ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-821

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

Sam Sims
University College London

Jake Anders
University College London

Matthew Inglis
Loughborough University

Hugues Lortie-Forgues
Loughborough University

Ben Styles
National Foundation for Educational Research

Ben Weidmann
Harvard University

Over the last twenty years, education researchers have increasingly conducted randomised experiments with the goal of informing the decisions of educators and policymakers. Such experiments have generally employed broad, consequential, standardised outcome measures in the hope that this would allow decisionmakers to compare effectiveness of different approaches. However, a combination of small effect sizes, wide confidence intervals, and treatment effect heterogeneity means that researchers have largely failed to achieve this goal. We argue that quasiexperimental methods and multi-site trials will often be superior for informing educators' decisions on the grounds that they can achieve greater precision and better address heterogeneity. Experimental research remains valuable in applied education research. However, it should primarily be used to test theoretical models, which can in turn inform educators' mental models, rather than attempting to directly inform decision making. Since comparable effect size estimates are not of interest when testing educational theory, researchers can and should improve the power of theory-informing experiments by using more closely aligned (i.e., valid) outcome measures. We argue that this approach would reduce wasteful research spending and make the research that does go ahead more statistically informative, thus improving the return on investment in educational research.

VERSION: August 2023

Suggested citation: Sims, Sam, Jake Anders, Matthew Inglis, Hugues Lortie-Forgues, Ben Styles, and Ben Weidmann. (2023). Experimental education research: clarifying why, how and when to use random assignment. (EdWorkingPaper: 23-821). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/6cyn-zn44

Experimental education research: clarifying 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[4]

[1] UCL
[2] Loughborough University
[3] National Foundation for Educational Research
[4] Harvard University

Over the last twenty years, education researchers have increasingly conducted randomised experiments with the goal of informing the decisions of educators and policymakers. Such experiments have generally employed broad, consequential, standardised outcome measures in the hope that this would allow decisionmakers to compare effectiveness of different approaches. However, a combination of small effect sizes, wide confidence intervals, and treatment effect heterogeneity means that researchers have largely failed to achieve this goal. We argue that quasi-experimental methods and multi-site trials will often be superior for informing educators' decisions on the grounds that they can achieve greater precision and better address heterogeneity. Experimental research remains valuable in applied education research. However, it should primarily be used to test theoretical models, which can in turn inform educators' mental models, rather than attempting to directly inform decision making. Since comparable effect size estimates are not of interest when testing educational theory, researchers can and should improve the power of theory-informing experiments by using more closely aligned (i.e., valid) outcome measures. We argue that this approach would reduce wasteful research spending and make the research that does go ahead more statistically informative, thus improving the return on investment in educational research.

Randomised 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 randomized experiments since at least 1918 (Hedges & Schauer, 2018). In the last 20 years, there has been a marked increase in the frequency with which experiments have been conducted in education. In the US, organizations such as the National Center for Education Research (NCER) and the National Center for Education Evaluation and Regional Assistance (NCEE) have commissioned more than 350 randomized experiments since 2002 (Hedges & Schauer, 2018). Likewise, the Education Endowment Foundation (EEF) in England has commissioned more than 157 randomized experiments since 2010. These experiments are expensive, with the average cost for an EEF trial estimated at around £500,000 (Lortie-Forgues & Inglis, 2019).

As educational researchers have conducted more experiments, so they have learned more about the challenges involved. We now know more about artefactual variation in Cohen's *d* effect sizes (Cheung & Slavin, 2016; Ost et al., 2017; Wolf & Harkbatkin, 2023). Partly as a result of this, typical effect sizes in experiments run by the NCEE and EEF have turned out to be much smaller than was suggested by prior empirical research (see Kraft, 2020). It has proven difficult to increase the precision of experiments in response (Spybrook et al., 2016). As a result, many experiments are uninformative, in the technical sense that findings are consistent with the intervention being either effective or ineffective (Lortie-Forgues & Inglis, 2019) (that said, we acknowledge that such studies taken as a whole are often informative in the broader sense of this term and, to avoid confusion with this broader concept, we use 'statistically (un)informative' throughout this paper). Furthermore, even those that initially appear to show promising results substantially exaggerate the true effect size (Sims et al., 2022). Moreover, there appears to be substantial heterogeneity in effect sizes across settings, meaning that 'what works' in one school may not work in another (Bloom et al., 2017). In short, all is not well in experimental education research.

This paper takes stock of what we have learned and reconsiders why, how and when education researchers should conduct experiments. In the first section, we develop a simple framework distinguishing two broad goals (the *why*) of experimental research and the methodological choices (the *how*) that would support each of these goals. In the second section, we look at how educational experiments are conducted in practice and evaluate how well they address these two broad goals. In the final section, we use our framework to synthesise what we have learned about experimental research over the last 20 years. This

allows us to clarify *when* educational experiments should (or should not) be conducted, as well as provide recommendations as to *how* researchers should go about conducting experiments to make them more statistically informative.

Clarifying the goals of experiments in education research

There are at least two broad goals that an educational 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). These two goals sit within different philosophical frameworks, aim at estimating different quantities, and therefore suggest different research methods. We consider each in turn.

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 theoretical conjectures and then trying to refute them. If the newly observed data refute the theoretical conjecture, the researchers' theoretical understanding must be modified appropriately; if it does not, then the conjecture can be tentatively accepted, at least until new data emerge to falsify it. The intention is that, 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 theory 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*.

The desire to inform theory, coupled with an endorsement of Popper's methodological falsification, leads to the adoption of certain typical research practices. To test the ordinal predictions made by theory, educational researchers often conduct experiments in which

participants are randomly assigned to receive an intervention manipulating a variable e.g., growth mindset. The researcher aims to conduct a tightly controlled experiment, perhaps in a lab setting, in order to provide a valid test of the theory. Across many experiments, randomisation guarantees that the two groups have equivalent characteristics besides receipt of the intervention. This implies that, so long as the researcher validly measures the construct(s) that are specified in the theory, the direction of the estimated effect allows the researcher to check whether the results falsify the ordinal prediction or not (Tunc, Tunc, & Laken, 2021). The experimental test should, of course, be set up so as to maximise the chances of it falsifying the theory, if indeed the theory is incorrect (Mayo, 2018).

Randomisation 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). (For more on randomisation inference, and how it differs from inference to a different set of units, see: Keel et al., 2012 and Rosenberger et al., 2019.) If the randomisation inference 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. Other results, in particular the point estimate (or effect size) are of interest only in the instrumental sense that they determine the centre of the confidence interval. The theory makes no prediction about them and they therefore cannot falsify the theory. Note that experiments informing theory do not need to statistically extrapolate the findings to a population of units outside the study sample. As long as the experimental sample is composed of units within the boundary conditions of the theory (Trafimow, 2022), the randomisation inference confidence interval can still potentially falsify the theory (Mook, 1983).

Experiments directly informing decisions

A researcher who aims to directly inform decisions will typically, though often implicitly, adopt a philosophical framework derived from welfare economics (Cohn, 2003; Ng, 2003). This approach to social science progresses by estimating the quantitative inputs necessary to inform a cost-benefit analysis (Cullis & Jones, 2009, ch. 6). 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 typically leads to the adoption of certain research practices. Costs can be quantified through careful accounting exercises (Belfield & Bowden, 2019). To quantify the benefits, educationalists 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'. In the interests of realism, the intervention is usually delivered by those who would deliver it in practice e.g., teachers. 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.

The result of interest in a directly decision informing experiment differs depending on whether there are known competing alternatives. A known competing alternative exists if the research community is unsure as to whether the opportunity cost (the benefit of the next best alternative foregone) is smaller or larger than the option under consideration. If there are no known competing alternative courses of action, then the decision maker is interested in whether the confidence interval is entirely above zero. If so, then the decision maker should pursue this course of action. This applies to many decisions made by classroom teachers. For example, consider a teacher who needs to produce a new learning resource and they are considering whether to place labels on the diagram or away from the diagram. However, if there are known competing alternatives (KCA) then the decision maker is interested in the point estimate (or effect size), as well as the confidence interval around it. This applies to decision where there are multiple evidence based approaches. It also applies to any decisions that incur financial costs, in that there are always KCA for spending money. For example, when a school principal is deciding whether to purchase curriculum resources or professional development programmes, they could instead use the money to purchase one-to-one tutoring for pupils.

Regardless of the KCA, the decision maker cares about the confidence interval *in the population of interest* i.e., the pupils with whom they work. There are therefore two inferences that must be made for this sort of experiment (Abadie et al., 2020). The first is whether the random treatment assignment genuinely found an effect in the sample

(randomisation inference). The second is whether any effect found in the randomly drawn sample would generalize to the population of interest (sample/population inference) – those who would receive the intervention should the decision maker select it.

Summary

Table 1 summarizes the goals (the *why*) and methods (the *how*) of the two types of experimental research outlined above. Experiments informing theory aim to test predictions deduced from theory, for which the quantity of interest is the confidence interval (or *p* value) within the study sample. Methodologically, the relevant dependent variables are the constructs specified in the theory and the measurement instruments are chosen based on the validity with which they capture this construct. By contrast, experiments informing decision making directly evaluate interventions, and aim to estimate comparable effect sizes as well as confidence intervals (or *p* values) for the target population. Methodologically, it is preferable that the intervention be implemented by educators and the dependent variables should be based on widely used test instruments (to enhance comparability) and be broad enough to matter for pupils' life chances (so that it can be interpreted as the social benefit of the intervention).

Table 1. Ideal features of theory-informing and decision-informing experiments

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

One further clarification is necessary at this point. Experiments informing theory can still be of use to educators making decisions about how to teach. However, this will be an indirect process in

which the experiment informs theory development, which then informs educators mental models, which then informs their decisions. For example, experiments testing growth mindset theory might change teachers minds about the importance of growth mindset, which might influence the way they communicate with pupils. By contrast, what we have called directly decision informing experiments are intended to provide information which is directly informative about the value of different options, without the mediating step of influencing educators' mental models. We refer to the latter *directly* decision informing in order to emphasise this point. In sum, both types of experiments are potentially valuable for improving educational outcomes, though in different ways.

Experimental design in practice

In the previous section we distinguished two broad goals of experimental research, each of which suggest a certain set of research methods. In this section, we turn to evaluating how experiments are actually conducted in education research. Our argument will be that, over the last 20 years, education researchers have tried but largely failed to conduct experiments capable of directly informing decisions. As a result, a series of compromises have made around *how* decision-informing experiments are designed, which has resulted in confusion about *why* those experiments are being conducted.

Experiments informing theory

There are currently many experiments conducted in education that attempt to inform theory. For example, consider Hodds, Alcock and Inglis's (2014) investigation of the efficacy of self-explanation prompts as a technique for improving mathematics undergraduate students' comprehension of mathematical proofs. Hodds et al. were attempting to build on the literature that shows self-explanation could improve students' reading comprehension in scientific domains (e.g., Ainsworth & Burcham, 2007). In their first experiment 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, $p < .001$). In two follow-on experiments, Hodds et al. investigated how their training materials influenced students' eye movements while reading proofs and whether the materials could be successful in real classroom settings.

Hodds et al.'s studies are aligned with the goals and methods listed in the informing theory column of Table 1 and thus allow them to draw theoretical conclusions – they do not reject the claim that generating self-explanations is causally linked to higher levels of

understanding when reading mathematical proofs for undergraduate students. However, as a result of adopting these design features, they do not allow them to 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. In sum, their work informs theory but does not directly inform decision making.

Experiments directly informing decisions

The vast majority of EEF or NCEE funded experiments are aimed at informing decisions. For example, the EEF state that one of their main aims is "Supporting education practitioners… to use evidence in ways that improve teaching and learning."[1] Likewise, the NCEE states that it is "transforming education into a field in which decisionmakers routinely turn to evidence to inform policies and practices that affect students."[2] Having said that, 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. In particular, there are two places in which experiments tend to diverge from this ideal.

First, such experiments deviate from the methods set out in the informing decisions column in that it is extremely rare for them to employ representative samples. This is presumably because representativity 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 – the Upward Bound high school enrichment programme (Myers & Schirm, 1999), the Head Start Impact Study (Puma et al., 2010), and an evaluation of the Tennessee statewide Pre-K programme (Durkin et al., 2022) – are all government funded programmes, in which participation in the experiment could be legally mandated. Even in these three cases, it was only possible to use a representative sample of *oversubscribed* sites, which is not likely to be the true population of interest for decision makers. 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 a convenience sample may be a poor proxy for effect sizes in the population of interest (Olsen et al., 2023; Weiss et al., 2017).

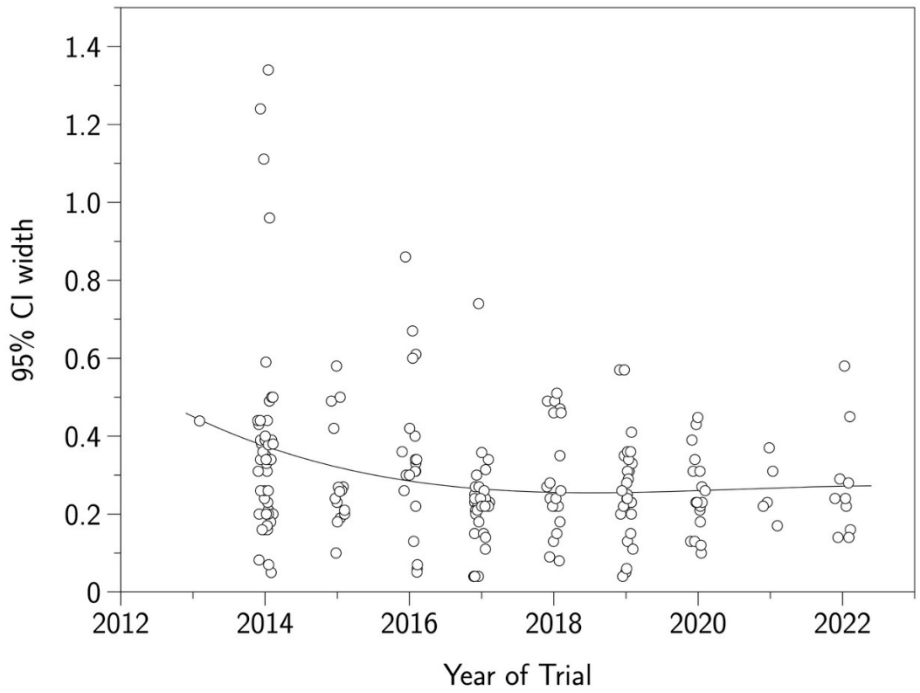
[1] https://educationendowmentfoundation.org.uk/about-us/how-we-work

[2] https://ies.ed.gov/ncee/pdf/ncee_brochure.pdf

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 RCTs conducted by the EEF and NCEE was 0.24 SDs. Figure 2 updates this analysis to include EEF trials published up to and including 2022 and shows that there has been no appreciable improvement in the precision of trials published by this particular funder over this period, although we note that, given the lag between commissioning and publication, this may not reflect recent efforts to address the issue. 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. This implies that we are learning very little about the relative impact of different interventions.

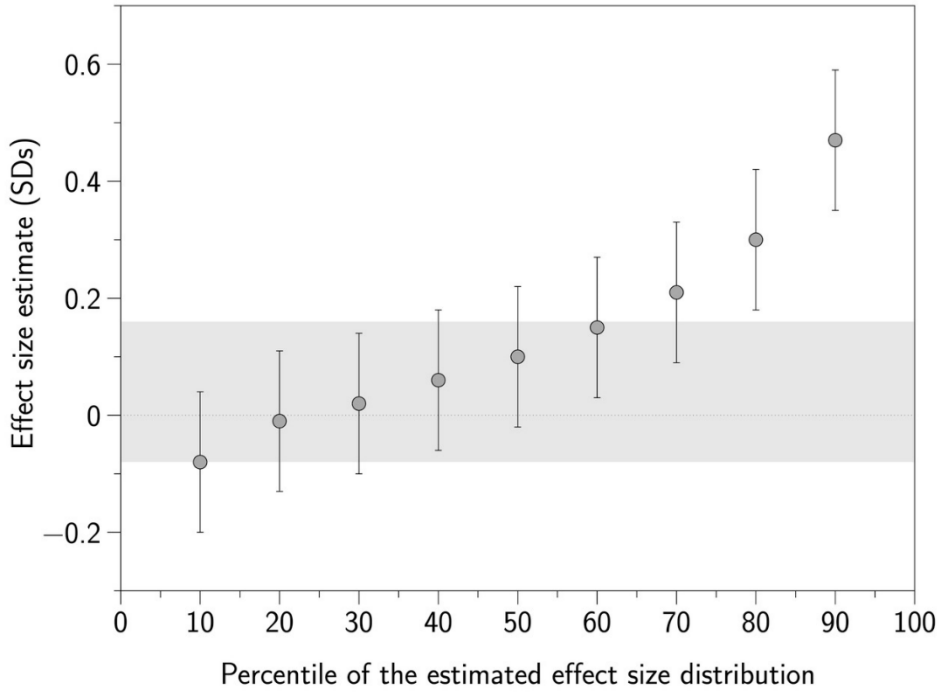
To illustrate this, Figure 2 plots the deciles of the distribution of effect sizes of experiments using broad outcome measures, as reported by Kraft (2020). For each decile, we have also super-imposed the mean confidence interval from NCEE and EEF trials, as reported in Lortie-Forgues and Inglis (2019). 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 figure makes it clear that the median effect size estimate cannot be distinguished from effect size estimates with magnitudes ranging from the 10^{th} to the 80^{th} percentile, in that the confidence intervals overlap. Only effect size estimates from the extreme ends of the distribution can be distinguished from each other and these extreme estimates are, by nature, likely to be exaggerated (Sims et al., 2022).

Figure 1. Change in width of 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, plotted by the year in which the trial reports were published. Cubic line of best fit.

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 & Inglis (2019). The grey region shows the 95% confidence interval around the precision-weighted mean.

Even in cases where there are no KCA, meaning that decision makers are largely interested in whether the confidence interval excludes zero, the combination of small average effect sizes and large confidence intervals renders many (but not all) educational experiments statistically uninformative. 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 (see the right-hand plot in Figure 2). 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).[3]

Experiments that fall somewhere between informing decisions and informing theory

[3] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/embedding-formative-assessment

Due to the difficulties in conducting the ideal decision-informing experiment, researchers have in practice made a number of methodological compromises. First, almost all experiments are conducted using a convenience sample constrained by a set of eligibility criteria. 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 purposes of the research project). Third, instead of focusing on broad, end-of-year, standardized tests, researchers sometimes opt instead for tests that are more aligned with the content covered in the intervention, which are also often captured more proximally to the intervention. Fourth, researchers sometimes opt to implement the experiments themselves, sometimes outside of an authentic school setting.

A good example of an experiment making many of these compromises is Rohrer, Dedrick, Hartwig & Cheung's (2019) study of interleaving in mathematics. They conducted a large-scale cluster RCT where 54 mathematics classes were assigned either blocked (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 with their own pupils. Participants comprised a convenience sample of grade 6-8 schools within a given travel distance of the researchers' university. A month after the end of the intervention, all participants took an unannounced test, which was developed by the researchers to cover the exact content covered in the blocked/interleaved worksheets. Pupils in the classes allocated to interleaved assignments scored higher than those who received blocked assignments (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 (as highlighted in the abstract) was the magnitude of the standardized 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. Our intention here is not to criticise the Rohrer et al. study (see next section), only to use it to illustrate common methodological compromises made in an apparently decision-informing experiment.

These compromises are summarized in Table 2 below. Note that there is no analogous set of compromises for experiments informing theory because researchers typically find the methods in that column straightforward to adopt.

Table 2. Features of theory-informing and decision-informing experiments in practice

	Experiments informing theory	**Experiments directly informing decisions**	
	Ideal	**Common compromise**	**Ideal**
Aims (*why*):			
Testing:	Deduced hypotheses	Interventions / Policies	Interventions / Policies
Quantities of interest:	Confidence interval / *p* value	Sample Average Treatment Effect	Population Average Treatment Effect
Methods (*how*):			
Sampling:	Purposeful	Convenience	Representative
Dep. Var.:	Relevant construct	Socially desirable	Consequential outcome
Instruments:	Maximally valid	Narrow/proximal	Broad standardized
Implementer:	Researcher	Developer	Educator

Rethinking the design of experiments in education

In this final section, we use the framework in Table 2 to synthesise what we have learned about conducting experiments in education over the last 20 years. Our discussion is structured around a series of recommendations for why, how and when education researchers should (not) use experiments.

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

Our first recommendation is that experimental researchers should use narrower, more aligned outcome measures, particularly where there are no KCA. This is illustrated nicely by the Rohrer et al. (2020) interleaving experiment summarised in the last section. We believe that many researchers and funders of experiments in the decision-informing tradition 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 tests designed by the researcher to assess the content covered in the intervention tended to find larger effects. Based on this, the authors recommend that researchers *always* include standardized test outcome measures because "they have the great advantage that they facilitate a more reliable comparison of multiple intervention effects than do unstandardized measurements." Likewise, Slavin (2019) recommends that the results on researcher designed outcome measures should *never* be emphasized in research reports. Citing these arguments, the Institute of Education Sciences (IES) recently announced its intention only to fund experiments that use standardized tests as outcome measures, on the grounds that "without common measures, we have little ability to look across interventions for what works and what is most cost effective" (Schneider, 2020, p.1). We believe that the severe lack of precision in educational experiments that use broad outcome measures (Figure 1 and 2) means that this is likely to be poor advice. Education experiments simply do not have the precision to compare the effectiveness of different interventions. Aligning the outcome measures to enhance comparability in experimental findings is therefore of no 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, thus making it more statistically informative. An instructive example of an experiment employing narrow, more aligned outcomes is the Nuffield Early Language Intervention (Dimova et al., 2020). This is a scripted language programme delivered by teaching assistants to small groups of pupils entering primary (elementary) school, costing £58 per pupil. The primary outcome was a 'language skills' score derived from four tests, each chosen to be closely aligned with the intervention. Two separate experiments in England found positive impact on these closely-aligned tests (d=0.27, 95% CI 0.07-0.46; d=0.26, 95% CI 0.17-0.35) and the intervention has now been rolled out nationally (Sibieta et al., 2016; Dimova et al., 2020). Using a broad outcome measure would have diluted the effect size (Kraft, 2020; Wolf & Harbatkin, 2023) and would likely have led to a null finding, resulting in no national roll out. The NELI evaluations are fairly unusual in having such narrow primary outcomes measures for their trials, which makes it particularly instructive that the Nuffield Early Language Intervention is one of just two EEF trials to have shown evidence of impact in both an initial efficacy trial and a second (scaled-up) effectiveness trial, the other being the Abracadabra (ABRA) programme.

Why: Theory-informing experiments conducted by researchers can help cut research waste on statistically uninformative decision-informing experiments

Our next recommendation rests on two observations. First, theory-informing experiments tends to be substantially cheaper than directly decision-informing experiments. For example, the Hodds et al. (2014) series of 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 and therefore more statistically informative than directly decision-informing experiments (Lortie-Forgues & Inglis, 2019). It follows from these two observations 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 implied by the interventions (Deaton & Cartwright, 2018). Where a prior theory-informing experiment finds positive results, this increases the warrant for conducting a subsequent decision-informing experiment. By contrast, 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, which is unlikely to provide a statistically informative finding (Ioannidis et al., 2014).

An instructive example here is the Helping Handwriting Shine trial (Stone et al., 2020). This has many of the features that one would expect from a directly 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 primary outcome measure was a broad assessment of general writing skills, with the ultimate aim of improving reading and writing in high stakes tests at age 11 (Stone et al., 2020, p.12). The theory embedded within this intervention was that children who have slow and effortful handwriting use scarce cognitive resources on the act of writing, which diverts them from the substance their writing, thus harming their academic work. Improving handwriting automaticity would free up cognitive capacity and thereby improve writing. The intervention is estimated to cost around £180 per targeted child. A pair of trials in England found null results for both 6-7 year olds (d = -0.02, p = 0.77) and 9-10 year olds (d = 0.12, p = 0.16). Crucially, the trial also found null results on the secondary outcome measure, writing speed, suggesting that the intervention didn't even succeed in improving handwriting automaticity. A theory-informing experiment focused on writing speed and other

measures of cognitive capacity could have established this much faster and would have saved the considerable expense of conducting two large field experiments yielding null findings.

Why: Some applied research questions can only be answered through theory-informing experiments

Our next recommendation relates to the value of theory-informing experiments for answering a certain set of research questions in education. Figure 1 and 2 showed the marked lack of precision/power in education experiments, even when these are testing sustained, broad-based interventions, such as curriculum reform covering an entire maths course. However, there is a whole class of questions that educationalists are interested in that involve less sustained or broad-based change. This is well illustrated by the Hodds et al. self-explanation study introduced above, which focused on the specific practice of self-explanation. Since these small grain size questions are (other things equal) likely to have correspondingly smaller effect sizes, it is very unlikely that they can be addressed by experiments adopting decision-informing goals and methods. 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 mathematical proofs) to detect any effects. This leaves theory testing experiments as the only suitable method for providing a causal test of such small grain size changes.

Another illustrative example here is teacher professional development (PD). There are hundreds of experimental evaluations of months-long teacher professional development programmes aimed at bringing about broad changes in pedagogy (Lynch et al., 2019; Sims et al., 2021). However, teacher educators are often interested in smaller grain size questions, such as whether to add modelling (observable examples of specific teaching techniques) to an existing professional development programme. Sims et al. (2023) therefore conducted a classroom simulator experiment to test a number of theoretically informed hypotheses about whether and how modelling improves teacher professional development, finding that it did indeed improve teacher's 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. However, it is unlikely that a directly decision-informing experiment could have detected effects of a granular change such as adding/removing modelling from some PD.

How: Theory-informing experiments can be conducted more efficiently using sequential analysis

We have already noted that theory-informing experiments can often be conducted at lower costs than decision-informing experiments. The value for money of theory-informing research can be improved further still by using sequential analysis. This differs from traditional experiments in which the sample size is fixed in advance or where there is a cut-off date for recruitment. In sequential analysis, the data is instead analysed in successive steps and the sample grows until the evidence satisfies some criteria. This approach enables evaluators to terminate a trial early if convincing evidence of an effect is found or if the intervention tested appears highly unlikely to be effective (Wald, 2004). The alpha level at each step is adjusted to maintain control of Type 1 error rates. The advantage of sequential analysis is that the additional flexibility around when to stop recruitment often leads to a reduction in the number of participants required in a trial (Lakens, 2014). The disadvantage is that the effect size estimate is likely to be upwardly biased and there is no consensus on how to correct this. However, this is not a problem in theory-informing education experiments, which are not (directly) concerned with estimating effect sizes anyway. For an example of an education evaluation using sequential analysis, see Worth et al. (2018). Accessible tutorials on sequential analysis can be found in Miller and Ulrich (2021) and in Lakens, Pahlke and Wassmer (2021).

When: When evaluation is needed to inform decisions directly and programmes are already in widespread use, quasi-experimental methods will often be more statistically informative than experiments

A considerable proportion of experimental education research aims at testing the effectiveness of commercially available programmes that are already in widespread use in schools. For example, recent experiments in England have tried to evaluate the impacts of particular systematic synthetic phonics programmes (Molotsky et al., 2022). Since these programmes have several KCA, we would ideally like to run a decision-informing experiment to evaluate them, using broad high-stakes outcomes of the sort that are captured in secondary data. However, the persistent lack of precision of such experiments (see Figure 1 and 2) suggests that experiments are unlikely to provide statistically informative findings about such programmes. This is particularly true where programmes are already in widespread use because the number of schools who can be persuaded to switch to using them is necessarily constrained. Furthermore, since there can be considerable cross-site (between

school) variation in the effects of education interventions (Weiss et al., 2017), the convenience samples that such experiments inevitably end up using may not be informative about the effect size in schools outside of the trial.

In such cases, we would recommend moving away from random assignment unless it is possible to run very large trials, perhaps through governments mandating participation or when pupil-level randomisation is possible. In all other cases, quasi-experimental methods applied to naturally occurring variation are likely to provide more statistically informative findings, something which research funders such as the EEF are already moving towards. Experimentalists will likely object to this on the grounds that quasi-experimental methods require stronger assumptions to identify causal effects. However, many within-study comparisons (including work funded by the EEF as part of its work on supporting a greater diversity of evaluation methods) have now been conducted showing that quasi-experimental methods actually provide very similar point estimates (effect sizes) to experimental evaluations of the same education programmes. Indeed, meta-analyses have shown that regression discontinuity designs (Chaplin et al., 2018), propensity score matching (Weidmann & Miratrix, 2021), and comparative interrupted time series designs (Coopersmith et al., 2022; Sims et al. 2022) all show very small mean absolute bias relative to experimental benchmarks.

Where many schools are already receiving the intervention of interest, quasi-experiments can leverage this naturally occurring variation in exposure to achieve greater precision, faster, and at lower costs than in an experiment – so long as the relevant outcomes are captured in secondary data. Of course, this will not always be the case, and, more broadly, our argument should not be taken as suggesting quasi-experimental trials are not without their own methodological and practical challenges. Our argument is towards greater consideration of these as an option, recognising the challenges of both experimental and quasi-experimental approaches.

This approach would also improve the potential for ranking the effect sizes of, e.g., alternative phonics programmes, in a way that can better inform the decisions of budget holders. Furthermore, such large sample size quasi-experiments can potentially provide impact estimates for sub-groups of treated units, for example by matching and comparing schools with higher or lower levels of deprivation. This allows researchers to explore heterogeneity in treatment effects, which is valuable for decision makers, who want to know about likely effect sizes for pupils like those in their settings (Jaciw, 2023).

An important caveat here is that a switch to using more such quasi-experimental evaluations should not involve a move away from the pre-registration practices that have become almost standard in experimental research. Wherever possible, such quasi-experiments should carefully pre-register their analyses in order to avoid the bias that is known to results from selective reporting when pre-registration is absent (Brodeur et al., 2022; Dreber et al., 2023).

When: When evaluation is needed to inform decisions directly and programmes can be implemented at the pupil level without contamination, multi-site trials will often be a useful approach

When a programme can be successful implemented at the pupil level without risk of contamination and related challenges, which we acknowledge represent a significant barrier in many such circumstances, we should aim to randomise at the pupil level. This is crucial to address pupil-level selection bias which is considerably harder to control for using quasi-experimental designs (e.g., Poet et al., 2022). Pupil-level randomisation often allows for adequate precision to detect the kinds of effect sizes seen in education trials with distal outcomes. Furthermore, randomising pupils within multiple separate schools (a multi-site trial) allows researchers to measure the site-by-treatment interaction, thus revealing the extent of between-site treatment heterogeneity. For examples of such multi-site trials in education, see Weiss et al. (2017), Olsen et al. (2023), and Lord et al. (2021).

Conclusion

Experiments promise unbiased causal inference with few additional assumptions. Education researchers have therefore conducted a large number of such experiments over the last twenty years in an attempt to better inform decision making. However, a combination of small effect sizes, wide confidence intervals, and treatment effect heterogeneity means that many such experiments have contributed little to educators' or policymakers' understanding of which course of action they should pursue. In this paper, we have developed a new framework for distinguishing different types of educational experiments. By reviewing what we have learned about experimental research over the last twenty years through this lens, we believe that this paper makes three original contributions.

First, we have made the case that, when evaluation is needed to inform decisions directly, quasi-experimental methods will often be superior to experiments on the

grounds that they can often achieve more precise estimates of the treatment effects of interest and better assess heterogeneity in the magnitude of these treatment effects. This is particularly true when interventions are already in widespread use. This conclusion is now much better warranted than it was ten years ago because we have better empirical evidence on 1) the likely (small) magnitude of effect sizes, 2) the likely (considerable) heterogeneity of treatment effects, 3) the (generally quite strong) internal validity of quasi-experimental methods. The exception to our recommendation that quasi-experimental methods should be used more often in place of experiments for informing decisions is when interventions can be randomised at the pupil level, which allows for more precise estimates compared to cluster randomisation and allows multi-site designs to be used to assess heterogeneity.

Second, we have made the case that theory-informing experiments should be conducted more often in education research. There are many important questions faced by educators that simply cannot be addressed using either quasi-experimental methods or (given what we have learned about effect sizes) directly decision informing experiments. The findings from such decision informing experiments are valuable in applied education research in that they can help educators' to refine their mental models and thus make better decisions. Since theory informing experiments can also be conducted quicker and at lower cost than decision making experiments, they should also be used more often to efficiently falsify intermediate links in the causal chain embedded within large programmes. This would prevent research waste.

Third, we believe we have helped to resolve a debate about which outcome measures should be used in education experiments. In theory informing experiments, outcome measures should be aligned with the constructs specified in the relevant theory. Calls for more distal, consequential outcome measures simply misunderstand why the research is being conducted. Researchers conducting theory informing experiments can help to avoid this misunderstanding by clearly stating their theory-informing goals. In directly decision informing experiments, standardised and consequential outcomes remain the ideal, in that they can be used to compare the benefits of alternative courses of action. However, recent empirical evidence shows that effect sizes on such outcomes are likely so small that experiments usually lack the statistical precision to make such comparisons anyway. Using outcome measures that are more aligned with the

programmes being tested is therefore likely to result in more, not less statistically informative experiments.

Rich nations have spent billions of pounds on experimental education research in the last twenty years, however much of it has proven statistically uninformative (Lortie-Forgues & Inglis, 2020). We believe that experiments should continue to play an important role in education research, but only if the field makes significant changes in why, when and how experiments are conducted. Funders and researchers have already started to make some of the changes that we propose, but we argue that further systematic changes in approach are needed if we want governments to continue supporting experimental research in education for the next twenty years and beyond.

Acknowledgements

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 and enriching this work.

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.

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.

Brodeur, A., Cook, N., Hartley, J., & Heyes, A. (2022). Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking and Publication Bias?. GLO Discussion Paper, No. 1147, Global Labor Organization (GLO), Essen.

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.

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.

Cohn, E. (2003). Benefit-cost analysis: a pedagogical note. *Public Finance Review, 31*(5), 534–549.

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.

Dreber, A., Johannesson, M., & Yang, Y. (2023). Selective reporting of placebo tests in top economics journals. s, I4R Discussion Paper Series, No. 31, Institute for Replication (I4R), s.l.

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.

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., & Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*(9912), 166–175.

Jaciw, A. P. (2023). 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*, 10982140231160561.

Keele, L., McConnaughy, C., & White, I. (2012). Strengthening the experimenter's toolbox: Statistical estimation of internal validity. *American Journal of Political Science*, *56*(2), 484-499.

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

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710.

Lakens, D., Pahlke, F., & Wassmer, G. (2021). Group sequential designs: A tutorial. https://doi.org/10.31234/osf.io/x4azm

Lord, P., Rennie, C., Smith, R., Gildea, A., Tang, S., Miani, G., Styles, B. & Bradley, C. (2021) *Randomised Controlled Trial Evaluation of Families Connect.* Nuffield Foundation.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260-293.

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

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115.

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.

Miller, J., & Ulrich, R. (2021). A simple, general, and efficient method for sequential hypothesis testing: The independent segments procedure. *Psychological Methods*, *26*(4), 486.

Molotsky, A., Dias, P. & Nakamura, P. (2022). *Read Write Inc. Phonics and Fresh Start: Evaluation Report*. Education Endowment Foundation.

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.

Noyes, A., Adkins, M. & Taneva, S. (2022). *Glasses in Classes: A two-armed cluster randomised trial*. Education Endowment Foundation.

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.

Poet, H., Lord, P., Styles, B., Oppedisano, V., Zhang, M. & Dorsett, R. (2022) *Evaluation of year 1 of the Tuition Partners Programme: Impact evaluation for primary schools*. Education Endowment Foundation.

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.
Rosenberger, W. F., Uschner, D., & Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, *38*(1), 1-12.
Roth E. A., Kagel H. J. (1995). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy, 32*(4), 450-466.
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., Inglis, M., & Lortie-Forgues, H. (2022). Quantifying "Promising Trials Bias" in Randomized Controlled Trials in Education. *Journal of Research on Educational Effectiveness*, 1-18.
Sims, S., Anders, J., & 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), 876-897.
Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., ... & Anders, J. (2022). Effective Teacher Professional Development: New Theory and a Meta-Analytic Test. EdWorkingPaper No. 22-507. *Annenberg Institute for School Reform at Brown University*.
Sims, S., Godfrey-Fausset, T., Fletcher-Wood, H., Meliis, S., & Mccrea, P. (2023). Modelling in initial teacher education: causal effects on skills, knowledge and self-efficacy. Ambition Institute.
Slavin, R. (2019). Developer and researcher made measures [blog]. https://robertslavinsblog.wordpress.com/2019/10/24/developer-and-researcher-made-measures/
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.
Stone, G., Andrade J., Martin, K. and Styles, B. (2020). Helping Handwriting Shine Evaluation Report. Education Endowment Foundation.
Trafimow, D. (2022). A New Way to Think About Internal and External Validity. *Perspectives on Psychological Science*, 17456916221136117.
Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2021). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, 09593543231160112.
Wald, A. (2004). *Sequential analysis*. Courier Corporation.
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.

Worth, J., Nelson, J., Harland, J., Bernardinelli, D. & Styles, B. (2018*) GraphoGame Rime Evaluation Report*. Education Endowment Foundation.