Experimental education research: rethinking why, how and when to use random assignment

Sam Sims[1], Jake Anders[1], Matthew Inglis[2], Hugues Lortie-Forgues[2*], Ben Styles[3], Ben Weidmann[1]

[1] UCL
[2] Loughborough University
[3] National Foundation for Educational Research
* Corresponding author

Abstract

Over the last twenty years, education researchers have increasingly conducted randomized experiments. Such experiments are usually designed in the hope of supporting comparisons of cost effectiveness across interventions, for example by using broad, standardised outcome measures. However, small effect sizes, wide confidence intervals, and heterogeneous treatment effects have limited their practical usefulness. We argue that, rather than directly informing decisions, randomized experiments should primarily play the role of testing theoretical models, which can in turn inform educators' mental models. Since comparable effect sizes are not essential when testing theory, researchers can improve the power of such studies in various ways, including through the use of outcome measures more closely aligned with the theoretical constructs of interest. We also suggest that quasi-experimental methods may often be more suitable for informing education decisions, on the grounds that they can achieve greater precision at lower cost, with relatively moderate risks to internal validity.

Keywords: randomized experiments, decision-making, theory testing, educational research

Sam Sims, UCL, Gower St, London WC1E 6BT s.sims@ucl.ac.uk
Jake Anders, UCL, Gower St, London WC1E 6BT jake.anders@ucl.ac.uk
Matthew Inglis, Loughborough University, Epinal Way, Loughborough, LE11 3TU m.j.inglis@lboro.ac.uk
Hugues Lortie-Forgues, Loughborough University, Epinal Way, Loughborough, LE11 3TU h.lortie-forgues@lboro.ac.uk
Ben Styles, NfER, The Mere, Upton Park, Slough SL1 2DQ b.styles@nfer.ac.uk
Ben Weidmann, UCL, Gower St, London WC1E 6BT b.weidmann@ucl.ac.uk

Acknowledgements and funding
Ben Styles' contribution was funded by the National Foundation for Educational Research. Anders' and Sims' time was supported by the UCL Centre for Education Policy and Equalising Opportunities. This work was partially supported by Research England, via an Expanding Excellence in England grant to the Centre for Mathematical Cognition, and the Economic and Social Research Council [grant number ES/W002914/1]. This work was partially supported by UKRI Economic and Social Research Council [grant number ES/W002914/1]. The authors would like to acknowledge the contributions of their colleagues who participated in a seminar series on RCTs in Education, made possible by the funding of Research England. Their collaborative efforts were instrumental in shaping this work.

Randomized experiments (henceforth, experiments) allow researchers to identify the effects of causes, with few additional assumptions. This makes them useful for education researchers, who have been running experiments since at least 1918 (Hedges & Schauer, 2018). Indeed, there has been a marked growth in the use of experiments in education in the last two decades. In the US, the Institute for Education Sciences (IES) has supported more than 350 experiments since 2002 (Hedges & Schauer, 2018). Likewise, the Education Endowment Foundation (EEF) in England has commissioned more than 150 experiments since 2010.

As education researchers have conducted more experiments, so we have learned more about the challenges involved. We now know more about artefactual (unrelated to the intervention) variation in effect sizes (Cheung & Slavin, 2016; Ost et al., 2017; Wolf & Harkbatkin, 2023). Typical effect sizes in educational experiments have turned out to be much smaller than was suggested by prior empirical research and it has proven difficult to increase the precision of experiments in response (Spybrook et al., 2016). Consequently, many experiments are uninformative for decision making, in the sense that findings are close to equally consistent with the intervention being either effective or ineffective (Lortie-Forgues & Inglis, 2019).

This paper takes stock and reconsiders the use of experiments in education. In the first section, we begin by developing a simple framework distinguishing two idealized goals of experimental research (theory informing and decision informing) and the research decision choices aligned with each of them. In the second section, we use this framework to synthesize what we have learned about experiments, arguing that researchers have struggled to implement decision-informing experiments. Building on this, the third section concludes with recommendations for why, how and when education researchers should conduct experiments.

Clarifying the goals of experiments

There are at least two broad goals that an education researcher might have in mind when considering conducting an experiment. We call these the *informing theory* goal and the *directly informing decisions* goal (cf. Roth & Kagel, 1995).

Experiments informing theory

A researcher who aims to inform theory will typically, though often implicitly, adopt a philosophy of science based on Popper's (1958, 1962) version of methodological falsification. Because no empirical study can ever prove a universal claim to be true, Popper argued that science progresses by researchers putting forward predictions derived from explanatory theories and then trying to refute them. If the newly observed data refute the prediction, the researchers' theoretical understanding must be modified appropriately; if it does not, then the prediction survives, at least until new data emerge to falsify it. Across many studies, only true theories will be left unrefuted.

In the natural sciences, mathematical theories will often provide a quantitative prediction for empirical phenomena. For example, Columb's law states that the *magnitude* of the electrostatic force of attraction/repulsion between two point charges is directly *proportional* to the product of the *magnitudes* of charges. This prediction can be falsified by finding either a lack of a relationship in the theorized direction or a relationship that is not of the theorized magnitude. By contrast, in the social sciences, including education, we generally rely on verbal theories that make only ordinal predictions (Broers, 2021; Meehl, 1978). For example, when faced with challenging learning material, pupils with *more* of a growth mindset will learn *more*. This theory makes no predictions about the magnitude of this relationship and can therefore only be falsified by failing to find a relationship in the theorized *direction*. This reflects the aim of social science theories to be broadly applicable:

they are not tied to specific contexts, and because effect sizes vary across settings and measures, precise quantitative predictions are generally not possible.

The desire to inform theory, coupled with an endorsement of Popper's methodological falsification, suggests certain aligned research practices. To test the ordinal predictions made by theory, education researchers often conduct experiments in which participants are randomly assigned to receive different versions of an intervention, for example one that includes a growth mindset message and one that does not. Researchers will recruit a sample of participants to whom the theory is thought to apply. For example, growth mindset theory is posited to apply those engaged in difficult learning tasks. The sample need not be representative of a larger population, provided it includes individuals for whom the theoretical prediction is expected to hold (Calder et al., 1982; Stroebe et al., 2018). Researchers will then aim to conduct a tightly controlled experiment, perhaps in a lab setting, to provide a valid test of the theory. Randomization removes systematic allocation bias, so that differences in outcomes can be attributed to the intervention rather than to pre-existing differences between groups. Randomization inference can be used to account for the fact that repeating the experiment with the same set of units, but a different set of random treatment assignments for each unit, would have given different results (Rosenbaum, 2017)[1].

If the confidence interval for a well-powered experiment excludes an effect in the *direction* predicted by theory, then the theory is said to be falsified. For example, if the pupils with more of a growth mindset learn less, then this is said to falsify growth mindset theory[2].

---

[1] Unlike traditional statistical inference, randomization inference does not require assuming that the experimental units are sampled from a larger population. Inference is based on the distribution of possible outcomes that could have arisen under different treatment assignments within the same set of units (Rosenbaum, 2017).

[2] In practice, scientists rarely follow strict Popperian falsification. Rather, as Lakatos (1976) argued, theories are part of broader research programmes which are able to accommodate anomalies through their 'protective belts'. A failed prediction does not automatically falsify a programme: its effect depends on factors such as the severity of the test it provides (in Mayo's (2008) sense). In particular, minor anomalies (e.g., non-significant results in the theorised direction) might be ignored, while serious anomalies (e.g., significant effects in the opposite direction) might require modifications to the programme's protective belt. Only after many such ad hoc modifications to accommodate anomalies is a programme likely to be abandoned.

Other quantities, such as the point estimate, are of interest only in the instrumental sense that they determine the center of the confidence interval. The theory makes no prediction about them and they cannot, therefore, falsify the theory. If the theory is not falsified after a number of severe tests, then perhaps it should inform the mental models and decisions of educators. For example, growth mindset might inform the way that educators communicate with their pupils.

Experiments directly informing decisions

A researcher who aims to directly inform decisions will typically, though often implicitly, be working within a framework concerned with comparing the consequences of alternative actions, an approach closely related to welfare economics (Ng, 2003).This approach to social science progresses by estimating the quantitative inputs necessary to inform a cost-benefit analysis (Cullis & Jones, 2009). The costs include the financial, human and other resources necessary to pursue a given course of action and the benefits comprise the socially desirable causal effects of the course of action, ideally converted into a commensurable monetary value. When combined, this information is informative about the net benefit of a given action. The end goal here, across many studies, is to rank the net benefit of a set of options and thereby support decision makers to pick the one that maximizes social value.

The desire for researchers to inform decisions suggests certain aligned research practices. Costs can be quantified through careful accounting exercises (Belfield & Bowden, 2019). To quantify the benefits, researchers often codify a course of action in a manualized intervention and conduct experimental research in which participants are randomly assigned to either receive the intervention or to continue with 'business as usual'. The researcher would ideally recruit a representative sample of participants from the population of interest, so that their study is well aligned with the choice faced by the ultimate decision maker. For

the same reasons, the intervention would also be delivered by those who would deliver it in practice, e.g., teachers. The researcher is interested in both randomization inference, for internal validity, and sample-population inference, for external validity (Abadie et al., 2020).

So long as the researcher measures socially valuable outcomes (e.g., exam grades at the end of school), comparing these outcomes between these two groups provides an estimate of the benefit. Comparing these benefits to the costs provides an estimate of the net benefit. Results from multiple studies can then be combined in 'toolkits' or 'warehouses' ranking the net benefits of different courses of action. Decision makers can then directly draw on these rankings to inform their choices.

Summary

Table 1 summarizes the goals (the *why*) and methods (the *how*) of these two types of experiments. Those informing theory aim to test predictions deduced from theory. The primary quantity of interest for this is the direction of the treatment effect within the study sample (and the associated *p*-value) because this provides information about whether the ordinal predictions made by the theory are observed in the data or not. Methodologically, the relevant dependent variables are the constructs specified in the theory and measurement instruments are chosen based on the validity with which they capture these constructs. By contrast, experiments informing decisions directly evaluate interventions, and aim to estimate comparable effect sizes as well as confidence intervals (or *p*-values) for the target population. Methodologically, the intervention should be implemented by the intended user (educators),

and the dependent variables should be captured using comparable standardized tests and be broad enough to matter for pupils' life chances so as to capture the social benefits.[3]

The two positions can be further clarified by contrasting them directly. Experiments informing theory do not aim to test interventions per se because any given intervention represents only one of many possible instantiations of the theory. Conversely, experiments informing decisions do not aim to test hypotheses deduced from theory in that they aim to evaluate different courses of action, regardless of the extent to which these reflect theory. Experiments informing theory are typically not interested in estimating the (relative) magnitudes of effects, since educational theories typically make ordinal rather than quantitative predictions. Experiments informing theory do not need their findings to generalize beyond the sample, in the sense that there is no intended end user (e.g., a subsequent national roll out of the programme). For the same reason, theory informing experiments do not require the interventions to be implemented by end users. We acknowledge that both types of experiments would aim to use valid outcome measures, but the meaning of validity is different. In theory informing experiments validity reflects the correspondence between a construct within the theory and the corresponding measure. By contrast, in decision informing experiments there is no need for a theory specifying the relevant constructs. An effect on any one consequential outcome measure could be replaced by an effect on another consequential outcome measure without harming the success of a decision informing experiment.

[3] Readers familiar with 'What Works' methodology may argue that decision-making should be guided by meta-analyses and systematic reviews rather than by an individual experiment. We agree that individual experiments are unreliable guides for decision-making. However, experiments are often implicitly designed with decision-making in mind. For example, experiments funded by the EEF and the National Center for Education Evaluation (NCEE) include power analyses, implying that the goal of the experiment is to provide information independently of the wider literature. If contributing to a meta-analytic effect size was the goal of running an experiment, sample size should be based on the existing level of uncertainty in the meta-analytic effect size.

Table 1. Comparison of theory-informing and decision-informing experiments

	Experiments informing theory	**Experiments informing decisions**
Goals (why):		
Testing:	Deduced hypotheses	Interventions
Quantities of interest:	Direction of Sample Average Treatment Effect & *p*-value	Magnitude of Population Average Treatment Effect & *p*-value
Methods (how):		
Sampling:	Purposeful	Representative
Dependent Var.:	Relevant construct	Consequential outcome
Instruments:	Maximally valid	Broad standardized test
Implementer:	Researcher or educator	Educator

Experiments in practice

We now turn to evaluating how experiments are actually conducted in education research. Our argument will be that education researchers have tried but largely failed to conduct experiments capable of directly informing decisions.

Experiments informing theory

There are currently many experiments conducted in education that attempt to inform theory. For example, consider Hodds et al.'s (2014) investigation of the efficacy of self-explanation prompts as a technique for improving mathematics undergraduate students' comprehension of mathematical proofs. This study was motivated by literature showing that self-explanation could improve students' comprehension in scientific domains (Ainsworth & Burcham, 2007). They conducted a lab study where a convenience sample of undergraduate students were either given some self-explanation materials or a control activity and then

asked to individually read a mathematical proof and complete a researcher-designed comprehension test. They found that students in the experimental group scored higher than students in the control group ($d = 0.95$, $p < .001$).

Hodds et al.'s (2014) goals and methods are aligned with the informing theory column of Table 1 and thus allow them to draw theoretical conclusions: they subjected the claim that generating self-explanations improves understanding when reading mathematical proofs for undergraduates to a risky test. However, as a result of adopting these design features, they cannot conclude that self-explanation training is a better way of helping students understand mathematical proofs than any other intervention, or that the results merit the (opportunity) costs of the intervention. Their research informs theory but does not directly inform decision making.

Experiments directly informing decisions

The ways in which the EEF or the National Center for Education Evaluation (NCEE) present and summarize findings suggest that their experiments are aimed at informing decisions.[4] However, it is hard to point to any examples of experimental research from these funders that really fit into the informing decision column of Table 1. There are two places in which experiments tend to diverge from this ideal.

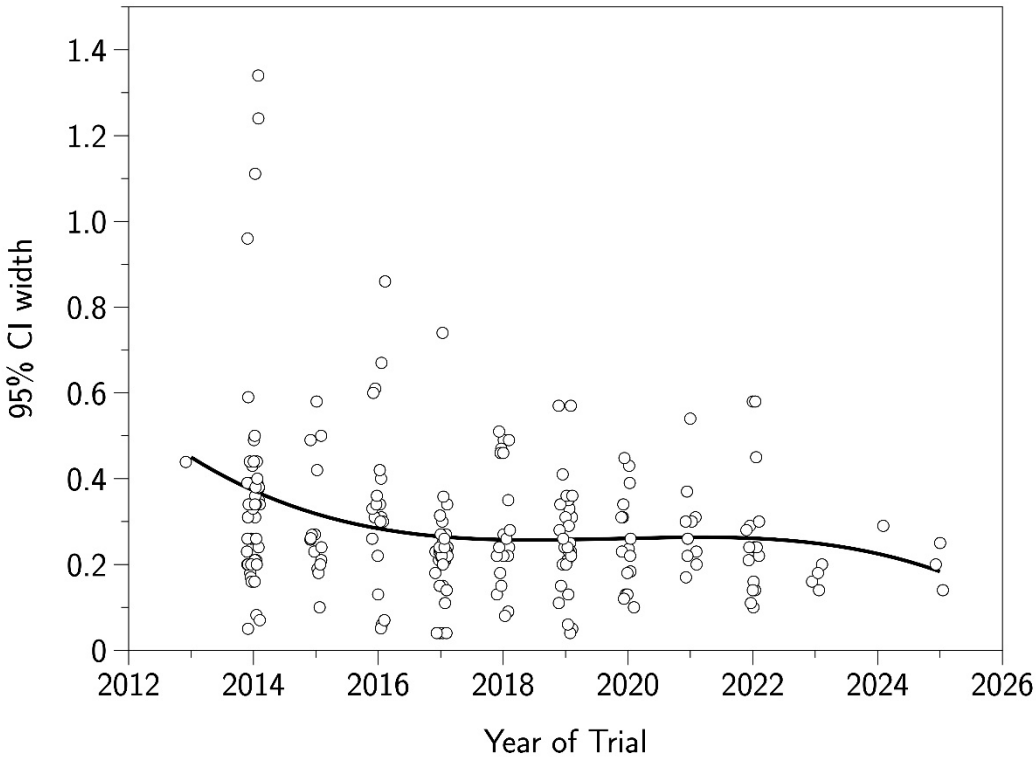
First, such experiments rarely employ representative samples. This is presumably because representativeness can so easily be undermined by sampled units not consenting to take part in the experiment. Consistent with this, the few examples of educational experiments using approximately representative samples – evaluations of the Upward Bound high school enrichment program (Myers & Schirm, 1999), Head Start (Puma et al., 2010),

[4] The National Center for Education Evaluation and Regional Assistance (NCEE) is part of the Institute of Education Science (IES) and is responsible for conducting large-scale field evaluations of education programs in the United States, including randomized experiments. In the remainder of the paper we focus on the NCEE and the EEF (which performs a similar role in education research grant-making in the UK).

and Tennessee statewide Pre-K (Durkin et al., 2022) – are all government-funded programs, in which participation could be legally mandated. Even in these three cases, it was only possible to use a representative sample of *oversubscribed* sites. Recent empirical estimates suggest that treatment effect heterogeneity across sites (e.g., schools) in educational interventions can be large, which means that effect sizes in such convenience sample may be a poor proxy for effect sizes in the population of interest (Olsen et al., 2023; Weiss et al., 2017).

Second, such experiments fall short of the goals in the informing decisions column in that they provide very little information about effect sizes. Lortie-Forgues and Inglis (2019), for example, showed that the median width of the confidence intervals in the experiments conducted by the EEF and the NCEE was 0.24 SDs. Figure 1 updates this analysis to include EEF trials published up to and including 2025 and shows that there has been no appreciable improvement in the precision of EEF trials over this period. This lack of precision is particularly problematic given that experiments informing decisions typically use broad standardized test score outcome measures. Typical effect sizes are noticeably smaller on broad tests (median 0.10 SD) than on narrow tests (0.17 SD), presumably because the latter are better aligned with the intervention (Wolf & Harbatkin, 2023). In sum, the typical 95% confidence interval in education experiments is 2.4 times larger than the typical effect size, implying that we are learning very little about the relative impact of different interventions.

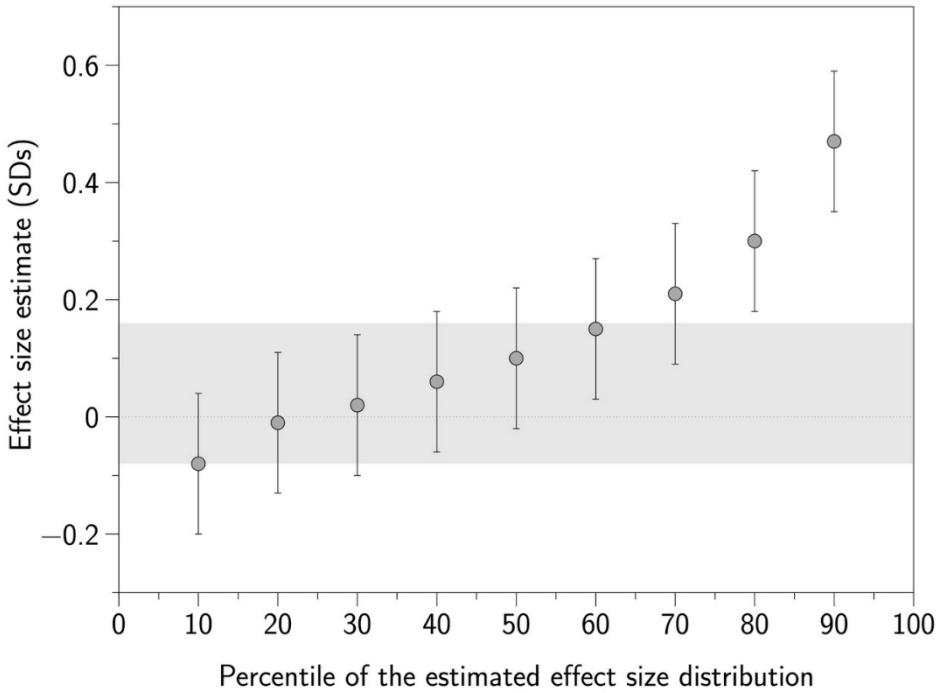
Figure 1. Change in 95% confidence intervals over time



Note. Width of the 95% confidence interval (in SDs) associated with the effect sizes (n = 207) of 119 distinct trials commissioned by the Education Endowment Foundation, by year of publication. Cubic line of best fit.

To illustrate this, Figure 2 plots the deciles of the distribution of effect sizes of experiments using broad outcome measures, as reported by Kraft (2020). For each decile, we have also superimposed the mean confidence interval from NCEE and EEF trials, as reported by Lortie-Forgues and Inglis (2019). Combining these two sources of information gives a sense of how likely it is that the estimated effect sizes from such experiments can be confidently ranked in terms of their magnitude. The median effect size estimate cannot be distinguished from effect size estimates with magnitudes ranging from the 10th to the 70th percentile.

Figure 2. Are effects size estimates from experiments using broad outcome measures statistically distinguishable?



Note. Shows the distribution (deciles) of the estimated effect sizes from experimental education research using broad outcome measures, as reported in Kraft (2020). For each of the deciles, we have superimposed the mean 95% confidence interval from EEF and NCEE trials, as reported in Lortie-Forgues and Inglis (2019). The grey region shows the 95% confidence interval around the precision-weighted mean.

Even taking the minimal aim of determining whether the confidence interval of a study excludes zero, the combination of small average effect sizes and large confidence intervals renders many (but not all) educational experiments statistically uninformative (Lortie-Forgues & Inglis, 2019). The precision-weighted mean effect size from experimental research on broad outcome measures is 0.04 (Kraft, 2020; Lortie-Forgues & Inglis, 2019). An effect size would need to be three times larger than this to be confidently distinguished from zero with the average 95% confidence interval of width of 0.24. For example, the Embedding Formative Assessment trial estimated an effect size of 0.1, which is twice the weighted average effect size estimate (Anders et al., 2022). However, even with this relatively large effect size estimate, the confidence interval still does not quite exclude zero (95% CI: -0.01, 0.21).

Common compromises

Due to the difficulties in conducting decision-informing experiments, researchers have made a number of methodological compromises. First, almost all experiments are conducted using a convenience sample. Second, instead of focusing on a maximally consequential outcome (such as grades in school-leaving exams), researchers have sometimes opted instead for less consequential outcome measures (such as scores on tests administered only for the specific research project). Third, instead of focusing on broad, standardized tests, researchers sometimes opt instead for tests that are more aligned with the content covered in the intervention, or captured more proximally to the intervention. Fourth, researchers sometimes opt to implement the interventions themselves, sometimes outside of an authentic school setting.

An example making many of these compromises is the study of interleaving in mathematics by Rohrer et al. (2019). They conducted a large-scale experiment where 54 mathematics classes were assigned to blocked (studied together) or interleaved (separated by gaps) mathematics assignments over a four-month period. These assignments came in the form of worksheets developed by the researchers but used by teachers. Participants comprised a convenience sample of grade 6-8 schools close to the researchers' university. A month after the intervention, all participants took an unannounced test, developed by the researchers to cover the exact content covered in the worksheets. Pupils in the interleaved group scored higher (d=0.83, $p<.001$), an effect which Rohrer et al. described as "large".

As is common in decision-informing research, Rohrer et al.'s intervention was implemented by teachers in real classrooms, and the primary finding (highlighted in the abstract) was the magnitude of the effect size, rather than a theoretical conclusion. However, like much theory-informing research, Rohrer et al. used a non-representative (convenience) sample and employed a low-stakes outcome measure that was maximally aligned to the

content covered in the intervention. As a result, the study's reported effect size is difficult to compare with those derived from other decision-informing experimental studies. Had Rohrer et al. evaluated the theory differently by, for example, using a different outcome measure the resulting effect size could have been substantially different.

Rethinking experiments in education

In this section, we draw on the analysis above to make recommendations for why, how and when education researchers should use experiments.

How: using outcomes measures aligned with the intervention will likely generate more useful evidence

Experimental researchers should use narrower, more theoretically aligned outcome measures. This is illustrated nicely by the Rohrer et al. (2020) interleaving experiment summarized in the last section. Many education researchers would criticize the outcome measures used in this study. For example, de Boer et al. (2014, p.538) found that effect sizes in studies using researcher designed tests tended to find larger effects and therefore recommend that researchers *always* include standardized test because "they have the great advantage that they facilitate a more reliable comparison of multiple intervention effects."

However, the severe lack of precision in educational experiments that use broad outcome measures (Figure 1 and 2) means that it is usually not possible to compare the effectiveness of different interventions. Aligning the outcome measures to enhance comparability in experimental findings is therefore of limited value. On the contrary, the fact that the outcome measures used in Rohrer et al. tend to increase effects sizes (de Boer et al., 2014; Cheung & Slavin, 2015; Wolf & Harbatkin, 2023) increases the power of their experiment. It is notable that the Nuffield Early Language Intervention, which is unusual in

using theoretically aligned outcome measures, is also one of just two EEF interventions to have shown consistent impact (Sibieta et al., 2016; Dimova et al., 2020; Smith et al., 2023).

Why: Use theory-informing experiments to cut research waste

Our next recommendation rests on two observations. First, theory-informing experiments tends to be substantially cheaper than decision-informing experiments. For example, the Hodds et al. (2014) experiments on self-explanation had a cost to funder of approximately £27,000, whereas the average EEF trial costs around £500,000 (Lortie-Forgues & Inglis, 2019). Second, theory-informing research tends to be better powered than directly decision-informing experiments[5]. It follows from this that, even where the end goal is informing decisions, there will often be a case for first conducting a theory-informing experiment on an intermediate part of the causal chain (Deaton & Cartwright, 2018). Where the prior theory-informing experiment falsifies the intermediate part of the causal chain, this could prevent 'research waste' on a considerably more expensive decision-informing experiment.

An instructive example here is the Helping Handwriting Shine trial (Stone et al., 2020), which has many features of a decision-informing experiment. It set out to test the effects of an eight-week, school-based intervention, in which a trained educator led pupils through a series of fine motor control practice tasks. The rationale for this was that children who have slow and effortful handwriting use scarce cognitive resources on the act of writing. Improving handwriting automaticity would free up cognitive capacity and thereby improve writing. A pair of experiments in England found null results on a broad writing skills outcome (d = -0.02, p = 0.77; d = 0.12, p = 0.16). Crucially, the trial also found null results

[5] Several factors contribute to this. For example, theory-informing studies are often conducted with more homogeneous samples (reducing extraneous variability), can often be carried out in controlled lab settings (where distractions are minimised), and, as noted above, can use outcome measures that closely align with the theoretical constructs.

on the intermediate measure of writing speed (automaticity). A theory-informing experiment focused on writing speed could have established this faster and at lower cost, thus avoiding waste.

Why: Use theory informing experiments to answer (some) applied research questions

Figure 1 and 2 showed the marked lack of precision in education experiments, even when testing sustained, broad-based interventions. However, there is a whole class of questions that education researchers are interested in that involve less sustained or broad-based change (e.g., self-explanation) and are therefore likely have smaller effect sizes. It is very unlikely that these can be addressed using decision informing experiments. Likewise, quasi-experimental methods are not suited to answering such questions because secondary datasets are unlikely to capture variation in the exposure (e.g., self-explanation) or to capture outcome measures narrow and proximal enough (e.g., comprehension of proofs) to detect any effects.

This leaves theory informing experiments as the only suitable method for causally testing such small grain size changes. An illustrative example here is teacher professional development (PD). There are hundreds of experimental evaluations of months-long teacher professional development programs aimed at bringing about broad changes in pedagogy (Sims et al., 2024). However, teacher educators are often interested in more granular questions, such as whether to add modelling (observable examples of teaching techniques) to an existing PD program. Sims et al. (2024) therefore conducted a classroom simulator experiment to test a number of theoretically informed hypotheses about modelling, finding that it did indeed improve teachers' use of evidence-based practices. This finding is likely to be of interest to teacher educators, who can use it to refine their mental models of how best to design PD.

When: Quasi-experiments will often be more suitable for informing decisions about widely used interventions

Many experiments aim to test the effectiveness of programs that are widely used in schools, e.g., phonics programs. In such cases the primary research goal is to estimate the program's population average treatment effect, ideally using a broad, high-stakes outcome. Pre-registered quasi-experimental studies have several practical advantages in achieving this goal. First, they can provide evidence more rapidly, enabling more timely decision-making. Second, they can be much less costly in terms of money, researcher time, and teacher time. The advantages of speed and cost are especially pronounced when researchers use retrospective designs and can access administrative outcome data. A third potential advantage of quasi-experiments is precision. In situations where a program of interest is widely used, quasi-experimental analyses can benefit from much larger samples than experiments (whose sample sizes are limited by both cost and the ability to recruit participants). The ability to analyze large samples helps address one of the biggest practical weaknesses of experiments, i.e., a lack of precision and a high likelihood that studies are uninformative (see Figures 1 and 2).

Researchers trained in causal inference may reasonably object that quasi-experimental methods require stronger assumptions to identify causal effects. This is doubtless true. However, emerging empirical evidence suggests that these assumptions can be plausible in education settings (Wong et al., 2017). For example, researchers may be concerned that schools who adopt a particular program may differ in unobserved ways from comparison schools, resulting in selection bias. Weidmann and Miratrix (2021) examined the magnitude of this bias for 14 primary school programs in England. For each program they compared the results from an experimental analysis with the results from a quasi-experimental analysis based on propensity-score matching. In a meta-analysis of 42 outcome measures they found

that the typical selection bias was 0.03SD. Analogous research by Sims et al., (2022) compared estimates using comparative interrupted time series with experimental benchmarks and found very similar results. Across four interventions (two in primary schools, two in secondary schools) the typical bias was 0.01SD.

Part of the reason for the success of quasi-experimental methods in education – compared to other contexts such as job-training – is the availability of high quality data. In particular, quasi-experimental methods are strengthened by reliable outcome measures from multiple time points, both before and after program implementation (Steiner et al., 2015). Such data is often available to education researchers – especially those focused on primary and secondary schools. For example, in the UK researchers can access the National Pupil Database which includes longitudinal achievement data, along with linkages to a wide range of other variables describing pupils, schools and neighborhoods. With access to such data, evidence suggests that the biases associated with quasi-experimental methods may be relatively small[6] (Coopersmith et al., 2022; Jaciw 2024; Sims et al., 2022; Weidmann and Miratrix, 2021; Wong et al., 2017; Wong et al., 2018). In sum, given the potential advantages of quasi-experiments, there is a strong case to be made for expanding their use to help inform decisions – especially if programs or practices are widespread and researchers can access administrative data that include longitudinal outcome variables.

Conclusion

We have developed a new framework for distinguishing the goals and methods of two types of educational experiments. By reviewing what we have learned about experiments through this lens, we have made three arguments that run counter to the prevailing practices. First, when evaluation is needed to inform decisions directly, quasi-experimental methods will

[6] This applies specifically to school-level selection bias, which is relevant to the vast majority of decision-informing experiments. Pupil-level selection bias is considerably harder to control for without randomisation.

often be superior to experiments on the grounds that they are more economical and precise. This is supported by recent empirical evidence suggesting that when researchers have access to high-quality data, quasi-experiments do a good job of replicating results from randomized experiments. Second, theory-informing experiments should be used more often, both because they can inform educators mental models and help to avoid research waste. Third, researchers should use theoretically aligned outcome measures more often in experimental research on the basis that they make experiments informative but no less comparable. We are at a point where change in how randomized experiments are used in education is necessary; without it, we risk continuing to produce studies of little practical value. Getting value from experiments requires changes to why, how and when they are used by educational researchers.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, *88*(1), 265-296.

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17*(3), 286-303.

Anders, J., Foliano, F., Bursnall, M., Dorsett, R., Hudson, N., Runge, J., & Speckesser, S. (2022). The Effect of Embedding Formative Assessment on Pupil Attainment. *Journal of Research on Educational Effectiveness*, *15*(4), 748–779. https://doi.org/10.1080/19345747.2021.2018746

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603-617.

Belfield, C. R., & Bowden, A. B. (2019). Using resource and cost considerations to support educational evaluation: Six domains. *Educational Researcher*, *48*(2), 120-127.

Bell, K., Ainsworth, H., Elliott, L., Fairhurst, C., Gascoine, L., Hanley, P., Haynes, A., Hewitt, C., Mitchell, A., Torgerson, C., Torgerson, D. & Whiteside, K. (2022). *ABRACADABRA (ABRA) and Reading and Understanding in Key Stage 1 (RUKS).* Evaluation Report, Education Endowment Foundation. September 2022.

Broers, N. J. (2021). When the numbers do not add up: The practical limits of stochastologicals for soft psychology. *Perspectives on Psychological Science*, *16*(4), 698-706.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of consumer research*, *9*(3), 240-244.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, *37*(2), 403-429.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283-292.

Coopersmith, J., Cook, T. D., Zurovac, J., Chaplin, D., & Forrow, L. V. (2022). Internal and external validity of the comparative interrupted time-series design: A meta-analysis. *Journal of Policy Analysis and Management*, *41*(1), 252-277.

Cullis, J., & Jones, P. (2009). *Public finance and public choice: analytical perspectives.* Oxford University Press.

de Boer, H., Donker, A. S., & van der Werf, M. P. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*(4), 509-545.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology, 58*(3), 470–484.

Dimova, S., Ilie, S., Brown, E. R., Broeks, M., Culora, A., & Sutherland, A. (2020). The Nuffield early language intervention. Education Endowment Foundation.

Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265-275.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education, 45*(1), 62-101.

Jaciw, A.P., 2024. Do social programs help some beneficiaries more than others? Evaluating the potential for comparison group designs to yield low-bias estimates of differential impact. *American Journal of Evaluation*, *45*(3), pp.361-392.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge University Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379.

Myers, D. E., & Schirm, A. L. (1999). *The impacts of Upward Bound final report for phase I of the national evaluation*. DIANE Publishing.

Ng, Y. K. (2003). *Welfare economics: towards a more complete analysis*. Springer.

Olsen, R. B., Orr, L. L., Bell, S. H., Petraglia, E., Badillo-Goicoechea, E., Miyaoka, A., & Stuart, E. A. (2023). Using a Multi-site rct to predict impacts for a single site: do better data and methods yield more accurate predictions?. *Journal of Research on Educational Effectiveness*, 1-27.

Ost, B., Gangopadhyaya, A., & Schiman, J. C. (2017). Comparing standard deviation effects across contexts. *Education Economics*, *25*(3), 251-265.

Popper, K. R. (1958). *The Logic of Scientific Discovery*. Basic Books.

Popper, K. R. (1962). *Conjectures and Refutations*. Basic Books.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., ... & Spier, E. (2010). Head Start Impact Study. Final Report. *Administration for Children & Families*.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C. N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology, 112*(1), 40-52.

Rosenbaum, P. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Roth E. A., Kagel H. J. (1995). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Schneider, M. (2020). Making common measures more common [blog]. https://ies.ed.gov/director/remarks/5-05-2020.asp

Sibieta, L., Kotecha, M., & Skipp, A. (2016). Nuffield Early Language Intervention: Evaluation Report and Executive Summary. *Education Endowment Foundation*.

Sims, S., Anders, J. and Zieger, L., 2022. The internal validity of the school-level comparative interrupted time series design: Evidence from four new within-study comparisons. *Journal of Research on Educational Effectiveness*, *15*(4), pp.876-897.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., ... & Anders, J. (2024). Effective teacher professional development: New theory and a meta-analytic test. *Review of Educational Research*, 00346543231217480.

Sims, S., Fletcher-Wood, H., Godfrey-Faussett, T., Mccrea, P., & Meliss, S. (2023). *Modelling evidence-based practice in initial teacher training: causal effects on teachers' skills, knowledge and self-efficacy* (CEPEO Working Paper No. 23-09). UCL Centre for Education Policy and Equalising Opportunities.

Smith, A., Staunton, R., Sahasranaman, A. & Worth, J. (2020). Impact Evaluation of Nuffield Early Language Intervention (NELI) Wave Two. Education Endowment Foundation.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255-267.

Steiner, P. M., Cook, T. D., Li, W., & Clark, M. H. (2015). Bias reduction in quasi-experiments with little selection theory but many covariates. *Journal of Research on Educational Effectiveness*, 8, 552–576.

Stone, G., Andrade, J., Martin, K., & Styles, B. (2020). *Helping Handwriting Shine: Evaluation Report*. National Foundation for Educational Research.

Stroebe, W., Gadenne, V., & Nijstad, B. A. (2018). Do our psychological laws apply only to college students?: External validity revisited. *Basic and Applied Social Psychology*, *40*(6), 384-395.

Weidmann, B., & Miratrix, L. (2021). Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management*, *40*(3), 964-986.

Weiss M. J., Bloom H. S., Verbitsky-Savitz N., Gupta H., Vigil A. E., Cullinan D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*, 843–876.

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134-161.

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*, 10, 207–236.

Wong, V. C., Steiner, P. M., & Anglin, K. L. (2018). What can be learned from empirical evaluations of nonexperimental methods? *Evaluation Review*, 42, 147–175.