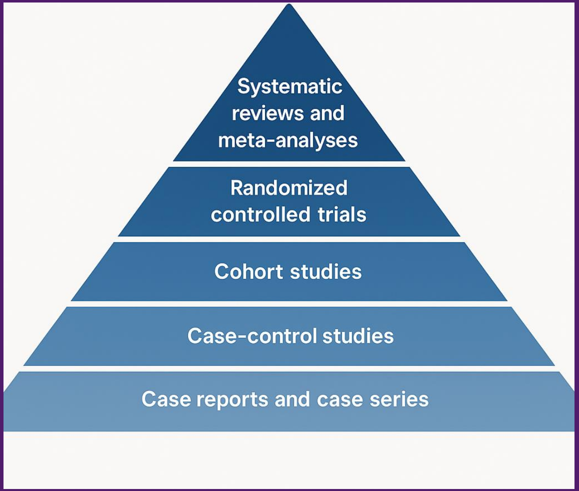
Experimental Design and the Standards of Evidence in Education


Systematic reviews and meta-analyses
Randomized controlled trials
Cohort studies
Case-control studies
Case reports and case series

Drs Tom Perry[1] and Matt Horton[2]

[1]Associate Professor (Reader), University of Warwick
[2]Evaluation Manager – University of Wolverhampton

NERUPI evaluation series
10th June 2025

tom.perry@warwick.ac.uk @TWPerry1 https://thomasperry.education/

WARWICK

Dr Tom Perry

Associate Professor (Reader), University of Warwick

Images: OpenAI

tom.perry@warwick.ac.uk @TWPerry1 @twperry1.bsky.social https://thomasperry.education/

Dr Matt Horton

Evaluation Manager University of Wolverhampton (APP, TEF & REC)

- Two decades of experience in impact evaluation. Led research across education, health, and social care sectors, (e.g., Sure Start, Mencap, LAs) with a focus on disadvantaged communities, improving outcomes and social mobility.
- 2008–2020 - Evaluation Manager for the Aimhigher West Midlands partnership (OFFA, APP, UCP, NCOP).
- PhD thesis in Widening Participation (impact of Summer Sch. + mentoring)
- Current - lead pre- and post-entry WP evaluation at the University of Wolverhampton. Interests include evaluation methods, toolkit validation, and demonstrating meaningful impact.

e: Matthew.horton@wlv.ac.uk LinkedIn: (2) Dr Matthew Horton | LinkedIn

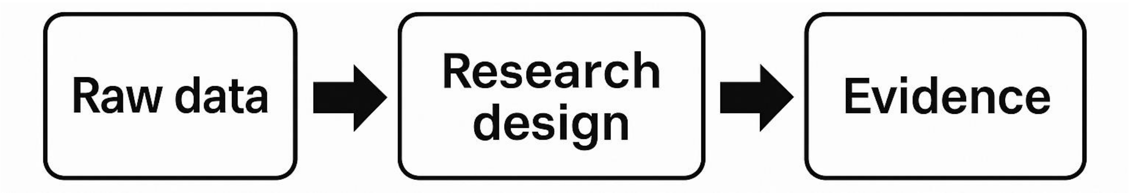
This Session

- A critical look at experimental design and the standards of evidence we use in education
- We’ll explore three key questions:
  - What do we mean by standards of evidence in education? (e.g., OfS APPs)
  - Why are experimental designs – especially RCTs – placed at the top of evidence hierarchies?
  - What does this mean for how we design and evaluate intervention strategies and write Access & Participation Plans in HE?

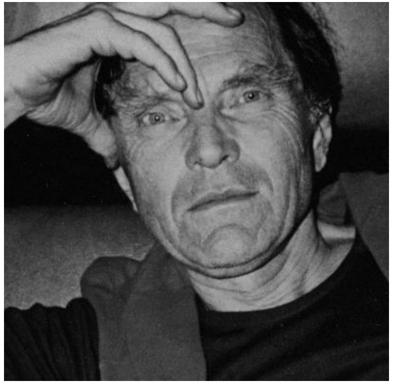
Experimental Design and Standards of Evidence

WARWICK

From data to evidence


Raw data → Research design → Evidence

- Different research designs solve different inference problems
- In open research there are no fixed rules
- In evaluation and “what works” contexts, we often narrow to specific types of questions:
  - Does this intervention work? What impact does it have?
- This drives the demand for standards of evidence – especially around causation

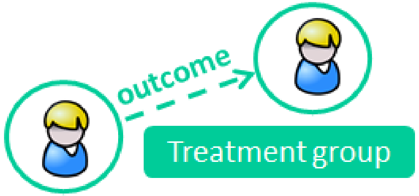


“anything goes!”
(Feyerabend)

Empirical & Causal Evaluation

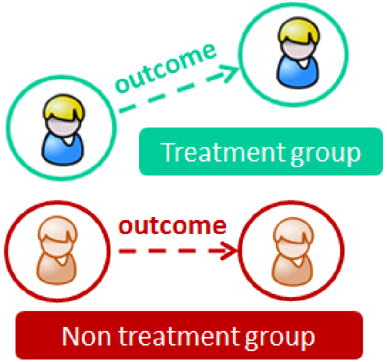
Experimental methods – what would have happened without the intervention?

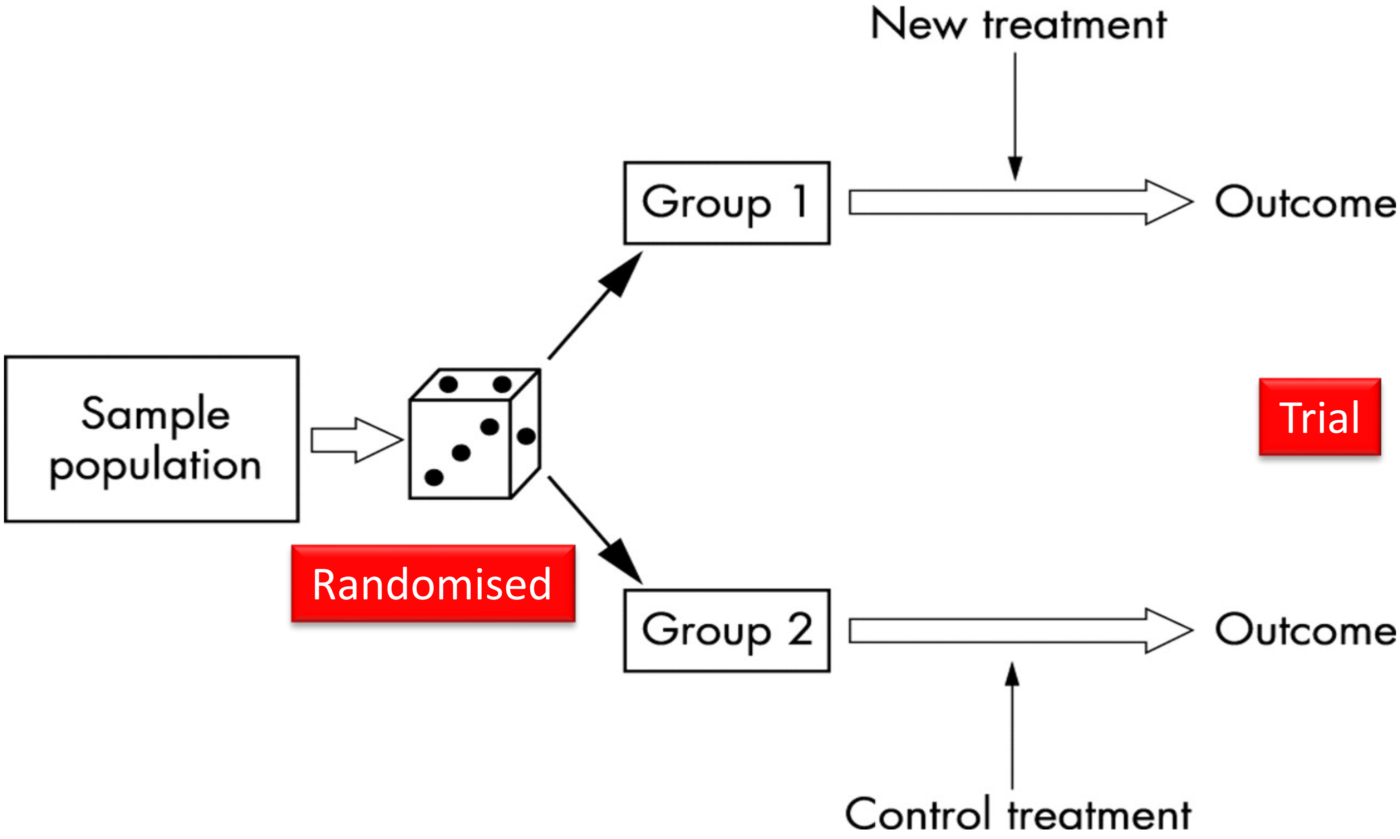
1 An improvement after taking part compared to before the activity (e.g. via a pre and post event questionnaire / assessment data). Sample includes those engaged in intervention only.



2 **Quasi-exp:** difference in outcomes between treatment & comparison group. Groups matched in terms most imp. variables affecting outcome.

3 **RCTs or Regression discontinuity design (RDD)** – compares outcomes between treatment & control group. More able to provide causal findings. Randomisation to groups reduces selection bias of known & unknown variables affecting outcome.





New treatment

Group 1

Outcome

Sample population

Trial

Randomised

Group 2

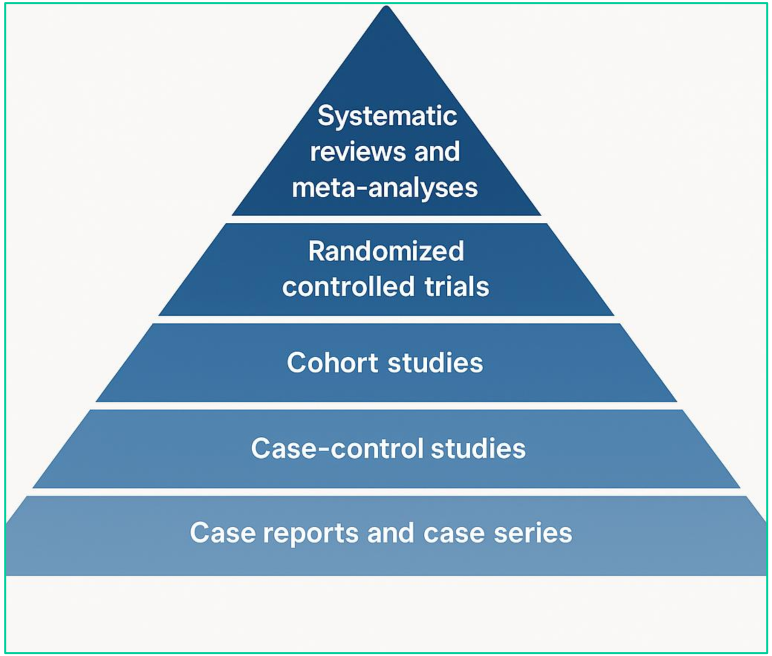
Outcome

Control treatment

Controlled

Leading to…

- Evidence hierarchies rank methods by their ability to make **causal claims**
- RCTs are placed at the top because they:
  - Reduce bias through random assignment
  - Provide clean comparisons between treated and control groups
  - Are seen as objective and generalisable
- Systematic reviews and meta-analyses of RCTs are ranked highest



Systematic reviews and meta-analyses

Randomized controlled trials

Cohort studies

Case-control studies

Case reports and case series

Standards of Evidence (OfS)

	Description of impact	Evidence	Claims you can make
Type 1: Narrative	Evaluation provides a narrative / coherent ToC / logic model to motivate selection of activities in the context of a coherent strategy	Refer to evidence – needs of target groups: • Data (gaps in outcomes) • Yours/others existing evidence • Research literature & theory (how changes = improvement).	We have a coherent explanation of **what** we do & **why** based on evidence. Claims are research-based. **ToC: minimum for all projects**
Type 2: Empirical enquiry	Evaluation evidence = those receiving an intervention have better outcomes.	Quant. &/qual. evidence of a pre/post intervention change or a difference compared to what might otherwise have happened	Interventions ***associated*** with beneficial results. Not causal. **Type 2: high cost projects**
Type 3: Causality	The impact evaluation methodology provides evidence of a causal effect of an intervention (e.g. RCT / Quasi exp)	Quant. &/or qual. evidence of a pre/post treatment change on participants relative to an control/comparison group who did not take part in the intervention	Intervention ***causes*** improvement demonstrated via a control / comparison group (no selection bias). **Type 3: high cost / pilots**

Resources and detailed guidance see: OfS and TASO websites

New Access & Participation Plans (APP)

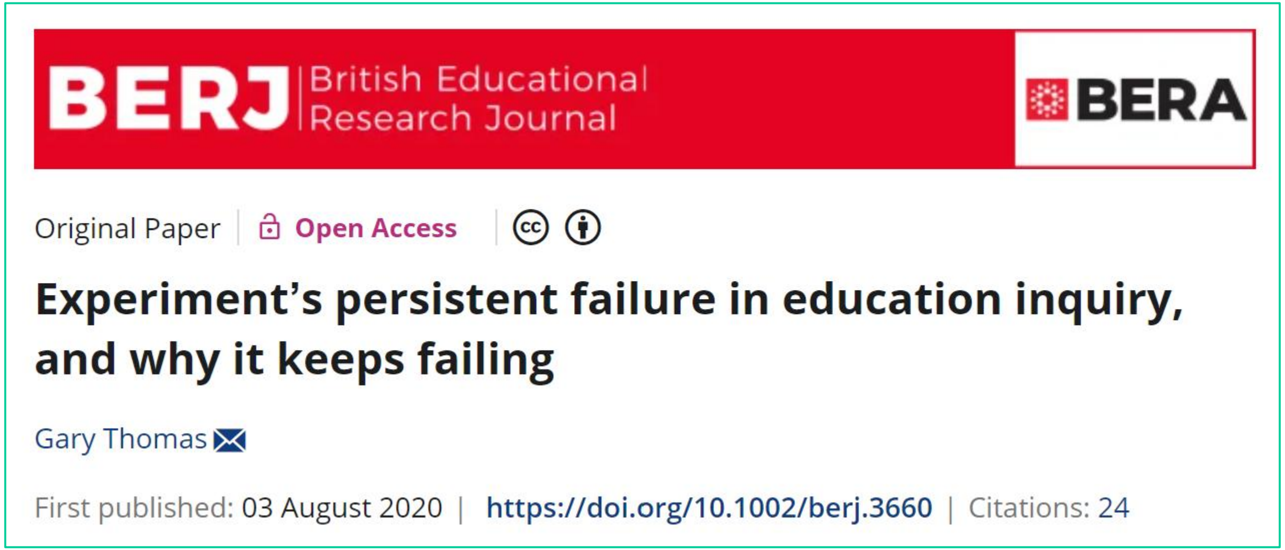
- HEPs charging higher tuition fee rate required to submit APP to the OfS.
- APPs set out how HEPs will close gaps in student outcomes across access + beyond.
- New APPs provide a stronger emphasis on evaluation to understand what works. Key expectations include to improve the:
  - Quantity of evaluation
  - Quality of evaluation with reference to OfS standards of evidence
  - Publication of findings

Failure to improve student outcomes across access and the student lifecycle can lead to OfS sanctions

How helpful are experimental designs (and RCTs specifically) for educational evaluation?

What are they good/not so good for?

Critiques of Experimental Design (/RCTs)



BERJ British Educational Research Journal

BERA

Original Paper | Open Access

Experiment's persistent failure in education inquiry, and why it keeps failing

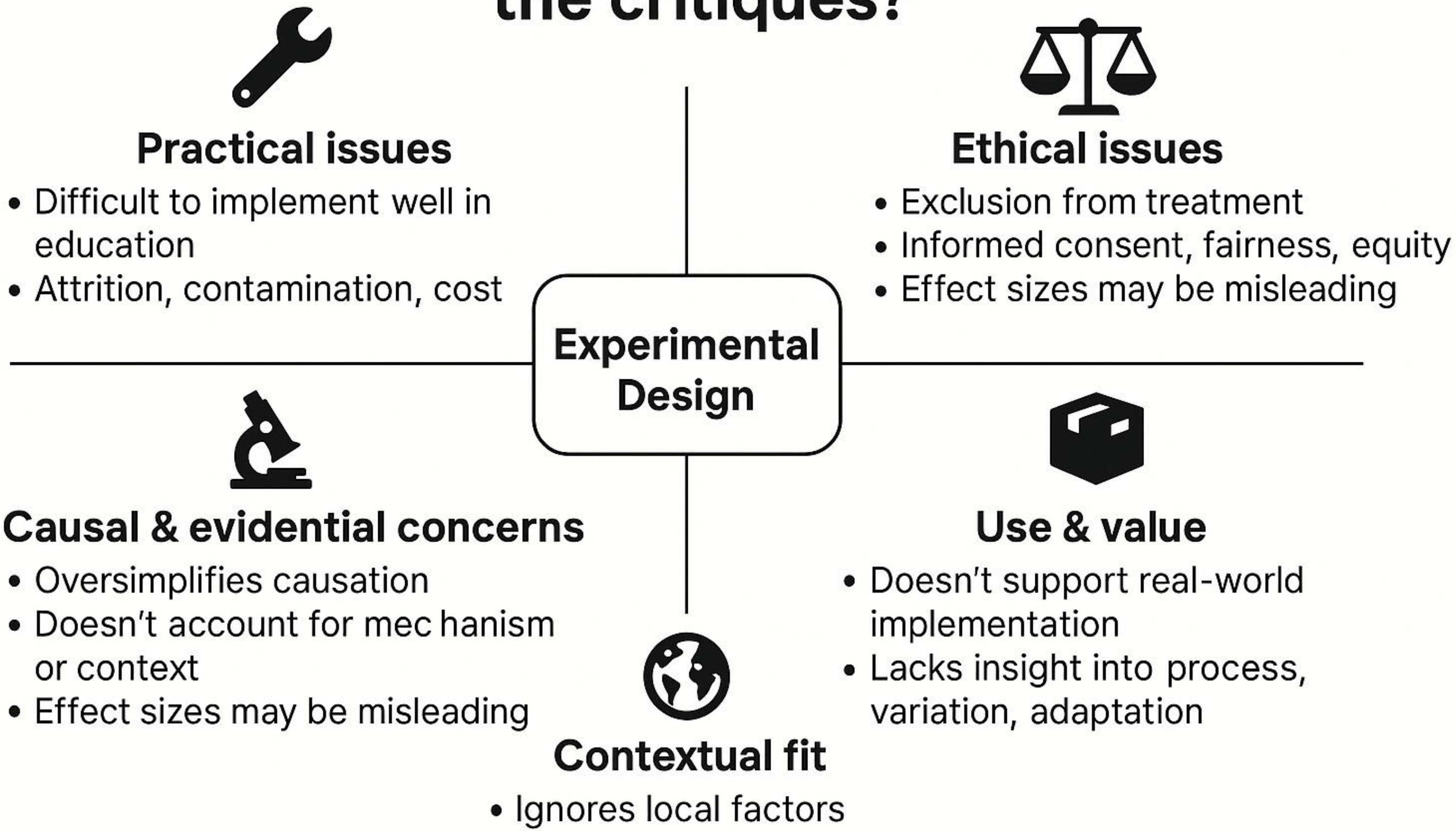
Gary Thomas

First published: 03 August 2020 | https://doi.org/10.1002/berj.3660 | Citations: 24

https://bera-journals.onlinelibrary.wiley.com/doi/10.1002/berj.3660

> Daddy Pig: *We’ll start by doing an experiment*.
>
> Peppa: *What’s an experiment?*
>
> Daddy Pig: *It’s a way to find out something we don’t know—like how many children does it take to lift Mme Gazelle*.
>
> Children [all at once]: *One, a hundred, six …*
>
> Daddy Pig: *You’re all guessing*.
>
> Danny Dog: *What’s the answer?*
>
> Daddy Pig: *I don’t know … but we can use an experiment to find out. Who wants to try to lift Mme Gazelle?*
>
> Peppa: *Me!* [Tries to lift Mme Gazelle] *I can’t lift her*.
>
> Daddy Pig: *Let’s try two children*. [Two try but they can’t lift Mme Gazelle] *Let’s try three children*. [Mme Gazelle rises]

Experimental design: What are the critiques?



Experimental Design

Practical issues

- Difficult to implement well in education
- Attrition, contamination, cost

Ethical issues

- Exclusion from treatment
- Informed consent, fairness, equity
- Effect sizes may be misleading

Causal & evidential concerns

- Oversimplifies causation
- Doesn’t account for mec hanism or context
- Effect sizes may be misleading

Use & value

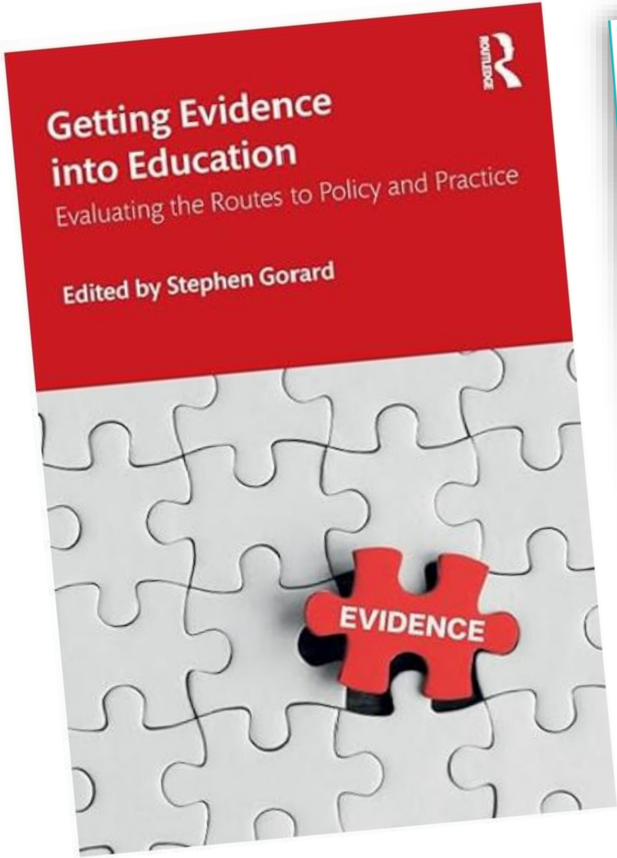
- Doesn’t support real-world implementation
- Lacks insight into process, variation, adaptation

Contextual fit

- Ignores local factors

Three (related) assumptions:

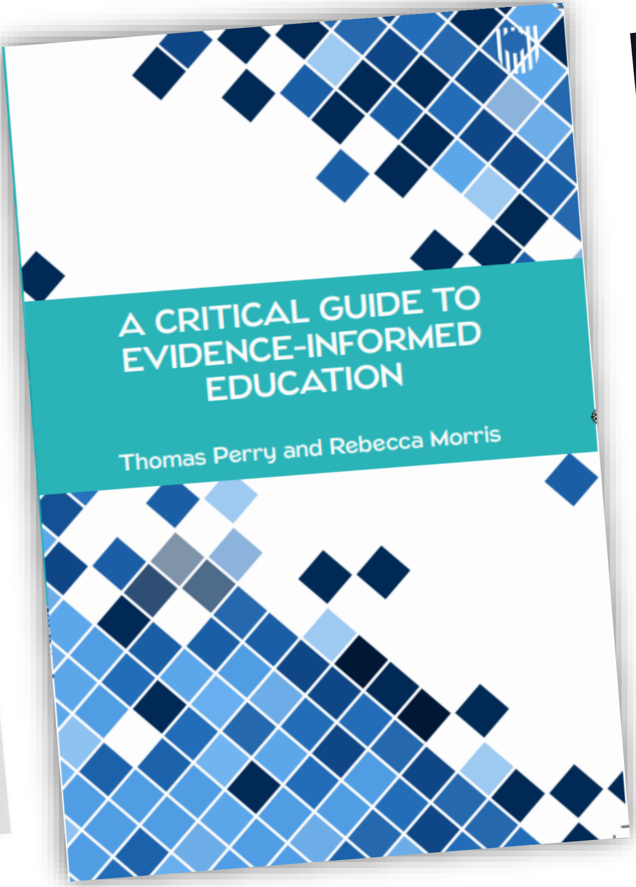
- **Causation** – assumes change is measurable, linear, singular, easily isolated and context independent
- **Evidence** – assumes that the evidence quality can be independent of interpretation and use, and must centre on internal validity (/rigour and robustness)
- **Evidence use** – assumes that adoption and implementation of evidence-based interventions is an effective form of educational improvement



Getting Evidence into Education

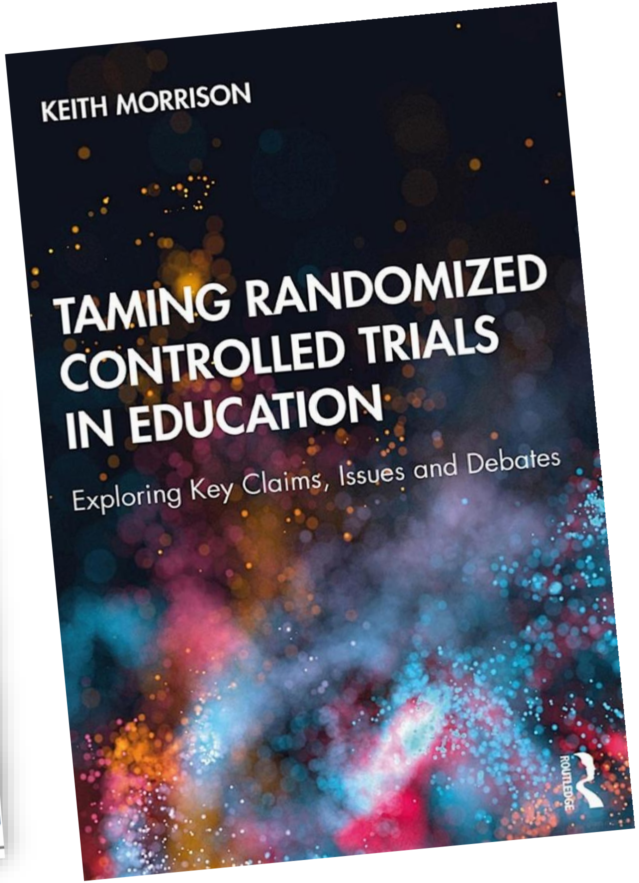
Evaluating the Routes to Policy and Practice

Edited by Stephen Gorard



A CRITICAL GUIDE TO EVIDENCE-INFORMED EDUCATION

Thomas Perry and Rebecca Morris



KEITH MORRISON

TAMING RANDOMIZED CONTROLLED TRIALS IN EDUCATION

Exploring Key Claims, Issues and Debates

Causation

WARWICK

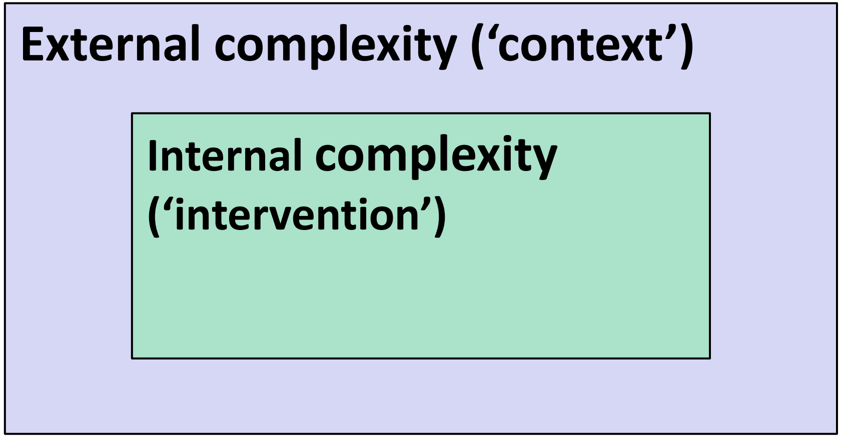
Causation is rarely linear or singular (‘simple’)

Initial Condition	Books B	Motivation M	Time T	Intended outcome after B?	Did B improve outcome?
1	✓	X	X	No	No
2	X	✓	X	No	No
3	X	X	✓	No	No
4	✓	✓	X	No	No
5	✓	X	✓	No	No
6	X	✓	✓	Yes	Yes
7	✓	✓	✓	Yes	No
8	X	X	X	No	No

Adapted from Befani (2012, p. 12)

Context (the intervention isn’t the only active ingredient)

- Social interventions are ‘complex systems thrust amid complex systems’ (Pawson, 2006, p. 168)



External complexity (‘context’)

Internal complexity (‘intervention’)

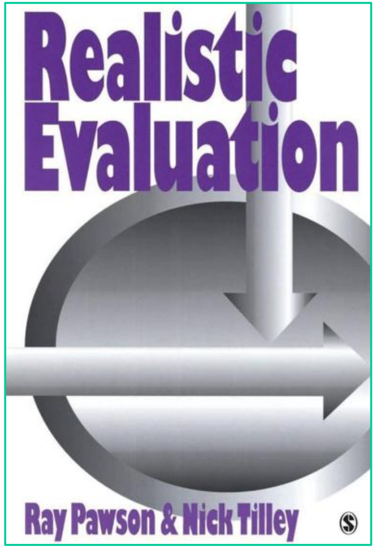
<u>Uninformative Trials</u>

Mean effect size of large-scale educational RCTs
= 0.06sd (CI = 0.3sd)

Lortie-Forgues and Inglis (2019)

Realistic Evaluation

- From interventions to mechanisms: The CMO approach (Realistic Evaluation, Pawson and Tilley)
- Not just ‘Does it work?’ but: ‘What works, for whom, in what circumstances, and why?’
- **The CMO Model:**
  - **C**ontext: the social, organisational, and cultural setting
  - **M**echanism: the process or reasoning that is triggered
  - **O**utcome: the result, produced *if* the mechanism fires *in* that context



Causation

Nancy Cartwright

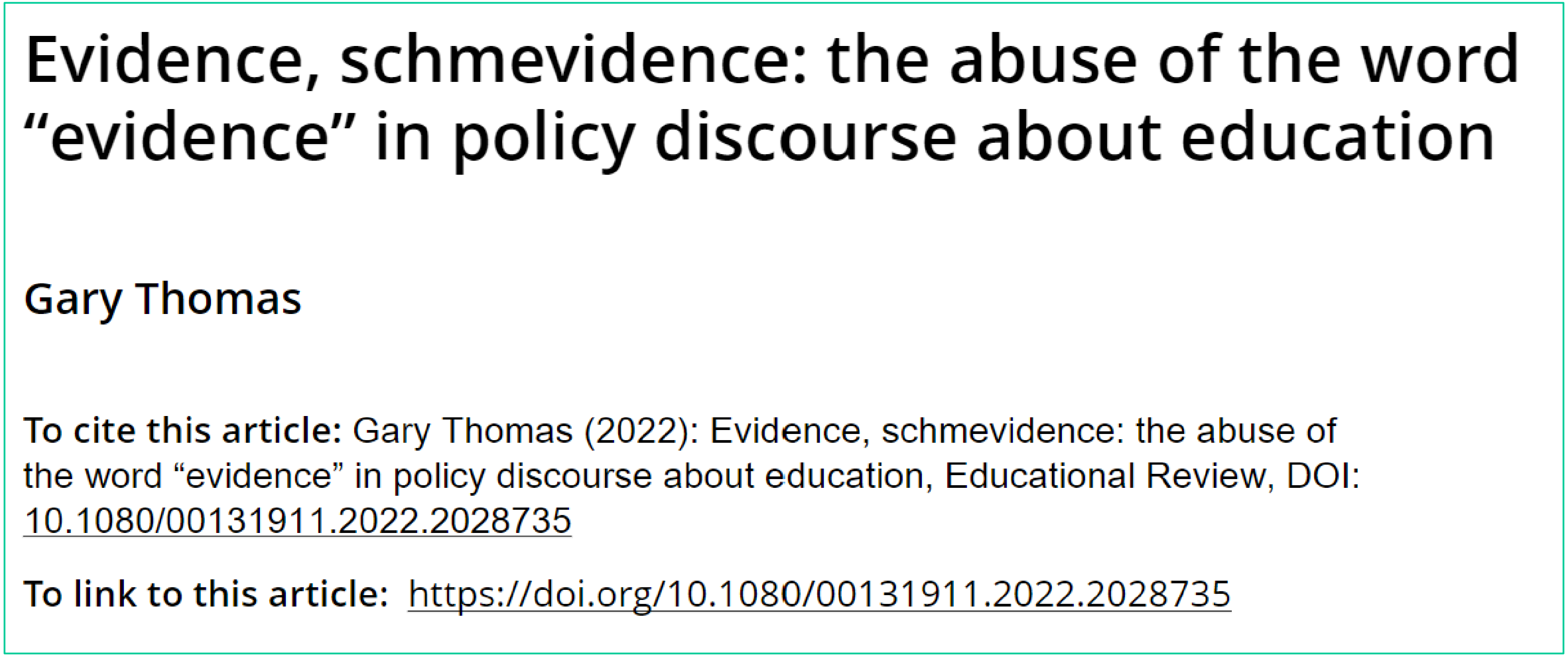
- Causal relationships are **fragile** – they depend on **supporting factors** (“*causal capacities*” or “*causal support factors*”)
  - RCTs control away context, but in doing so, they hide the very **conditions** that made the intervention work.
  - As a result, their findings are often **not portable** to other settings without further theory and contextual analysis.
- Cartwright warns against **methodological fundamentalism** – the idea that RCTs are inherently superior. She argues:
  - The **relevance** and **reliability** of evidence depend on whether the method fits the question, not on a hierarchy.
  - “No method is a gold standard in general. Fit-for-purpose is what matters.”

Theory → Questions → Designs → Evidence

If your theory says/question assumes...	**...then you might need:**
Outcomes depend on combinations of factors (configurations)	QCA (Qualitative Comparative Analysis)
Mechanisms matter, and are triggered by context	Realist Evaluation (CMO: Context, Mechanism, Outcome)
Effects vary by starting point or threshold	Nonlinear/moderating/interaction designs
Impact depends on how things are implemented	Process evaluation, case studies
Change is driven by long-term improvement	Theory-building studies, multiple linked evaluations
There’s one clear cause-effect relationship to isolate	RCTs, if feasible and ethical

Evidence

WARWICK



Evidence, schmevidence: the abuse of the word "evidence" in policy discourse about education

Gary Thomas

To cite this article: Gary Thomas (2022): Evidence, schmevidence: the abuse of the word "evidence" in policy discourse about education, Educational Review, DOI: 10.1080/00131911.2022.2028735

To link to this article: https://doi.org/10.1080/00131911.2022.2028735

Useful qualities of evidence

- **Rigour** – credible, well-designed, and methodologically sound
- **Relevance** – fits the issue or decision at hand (including context sensitivity e.g., the needs and capacities of the setting)
- **Usability** – accessible, understandable, timely
- **Coherence** – fits with other evidence and knowledge
- **Cogency** – logical and persuasive

What is evidence for?

What qualities should we look for in it?

Research produces a narrow (but potentially powerful) form of knowledge. Others are needed for impact (1)

Research produces a narrow (but potentially powerful) form of knowledge. Others are needed for impact (2)

Social knowledge ‘distillation’

Social reality

Practical knowledge (inc. tacit, experiential)

Codified knowledge

Robust knowledge

Intransitive generative mechanisms in the social world (Bhaskar, 2013)

Basic science is poorly applied. Applied science is under-valued.

Basic/pure research	Applied research
Focused	Holistic
Controlled	Interacting
Static	Dynamic
Fundamental	Realistic
General(isable)	Specific/contextualised
Prestigious (well-funded)	Perceived inferior (poorly funded)

Using Evidence

Different users, different needs

User	Need	Evidence quality depends on…
University leader	Strategic decisions	Timeliness, relevance, clarity, credibility
Teacher trainee	Learning and classroom practice	Usability, clarity, strong pedagogic framing
Policy advisor	Funding or scaling decisions	Causal clarity, cost-effectiveness, generalisability
Evaluator	Judging effectiveness	Detail, theory, robustness, mixed methods

The ‘Pipeline’

Find what works → Package it → Implement it → Improvement happens

- Interventions are discrete, transferable, and scalable
- Improvement comes from adopting proven solutions
- Users role is to implement
- Encourages “what works” toolkits, but not necessarily better practice

The ‘Ecosystem’

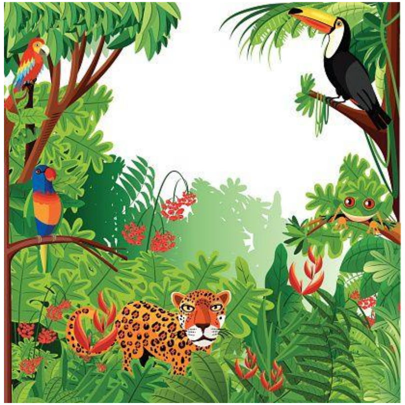
Evidence use is interactive, situated, and developmental

- Use is shaped by beliefs, relationships, routines, and capacities
- Research is just one form of knowledge among many
- Improvement requires dialogue, experimentation, and collective sense-making
- Building research literacy and professional capacity
- Embedding evidence in real decision-making and development processes
- Valuing different forms and sources of evidence

- Users
- Roles
- Practices
- Processes
- Contexts
- Capacities
- Leadership
- Decisions
- Purposes
- Questions
- Needs
- Values
- Interpretation
- Use
- Design
- Culture
- Collaboration
- Infrastructure
- Trust
- Professional judgement
- Translation
- Implementation
- Feedback loops
- Improvement cycles
- Local knowledge
- Research literacy
- Knowledge mobilisation

What does a research-informed education system look like?

- Are we nearly there yet? Strengths? Weaknesses?



What does a healthy evidence ecosystem look like?

- **Multiple roles:** researchers, teachers, leaders, analysts, facilitators
- **Diverse purposes:** enquiry, improvement, accountability, design, strategy
- **Embedded processes:** evaluation, inquiry, professional dialogue
- **Situated use:** shaped by context, values, capacity and culture
- **Infrastructure and resources:** time, tools, trust, training
- **Dynamic:** feedback loops, iterative learning, adaptation
- **Beyond adoption:** evidence as a resource, not a script

Recap

Assumption	Challenge	Implication
Causation	Change is not always linear, singular or ‘simple’	Designs must match theory and context
Evidence quality	Value depends on purpose, interpretation use	Evidence must be fit for purpose and user
Evidence use	Adoption/implementation isn’t usually improvement	Strengthen ecosystems, not just interventions

Validated

TAPE

Toolkit for Access & Participation Evaluation

Dr Matt Horton - University of Wolverhampton
Dr Thomas Perry - University of Warwick
Dr Tracy Whatmore - University of Birmingham

Published in the
Educational Review

Background & Context to the Research

PhD UoB: evaluated impact of **Aimhigher WM Access Programme**. Collab. 5 WM HEIs delivering outreach activities **(Summer Schools & Mentoring)** to schools and FE colleges.

Wider aims of evaluation

1. **QED evaluation** of the impact of interventions on pupils **attain., non-cognitive factors and HE entry**. Compared outcomes between matched T and NT groups.
2. **Do AABs mediate pupils HE entry outcomes?** Key area to investigate as HEIs spend millions £ per year on improving AABs. (E.g., HE expectations/attitudes). No robust evidence to show if high AABs are associated to increased likelihood of entering HE.
3. **Gap lack of consistency & validation in surveys to measure pre–post shifts in AABs** Lack of validation is hampering efforts to improve student outcomes > measuring 'what might work' rather than 'what works'!

Research conducted from 2012-2022. **4440 Pupils were longitudinally tracked** from year 9 to HE entry.

TAPE

Toolkit for Access & Participation Evaluation

Validation

TAPE Method & Validation Procedure

Toolkit validated: 100+ West Mids. secondary schools from 2012-22. Yr Grps. 11-13. Survey measures five dimensions: HE knowledge, expectations, attitudes, academic motivation and self-efficacy.

Repeated measures: Pupils completed the same standardised survey at two separate points – 1 year apart

Over 1000 pupils sampled.

Battery of validation tests:

- **Content** (WP practitioners, review of the lit. AABs and attain. + HE entry)
- **Face** (feedback from pupils and tested for readability age – SMOG)
- **Predictive** (can baseline scores predict HE outcomes for NT)
- **Test-retest reliability** (survey completed by same pupils e.g., Jan 17 + Jan 18

TAPE: Validity Testing

Predictive validity

Students who scored higher on each construct were between

44% to 81%

more likely to enter HE than, students with lower scores.

Test-retest reliability

All items were found to be reliable. Scores on re-testing remained largely consistent for pupils who did not engage in WP interventions (non-treatment group).

TAPE

Self-efficacy and academic motivation items had smaller sample frames and were not included in the validation testing beyond face and content validity.

How do TAPE & the ASQ compare?

Validation

What constructs do TAPE & the ASQ measure?

Constructs		TAPE	ASQ
🎓	HE knowledge	✓	✓
🎯	HE expectations	✓	✓
♥	HE attitudes	✓	✓
🧗	Academic motivation	✓	✗
🦸	Academic self-efficacy	✓	✓
⚙	Cognitive strategies	✗	✓

Validation by age (years)	
TAPE	9-13
ASQ	7-13

Validity Testing

	TAPE	ASQ
Face validity	✓	✓
Content validity	✓	✓
Reliability	✓	✗
Predictive validity	✓	✗
Construct validity	✗	✓

Only TAPE has test re-test reliability & predictive validity for most items.

Limitations

Both toolkits validated at aggregate level (e.g., across all HE knowledge items).

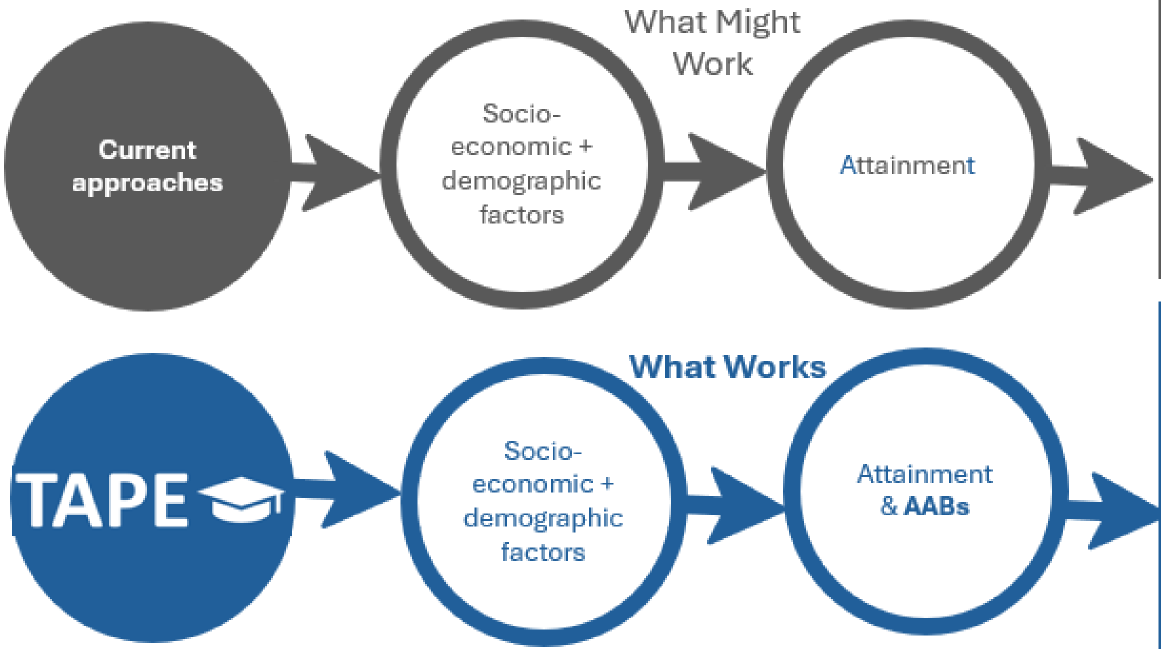
Is TASO's ASQ TOAST ?

- content, face & construct validity
- ASQ could not predict prior attain.
- more testing needed to combine scales (ASQ & TAPE) & support sector consistency

How can TAPE Improve the Impact of WP Outreach?

TAPE can improve the impact of WP outreach activities through employing the toolkit to improve targeting and a more needs led and preventative approach to supporting disadvantaged pupils.

WP Targeting (schools + pupils)



What Might Work

Current approaches → Socio-economic + demographic factors → Attainment →

What Works

TAPE → Socio-economic + demographic factors → Attainment & **AABs** →

Limitations

- Lacks context – what are schools / pupils needs in terms of AABs
- Deadweight/usual suspects those who engage may be those already on HE trajectory (= wastage of resource).

Strengths

- Measuring AABs at baseline provides a better understanding of who & what the content of activities should focus on.
- For example, more intensive interventions could focus on those who are not sure/definitely not considering HE participation (reduces deadweight) rather than low hanging fruit (e.g., those going to HE).

TAPE Summary

1. Lack of validation & consistency hampering the sectors progress in understanding to ‘what might work’ rather than ‘what works’
2. TAPE = most robust toolkit to understand ‘what works’. 1st access toolkit be thoroughly validated & peer reviewed.
3. 115+ providers using TAPE to strengthen evidence/APP commitments
4. AABs mediators of HE entry for certain students (disadvantaged but higher levels of attain). Means HEI annual spend (£180 m) on access+ AABs is not mis-directed as outlined in some research.
5. Predictive validity: use beyond measuring impact as TAPE = a preventative tool to identify need earlier, target support, & improve outcomes.

To understand **what works** in improving learner outcomes we need to employ robust & **validated toolkits.**

Resources

Both TAPE & ASQ toolkits can be accessed via the: TASO website & HEAT database.

If you would like to hear more about TAPE: ✉ matthew.horton@wlv.ac.uk

- TAPE LinkedIn results summary
- TAPE Educational Review Journal publication (Horton, Perry & Whatmore, 25)
- Impact of Multi-Intervention Access Programmes (Burgess, Horton & Moores, 21)
- Full thesis: Evaluating Impact of Aimhigher on AABs & HE Entry (Horton, 23)

covers who is underrepresented in HE, importance of AABs, attainment, review of what works in access, QED, TAPE & impact of summer schools and mentoring.

TAPE

Toolkit for Access & Participation Evaluation

Q&A

Statements	Construct	Response format (coding)
To what extent do you agree or disagree with the following statements:		
I am planning/considering going to higher education before I am 30 years old	HE intentions/expectations and academic motivation	Definitely Probably Not Sure Probably Not Definitely Not
I understand what student life would be like in higher education	HE Knowledge	
I know enough about higher education to decide whether to go or not		
I understand how to apply to higher education		
I know the qualifications that I will need to be able to go to higher education		
I know the grades that I will need to be able to go to higher education		
I am clear on which higher education course/subject to apply for		
I am clear on which higher education institutions I want to apply for		
I understand how the UCAS application process works (UCAS is the organisation responsible for managing applications to higher education courses)		
University is for people like me	HE Attitudes	

Block B questions (see page 2) focus on pupils' concerns/barriers. These questions are routed and should not be completed by all pupils. The routing is based on the response to the following question:

I am planning/considering going to higher education before I am 30 years old?

Responses	*Routing*
Not sure, probably not, and definitely not	Complete block 1 questions and then move on to block 2.
Definitely/probably	Complete block 1 questions but not block 2.

Do you have any concerns about going to higher education? If yes, please outline the extent to which you agree or disagree with the following statements:

Statements	Construct	Response format (coding)
I can't afford to continue into higher education because I am worried about getting into debt	HE Attitudes	Strongly Agree Agree Not Sure Disagree Strongly Disagree
It is not worthwhile continuing with education	HE Attitudes and academic motivation	
I'm not interested in education		
I will not get the required grades to go into higher education	HE intentions/expectations /confidence in academic ability	
I do not feel confident in my ability to cope with learning in higher education	HE Attitude/confidence in academic ability	
other reason (please specify)		