007. betterevaluation.org/tools-resources/c...

CW 07 Carol H. Weiss

Chapter on p. 152

KEY WORKS Evaluation Research: Methods of Assessing Program Effectiveness (1972); The Many Meanings of Research Utilization (1979); Evaluation: Methods for Studying Programs and Policies (1998).

1 **HGSE remembers Professor Carol Weiss.** Harvard Graduate School of Education, 2013. gse.harvard.edu/news/13/01/hgse-rememb...

2 **Carol Weiss obituary.** Boston Globe via Legacy.com, 2013. legacy.com/us/obituaries/bostonglobe/n...

3 **Weiss, Carol Hirschon.** SAGE Encyclopedia of Evaluation, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

4 **Interview with Carol H. Weiss, by Karen Horsch.** The Evaluation Exchange, Harvard Family Research Project, 1998. archive.globalfrp.org/evaluation/the-e...

5 **The Many Meanings of Research Utilization (summary).** AcaWiki. acawiki.org/The_Many_Meanings_of_Resea...

6 **TOC origins.** Center for Theory of Change. theoryofchange.org/what-is-theory-of-c...

7 **AEA365 post on Carol Weiss and theory-based evaluation.** American Evaluation Association. aea365.org/blog/?p=19062

8 **Chapter on Weiss's Evaluation Research, by Kathryn Newcomer.** The Oxford Handbook of Classics in Public Policy and Administration, 2015. academic.oup.com/edited-volume/27993/c...

9 **Past award winners.** American Evaluation Association. eval.org/About/Awards/Past-Award-Winners

HC 08 Huey-tsyh Chen

Chapter on p. 174

KEY WORKS The multi-goal, theory-driven approach to evaluation (1980), With Peter H. Rossi, Social Forces; Theory-Driven Evaluations (1990); Practical Program Evaluation (2005), Second edition 2015.

1 **Huey Chen biography.** Mercer University College of Health Professions. chp.mercer.edu/bio/huey-chen

2 **Huey T. (Tsyh) Chen author page.** SAGE Publishing. sagepub.com/explore-our-content/author...

3 **Chen, Huey-Tsyh (Encyclopedia of Evaluation entry).** SAGE Knowledge, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

4 **The Multi-Goal, Theory-Driven Approach to Evaluation (Social Forces 59:1).** Oxford Academic, 1980. academic.oup.com/sf/issue/59/1

5 **Evaluating With Sense.** IDEAS/RePEc (Evaluation Review), 1983. ideas.repec.org/a/sae/evarev/v7y1983i3...

6 **Development of a National Evaluation System to Evaluate CDC-Funded Health Department HIV Prevention Programs.** American Journal of Evaluation, 2001. journals.sagepub.com/doi/10.1177/10982...

7 **Review of Practical Program Evaluation.** CDC, Preventing Chronic Disease, 2006. cdc.gov/pcd/issues/2006/jan/05_0121.htm

8 **The bottom-up approach to integrative validity.** IDEAS/RePEc (Evaluation and Program Planning), 2010. ideas.repec.org/a/eee/epplan/v33y2010i...

9 **Evaluation Practice vol. 15 no. 1 contents.** KB National Library of the Netherlands, 1994. ibl.kb.nl/articles/0000000011209/15/1

10 **The Baggaging of Theory-Based Evaluation (Davidson).** Journal of MultiDisciplinary Evaluation, 2006. jmde.journals.publicknowledgeproject.o...

11 **Using Theory to Improve Program and Policy Evaluations.** Bank Square Books (Praeger listing), 1992. banksquarebooks.com/book/9780313283468

12 **Past Award Winners.** American Evaluation Association. eval.org/About/Awards/Past-Award-Winners

13 **Theory-Driven Evaluations (book page and table of contents).** SAGE Publishing, 1990. sagepub.com/shop/buy-a-book/theory-dri...

14 **Theory-driven evaluation (Encyclopedia of Evaluation entry).** SAGE Knowledge, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

15 **Practical Program Evaluation: Foundations of the Field and Current Advancements, third edition.** SAGE College Publishing, 2026. collegepublishing.sagepub.com/products...

MP 09 Michael Quinn Patton

Chapter on p. 204

KEY WORKS Utilization-Focused Evaluation (1978), First edition; fifth edition 2021 with Charmagne E. Campbell-Patton; Getting to Maybe (2006), With Frances Westley and Brenda Zimmerman; Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use (2010 or 2011), Sources differ on the year.

1 **Patton, Michael Quinn (Encyclopedia of Evaluation entry).** SAGE Knowledge, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

2 **Our Team.** Utilization-Focused Evaluation. utilization-focusedevaluation.org/our...

3 **Why Blue Marble Evaluation?** Utilization-Focused Evaluation. utilization-focusedevaluation.org/blue...

4 **Blue Marble Evaluation: Premises and Principles.** Routledge (Guilford Press title), 2019. routledge.com/Blue-Marble-Evaluation-P...

5 **Current Empirical Research on Evaluation Utilization (Cousins and Leithwood).** Review of Educational Research, via Western Michigan University, 1986. files.wmich.edu/s3fs-public/attachment...

6 **Facilitating Evaluation, chapter 4 (sample chapter).** SAGE, 2018. edge.sagepub.com/system/files/ch07_fac...

7 **The Dr. Seuss Approach to Evaluation.** Harvard Family Research Project, The Evaluation Exchange, 1994 talk. archive.globalfrp.org/evaluation/the-e...

8 **Developmental Evaluation: The Experience and Reflections of Early Adopters (Cabaj).** University of Waterloo, 2011. dspacemainprd01.lib.uwaterloo.ca/serve...

9 **Utilization focused evaluation: a primer for evaluators.** IDRC Digital Library, 2013. idl-bnc-idrc.dspacedirect.org/items/80...

10 **American Journal of Evaluation, vol. 17 no. 2 contents.** SAGE Journals, 1996. journals.sagepub.com/toc/ajeb/17/2

11 **Blue Marble Evaluation webinar.** CECAN, 2021. cecan.ac.uk/?p=4834

12 **Past Award Winners; Annual Conference History.** American Evaluation Association. eval.org/About/Awards/Past-Award-Winners

13 **Essentials of Utilization-Focused Evaluation: checklist chapter.** SAGE, 2011. edge.sagepub.com/system/files/ch16_ess...

14 **Essentials of Utilization-Focused Evaluation (product page).** SAGE College Publishing, 2011. collegepublishing.sagepub.com/products...

15 **Utilization-focused evaluation: the new century text (third edition, catalogue record).** BRAC University Library, 1997. library.bracu.ac.bd/vufind/Record/4187

16 **Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use.** Guilford Press, 2010. guilford.com/books/Developmental-Evalu...

17 **Developmental Evaluation Exemplars: Principles in Practice.** Guilford Press, 2015. guilford.com/books/Developmental-Evalu...

18 **Using developmental evaluation to enhance continuous reflection, learning and adaptation (table of DE principles).** Health Research Policy and Systems, BioMed Central, 2020. health-policy-systems.biomedcentral.co...

HP 10 Hallie Preskill

Chapter on p. 226

KEY WORKS Evaluative Inquiry for Learning in Organizations (1999), With Rosalie T. Torres; Evaluation in Organizations: A Systematic Approach to Enhancing Learning, Performance and Change (2001), With Darlene Russ-Eft; second edition 2009; Reframing Evaluation through Appreciative Inquiry (2006), With Tessie Catsambas.

1 **Preskill, Hallie (Encyclopedia of Evaluation entry).** SAGE Knowledge. sk.sagepub.com/ency/edvol/embed/evalua...

2 **Hallie Preskill (author page).** SAGE. sagepub.com/explore-our-content/author...

3 **Hallie Preskill.** FSG. fsg.org/?p=4948

4 **Past award winners.** American Evaluation Association. eval.org/About/Awards/Past-Award-Winners

5 **Building capacity for organizational learning through evaluative inquiry.** Evaluation 5(1), 1999. vibrantlearning.sd44.ca/wp-content/upl...

6 **Evaluative Inquiry for Learning in Organizations.** Woodslane (SAGE distributor), 1999. woodslane.com.au/products/9780761904540

7 **Reframing Evaluation Through Appreciative Inquiry.** SAGE, 2006. sagepub.com/shop/buy-a-book/reframing...

8 **A multidisciplinary model of evaluation capacity building.** American Journal of Evaluation, via ALNAP, 2008. alnap.org/help-library/resources/a-mul...

9 **Building Evaluation Capacity: Activities for Teaching and Training, 2nd ed.** SAGE, 2015. sagepub.com/shop/buy-a-book/building-e...

10 **Evaluating Social Innovation.** Center for Evaluation Innovation, 2012. evaluationinnovation.org/publication/e...

11 **Guide to Evaluating Collective Impact.** FSG, 2014. fsg.org/?p=9262

12 **Appreciative inquiry.** BetterEvaluation. betterevaluation.org/node/52

13 **Guide to Appreciative Inquiry.** UNITAR, 2015. learningwiki.unitar.org/images/f/f5/Gu...

14 **Using Appreciative Inquiry in Evaluation: New Directions for Evaluation, Number 100 (listing).** Hatchards, 2003. hatchards.co.uk/book/using-appreciativ...

15 **Resonance, stickiness and the value propositions of evaluation capacity building.** New Directions for Evaluation (Montclair State University copy), 2021. montclair.edu/ryte-institute/wp-conten...

16 **Communities of practice as an effective ECB approach for nonevaluation specialists.** New Directions for Evaluation (UTS repository), 2024. opus.cloud1.lib.uts.edu.au/bitstream/1...

SS 11 Sanjeev Sridharan

Chapter on p. 246

KEY WORKS Ten steps to making evaluation matter (2011), With April Nakaima, Evaluation and Program Planning; Valuing and embracing complexity (2020), With April Nakaima, Evaluation and Program Planning; From Silos to Ecologies: Towards an Ecology of Evidence to Navigate Solutions (2026), Co-edited with Frans L. Leeuw; release listed for 2026.

1 **Sridharan, Sanjeev (alumni record).** UC Irvine School of Social Ecology. grads.soceco.uci.edu/alumni/sridharan...

2 **Rethinking the role of complexity in global health interventions.** UCSF events calendar, 2019. calendar.ucsf.edu/event/rethinking_the...

3 **Sanjeev Sridharan, speaker.** Asian Evaluation Week, Asian Development Bank. adb.eventsair.com/aew/sanjeev-sridharan

4 **Understanding what evaluation is and why it matters (webinar).** Pan American Health Organization, 2021. paho.org/en/events/understanding-what...

5 **American Journal of Evaluation 27(2), table of contents.** SAGE, 2006. sage.cnpereading.com/toc/AJE/27/2

6 **Ten steps to making evaluation matter.** Evaluation and Program Planning (IDEAS/RePEc record), 2011. ideas.repec.org/a/eee/epplan/v34y2011i...

7 **The Top 10 Questions: A Guide to Evaluating Place-Based Initiatives.** Policy Horizons Canada, 2011. ccednet-rcdec.ca/sites/ccednet-rcdec.c...

8 **Valuing and embracing complexity.** Evaluation and Program Planning (IDEAS/RePEc record), 2020. ideas.repec.org/a/eee/epplan/v80y2020i...

9 **Learning from experiences of evaluators implementing theory-driven evaluations in diverse settings.** Evaluation and Program Planning (IDEAS/RePEc record), 2023. ideas.repec.org/a/eee/epplan/v97y2023i...

10 **From Silos to Ecologies: Towards an Ecology of Evidence to Navigate Solutions.** Routledge, 2026. routledge.com/From-Silos-to-Ecologies...

11 **Towards a transformative view of evaluation (speaker series).** Institute for Work and Health. iwh.on.ca/node/527

12 **Ten steps to making evaluation matter (full text).** Meghalaya Evaluation School (hosted PDF), 2011. meghalayaevaluationschool.com/2022/wp...

PR 12 Patricia J. Rogers

Chapter on p. 268

KEY WORKS Purposeful Program Theory: Effective Use of Theories of Change and Logic Models (2011), With Sue C. Funnell; Developing a research agenda for impact evaluation in development (2014), With Greet Peersman, IDS Bulletin 45; When and how to develop an impact-oriented monitoring and evaluation system (2016), With Peersman, Guijt, Hearn, Pasanen and Buffardi, ODI.

1 **Patricia Rogers in conversation.** Australian Evaluation Society blog, 2019. aes.asn.au/aes-blog/patricia-rogers-in...

2 **Patricia Rogers.** ANZSOG. anzsog.edu.au/faculty-and-expert-contr...

3 **Patricia Rogers.** World Bank Independent Evaluation Group. ieg.worldbankgroup.org/node/56768

4 **Patricia Rogers, speaker.** Asian Evaluation Week, ADB, 2023. asianevaluationweek.adb.org/speakers/p...

5 **Using programme theory to evaluate complicated and complex aspects of interventions.** Evaluation 14(1), via BetterEvaluation, 2008. betterevaluation.org/sites/default/fil...

6 **Purposeful Program Theory (listing).** MandE NEWS, 2011. mande.co.uk/2011/uncategorized/purpose...

7 **AEA and BetterEvaluation partnership.** AEA365, 2013. aea365.org/blog/?p=8785

8 **Better Evaluation Rainbow.** The Change Agency, 2015. thechangeagency.org/better-evaluation...

9 **BetterEvaluation (history).** BetterEvaluation. betterevaluation.org/node/2937

10 **Better Evaluation.** ANZSOG legacy site. legacy.anzsog.edu.au/resource-library...

11 **Theory of Change: Methodological Briefs, Impact Evaluation No. 2.** UNICEF Office of Research (IDEAS/RePEc record), 2014. ideas.repec.org/p/ucf/metbri/innpub747...

12 **Footprint Evaluation (AEA365 post).** AEA365, 2022. aea365.org/blog/?p=26384

13 **Purposeful Program Theory (sample chapter and contents).** Jossey-Bass via e-bookshelf, 2011. download.e-bookshelf.de/download/0000...

RC 13 Robert Chambers

Chapter on p. 296

KEY WORKS Rural Development: Putting the Last First (1983); Whose Reality Counts? Putting the First Last (1997); Handbook on Community-Led Total Sanitation (2008), With Kamal Kar.

1 **The origins and practice of participatory rural appraisal (World Development 22(7)).** University of Hohenheim (reading copy of the Elsevier article), 1994. entwicklungspolitik.uni-hohenheim.de/u...

2 **Interview with Robert Chambers (Development and Change).** Erasmus University Repository, 2013. repub.eur.nl/pub/41116/Metis_189703.pdf

3 **Robert Chambers, Honorary Fellow.** International Institute of Social Studies. iss.nl/en/media/robertchambershonorary...

4 **Robert Chambers author page.** Practical Action Publishing. practicalactionpublishing.com/book-aut...

5 **Revolutionizing Development: Reflections on the Work of Robert Chambers.** Institute of Development Studies, 2011. ids.ac.uk/?p=23115

6 **IDS Bulletin: reflecting on 50 years of Robert Chambers' work.** Institute of Development Studies. ids.ac.uk/?p=90013

7 **From an article to a world view: a lifetime relationship with participation.** Participatory Methods (IDS), 2026. participatorymethods.org/2026/01/14/fr...

8 **Putting the Last First (book notes).** Logan Cochrane, Thought Provokers. logancochrane.com/blog/putting-the-las...

9 **Handbook on Community-Led Total Sanitation.** SuSanA (Plan International UK and IDS), 2008. susana.org/knowledge-hub/resources?id=...

10 **Rapid and participatory rural appraisal: notes on practical approaches and methods.** IDS OpenDocs, 1990. opendocs.ids.ac.uk/articles/online_res...

G&L 14 Egon G. Guba and Yvonna S. Lincoln

Chapter on p. 318

KEY WORKS Naturalistic Inquiry (1985), Lincoln and Guba; But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation (1986), Lincoln and Guba; Fourth Generation Evaluation (1989), Guba and Lincoln.

1 **Egon Guba obituary.** Hillier Funeral Home, 2008. hillierfuneralhome.com/obituaries/egon...

2 **Guba, Egon G.** SAGE Encyclopedia of Evaluation, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

3 **Lincoln, Yvonna S.** SAGE Encyclopedia of Evaluation, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

4 **Yvonna S. Lincoln, author page.** SAGE Publishing. sagepub.com/explore-our-content/author...

5 **Criteria for Assessing Naturalistic Inquiries as Reports (ERIC ED297007), by Lincoln and Guba.** ERIC via Internet Archive, 1988. archive.org/download/ERIC_ED297007/ERI...

6 **Guidelines and Checklist for Constructivist (a.k.a. Fourth Generation) Evaluation, by Guba and Lincoln.** SAGE (study resources), 2001. study.sagepub.com/sites/default/files...

7 **Trustworthiness.** Research Basics, University of Connecticut. researchbasics.education.uconn.edu/tru...

8 **Fourth Generation Evaluation (publisher page).** SAGE Publishing, 1989. collegepublishing.sagepub.com/products...

9 **Toward a Methodology of Naturalistic Inquiry in Educational Evaluation.** CRESST, UCLA, 1978. cresst.org/publication/toward-a-method...

10 **AEA365 post by Jennifer Greene on Guba and Lincoln.** American Evaluation Association. aea365.org/blog/?p=15603

11 **Founders: Ohio State.** The Evaluation Center, Western Michigan University. wmich.edu/evaluation/about-0/founders...

12 **Founders.** The Evaluation Center, Western Michigan University. wmich.edu/evaluation/about-0/founders

JG 15 Jennifer C. Greene

Chapter on p. 340

KEY WORKS Mixed Methods in Social Inquiry, Sole author; Disrupting Program Evaluation and Mixed Methods Research for a More Just Society: The Contributions of Jennifer C. Greene (2023), Tribute volume about her work, edited by Jori Hall, Ayesha Boyce and Rodney Hopson.

1 **Jennifer Caroline Greene profile.** University of Illinois Experts. experts.illinois.edu/en/persons/jennif...

2 **Keynote speakers, Münster 2009.** DeGEval archive, 2009. archiv.degeval.org/tagungsarchiv/jahre...

3 **Theorists' Theories of Evaluation: A Conversation with Jennifer Greene (Tarsilla).** Journal of MultiDisciplinary Evaluation, 2010. jmde.journals.publicknowledgeproject.o...

4 **Toward a Conceptual Framework for Mixed-Method Evaluation Designs.** ALNAP, 1989. library.alnap.org/node/57956

5 **Advancing Mixed-Method Evaluation.** Harvard Family Research Project, The Evaluation Exchange. archive.globalfrp.org/evaluation/the-e...

6 **Mixed Methods in Social Inquiry (author information and excerpt).** Wiley / Jossey-Bass, 2007. catalogimages.wiley.com/images/db/pdf...

7 **The SAGE Handbook of Evaluation.** SAGE Publishing, 2006. in.sagepub.com/en-in/sas/the-sage-hand...

8 **Using a values-engaged educative evaluation approach (Tillman dissertation).** University of Illinois IDEALS, 2015. ideals.illinois.edu/items/73061

9 **The Mystery of Mixing Methods (Vaessen).** World Bank Independent Evaluation Group, 2017. ieg.worldbankgroup.org/comment/3306

10 **Lecture series brings renowned professor of education psychology to ASU.** ASU News, 2010. news.asu.edu/content/lecture-series-br...

11 **Disrupting Program Evaluation and Mixed Methods Research for a More Just Society.** University of Illinois Experts, 2023. experts.illinois.edu/en/publications/d...

12 **Past Award Winners; Annual Conference History.** American Evaluation Association. eval.org/Events/Evaluation-Conference...

D&D 16 Rick Davies and Jess Dart

Chapter on p. 362

KEY WORKS A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique (2003), Dart and Davies, American Journal of Evaluation; The Most Significant Change (MSC) Technique: A Guide to Its Use (2005), Davies and Dart.

1 **The Most Significant Change (MSC) Technique: A Guide to Its Use.** Davies and Dart (copy hosted by UC Agriculture and Natural Resources), 2005. oc4h.ucanr.edu/sites/default/files/202...

2 **Most Significant Change (MSC) special issue page.** Monitoring and Evaluation NEWS. mande.co.uk/?p=27

3 **Order and Diversity: Representing and Assisting Organisational Learning in Non-Government Aid Organisations.** Monitoring and Evaluation NEWS, 1998. mande.co.uk/2017/uncategorized/order-a...

4 **Stories, Evaluation and Learning: Target 10 trial (PLA Notes 38).** IIED, via BetterEvaluation, 2000. betterevaluation.org/sites/default/fil...

5 **A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique.** ALNAP library (American Journal of Evaluation), 2003. library.alnap.org/node/57977

6 **Rick Davies profile.** CEDIL. cedilprogramme.org/?p=4356

7 **Jess Dart speaker profile.** Australian Evaluation Society conference, 2026. aes26.sched.com/jessdart1

8 **Most significant change technique (chapter on dialogue methods).** ANU Press. press-files.anu.edu.au/downloads/press...

9 **Stories of Significance: Redefining Change.** Monitoring and Evaluation NEWS, 2007. mande.co.uk/2007/uncategorized/stories...

DM 17 Donna M. Mertens

Chapter on p. 384

KEY WORKS Transformative Research and Evaluation (2008); Indigenous Pathways to Social Research (2013), Edited with Fiona Cram and Bagele Chilisa; Mixed Methods Design in Evaluation (2018).

1 **Donna M. Mertens author page.** SAGE Publishing. sagepub.com/explore-our-content/author...

2 **Mertens, Donna M. (Encyclopedia of Evaluation entry).** SAGE Knowledge, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

3 **Transformative Research and Evaluation (author and book details).** Guilford Press, 2008. guilford.com/books/Transformative-Rese...

4 **Program Evaluation Theory and Practice, third edition (authors).** Guilford Press, 2025. guilford.com/books/Program-Evaluation...

5 **Donna Mertens recognized for global impact with TEI's first Social Betterment Award.** Claremont Graduate University, 2025. cgu.edu/news/2025/05/donna-mertens-rec...

6 **Transformative Paradigm: Mixed Methods and Social Justice.** Journal of Mixed Methods Research, via SAGE study site, 2007. study.sagepub.com/sites/default/files...

7 **American Journal of Evaluation vol. 20 no. 1 contents.** SAGE Journals, 1999. journals.sagepub.com/toc/aje/20/1

8 **Inclusiveness and Social Justice in Evaluation: A Conversation with Donna M. Mertens (Tarsilla).** Journal of MultiDisciplinary Evaluation, 2010. jmde.journals.publicknowledgeproject.o...

9 **Indigenous Pathways into Social Research: Voices of a New Generation.** MIT Press Bookstore (Left Coast Press listing), 2013. mitpressbookstore.mit.edu/book/9781598...

10 **Avaliação Transformadora para mudanças transformadoras (keynote slides).** ENAP Repository, 2026. repositorio.enap.gov.br/jspui/bitstrea...

11 **Research and Evaluation in Education and Psychology, sixth edition.** Bokus (SAGE listing), 2024. bokus.com/bok/9781071853801/research-a...

12 **Past Award Winners; Annual Conference History.** American Evaluation Association. eval.org/Events/Evaluation-Conference...

RR 18 Ray C. Rist

Chapter on p. 410

KEY WORKS Ten Steps to a Results-Based Monitoring and Evaluation System (2004), With Jody Zall Kusek, World Bank; The Capacity to Evaluate: Why Countries Need It (2006), With Linda G. Morra Imas; The Road to Results: Designing and Conducting Effective Development Evaluations (2009), With Linda G. Morra Imas.

1 **Rist, Ray C. (Encyclopedia of Evaluation).** SAGE, 2005. sk.sagepub.com/ency/edvol/embed/evalua...

2 **Linking Evaluation Theory and Practice: Exploring Ray Rist's Enduring Legacy.** Journal of MultiDisciplinary Evaluation, 2025. jmde.journals.publicknowledgeproject.o...

3 **Linking Evaluation Theory and Practice (full text).** Journal of MultiDisciplinary Evaluation, 2025. jmde.journals.publicknowledgeproject.o...

4 **Capacity Development in Evaluation: The Role of IPDET.** Journal of MultiDisciplinary Evaluation (via ResearchGate), 2025. researchgate.net/publication/392184708...

5 **Ten Steps to a Results-Based Monitoring and Evaluation System.** World Bank, 2004. documents1.worldbank.org/curated/en/63...

6 **Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education.** Harvard Educational Review, 1970. harvardeducationalreview.org/content/4...

7 **The Urban School: A Factory for Failure (catalogue record).** National Institute for Educational Policy Research library, 1973. nierlib.nier.go.jp/opac/opac_link/bibi...

8 **Ray C. Rist author profile.** World Bank Open Knowledge Repository. openknowledge.worldbank.org/entities/p...

9 **Trustees.** Horowitz Foundation for Social Policy. horowitz-foundation.org/trustees

10 **The Clash of the Counter-bureaucracy and Development.** Center for Global Development, 2010. cgdev.org/sites/default/files/1424271...

11 **IPDET: International Program for Development Evaluation Training.** IPDET, 2026. ipdet.org

12 **IPDET Welcomes 171 Participants from 63 Countries to its 2023 On-Site Evaluation Training Program.** Global Evaluation Initiative, 2023. globalevaluationinitiative.org/node/63...

13 **The Capacity to Evaluate: Why Countries Need It (Capacity Development Briefs No. 17), Linda Morra-Imas and Ray C. Rist.** World Bank Open Knowledge Repository, 2006. openknowledge.worldbank.org/entities/p...

D&B 19 Esther Duflo and Abhijit Banerjee

Chapter on p. 432

KEY WORKS Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty (2011), Banerjee and Duflo; Handbook of Field Experiments, Volumes 1 and 2 (2017), Edited by Duflo and Banerjee; Good Economics for Hard Times (2019), Banerjee and Duflo.

1 **Abhijit Banerjee: Biographical.** NobelPrize.org, 2019. nobelprize.org/prizes/economic-science...

2 **Esther Duflo: Biographical.** NobelPrize.org, 2019. nobelprize.org/prizes/economic-science...

3 **Press release: The Prize in Economic Sciences 2019.** NobelPrize.org, 2019. nobelprize.org/prizes/economic-science...

4 **Field Experiments and the Practice of Policy (Nobel lecture).** NobelPrize.org, 2019. nobelprize.org/uploads/2019/10/duflo-l...

5 **Field Experiments and the Practice of Economics (Nobel lecture).** NobelPrize.org, 2019. nobelprize.org/uploads/2019/10/banerje...

6 **About J-PAL.** J-PAL. povertyactionlab.org/about-j-pal

7 **Balsakhi remedial tutoring in Vadodara and Mumbai, India.** J-PAL. povertyactionlab.org/evaluation/balsak...

8 **Improving Immunization Rates Through Regular Camps and Incentives in India.** J-PAL. povertyactionlab.org/evaluation/improv...

9 **Duflo and Banerjee to Join UZH.** University of Zurich, 2025. news.uzh.ch/en/articles/media/2025/Duf...

10 **Instruments, Randomization, and Learning about Development.** Journal of Economic Literature (author copy, Princeton), 2010. princeton.edu/~deaton/downloads/Instru...

11 **French academic becomes youngest ever winner of Nobel prize for economics.** The Local France, 2019. thelocal.fr/20191014/french-academic-b...

12 **Broadening the range of designs and methods for impact evaluations.** GSDRC (DFID Working Paper 38), 2012. gsdrc.org/document-library/broadening...

13 **Using Randomization in Development Economics Research: A Toolkit (NBER Technical Working Paper 333).** National Bureau of Economic Research, 2006. nber.org/papers/t0333

14 **Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of Teaching at the Right Level in India (NBER Working Paper 22746).** EdTech Hub library record, 2016. docs.edtechhub.org/lib/LEGZTSPV

ZO 20 Zenda Ofir

Chapter on p. 454

KEY WORKS Making Evaluation Our Own (statement from the 4th AfrEA conference, Niger) (2007), Session led with Sulley Gariba; Post on the International Evaluation Academy and evaluation for transformation (AEA365) (2022), With Scott Chaplowe.

1 **Zenda Ofir (fellow profile).** Robert Bosch Academy. robertboschacademy.de/en/zenda-ofir

2 **Zenda Ofir.** Claremont Evaluation Center New York. research.cgu.edu/cec-ny/people/zenda-o...

3 **Zenda Ofir.** Bounce Beyond. bouncebeyond.global/zenda-ofir

4 **Zenda Ofir, Evaluation Reference Group.** CGIAR Independent Advisory and Evaluation Service. iaes.cgiar.org/node/12271

5 **Zenda Ofir (professional profile).** LinkedIn. ch.linkedin.com/in/zendaofir

6 **Made in Africa Evaluation 1. The beginning.** Zenda Ofir, Tomorrow and Then blog, 2018. zendaofir.com/made-africa-evaluation-p...

7 **Tomorrow and Then (blog home page).** zendaofir.com. zendaofir.com

8 **CECAN seminar with Zenda Ofir.** CECAN, 2018. cecan.ac.uk/?p=519

9 **International Evaluation Academy week post (Ofir and Chaplowe).** AEA365, 2022. aea365.org/blog/?p=25179

10 **3 Horizons Evaluation Initiative.** International Evaluation Academy, 2024. ieac.global/?p=4416

11 **Research Quality Plus.** IDRC. idrc-crdi.ca/en/rqplus

12 **Embedding evaluation theory on African philosophies: An asset to evaluation transformation.** African Evaluation Journal 12(2), 2024. aejonline.org/index.php/aej/article/do...

13 **Research Quality Plus: a holistic approach to evaluating research (catalogue record).** IDRC Digital Library, 2016. idl-bnc-idrc.dspacedirect.org/items/a0...

14 **Research Quality Plus: a holistic approach to evaluating research.** IDRC, 2018. idrc-crdi.ca/en/stories/research-quali...

15 **A better measure of research from the global south (journal club summary of Lebel and McLean, Nature 2018).** Research Impact Canada, 2018. researchimpact.ca/archived/a-better-me...

16 **IDRC Research Quality Plus: A Holistic Approach to Evaluating Research.** On Think Tanks, 2016. onthinktanks.org/resource/idrcresearch...

17 **Three horizons: a pathways practice for transformation (Sharpe, Hodgson, Leicester, Lyon and Fazey), Ecology and Society 21(2).** University of Dundee Discovery, 2016. discovery.dundee.ac.uk/en/publications...

HW 21 Howard White

Chapter on p. 476

KEY WORKS Theory-based impact evaluation: principles and practice (2009), Journal of Development Effectiveness 1(3); Journal of Development Effectiveness, Founding managing editor; 3ie evidence gap maps (2015), First five maps published by 3ie.

1 **Professor Howard White (biography).** World Health Organization. who.int/about/people/biography/howard...

2 **Howard White.** Global Development Network. gdn.int/author/howard-white-director-e...

3 **Howard White (profile).** Guidelines International Network. g-i-n.net/?p=4897

4 **Howard White, presenter biography.** KTDRR 2019 conference, 2019. ktdrr.org/conference2019/presenters/ho...

5 **Theory-based impact evaluation: principles and practice (3ie Working Paper 3).** 3ie, 2009. cnxus.org/wp-content/uploads/2022/04/W...

6 **Q&A with Howard White, Head of New Impact Evaluation Entity, 3IE.** Center for Global Development. cgdev.org/node/1423609

7 **When Will We Ever Learn? Closing the Evaluation Gap.** Center for Global Development, 2006. cgdev.org/page/when-will-we-ever-learn...

8 **3ie Annual Report 2015.** 3ie, 2015. 3ieimpact.org/sites/default/files/2018...

9 **15 years of 3ie: reflecting on our contributions and the work ahead.** 3ie, 2023. 3ieimpact.org/blogs/15-years-3ie-refle...

10 **The twenty-first century experimenting society: the four waves of the evidence revolution.** Palgrave Communications (IDEAS/RePEc record), 2019. ideas.repec.org/a/pal/palcom/v5y2019i1...

11 **Guidance for producing a Campbell evidence and gap map.** Campbell Systematic Reviews (PMC), 2020. pmc.ncbi.nlm.nih.gov/articles/PMC8356343

12 **Some reflections on current debates in impact evaluation.** 3ie via ALNAP, 2009. alnap.org/help-library/resources/some...

BC 22 Bagele M. Chilisa

Chapter on p. 498

KEY WORKS Indigenous Research Methodologies (2012), Second edition 2019; Decolonising and Indigenising Evaluation Practice in Africa: Toward African Relational Evaluation Approaches (2016); Made in Africa Evaluation, African Evaluation Journal 12(2) (2024), Commissioned and funded by EvalPartners.

1 **Professor Chilisa appointed to UNDP Evaluation Advisory Panel.** University of Botswana, 2018. ub.bw/news/professor-chilisa-undp-eval...

2 **Mastercard Foundation funds Pan African Indigenous Evaluation Consortium project.** University of Botswana. ub.bw/news/mastercard-foundation-funds...

3 **Decoloniality and indigenous methods in global health evaluation.** London School of Hygiene and Tropical Medicine, 2024. lshtm.ac.uk/newsevents/events/decoloni...

4 **Indigenous Research Methodologies, second edition.** SAGE, 2019. in.sagepub.com/en-in/ind/indigenous-re...

5 **Review of Indigenous Research Methodologies (Van der Westhuizen).** African Evaluation Journal, 2013. aejonline.org/index.php/aej/article/vi...

6 **Review of Indigenous Research Methodologies (Hanson).** International Journal of Critical Indigenous Studies, 2012. ijcis.qut.edu.au/article/download/97/97

7 **Decolonizing and Indigenizing Evaluation Practice in Africa: Toward African Relational Evaluation Approaches.** Canadian Journal of Program Evaluation 30(3), 2016. ces.journals.uvic.ca/index.php/cjpe/ar...

8 **Indigenous Made in Africa Evaluation Frameworks (Chilisa and Mertens).** American Journal of Evaluation 42(2), 2021. journals.sagepub.com/doi/abs/10.1177/1...

9 **Towards defining and advancing Made in Africa Evaluation (Omosa et al.).** African Evaluation Journal 9(1), 2021. aejonline.org/index.php/aej/article/do...

10 **Embedding evaluation theory on African philosophies: An asset to evaluation transformation.** African Evaluation Journal 12(2), 2024. aejonline.org/index.php/aej/article/do...

11 **Bagele Chilisa.** Wikipedia. en.wikipedia.org/wiki/Bagele_Chilisa

12 **Made in Africa Evaluation blog series, part 3.** SAMEA, 2017. samea.org.za/?p=2809

13 **Research ethics principles (teaching slides citing Chilisa 2012 on the four Rs).** National Centre for Research Methods. repository.ncrm.ac.uk/resources/online...

MS 23 Marco Segone

Chapter on p. 516

KEY WORKS How to Design and Manage Equity-Focused Evaluations; Evaluation for Equitable Development Results; Evaluating the SDGs with a no one left behind lens.

1 **Marco Segone (speaker profile).** Asian Evaluation Week, Asian Development Bank. asianevaluationweek.adb.org/speakers/m...

2 **Peer Review Panel Members biographies 2020 to 2021.** UNEG, 2020. uneval.org/sites/default/files/file_up...

3 **Marco Segone (author profile).** Academia.edu. independent.academia.edu/MarcoSegone

4 **EvalPartners history.** EvalPartners. evalpartners.org/about/evalpartners-hi...

5 **Previous co-chairs.** EvalPartners. evalpartners.org/about/previous-co-cha...

6 **2015 declared as International Year of Evaluation.** EvalPartners, 2013. evalpartners.org/evalyear/2015-declare...

7 **Building capacity for the evaluation of development activities at the country level, A/RES/69/237.** UN Digital Library, 2014. digitallibrary.un.org/record/789573

8 **From Policies to Results: Developing Capacities for Country Monitoring and Evaluation Systems.** UNICEF, 2010. theoryofchange.org/wp-content/uploads...

9 **How to design and manage equity-focused evaluations.** ALNAP (UNICEF publication), 2011. alnap.org/help-library/resources/how-t...

10 **Eval4Action campaign concept note.** European Evaluation Society, 2020. europeanevaluation.org/wp-content/uplo...

11 **Annual report on the evaluation function 2020.** UNFPA, 2021. unfpa.org/sites/default/files/admin-re...

12 **Note verbale on recruitment of the Director, Independent Evaluation Office.** UNFPA, via SKNIS, 2026. sknis.gov.kn/wp-content/uploads/2026/0...

13 **How to design and manage Equity-focused evaluations (download page).** GIFE Sinapse, 2011. sinapse.gife.org.br/dlm_file/how-to-de...

SZ 24 Sonal Zaveri

Chapter on p. 536

KEY WORKS The Importance of Context in Participatory Evaluations: Reflections from South Asia (2014), Chapter in Making Evaluation Matter: Writings from South Asia, edited by Katherine Eve Hay and Shubh Kumar-Range; Real-Time Evaluation of Gender Integration in the UNICEF COVID-19 Response in South Asia (2020 to 2021), Evaluation led with UNICEF gender and M&E staff.

1 **Sonal Zaveri, Evaluation Reference Group.** CGIAR Independent Advisory and Evaluation Service. iaes.cgiar.org/node/12272

2 **Q and A: Sonal Zaveri on the power and potential of real-time evaluation.** CGIAR, 2023. cgiar.org/news-events/news/qa-sonal-za...

3 **Making Evaluation Matter: Writings from South Asia.** IDRC and SAGE India, 2014. idrc-crdi.ca/sites/default/files/opene...

4 **TWG1 Newsletter no. 1 2022 (article on the UNICEF real-time gender evaluation).** European Evaluation Society, 2022. europeanevaluation.org/wp-content/uplo...

5 **Practical Evaluation Capacity Development: Lessons from South Asia.** Global Evaluation Week, 2022. glocalevalweek.org/event/practical-eva...

6 **DECI final technical report.** BetterEvaluation (IDRC project), 2012. betterevaluation.org/sites/default/fil...

7 **GENSA: A Movement for Gender Equity in South Asia.** Southern Voice, 2025. southernvoice.org/?p=27624

8 **About Sonal Zaveri (author page).** Southern Voice, 2025. southernvoice.org/author/sonal-zaveri

9 **Making Evaluation Matter: Writings from South Asia (full text copy).** Tamil Nadu Agricultural University (hosted PDF of the SAGE and IDRC edition), 2014. agritech.tnau.ac.in/pdf/2014/Making_Ev...

Lee J. Cronbach, in one line

He taught us that a number is only as trustworthy as the argument behind it, and that evaluation serves everyone who shapes what happens next.

Picture credits

TOOLS AT THE BACK

Portraits are vector adaptations of the original photos; CC BY-SA adaptations are shared under the same licence. Every other figure is shown by an original HubForge emblem that stands for their big idea, not by a likeness.



Ralph W. Tyler

Photo of Ralph W. Tyler, University of Massachusetts Amherst faculty photographs, via Wikimedia Commons, CC BY-SA 4.0

Licence: **CC BY-SA 4.0** · Chapter on **p. 20**



Michael Quinn Patton

Photo by Omar David Sandoval Sida, EVALAC 2023, via Wikimedia Commons, CC BY-SA 4.0

Licence: **CC BY-SA 4.0** · Chapter on **p. 204**



Robert Chambers

Photo by Vysotsky, The Hague (ISS) 2015, via Wikimedia Commons, CC BY-SA 4.0

Licence: **CC BY-SA 4.0** · Chapter on **p. 296**



Esther Duflo and Abhijit Banerjee

Esther Duflo: photo by US Embassy Sweden, Nobel week 2019, via Wikimedia Commons, CC BY 2.0
Abhijit Banerjee: photo by Financial Times, FT Goldman Sachs Business Book of the Year Award 2011, via Wikimedia Commons, CC BY 2.0

Licence: **CC BY 2.0** · Chapter on **p. 432**

Michael Quinn Patton, in one line

Michael Quinn Patton taught a whole profession that an evaluation is only as good as the difference it makes in someone's hands.

All 24 emblems



Tyler



Cronbach



Scriven



Stake



Campbell



Stufflebeam



Weiss



Chen



Patton



Preskill



Sridharan



Rogers



Chambers



Guba and Lincoln



Greene



Davies and Dart



Mertens



Rist



Duflo and Banerjee



Ofir



White



Chilisa



Segone



Zaveri

LICENCES

CC BY-SA 4.0 and CC BY 2.0 are Creative Commons licences that allow reuse with credit; CC BY-SA adaptations must be shared under the same licence. The emblems and all other artwork in this book are by HubForge Global Impact Network Foundation.

An honest note on method

HOW THIS BOOK WAS MADE

Why we made this book

Every tool in your MEL kit was once somebody's answer to a hard question. We curated this book to celebrate the people who asked those questions first, to make their big ideas clear enough for anyone to use, and to inspire a new generation to join the work. Behind it sits one purpose: a stronger evaluation ecosystem, of practitioners, students, commissioners, governments and communities, using evidence well to help build a fairer and better world.

Who made it

The book was curated by Shashidhar S A, founder of HubForge Global, and is published by HubForge Global Impact Network Foundation. The curator chose the 24 figures, set the brief for every chapter, directed the writing and design through many rounds of revision, and reviewed and approved the final text. The choices of who to include, what to emphasise and how to tell each story are ours, and so is the responsibility for them.

How we chose the 24

We chose the 24 to show how the field grew, not to rank anyone. We looked for people whose ideas changed everyday practice, across different eras, regions, methods and roles, from university theorists to field practitioners and movement builders, with a deliberate place for voices from Africa and South Asia. Many others shaped the field, and we hope future editions will tell more of their stories.

Where the facts come from

Each profile was built from a research fact sheet compiled from public sources: university and institutional pages, publishers' author pages, professional association records such as the American Evaluation Association's award lists, encyclopaedia entries, obituaries and the person's own published work. More than 300 of these sources are listed in Notes and sources, from p. 569, as a starting point for your own reading rather than a complete bibliography. Where sources disagreed on a date or detail, we left it out or hedged it, for example by writing around 1990 or the mid-1960s. Where a fact rested on one weak source, or could not be confirmed, we left it out.

How we used AI

We used artificial intelligence openly in making this book. AI tools helped us to search and summarise public sources, draft and edit the text, check consistency and lay out the pages, always under the curator's direction. We did not let AI supply facts of its own: every fact in the book is meant to trace back to a listed public source. We ran automated checks that every number, date and name in the rewritten chapters already appeared in the sourced research, followed by a separate fact-checking pass that corrected errors and removed claims the sources did not support. AI can still make mistakes, and so can people, which is why we invite corrections.

How we tell the stories

The storytelling is ours. Chapter openings, passages that invite you to imagine a situation, the Picture it comparisons, the Your moves cards, the worked examples, templates, AI prompts and reflection questions are HubForge's own writing and interpretation. They are not claims about what any person said, believes or would recommend. Worked examples are invented for teaching unless a source is named. We did not invent scenes, dialogue, thoughts or feelings, and we deliberately avoided direct quotations, because quotations are often misremembered, trimmed or lifted out of context. Instead, we describe each person's

ideas in our own words and point you to their original work.

Portraits and emblems

Five figures are shown by portraits adapted from openly licensed photographs: cropped, traced into vector shapes and recoloured, and credited in Picture credits, from p. 590. Every other figure is shown by an original HubForge emblem that stands for their big idea. The emblems are not likenesses.

Respect for the people

We have not stated whether a person is living unless that was clearly confirmed, we have kept current job titles to a minimum, and we have kept to each person's public professional record. Criticisms are summaries of published debate, offered in good faith for learning, not personal judgements. Inclusion in this book is not an endorsement by the people profiled, their families or their institutions.

Corrections and limits

Facts were checked against public sources as of October 2026. If you spot an error, an omission or an outdated detail, or if you are a person profiled and want something corrected, write to hubforgeglobal@gmail.com and we will review it promptly and fix the next edition. This book is for learning and inspiration, not professional advice; please read the Legal notices, from p. 593, before relying on it.

OUR HOUSE RULES

Every fact traced to a public source. No invented scenes, dialogue or quotations. Hedged dates where sources disagree. No statement of living status unless confirmed. AI used openly, never as a source of facts. Facts checked as of October 2026. Spotted an error? Write to hubforgeglobal@gmail.com.

Legal notices

PLEASE READ BEFORE RELYING ON THIS BOOK

General information, not advice. This book is for general education and inspiration. It is not legal, financial, ethical, medical or other professional advice. Before you act on anything in it, use your own judgement, follow your organisation's policies and the law where you work (including data protection, safeguarding and research ethics rules), and take qualified advice where it matters.

Accuracy and no warranty. The profiles were compiled in good faith from public sources available up to October 2026, and are listed in Notes and sources. Public records can be incomplete or wrong, and dates, roles and details change. The book is provided as is, without any warranty, express or implied, of accuracy, completeness, fitness for a purpose or freedom from error.

Our interpretation, not theirs. Lessons, advice, worked examples, illustrations, templates, AI prompts, reflection questions and anything marked as our reading are HubForge's own interpretation. They are not statements by, and must not be attributed to, the people profiled or their institutions. Worked examples and illustrations are invented for teaching; any resemblance to a real programme, organisation or person is coincidental unless a source is given.

No endorsement or affiliation. Inclusion in this book does not mean that any person profiled, their family, estate, employer or publisher, or any organisation named, has reviewed, approved, endorsed, sponsored or is affiliated with this book or with HubForge, unless we say so explicitly.

Debates and criticism. Our accounts of debates, criticisms and limits summarise published scholarship and public discussion. They are offered fairly and in good faith for learning, and are not personal judgements about anyone's character, conduct or work.

Corrections and removal requests. If you find an error, or if you are a person profiled, a family member, an estate or a rights holder and want a detail corrected or a portrait or emblem changed or removed, write to hubforgeglobal@gmail.com. We will review every request promptly and in good faith, and correct the digital edition and the next print edition where needed.

Pictures and licences. The portraits of Ralph W. Tyler, Michael Quinn Patton, Robert Chambers, Esther Duflo and Abhijit Banerjee are adaptations of photographs published on Wikimedia Commons under Creative Commons licences: CC BY-SA 4.0 for Tyler, Patton and Chambers, and CC BY 2.0 for Duflo and Banerjee. We changed the originals by cropping, tracing them into vector shapes and recolouring them. The adapted portraits are shared under the same licence as their source photograph. The photographers and the people shown do not endorse this book. All other figures are represented by original HubForge emblems, which are not likenesses.

Third-party content and links. Links and references to external websites, publications and tools are provided for convenience. We do not control them and are not responsible for their content, availability or accuracy, and they may change after publication.

AI tools. The AI prompts are starting points. Outputs from AI tools can be wrong, biased or invented. Check every output against your evidence, do not paste confidential or personal data into tools that are not approved for it, and follow the terms of the tool you use.

Limitation of liability. To the fullest extent permitted by law, HubForge Global Impact Network Foundation, the curator and contributors are not liable for any loss, damage, claim or expense, direct or indirect, arising from the use of, or reliance on, this book or its contents. Nothing in these notices excludes any liability that cannot be excluded by law.

Trademarks and names. Names of people, organisations, publications, methods and products are used only to identify and describe them, and may be trademarks of their owners.

Governing law. These notices and any use of this book are governed by the laws of India, and the courts at Bengaluru, Karnataka have jurisdiction, unless the law where you live requires otherwise.

Licence
Copyright © 2026 HubForge Global Impact Network Foundation. Text, design and original artwork licensed under CC BY-NC-ND 4.0, except third-party material as stated in Picture credits. For permissions, translations or commercial use, write to hubforgeglobal@gmail.com.

THE FIVE GREAT DEBATES

Proof Here or Promise Elsewhere? p. 196
Should evaluation first prove that a programme caused an effect, or show whether the effect will travel?

Goals or No Goals? p. 104
Should a programme be judged against its own stated objectives, or against everything it actually does?

Use or Judgement? p. 290
Is an evaluation's first duty to be used by the people who need it, or to deliver an independent verdict on worth?

The Paradigm Wars p. 402
Do qualitative and quantitative evaluation rest on incompatible views of reality, or can they be combined?

The Randomistas and Their Critics p. 554
Are randomised trials the best route to knowing what works against poverty, or one tool among many?

Rick Davies and Jess Dart, in one line

Rick Davies and Jess Dart showed that a well chosen story, with its reasons written down, can carry a village's judgement all the way to the boardroom.

Acknowledgements WITH THANKS

This book stands on the work of many people. Our first thanks go to the 24 giants in these pages, and to the colleagues, students and communities who worked beside them. We have tried to tell their ideas fairly, and any errors are ours.

We are grateful to the universities, professional associations, publishers, journals, libraries and archives whose public records made the research possible, and to the photographers who shared their pictures under open licences.

Thank you to the evaluation community around the world, from national associations to global networks such as AfrEA, IOCE, EvalPartners, EvalYouth and Eval4Action, for keeping the question of how we would know alive in every region.

And thank you to every MEL practitioner in a district office, a village meeting or a ministry. This book is for you.

Index NAMES, IDEAS AND ORGANISATIONS

Page numbers in bold mark a chapter about the person. Notes and sources, from **p. 569**, list the public sources behind every page.

3

3ie, 116, 407, 477 to 482, 484, 486, 489, 492, 495 to 496

A

African Evaluation Association (AfrEA), 407, 455 to 458, 460 to 464, 466, 473 to 474, 500 to 502, 510, 513

Alkin, Marvin, 202, 291, 408, 504, 509, 558 to 559

American Evaluation Association, 8, 62 to 63, 84 to 86, 101, 137, 154, 176, 206 to 207, 227 to 229, 243, 270 to 271, 273, 282, 284

Appreciative inquiry, 227 to 231, 235, 237, 239, 241 to 244

B

Banerjee, Abhijit, 9, 19, 106, 111, 119, 122, 125, 129, 132 to 133, 190 to 191, 198, 203, **432 to 453**

BetterEvaluation, 269 to 273, 282 to 284, 286 to 287

Blue Marble Evaluation, 205 to 208, 210 to 211, 217 to 218, 220, 223

C

Campbell Collaboration, 116, 125, 478 to 479, 489 to 490, 492, 495

Campbell, Donald T., 9, 19, 43, 45, 48, 51, 53 to 55, 58 to 59, 106, 108 to 109, **112 to 133**

Campbell's law, 119, 125 to 127, 132

Caracelli, Valerie, 293, 341 to 343, 348 to 353, 403

Case study, 83 to 85, 87 to 89, 93, 95 to 102, 115, 120, 320, 328, 508, 567

Catsambas, Tessie, 227 to 229, 235, 237, 244

Chambers, Robert, 9, 19, 95, 106, 111, 128, 198, 203, 237, 292 to 295, **296 to 317**, 331

Chen, Huey-tsyh, 9, 19, 30, 106, 108 to 111, 159, 162, 168, 172 to 173, **174 to 195**, 196

Chilisa, Bagele M., 9, 19, 51, 55, 67, 73, 80 to 81, 98, 106, 111, 198, 203, 294, **498 to 515**

Christie, Christina, 104 to 105, 202, 291, 559

CIPP, 26, 36 to 37, 109, 135 to 140, 142 to 144, 146 to 151, 320, 567

CIRCE, 40, 83 to 84, 86, 91 to 92, 95, 101

Coefficient alpha, 17, 39 to 42, 44 to 46, 48, 51, 55 to 58, 567

Community of Evaluators South Asia, 407 to 408, 537 to 538, 540, 546, 551

Construct validity, 17, 39 to 41, 44 to 45, 48 to 51, 55, 57 to 58

Cook, Thomas, 113, 115 to 116, 122, 132, 197

Counterfactual, 119, 132 to 133, 181, 194 to 195, 436, 438 to 440, 444, 452 to 453, 482 to 486, 489, 491, 493 to 497, 562, 566

Cram, Fiona, 385 to 388, 399 to 400, 500, 502

Cronbach, Lee J., 9, 16 to 19, 33, **38 to 59**, 63, 68, 70, 77, 106, 111, 117, 119

D

Dart, Jess, 9, 19, 73, 89, 98, 102 to 103, 106, 111, 128, 198, 203, 237, 292, **362 to 383**

Davies, Rick, 9, 19, 73, 89, 98, 102 to 103, 106, 111, 128, 198, 203, 237, 292, **362 to 383**

Deaton, Angus, 437, 555, 558 to 559

Developmental evaluation, 69, 184, 201, 205 to 211, 214 to 215, 217, 220 to 221, 223 to 225, 230 to 231, 241, 253, 266 to 267, 277, 291

Duflo, Esther, 9, 19, 106, 111, 119, 122, 125, 129, 132 to 133, 190 to 191, 198, 203, **432 to 453**

E

Eight-Year Study, 17, 21 to 23, 25 to 26, 30 to 31, 33, 36, 39 to 40, 45, 58 to 59, 104

Enlightenment, 109, 155 to 159, 163 to 165, 169, 171, 291

Equity-focused evaluation, 249 to 250, 391, 400 to 401, 518, 521, 524 to 527, 531 to 534

Estes, Nolan, 135 to 136

Eval4Action, 518 to 521, 529 to 531, 533

EvalPartners, 407, 499, 502, 514, 517 to 521, 524, 528 to 531, 533 to 535, 538, 542, 552 to 553

Evaluation capacity building, 227 to 229, 231, 238, 240 to 241, 243 to 245

Evaluation Center, Western Michigan University, 62, 72, 76, 136 to 139, 144, 320 to 321, 323

Evaluative inquiry, 201, 211, 224 to 225, 227 to 234, 241 to 244

EvalYouth, 518 to 521, 528, 531, 533

Evidence gap map, 477 to 480, 489, 491, 494, 496, 565, 568

Experimenting society, 9, 19, 106, 108 to 109, 111, 113, 117, 119, 122 to 123, 125, 128, 132, 198, 202 to 203

External validity, 43, 109, 115 to 118, 120 to 122, 127, 188, 197, 280, 321, 324, 326, 328, 402, 438

F

Fiske, Donald, 51, 114, 118, 129 to 130

Formative evaluation, 42, 44, 53, 66, 68, 70, 78, 80, 215 to 216, 567

Fourth generation evaluation, 86, 89, 102 to 103, 293, 319 to 325, 328 to 329, 331, 334 to 336, 338, 402

Funnell, Sue, 201, 269, 271, 275, 278 to 279, 281, 286, 288

G

Gage, N. L., 115, 122, 293

Gariba, Sulley, 407, 455 to 457, 460 to 464, 472 to 474, 555, 560

Generalizability theory, 39, 44 to 48, 51, 55, 57

Glass, Gene, 61 to 62, 64 to 66

Goal-free evaluation, 17, 24, 26, 33, 36 to 37, 61 to 67, 70 to 71, 73, 76 to 77, 79 to 81, 104 to 105, 270, 275, 288 to 289

Greene, Jennifer C., 9, 19, 95, 106, 111, 198, 202 to 203, 292 to 295, 320, 322, **340 to 361**

Guba, Egon G., 9, 19, 65, 76, 89, 95, 98 to 99, 102 to 103, 106, 111, 136 to 137, **318 to 339**

H

Hindsman, Edwin, 137

I

IDEAS, 21, 25, 35, 39, 45, 55, 57, 61, 67, 79, 83, 89, 101, 113

IDRC, 171, 214, 457 to 460, 468 to 471, 473, 538, 542, 546

Institute of Development Studies (IDS), 269, 275, 288, 297 to 298, 300 to 301, 309, 477

Internal validity, 43, 53, 109, 115, 117 to 118, 120, 124, 188, 190, 196, 324, 326, 328

International Evaluation Academy, 65, 455 to 458, 460 to 461, 465, 467, 471, 473 to 474, 538

IOCE, 456 to 458, 473, 517 to 519, 524, 528 to 529

IPDET, 207, 220, 407, 412 to 414, 416, 424 to 426, 429

J

J-PAL, 407, 434 to 436, 438, 442, 445, 450 to 451

K

Kennedy, Robert, 8, 109

Keppel, Francis, 24

Key Evaluation Checklist, 61, 64 to 67, 73 to 76, 79 to 80

Kremer, Michael, 407, 434 to 435, 440, 442, 555

Kusek, Jody Zall, 407, 411 to 414, 418, 420, 423, 428 to 430

L

Lewin, Kurt, 155

Lincoln, Yvonna S., 9, 19, 65, 76, 89, 95, 98 to 99, 102 to 103, 106, 111, 141, 144, **318 to 339**

Logframe, 18, 24, 33, 37, 71 to 73, 78, 94, 105, 215, 373 to 375, 485

Logic model, 168, 177, 179, 185, 187, 190, 202, 221, 254, 263, 268 to 269, 271 to 272, 275 to 277, 279

M

Made in Africa Evaluation, 407, 455 to 462, 464, 466 to 467, 472 to 475, 499 to 503, 505, 511, 514 to 515, 555, 568

McCall, William, 114, 122

Meehl, Paul, 17, 39 to 42, 45, 49 to 51, 58

Mertens, Donna M., 9, 19, 106, 111, 147, 198, 203, 292 to 293, 295, 312, 328, 331, 334, **384 to 401**

Metaevaluation, 61, 64, 66, 74, 84, 135, 139 to 140, 143 to 147, 149

Mixed methods, 325, 338 to 339, 341 to 345, 347, 350 to 351, 353, 356, 358 to 360, 385 to 388, 391 to 392, 394 to 395, 397, 399 to 400, 477

Morra Imas, Linda, 407, 411 to 414, 420, 424, 426, 430

Most Significant Change, 64, 89, 102 to 103, 105, 293, 303, 316 to 317, 363 to 371, 373, 375, 378 to 382, 568

Mullainathan, Sendhil, 407, 434 to 435, 451

N

NAEP, 24, 26, 35

Nakaima, April, 201, 247 to 249, 251, 253, 256 to 257, 259 to 260, 262, 265 to 266

Naturalistic inquiry, 95, 293, 319 to 321, 324 to 326, 328, 331 to 335, 338, 402, 567

O

Ofir, Zenda, 9, 19, 106, 111, 198, 203, 220, 295, 404, 406 to 409, **454 to 475**, 505

P

Participatory rural appraisal, 293, 297 to 300, 302 to 303, 306 to 307, 311 to 315, 567

Patton, Michael Quinn, 9, 19, 45, 54, 58 to 59, 70, 77, 95, 106, 111, 138 to 139, 141, **204 to 225**

Personal factor, 201, 207, 210, 212, 220, 224, 290

Policy-shaping community, 41, 44 to 45, 52 to 55, 57, 197

Pratham, 407, 435, 437 to 438, 443 to 445, 449, 451 to 452

Preskill, Hallie, 9, 19, 106, 111, 165, 198, 200 to 203, 211, 214, 221, 224, **226 to 245**

Program Evaluation Standards, 140, 145 to 147

Program logic, 201, 269 to 270, 278, 281

Q

Quasi-experiment, 115, 117 to 120, 122, 125, 128, 132, 179, 555

R

Rainbow Framework, 269, 271, 273 to 274, 282 to 283, 285, 287

Randomised, 43, 54, 107, 119, 122 to 123, 125, 132 to 133, 182, 189, 199, 405, 407 to 408, 433 to 434, 436

Readiness assessment, 411, 413, 416 to 423, 426 to 429, 564

Real-time evaluation, 211, 221, 224 to 225, 313, 448, 470, 537 to 543, 548 to 549, 552 to 553

Regression-discontinuity, 116, 118, 120 to 121, 123, 129

Responsive evaluation, 24, 83 to 89, 91 to 93, 95, 98, 100 to 103, 105, 325, 331, 338 to 339, 509, 567

Rist, Ray C., 9, 19, 106, 111, 125, 128, 141, 144, 150 to 151, 165, 168, 198, 203, **410 to 431**

Rogers, Patricia J., 9, 19, 67, 80 to 81, 106, 111, 159, 162, 168, 172 to 173, 181, 184, **268 to 289**

Rossi, Peter H., 109, 175 to 178, 184, 193, 197, 558

S

Scriven, Michael, 9, 16 to 19, 24 to 26, 30, 33, 36 to 37, 54 to 55, **60 to 81**

Segone, Marco, 9, 19, 106, 111, 168, 198, 203, 295, 391, 400 to 401, 404, 406 to 407, **516 to 535**

Seva Mandir, 435, 442, 451

Shadish, William, 109, 197, 403

Shavelson, Richard, 40, 44, 48, 57

Sridharan, Sanjeev, 9, 19, 106, 111, 162, 168, 198, 200 to 203, 217, **246 to 267**, 275, 278

Stake, Robert E., 9, 16 to 19, 24, 33, **82 to 103**, 104 to 107, 109, 111, 198

Stanley, Julian C., 109, 113 to 117, 120 to 122, 132, 182, 196

Stern, Elliot, 437, 481, 555

Stufflebeam, Daniel L., 9, 19, 26, 30, 33, 36 to 37, 70, 76 to 77, 92, 99, 106, 108, **134 to 151**

Summative evaluation, 17, 61 to 63, 66 to 68, 77, 79 to 80, 180, 215 to 216, 230, 290, 567

Systematic review, 116, 138, 261, 478, 480, 482 to 483, 487, 489 to 491, 493 to 495, 565

T

Teaching at the Right Level, 433, 436, 438 to 439, 442 to 444, 448 to 449, 451 to 452

Theory-based evaluation, 26, 36 to 37, 153, 157 to 160, 165, 170 to 172, 181, 194 to 195, 275, 288 to 289, 479

Theory-driven evaluation, 109, 159, 172 to 173, 175 to 179, 181 to 182, 184, 187, 190, 192 to 193, 249, 483, 496 to 497, 567

Theory of change, 153, 156 to 159, 162, 169 to 170, 172 to 173, 178 to 179, 261, 271 to 272, 274, 277, 279 to 281, 283, 285 to 289, 377

Torres, Rosalie, 201, 227 to 229, 232 to 234, 244

Transformative paradigm, 294, 347, 360 to 361, 385 to 392, 395 to 400, 403, 534 to 535, 559, 568

Trustworthiness, 293, 319, 321 to 322, 324 to 326, 328, 333, 336 to 338, 402, 561, 564, 567

Tyler, Ralph W., 8 to 9, 16 to 19, **20 to 37**, 39 to 41, 45, 48, 54 to 55, 58

U

UNICEF, 271 to 272, 281, 500, 517 to 521, 524 to 525, 527 to 529, 532, 537 to 543, 547 to 549, 551 to 552

Utilization-focused evaluation, 201, 205 to 208, 210 to 214, 217, 219 to 220, 222 to 224, 244 to 245, 290, 567

UTOS, 44, 53 to 55, 57

W

Weiss, Carol H., 9, 19, 26, 30, 36 to 37, 54, 106, 108 to 111, 122, 129, **152 to 173**

White, Howard, 9, 19, 76, 106, 111, 119, 122, 125, 129, 132 to 133, 162, 168, 181, **476 to 497**

Wholey, Joseph, 109, 180, 559

World Bank, 136, 138, 154, 156, 171, 237, 298, 300, 302, 341, 344, 350, 357, 408

Y

Youker, Brandon, 61, 64 to 65, 71 to 73, 104 to 105

Z

Zaveri, Sonal, 9, 19, 106, 111, 168, 198, 203, 208, 211, 214, 221, 224 to 225, 295, **536 to 553**

Zimmerman, Brenda, 205, 208, 217, 224, 270 to 271, 276, 278

ABOUT HUBFORGE

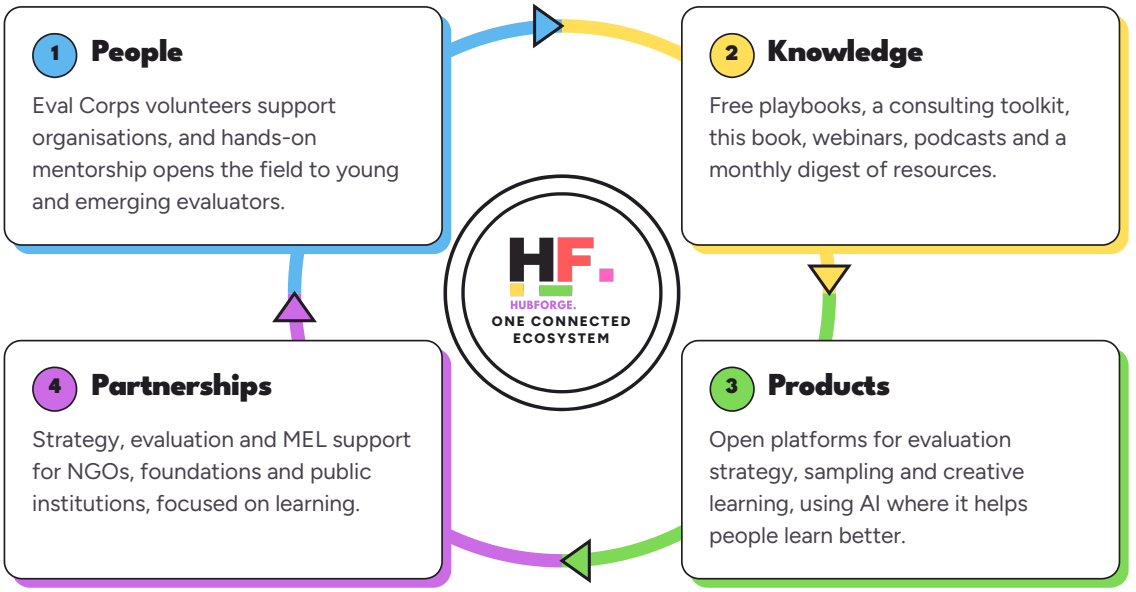
Strengthening the world's evidence ecosystem

MONITORING EVALUATION RESEARCH LEARNING

The giants in this book did not work alone. Their ideas spread because people taught them, tested them, argued over them and passed them on. HubForge exists to keep that going.

HubForge Global Impact Network Foundation is a not-for-profit organisation founded in India in October 2025, with a global purpose: a home for everyone who works to generate and use evidence for social good.

Rigorous evidence. Open to all. Made with joy.



Each part feeds the others. Volunteers learn by supporting real organisations, what they learn becomes open tools and resources, and income from services goes back into public goods for everyone.

ABOUT HUBFORGE

Open tools, built for the field

EvalCanvas

Build an evaluation strategy step by step, with an open MEL library, learning games and a marketplace of MERL partners.

EVALCANVAS.COM

Sampling·Shala

A free global hub to learn sampling, build sampling strategies and calculate sample sizes.

SAMPLINGSHALA.COM

Playbooks

by HubForge

Free playbooks, templates and workbooks for monitoring, evaluation, consulting and better decisions with data and AI.

PLAYBOOKS.HUBFORGEGLOBAL.COM

EvalStudio

An open creative space to turn ideas from everyday work into stickers, workshop materials and memes.

PLAYBOOKS.HUBFORGEGLOBAL.COM

The first year

AS OF OCTOBER 2026

68 countries reached by EvalCanvas

336 organisations on EvalCanvas

630+ people using EvalCanvas

4 free playbooks

10 volunteers

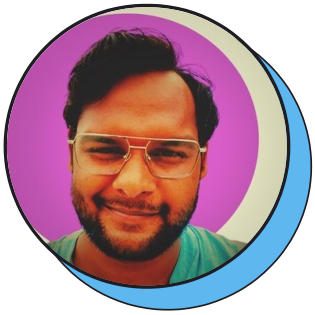
5 NGOs supported

3 webinars

Join in

If this book has sparked something, there is a place for you: as a volunteer, a mentee, a partner or a supporter.

hubforgeglobal@gmail.com · www.hubforgeglobal.com



ABOUT THE CURATOR

Shashidhar S A

Founder, HubForge Global. Driving evaluation ecosystems through people, process and product.

MERL & AI INNOVATOR · EN-INTRAPRENEUR · ECOSYSTEM BUILDER

WHY THIS BOOK

Shashidhar has long loved the lives, ideas, experiments and breakthrough moments of scientists, mathematicians and entrepreneurs: the stories behind original thinking, and the way those stories inspire people. He curated Giants of Evidence to bring that kind of story to monitoring, evaluation and learning, to celebrate the people who built the field, and to help a new generation ask better questions of its own.

An engineer by training, Shashidhar has spent about a decade in monitoring, evaluation and learning across India's social sector. His path runs through seven roles, from grassroots field work to state-level government partnerships and the design of MEL systems.

In the global evaluation community

- An active volunteer, contributing to **IOCE**, the International Organization for Cooperation in Evaluation
- Contributes to **Eval4Action** and works alongside the **EvalYouth Global Network**

What he builds

- **HubForge**, a not-for-profit strengthening the evidence ecosystem
- Open tools: **EvalCanvas**, **SamplingShala** and **EvalStudio**
- **Eval Corps** and the **HubForge Playbooks** library

HUBFORGE.

Write to the curator with corrections, ideas for the next edition and requests to use the book with your team: **hubforgeglobal@gmail.com** · **linkedin.com/in/shashidharsa**

Share this book freely for non-commercial use, with credit (CC BY-NC-ND 4.0).

ALSO FROM HUBFORGE

Take the ideas into your next assignment

This book shows where the field's ideas came from. The HubForge playbooks and toolkit show how to put them to work, with AI as a collaborator.

PLAYBOOK 01 · DESIGN

Designing Dashboards with AI

Make dashboards people can use to act. Follow the six-stage DESIGN framework, bring AI in as a collaborator and try each idea with practical exercises.

27-PAGE PDF

PLAYBOOK 02 · EVIDENCE

Claude Dashboard Playbook

Build evidence dashboards: start with a prompt, connect your data, explore results, trace numbers to their sources and share your work.

12-PAGE PDF

PLAYBOOK 03 · BETTER QUESTIONS

Designing Surveys and Forms with AI

Design surveys and forms with the six-stage SURVEY method, question design tools, a workbook and XLSForms.

PLAYBOOK AND WORKBOOK

PLAYBOOK 04 · CONSULT

Consulting in M&E with AI

Win, run and land monitoring and evaluation assignments with the seven-stage CONSULT method, 36 printable templates and terms of reference checks.

60-PAGE PDF

THE M&E CONSULTING TOOLKIT

From Zero to World-Class

36 templates and 36 concept guides: 21 tools across the seven CONSULT stages, nine documents every consultant writes and six tools for your brand and network.

TOOLKIT

READ AND DOWNLOAD

playbooks.hubforgeglobal.com

Every HubForge playbook, toolkit and template in one place.

Get in touch hubforgeglobal@gmail.com · www.hubforgeglobal.com

A HUBFORGE MEL BOOK

Every tool in your MEL kit was once somebody's answer to a hard question.

Meet the 24 giants who built monitoring, evaluation and learning, from the 1930s to today. Each chapter tells a life, the spark and the breakthrough, then sets out every big idea in full: the problem, how it works, a worked example and when to use it.

Then it turns those ideas into practice, with moves you can make this month, the great debates and a family tree of who shaped whom.

24	5	67	316
LIVES	ACTS	BIG IDEAS	PUBLIC SOURCES



The next giant of evidence could be you.

Curated by **Shashidhar S A**, founder of HubForge Global. For practitioners, students, commissioners and anyone who wants evidence to change lives.

HUBFORGE.

Published by **HubForge Global Impact Network Foundation**

Lessons, templates and prompts are HubForge's own reading of each person's ideas, not claims about what they said or would recommend. Inclusion is not an endorsement by the people profiled. Facts checked against public sources as of October 2026.