# Missing, presumed different:

## Quantifying the risk of attrition bias in education evaluations

We estimate the magnitude of attrition bias for 10 Randomized Controlled Trials (RCTs) in education. We make use of a unique feature of administrative school data in England that allows us to analyse post-test academic outcomes for nearly all students, including those who originally dropped out of the RCTs we analyse. We find that the typical magnitude of attrition bias is $0.015\sigma$, with no estimate greater than $0.034\sigma$. This suggests that, in practice, the risk of attrition bias is limited. However, this risk should not be ignored as we find some evidence against the common 'Missing At Random' assumption. Attrition appears to be more problematic for treated units. We recommend that researchers incorporate uncertainty due to attrition bias, as well as performing sensitivity analyses based on the types of attrition mechanisms that are observed in practice.

Keywords: education; evaluation; bias; attrition bias; missing-at-random; uncertainty; RCT

Corresponding author: Ben Weidmann; benweidmann@g.harvard.edu;
11 Bayston Road, Stoke Newington, London, UK, N16 7LU.

**Acknowledgements**

We are grateful to the UK Department for Education for providing access to the National Pupil Database, and to the Education Endowment Foundation for providing access to their archive of randomized controlled trials. This work reflects the views of the authors and neither of the aforementioned organizations.

## Section 1: Introduction and background

Attrition has been described as "the Achilles Heel of the randomized experiment" (Shadish et al., 1998 p.3). Attrition looms as a threat because it can undermine group equivalence, eroding the methodological strength at the heart of randomized evaluations. In short, attrition can cause bias.

Attrition bias is the focus of this paper. We define attrition bias as the difference between the expected Average Treatment Effect (ATE) estimate of the final analysis sample, and the ATE of the randomization sample. Our main goal is to quantify and explore the nature of attrition bias *in practice.* We focus on the context of education research, a field which has seen a large increase in the number of randomized experiments over the past two decades (Connolly et al., 2018).

The threat of attrition bias plays a significant role in assessing the quality of education evaluations. The What Works Clearinghouse (WWC) and the Education Endowment Foundation (EEF) – organisations responsible for setting evidence standards in the US and the UK – both have threshold rates of attrition (EEF, 2014; WWC, 2017). Beyond these thresholds, studies officially lose credibility. In the case of the EEF, for example, if attrition is greater than 50% then the results of the evaluation are largely disregarded.

Despite the awareness of attrition as a threat to the quality of education research, remarkably little scholarship has focussed on quantifying the magnitude of attrition bias (Dong & Lipsey, 2011). The reason is simple: estimating attrition bias requires outcome information from pupils who, by definition, are no longer participating in research.

In response to this fundamental empirical challenge, existing literature has largely focussed on simulation studies. These studies demonstrate scenarios for which attrition bias is larger than the typical effect sizes in education interventions (Dong & Lipsey, 2011; Lewis, 2013; Lortie-Forgues & Inglis, 2019; WWC, 2014). Equally, it is well-known that if attrition is unrelated to either treatment status or outcomes, then randomized experiments remain unbiased regardless of the level of attrition (Little & Rubin, 2019).

Whilst theory and simulation studies illustrate the potential for attrition bias to cause problems, they provide practitioners with limited guidance about the risk of attrition bias in practice. Deke and Chiang (2017) take the first step toward providing such guidance. They attempt to sidestep the fundamental challenge of estimating attrition bias by analysing pre-test academic achievement as proxies for post-test outcomes. Pre-tests are often completed by pupils who stay in the evaluation (responders) as well as those who ultimately drop out (attriters). Using pre-tests, Deke and Chiang estimate attrition bias for four experiments, in each case comparing the estimated Sample Average Treatment Effect (SATE) of the whole sample to the estimated SATE of responders.

While Deke and Chiang (2017) represents an important step forward, it has several limitations. First and foremost, attrition may be shaped by events that happen *after* randomization. For example, during the course of an evaluation, a school may experience a change of leadership. If the new leader decides that implementing a research intervention is a distraction, they may drop out of the study. The resulting attrition could lead to bias if the leadership change coincides with, or causes, a decline in academic attainment. This bias would not be captured in an analysis that used a pre-test as a proxy for post-test outcomes. Second, analysing attrition bias using pre-tests makes it difficult to know whether attrition bias is problematic conditional on predictive covariates as the pre-test – by far the most predictive covariate – is being used as the outcome. As Deke and Chiang note "after conditioning on the pre-test, the residual difference in the post-test between respondents and nonrespondents could be completely different from the observed pre-test difference" (p139). Finally, the study only looked at four interventions. This makes it difficult to describe the distribution of attrition bias across studies, and to estimate the typical value of attrition bias. The relative lack of cases also makes it hard to analyse some of the factors that might moderate attrition bias.

We avoid these limitations by utilizing a unique feature of English administrative school data. Specifically, we make use of an archive of Randomized Controlled Trials (RCTs) that can be linked to a census of pupil and school information. Ten of the RCTs in the archive met two crucial conditions: a) the original outcomes were subject to attrition; b) after the intervention, students had sat a compulsory achievement test. For the 10 RCTs in our sample, we were able to analyse *post-test* academic achievement outcomes for students who exited the original randomized experiments. By comparing these outcomes to the equivalent outcomes of responder students, we can estimate attrition bias.

Attrition rates in our sample of 10 RCTs were typical of education experiments. At the student level, the mean rate of attrition across the 10 studies was 19 percent. A broader analysis of education RCTs in the UK found mean student-level attrition of 19 percent (n=79 experiments, see Demack et al., forthcoming).

Our paper makes four contributions. First, we present novel estimates of attrition bias for 10 education RCTs, spanning 22 outcomes. This advances empirical scholarship by providing estimates of bias based on post-test outcomes. Using techniques from meta-analysis, we then estimate the typical magnitude of attrition bias across studies and outcomes.

Second, we present a framework for decomposing attrition bias. This illustrates that attrition bias is a function of four components: the rate of attrition in each treatment arm, and the association between attrition and outcomes in each arm. We quantify the magnitude of these components across 22 study-outcome pairs, and report parameter values that define how pernicious attrition mechanisms tend to be in practice.

Third, we examine the plausibility of the "Missing At Random" (MAR) assumption. In most real-world situations, this assumption is untestable. In our context, however, we are able to test MAR at the study-outcome level, as well as providing a global test across all the outcomes in our sample of evaluations. We find evidence against MAR.

Finally, we provide two substantive recommendations for researchers in the field: incorporate uncertainty from attrition bias, and check whether conclusions are sensitive to 'worst-observed case' attrition mechanisms. For both recommendations we offer simple techniques that can be used in applied research. We illustrate these techniques with the *REACH* evaluation, an RCT of a reading intervention in England (Sibieta, 2016).

The paper is organized as follows. Section 2 defines and decomposes attrition bias. Section 3 describes the data and the interventions that underpin our analyses. Section 4 illustrates our approach to estimating attrition bias and presents headline estimates across 22 study-outcome pairs. In section 5 we test the MAR assumption and examine some potential predictors of pernicious attrition mechanisms. Section 6 provides researchers with recommendations about dealing with attrition bias, and section 7 concludes.

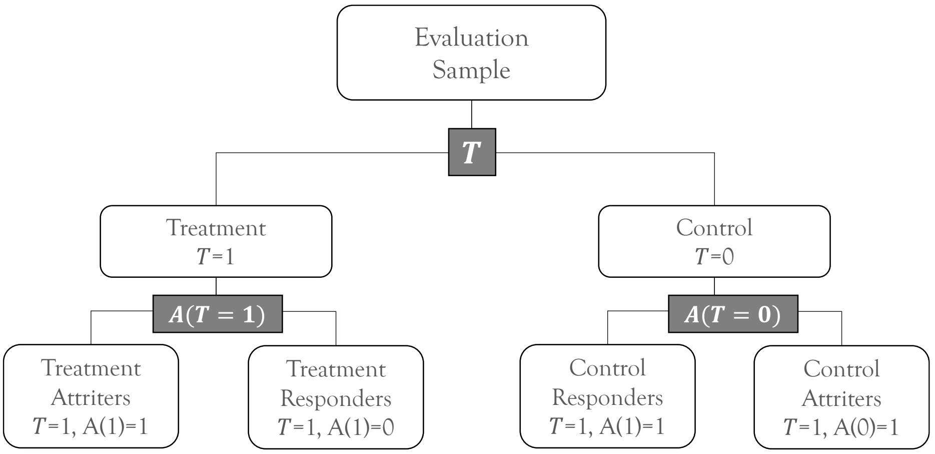
## Section 2: Conceptual Framework

### 2.1 Overview

Consider an evaluation with n students in the sample at randomisation, with $T_i$ as the randomly assigned binary treatment indicator for student $i$. We use potential outcomes notation in which $Y_i(t)$ denotes the post-treatment outcome when $T_i = t$. In our analyses, outcomes are standardized achievement tests across a range of domains including maths, reading, and science. The estimand of interest is the finite-sample Sample Average Treatment Effect. This is denoted by $\tau_{FULL} = E[Y(1) - Y(0)]$. The 'full' subscript indicates that we are interested in the average treatment effect for all the units in our original sample, before any attrition. The "$E[]$"denotes the simple average across the $n$ units in the sample.

Estimates of $\tau_{FULL}$ can be biased by non-random attrition. To formalize this, let $A_i(t)$ be a binary indicator of attrition under treatment assignment $t$. For example, $A_i(1) = 1$ describes a unit who left the evaluation after being assigned to treatment (a 'Treatment Attriter'). Figure 1 summarizes our setup. Our original treatment and control groups are directly comparable (up to imbalance caused by randomness in treatment assignment). Our evaluation sample consists of the Treatment Responders and the Control Responders.

Figure 1: Conceptual overview


Evaluation Sample
$T$
Treatment $T$=1
Control $T$=0
$A(T = 1)$
$A(T = 0)$
Treatment Attriters $T$=1, A(1)=1
Treatment Responders $T$=1, A(1)=0
Control Responders $T$=1, A(1)=1
Control Attriters $T$=1, A(0)=1

There are two concerns with an analysis of the evaluation sample. First, the Treatment Responders and Control Responders may not be directly comparable to each other if different types of students tended to attrit when exposed to treatment as compared to control. Consider an extreme case in which all struggling schools drop out of a particularly time-intensive treatment arm. The final treatment group would have the remaining high performers, and the control group a mix of both. This could bias impact estimates due to this systematic imbalance between the two groups. We call this "differential attrition bias".

There is a second, more subtle source of attrition bias. Even if we generate a correct impact estimate for our final evaluation sample, it may not generalize to the full sample. Consider the case where those most sensitive to treatment also are least likely to drop out of the study regardless of their assignment. In this case, even in the absence of differential attrition bias, our impact estimates would be systematically too high. We call this "generalizability attrition bias".

Attrition bias, as we define it, encompasses both differential and generalizability attrition bias. To state our definition more formally, let $\tilde{\tau}_{Responder}$ be the expected contrast between the 'treated responders' and 'control responders':

$$\tilde{\tau}_{Responder} = \tilde{\tau}_{\mathrm{R}} = E[Y(1)|A(1) = 0] - E[Y(0)|A(0) = 0] \quad (1)$$

The above is the average treatment outcome of all students who respond under treatment minus the average control outcome of all students who respond under control. This is not a causal effect estimand: while these two groups may share some portion of students, they do not necessarily share all of them, so our difference is not a well-defined treatment vs. control contrast on the same units. $\tilde{\tau}_R$ is, however, what is being estimated by contrasting the outcomes of treated and control units in the responder sample. This is almost always the contrast that applied researchers examine.

Using the above, we define "attrition bias" as the difference between the preferred estimand ($\tau_{FULL}$) and the typical focus of applied researchers ($\tilde{\tau}_R$):

$$\beta = \tau_{Full} - \tilde{\tau}_R \quad (2)$$

To better understand attrition bias, we decompose $\beta$ into four elements: the rate of attrition in each treatment arm ($P_T$,$P_C$) and the extent to which the attrition mechanism is associated with outcomes, in each arm ($\Delta_T, \Delta_C$). The next sub-section defines these terms and illustrates the decomposition.

### 2.2 Bias decomposition

We can write attrition bias in terms of how attrition impacts the estimated average outcome on the treatment side and the control side. Let $\beta_T$ be the attrition bias in the treatment arm:

$$\beta_T = E[Y(1)] - E[Y(1)|A(1) = 0] \quad (3)$$

$\beta_T$ represents the gap, in terms of average outcome under treatment, between the full sample and the 'treatment responders'.

Defining $P_T = P(A(1) = 1)$, and using the law of total probability, we have:

$$\beta_T = P_T \cdot E[Y(1)|A(1) = 1] + (1 - P_T) \cdot E[Y(1)|A(1) = 0] - E[Y(1)|A(1) = 0]$$

$$= P_T \cdot (E[Y(1)|A(1) = 1] - E[Y(1)|A(1) = 0]) \;\; (4)$$

Next, let $\Delta_T$ be the difference in expected outcomes between 'treatment attriters' and 'treatment responders':

$$\Delta_T = \; E[Y(1)|A(1) = 1] - E[Y(1)|A(1) = 0] \;\;\; (5)$$

Combining (5) and (6) we have:

$$\beta_T = P_T\Delta_T$$

Following the same logic on the control side, where $P_C = P(A(0) = 1)$, we have:

$$\beta_C = P_C \cdot (E[Y(0)|A(0) = 1] - E[Y(0)|A(0) = 0])$$

$$= P_C\Delta_C$$

Returning to our original definition of attrition bias (2), we have:

$$\beta = \tau_{FULL} - \tilde{\tau}_R$$

$$= (E[Y(1)] - E[Y(0)]) - (E[Y(1)|A(1) = 0] - E[Y(0)|A(0) = 0])$$

$$= (E[Y(1)] - \; E[Y(1)|A(1) = 0]) - (E[Y(0)] - E[Y(0)|A(0) = 0])$$

$$= \beta_T - \beta_C$$

$$= P_T\Delta_T - P_C\Delta_C \;\;\; (6)$$

Equation (6) shows attrition bias can be conceptualized as a function of four parameters: $P_T, P_C, \Delta_T$, and $\Delta_C$.

### 2.3 Covariate adjustment

Our parameter $\beta$ captures attrition bias if we make no attempt to account for attrition using measured pre-treatment covariates ($X$). We next discuss the extent to which covariate adjustment can help reduce attrition bias. Covariate adjustment is used in RCTs to account for chance imbalance of covariates that are predictive of outcomes. By conditioning on these covariates – for example, by using a regression model – researchers can address these imbalances and compare the treatment and control groups on a more equal footing. For example, if the treatment group had students with systematically higher baseline test scores, we would want to adjust for these scores as we might imagine the treatment group outcomes would be systematically higher regardless of treatment impact.

Differential attrition can also cause systematic differences in $X$ between the treatment and control groups. Including covariates in a regression adjustment can help redress these imbalances. Regression adjustment first estimates expected outcomes conditional on covariates, and then averages these across the shared distribution of covariates to get adjusted estimates for the treatment and control groups. Consider:

$$f_t(x) = E[Y(t)|A(t) = 0, X = x]$$

Our adjusted estimate is then:

$$\tilde{\tau}_R^X = E[f_1(X)] - E[f_0(X)]$$

$$= E\big[E[Y(1)|A(1) = 0, X = x]\big] - E\big[E[Y(0)|A(1) = 0, X = x]\big]$$

Our remaining attrition bias after adjustment is then:

$$\beta^X = \tau_{Full} - \tilde{\tau}_R^X$$

Importantly, regression adjustment can only correct for imbalance between the treatment and control groups in terms of *observed* covariates (and generally assumes a linear relationship between covariates and outcomes). If the treatment responders and control responders are different in unobserved ways correlated with outcomes, the evaluation will suffer from differential attrition bias. Regression adjustment also cannot repair any ‘generalizability attrition bias’, i.e. bias due to our evaluation sample not being representative of our full sample.

## Section 3: Data and interventions

Our analysis relies on a unique set of linked databases in England. The key data source is an archive of RCTs maintained by the Education Endowment Foundation (EEF). The crucial feature of the archive is that it can be linked to the National Pupil Database (NPD), a census of publicly-funded schools and pupils that represents over 90 percent of English school children (DfE, 2015). The NPD contains standardized achievement measures at multiple grade levels, along with information about student demographic characteristics. The outcomes and covariates available for analysis are summarized below.

Table 1 – Overview of covariates

| | | |
|---|---|---|
| Student achievement (national assessments) | Grade 2 (age 7); end of “Key Stage 1” | Maths, Reading |
| | Grade 6 (age 11); end of “Key Stage 2” | Maths, Reading, Writing |
| | Grade 11 (age 16); end of “Key Stage 4” | Maths, English, Science |
| Student demographics | Age | Months of age |
| | FSM | Free-school-meal status |
| | Female | Binary indicator of gender |

Notes: this table describes the outcomes and covariates available for our analyses. Data are described in NPD (2015).

These data provide an unusual opportunity to study attrition bias. Because the RCT archive is linked to a census that contains achievement data, we can examine post-randomization outcomes data for students who would normally be lost to research.

The outcomes of the original RCTs were researcher-administered achievement tests in literacy, mathematics and science. These outcomes were naturally subject to attrition. For each reported outcome we find an analogous outcome in the NPD. For example, the “Shared Maths” RCT used two maths modules of the Interactive Computerised Assessment System as the primary outcome (Lloyd et al., 2015); to estimate attrition bias, we use the total marks on the Key Stage 2 maths assessment (NPD, 2015).

We sought NPD tests that were administered as soon as possible after RCT intervention had finished. For many of the RCTs, there was a short delay between the original outcome measure and the NPD outcome. Across our ten RCTs, the median delay was seven months.

Finally, we note that despite the possible differences in the timing and content of the tests, there were strong correlations between the original evaluation outcomes and the outcomes we use in our attrition analyses. The mean correlation across 22 outcomes was $\bar{\rho}$ =0.72. This value is attenuated by measurement error and would be strictly less than one even if the tests measured exactly the same domain at the same time.

### 3.2 Interventions

The 10 interventions we analyse represent all the available randomized trials from the EEF archive that met two criteria: a) the original outcome was subject to attrition; and b) pupils subsequently sat a standardized national achievement test.

The interventions were quite diverse. While the overarching purpose of all 10 interventions was to raise academic achievement, programs pursued this mission in a variety of ways. Some were directly focused at achievement outcomes and targeted at low-achieving pupils. For example, the "LIT programme" provided struggling readers in grade 7 with small-group instruction for 3-4 hours each week over a period of 8 months. Other interventions were less direct. "Act, Sing, Play", for example, sought to raise achievement of whole classes by running music workshops for children in grade 2. There was also diversity in the ages of the pupils who participated in the 10 interventions, ranging from 6 to 15 years. Table 2 summarizes the interventions.

Table 2 – Summary of interventions

| Intervention | Brief description of intervention | Pupils | Attrition† | Outcomes (ES‡) | Reference |
|---|---|---|---|---|---|
| Act, Sing, Play | Music and drama workshops, in groups of 10, for students in grade 2, once a week for 32 weeks | 894 | 7.8 | Maths (0.00$\sigma$)<br>English (0.03$\sigma$) | Haywood et al. (2015) |
| Changing Mindsets (INSET) | Professional development course for primary school teachers in how to develop Growth Mindset in pupils | 1,035 | 10.8 | Maths (0.01$\sigma$)<br>English (-0.11$\sigma$) | Rienzo et al. (2015) |
| Changing Mindsets (Pupil) | 6-week course of mentoring and workshops for grade 5 students, with a focus on developing pupils' growth mindset | 195 | 8.2 | Maths (0.10$\sigma$)<br>English (0.18$\sigma$) | Rienzo et al. (2015) |
| Dialogic Teaching | Grade 5 teachers were trained in techniques to encourage dialogue, argument and oral explanation during class time | 4,918 | 21.4 | Maths (0.09$\sigma$)<br>English (0.15$\sigma$) | Jay et al. (2017) |
| LIT programme | Targeted literacy intervention for struggling readers in grade 7 using 'reciprocal teaching', for 3-4 hours per week for 8 months. | 5,286 | 19.0 | English (0.09$\sigma$) | Crawford & Skipp (2014) |
| Mind the Gap | Teacher training and parent workshops, over a 5 week period, to help grade 4 students be more 'meta-cognitive'. | 1,496 | 60.1 | Maths and Reading (-0.14$\sigma$) | Dorsett et al. (2014) |
| ReflectEd | Grade 5 pupils have weekly lessons, over a 6 month period, focused on strategies to monitor/manage their own learning | 1,843 | 15.4 | Maths (0.30$\sigma$)<br>Reading (-0.15$\sigma$) | Motteram et al. (2016) |
| Shared Maths | Cross-age peer math tutoring: older pupils (grade 6) work with younger ones (grade 4) for 20 mins per week for 2 years. | 3,119 | 14.1 | Maths (0.02$\sigma$)<br>Reading (§) | Lloyd et al. (2015) |
| Talk of the Town | Whole-school intervention to help support the development of children's speech, language and communication. | 1,512 | 14.8 | Reading (-0.03$\sigma$)<br>Maths (§) | Thurston et al. (2016) |
| Texting Parents | Parents of secondary school pupils sent text messages about homework, upcoming tests etc., over 11 months | 5,026 | 14.4 | English (0.03$\sigma$)<br>Maths (0.07$\sigma$)<br>Science (-0.01$\sigma$) | Miller et al. (2016) |

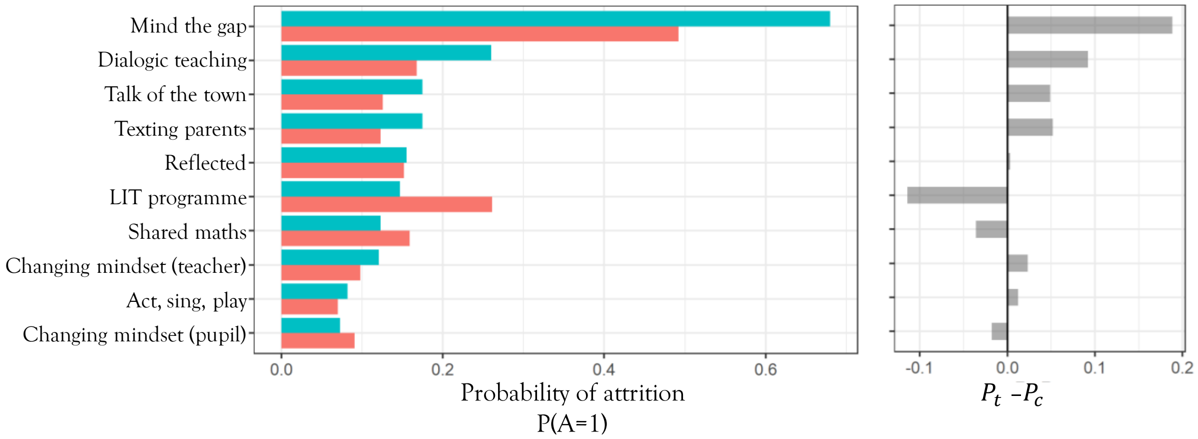
Notes: this table summarizes the 10 RCTs analysed in this paper; n=number of pupils, at randomization, with valid pupil identifiers in the National Pupil Database. †Pupil level attrition rate. ‡Effect size. §In Shared Maths, reading results were not reported due to missing data; for Talk of the Town, KS2 maths was a tertiary outcome, not reported in the original trial.

### 3.3 Attrition in the RCTs

Attrition probability varied across studies and treatment arms, with a mean rate of 19% (see Figure 2). This is typical of a broader set of 79 education RCTs in the UK, funded by the EEF, which also had a mean attrition rate of 19% (Demack et al., forthcoming).

An examination of the 10 original evaluation reports reveals that attrition was a prominent concern in the minds of the researchers. Each of the RCTs include a 'padlock rating', provided by EEF peer reviewers. This is intended as a measure of overall study quality. These ratings are based on five criteria: design, power, attrition, balance, and other threats to validity. The lowest score across these criteria defines the final rating. In five out of the 10 evaluations attrition was cited as the limiting factor. Two further trials listed 'imbalance' as the limiting factor, with the authors in both cases noting the role of differential attrition in creating imbalance. It is worth re-iterating that these 10 studies were not chosen because they were disproportionately affected by attrition. Rather, the prominence of attrition as a concern is a stark reminder that, from the point of view of researchers, attrition is arguably the biggest threat to generating unbiased experimental results (Deke & Chiang, 2017; Greenberg & Barnow, 2014).

Figure 2 – probability of attrition in RCTs



Notes: This figure summarizes the probability of student attrition across the 10 RCTs analysed in this paper. The left panel shows the raw rates of attrition in the treatment (blue) and control (red) arms. The right panel illustrates the difference between the rates of treatment attrition and control attrition.

# Section 4: Estimates of Attrition Bias

## 4.1 Estimating Attrition Bias

Our definition of attrition bias lends itself to a simple estimation procedure. Consider Model 1, in which $Y_{ijkw}$ is the outcome that student $i$ in school $j$ achieves in intervention $w$ for outcome $k$, and $T_{ij}$ is a binary treatment indicator

$$\begin{gathered} Y_{ijkw} = \alpha_j + \tau T_{ij} + e_{ij} \qquad (Model\ 1) \\ \alpha_j \sim N(\alpha_0, \sigma_\alpha^2) \\ e_{ij} \sim N(0, \sigma^2) \end{gathered}$$

For each intervention and outcome, we fit this model to two samples: the “responder sample” (students who provided data for the initial evaluation) and the “full sample” (all units, with valid pupil identifiers, who were recorded as being randomized). Note that in Model 1 the SATE estimand depends on the sample used in fitting the model. For example, when we fit Model 1 using the full evaluation sample, $\tau$ is defined as $\tau_{FULL}$. As such, generating an estimate of attrition bias for a particular study-outcome pair is a straightforward three step process:

(i) Fit Model 1 only using units from the ‘responder’ sample (all $i$ s.t. $A_i = 0$). As we are using the responder sample, the estimate of $\tau$ will be $\hat{\tau}_R$

(ii) Refit Model 1 using the full sample ($all\ i$). Here, the estimate of $\tau$ will be $\hat{\tau}_{FULL}$

(iii) Take the difference: $\hat{\beta} = \hat{\tau}_{FULL} - \hat{\tau}_R$

## 4.2 Attrition bias after covariate adjustment

$\beta^X$ measures attrition bias *after* using a linear model to condition on observed covariates $X$. Conditioning on X using a model aims to address any imbalance in observed characteristics due to attrition (or chance treatment assignment). Estimation of $\beta^X$ involves the same three-step process as estimation of $\beta$. The only change from Model 1 is the inclusion of covariates in the estimation model, represented by $\boldsymbol{X_{ij}}$:

$$\begin{gathered} Y_{ijkw} = \alpha_j + \boldsymbol{\delta X_{ij}} + \tau T_{ij} + e_{ij} \qquad (Model\ 2) \\ \alpha_j \sim N(\alpha_0, \sigma_\alpha^2) \\ e_{ij} \sim N(0, \sigma^2) \end{gathered}$$

There are two reasons to examine attrition bias in the context of covariate-adjusted impact estimates. First, we are interested in the magnitude of attrition bias in practice. As applied researchers generally adjust for covariate imbalance – and all 10 evaluations considered here adjusted for covariates – we follow this convention. Second, by estimating attrition bias both with and without covariate adjustment, we are able to examine the extent to which condition on covariates repairs attrition bias. As a final note, estimates of attrition bias may also benefit from precision gains if variation in outcomes is captured by covariates. This could improve our ability to detect attrition bias even when there is no differential attrition.

### 4.3 Estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\beta^X}$

Figure 3 presents initial estimates of $\beta$ and $\beta^X$. The plot also presents 95% confidence intervals for $\hat{\beta}^X$. These uncertainty estimates are based on a set of simulations described in Appendix A. We use simulation-based inference rather than conventional standard errors to allow for the dependence structure of $\beta^X$ across studies and outcomes.

There are two things to note about Figure 3. First, it appears as though conditioning on covariates lessens attrition bias. The estimates of $\hat{\beta}^X$ tend to be closer to zero than the $\hat{\beta}$ estimates. Second, 20 out of the 22 estimates of $\hat{\beta}^X$ have confidence intervals that include zero.

Next, we analyse the *distribution* of attrition bias across interventions and outcomes. The boxplots at the bottom of Figure 3 are a useful starting point in this endeavour. However, these data are over-dispersed due to measurement error. This may create a misleading impression about the typical magnitude of attrition bias. To see why the distributions underlying the boxplots are over-dispersed, consider an attrition mechanism $A_{null}$ that is completely random: $A_{null} \perp Y(1), Y(0), X$. If this mechanism were responsible for deleting data from each of our 10 evaluation samples, estimates of attrition bias would be non-zero even though no bias had been introduced. In other words, the raw estimates of bias presented in the boxplots include both underlying attrition bias and ‘attrition sampling variation’ – that is, uncertainty about which units will leave the study.

We account for this error using tools from meta-analysis. This approach addresses two overlapping goals: to present estimated distributions for $\beta$ and $\beta^X$ that are not over-dispersed due to sampling variation, and to estimate the typical value of underlying attrition bias for our setting. The aim in both cases is to help education researchers and funders understand the typical magnitude of attrition bias in typical RCTs.

Observed attrition bias estimates $\hat{\beta}_{kw}$ are assumed to be made up of several components:

$$\hat{\beta}_{kw}|\beta_{kw} \sim N(\beta_{kw}, \sigma^2_{kw})$$
$$\beta_{kw} \sim N(\nu, \eta^2)$$

Where:

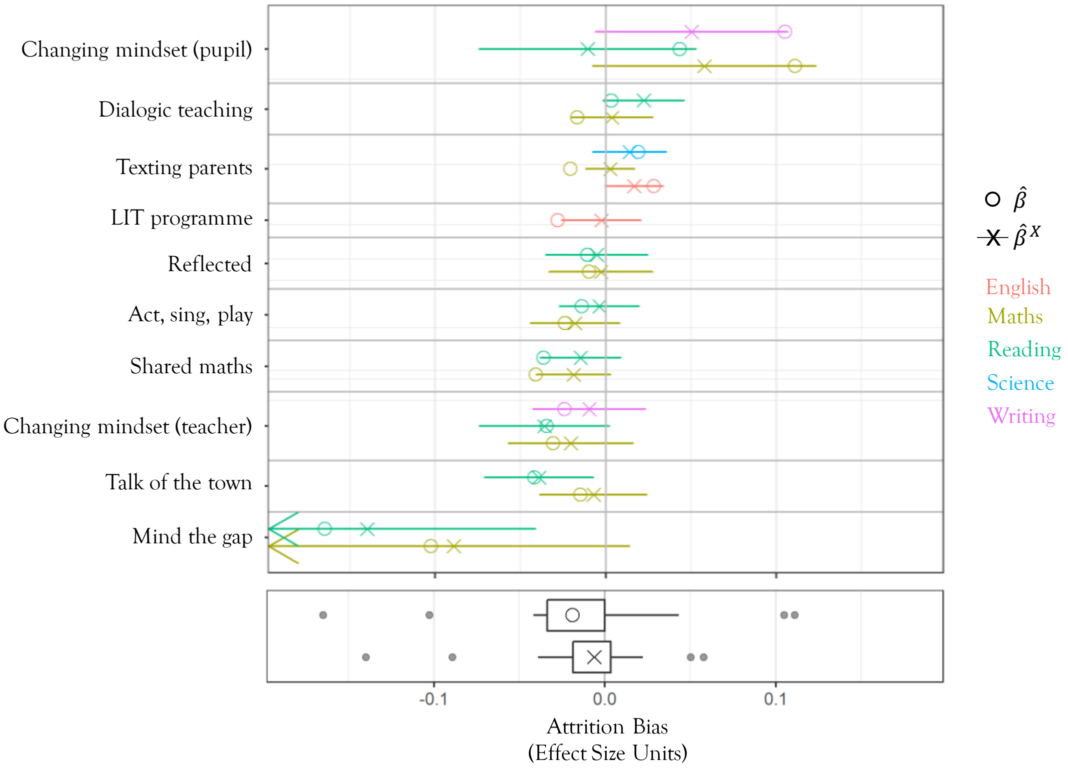
(i) $\nu =$ the mean attrition bias across all interventions and outcomes

(ii) $\beta_{kw} =$ the underlying attrition bias for outcome $k$ in intervention $w$. This has a variance of $\eta^2$ reflecting the fact that not all interventions will have the same attrition bias. $\beta_{kw}$ could change due to context, the nature of the treatment, the outcome, and so on.

(iii) $\hat{\beta}_{kw} =$ observed bias. This deviates from underlying bias $\beta_{kw}$ with a variance of $\sigma^2_{kw}$, which is largely determined by the level of attrition.

Appendix B provides details of our approach to estimating these parameters, which draws heavily on random effects meta-analysis (Higgins et al., 2009). For each intervention-outcome pair we calculate a constrained empirical Bayes’ estimate of attrition bias, $\tilde{\beta}_{kw}$ (Weiss et al., 2017). The estimated distributions of $\tilde{\beta}$ and $\tilde{\beta}^X$ are presented in Figure 4.

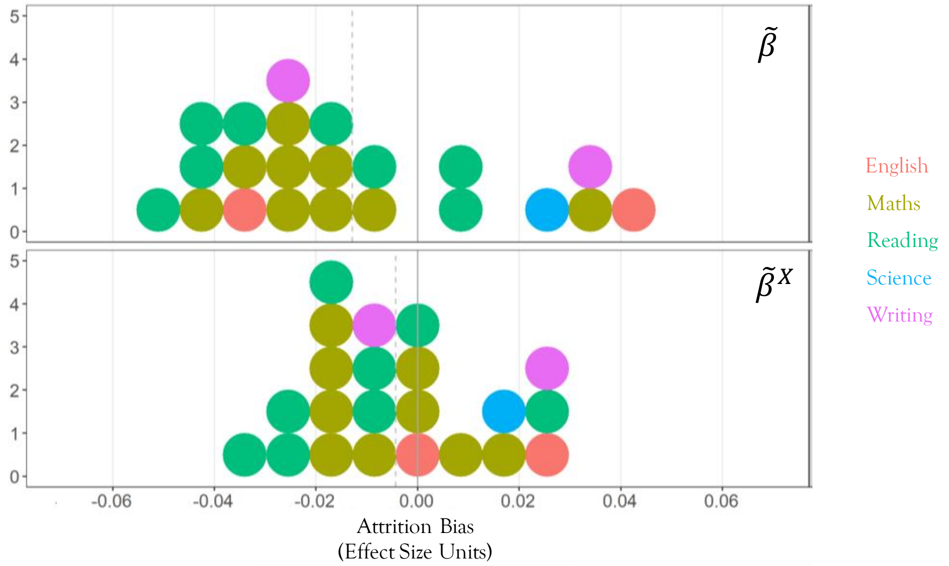
Figure 3 – Estimates of attrition bias, before conditioning ($\hat{\beta}$) and after ($\hat{\beta}^X$)



Note: this figure represents initial point estimates of attrition bias before conditioning on covariates ($\hat{\beta}$) and afterwards ($\hat{\beta}^X$). Estimates of $\hat{\beta}^X$ have 95% CIs, derived from simulations described in Appendix A.

The top panel of Figure 4 represents our best guess at the distribution of attrition bias in cases where researchers do not adjust for observed differences between treated and control units. The mean of the estimated $\tilde{\beta}$ distribution is $\hat{\nu} = -0.013\sigma$, with mean absolute value of $0.026\sigma$. Where we control for covariates, including a pre-test, the estimated distribution of $\tilde{\beta}^X$ has a mean of $\hat{\nu} = -0.004\sigma$, and a mean absolute value of $0.015\sigma$. No values of $\tilde{\beta}^X$ have a magnitude greater than $0.034\sigma$. This suggests that, in practice, the typical magnitude of attrition bias is small, particularly when researchers have access to predicative covariates.

Figure 4 – Final estimates of underlying attrition bias ($\tilde{\beta}, \tilde{\beta}^X$)



Notes: the top panel shows estimates of constrained empirical Bayes estimates of $\tilde{\beta}$ for 10 studies and 22 outcomes. The bottom panel is the equivalent plot after conditioning on covariates $\tilde{\beta}^X$. The estimated mean is shown with a grey dotted line.

### 4.4 Contextualising attrition bias

To put the magnitude of these attrition bias estimates into context, we offer three points of reference. First, note that the What Works Clearinghouse set a threshold for problematic bias at $0.05\sigma$ (WWC, 2014). The EEF takes a similar position, and views $0.05\sigma$ as a threshold beyond which bias becomes a substantial concern. None of the estimates of attrition bias presented above are greater than this threshold – including the estimates that do not condition on covariates.

Second, a recent meta-analysis of similar interventions found that the typical value of selection bias due to non-random assignment was $0.15\sigma$ without covariates, and $0.03\sigma$ after controlling for observable characteristics (Weidmann and Miratrix, forthcoming). This study used a set of 14 interventions that are very similar to those studied here. The findings show that, in the absence of predictive covariates, typical selection bias due to non-random assignment is roughly six times larger than typical attrition bias. When researchers have access to predictive covariates, typical selection bias is roughly twice as large as typical attrition bias.

Third, we draw readers attention to initial estimates of external validity bias due to non-random sampling. To our knowledge such estimates only exist for one program (Reading First), and suggest that the mean bias due to non-randomly selected study samples is $0.1\sigma$ (Bell et al., 2016). Given the scarcity of evidence, we draw no firm conclusions. However, if an evaluation has the policy-relevant goal of estimating a population average treatment effect, it is plausible that the risk of external validity bias overshadows internal-validity risks. In light of this, we argue that it is an urgent priority to develop a stronger understanding of the risk of external validity bias.

## Section 5: Decomposing and understanding attrition mechanisms

### 5.1 Decomposing the elements of attrition bias

Section 2 shows that attrition bias can be viewed as a function of four elements: the rate of attrition in each arm ($P_t, P_c$) and the extent to which attrition is associated with outcomes ($\Delta_t, \Delta_c$). Here, we explore the estimated magnitude of $\Delta$ parameters, beginning with $\Delta_T$ and $\Delta_C$:

$$\Delta_T = E[Y(1)|A(1) = 1] - E[Y(1)|A(1) = 0]$$

$$\Delta_C = E[Y(0)|A(1) = 1] - E[Y(0)|A(1) = 0]$$

These parameters can be estimated using a model that accounts for the clustering of students ($i$) within schools ($j$):

$$Y_{ij} = \alpha_j + \gamma A_{ij} + \lambda_1 T_{ij} + \delta A_{ij} T_{ij} + e_{ij} \qquad (Model\ 3)$$
$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$
$$e_{ij} \sim N(0, \sigma^2)$$

With this setup $\widehat{\Delta}_C = \hat{\gamma}$, and $\widehat{\Delta}_T = \hat{\gamma} + \hat{\delta}$. Values of $\Delta_a$ closer to zero suggest that the attrition mechanism in treatment arm $a$ is less associated with outcomes.

Because education researchers frequently have access to predictive covariates – as we do for all 10 of the randomized experiments in our dataset – we also consider versions of the $\Delta$ parameters that condition on covariates ($\Delta_C^X$ and $\Delta_T^X$). To estimate these quantities we fit Model 4, which adds covariates and their interactions to Model 3:

$$Y_{ij} = \alpha_j + \gamma_2 A_{ij} + \lambda_2 T_{ij} + \delta_2 A_{ij} T_j + \boldsymbol{\nu_C X_{ij}} + \boldsymbol{\nu_T X_{ij}} T_{ij} + e_{ij} \qquad (Model\ 4)$$
$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$
$$e_{ij} \sim N(0, \sigma^2)$$

In Model 4, $\widehat{\Delta}_C^X = \hat{\gamma}_2$. The parameter $\Delta_C^X$ represents the mean difference in outcomes between 'control responders' and 'control attriters' who have the same covariate values. $\widehat{\Delta}_T^X = \hat{\gamma}_2 + \hat{\delta}_2$, and is the equivalent parameter on the treatment side. While estimates of these parameters contain useful information about the nature of attrition mechanisms, our primary focus is on the distribution of the $\Delta$ parameters across studies and outcomes. To that end, note that the issue of over-dispersal, discussed above with reference to $\hat{\beta}$ and $\hat{\beta}^X$, also applies to the $\widehat{\Delta}$ estimates. We therefore rely on the same meta-analytic tools to model the underlying distributions of the $\Delta$ parameters.

As an example of our modelling approach consider $\widehat{\Delta}_{C,kw}$: the difference in the average outcomes between control attriters and control responders, for outcome $k$ in intervention $w$. We model this as follows:

$$\widehat{\Delta}_{C,kw}|\Delta_{C,kw} \sim N(\Delta_{C,kw}, \psi_{kw}^2)$$

$$\Delta_{C,kw} \sim N(\phi_C, \theta_C^2)$$

Where:

(i) $\phi_C$ = the mean difference, across all studies, in the average outcomes of ‘control attriters’ and ‘control responders’.

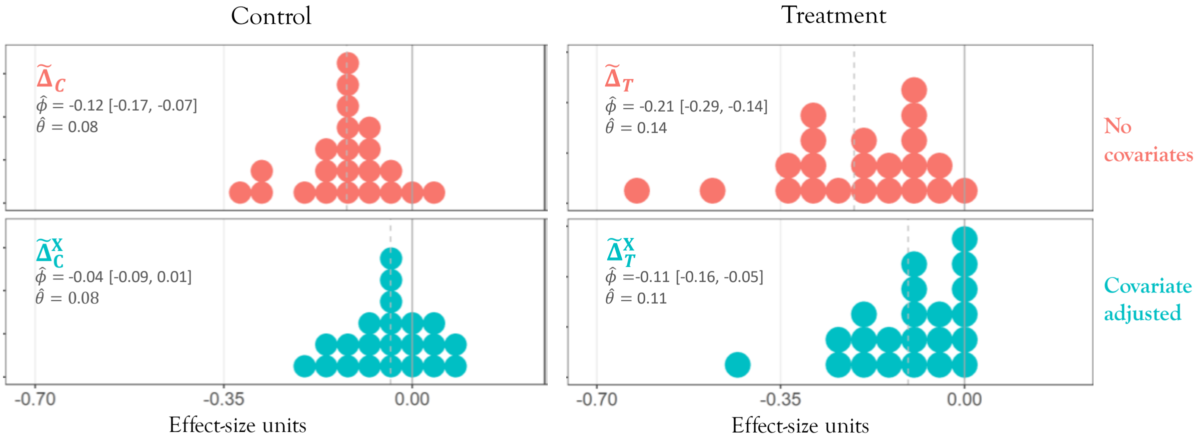
(ii) $\Delta_{C,kw}$ = the underlying difference, for intervention $w$ and outcome $k$, between the mean outcome of ‘control attriters’ and ‘control responders’. This has a variance of $\theta_C^2$ across intervention-outcome pairs.

(iii) $\widehat{\Delta}_{C,kw}$ = the observed mean difference in the average outcome of ‘control attriters’ and ‘control responders’. This has sampling variance of $\psi_{kw}^2$.

For analyses in which we condition on covariates, equivalent parameters have an X superscript. For example, $\Delta_{T,kw}^X$ is the difference in mean outcomes for ‘treatment attriters’ and ‘treatment responders’, after conditioning on covariates with a linear model.

We conduct four separate meta-analyses, one each for $\Delta_C, \Delta_T, \Delta_{\mathrm{C}}^{\mathrm{X}}$ and $\Delta_T^X$. Once again we compute a set of 22 constrained empirical Bayes estimates for each parameter: $\widetilde{\Delta}_C, \widetilde{\Delta}_T, \widetilde{\Delta}_{\mathrm{C}}^{\mathrm{X}}$ and $\widetilde{\Delta}_T^X$ (Weiss et al., 2017). The results, along with core parameter estimates from the meta-analyses, are presented in Figure 5.

Figure 5 – Estimates of $\widetilde{\Delta}$ distributions



Note: the estimation approach for the meta-analyses is described in Appendix B. $\hat{\phi}$ represents the estimated mean of each distribution; $\hat{\theta}$ is the estimated standard deviation. Due to the small number of interventions we could not use the estimated variance of $\hat{\theta}^2$ as the basis of a confidence interval (Higgins et al., 2009).

Examination of these plots emphasizes two features of our data. First, conditioning on covariates substantially reduces the perniciousness of attrition mechanisms. The distributions of $\widetilde{\Delta}_C^X$ and $\widetilde{\Delta}_T^X$ are centered much closer to zero than their unadjusted counterparts. Second, it appears as though there is a systematic relationship between attrition and outcomes. Units who leave studies appear to have worse outcomes than responders, even after conditioning on covariates. This is particularly true on the treatment side. Almost all the estimates of $\widetilde{\Delta}_T$ and $\widetilde{\Delta}_T^X$ are negative, or zero (<0.005$\sigma$). A similar effect is present on the control side, but it much less pronounced.

To emphasize the potential difference between the treatment and control side, we perform a non-parametric test of the hypothesis that the mean value of $\widetilde{\Delta}$ is the same for both arms, $H_0$: $\phi_C = \phi_T$. To generate a draw under the null, we permute treatment status within each study-outcome pair and calculate mean($\widetilde{\Delta}_C$)− mean($\widetilde{\Delta}_T$). We compare the observed value of mean($\widetilde{\Delta}_C$)− mean($\widetilde{\Delta}_T$) with 10000 draws under the null, and find evidence against the hypothesis that the means are equivalent (p=0.01). We then test the analogous hypothesis for $\widetilde{\Delta}^X$ and similarly find that the difference between the means is significant (p=0.008). In both cases – with and without conditioning on covariates – attrition appears to be more problematic on the treatment side.

The meta-analyses underlying Figure 5 can also be viewed as tests of two common assumptions: "Missing At Random" (MAR) and "Missing Completely At Random" (MCAR). Specifically, if MCAR holds across our studies then the distributions of $\Delta_T$ and $\Delta_C$ will have a mean of zero. This is clearly not the case. $\hat{\phi}_T$ and $\hat{\phi}_C$ are negative and their 95% confidence intervals do not include zero (both p-values are < 0.001).

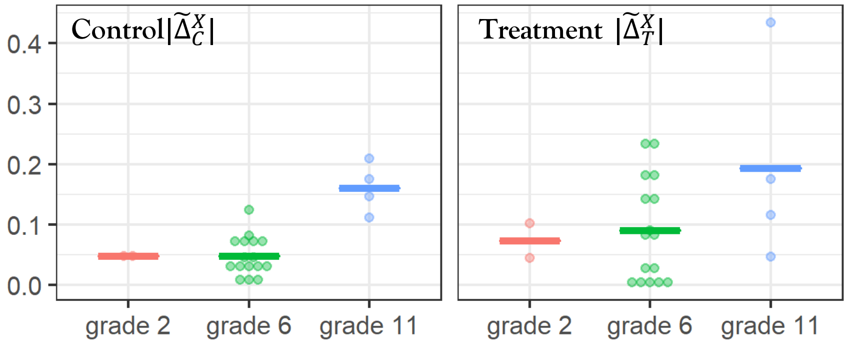
There is also some evidence that attrition mechanisms do not meet the MAR assumption. In particular, note that $\hat{\phi}_T^X = -0.11$ and is significantly different from zero (p=0.002). The picture on the control side is slightly less clear. After adjusting for covariates, the mean difference between attriters and responders is negative ($\hat{\phi}_C^X = -0.04$) but not significantly different from zero ($p = 0.11$). Overall, however, these results cast substantial doubt on MAR being a plausible assumption in our context.

We see further evidence against the MAR assumption when we examine the 22 individual study-outcome pairs in our data. For each pair we use a likelihood ratio test to examine the hypothesis that $\delta = \gamma = 0$. When applied to Model 4, this hypothesis test is a test of MAR. After correcting for multiple comparisons (Hochberg, 1988) we find 5 out of 22 reject MAR. Full results of these hypothesis tests are presented in Appendix C. In sum, these findings reinforce the conclusions of the distributional analyses and suggest that, in our context, researchers cannot safely assume that attrition mechanisms will meet the MAR assumption.

### 5.2 What predicts pernicious attrition?

Are there situations in which problematic attrition mechanisms are more likely? Here we examine three possible predictors of non-random attrition. First: school grade. Seven of the 10 trials focused on outcomes from grade six (age 11-12) which limits our ability to draw conclusions about the effect of grade on attrition mechanisms. However, we note that two of the trials with the largest values of $\widetilde{\Delta}^X$ were in grade 11 (see Figure 6). Moreover, examination of the MAR tests at the study-outcome level shows that four of the five rejections of the MAR hypothesis came from the two trials with outcomes at grade 11. As we only have two cases, we are tentative in drawing conclusions. However, we suggest that researchers working with outcomes for later grades should be especially wary of attrition.

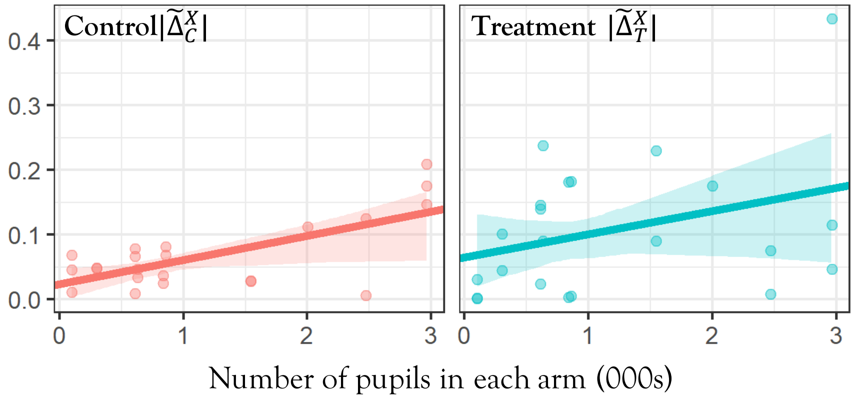
Figure 6 – $|\widetilde{\Delta}^X|$ by grade



Notes: estimates of $|\widetilde{\Delta}^X|$ (the magnitude of the association between attrition and outcomes, conditional on covariates) by grade. Each panel shows results from 10 RCTs and 22 outcomes. The left panel presents estimates from control students, the right panel from treated students.

Second, we examine whether there is any association between the size of a study and the nature of attrition. We find a positive association between $|\widetilde{\Delta}^X|$ and trial size:

Figure 7 – $|\widetilde{\Delta}^X|$ by number of pupils



Notes: $|\widetilde{\Delta}^X|$ is the magnitude of the association between attrition and outcomes, conditional on covariates. Each panel shows results from 10 RCTs and 22 outcomes. The left panel presents estimates from the control side, the right panel from the treatment side. 95% confidence intervals for the predicted linear relationship are based on bootstrapped samples.

A simple regression analysis found that an increase of 1,000 students in a trial arm was associated with an increase of $0.04\sigma$ in $\widetilde{\Delta}^X$ on both the control and treatment side. The association is statistically different from zero on the control size (p<0.001) but not on the treatment size (p=0.102; see Appendix D for details). This suggests that there may be a positive association between bigger trials and bias-inducing attrition mechanisms.

Finally, we tested the hypothesis that the rate of attrition is associated with its perniciousness. We find no evidence that attrition mechanisms are more problematic in cases where attrition is high. The correlation between $P_a$ and $\widetilde{\Delta}_a$ was not significantly different from zero for treated or control arms.

## Section 6: recommendations for applied researchers

Evidence presented in section 5 suggests that, in our context, outcome data cannot safely be assumed to be Missing At Random, let alone Missing Completely At Random. We argue that researchers should respond in three ways. First, adjust for covariate differences in estimating treatment effects. This is already common practice, so we do not discuss it further. Second, incorporate ‘attrition bias uncertainty’ into inferences. Third, perform sensitivity analyses to see whether core findings are robust to the types of attrition mechanisms we observe in practice. This section describes the latter two recommendations, culminating in a worked example using the evaluation of *REACH* program (Sibieta, 2016).

### 6.1 Incorporating attrition bias into uncertainty estimates

The threat of attrition bias is a source of uncertainty in estimating the Sample Average Treatment Effect (SATE) on the full sample ($\tau_{FULL}$). This uncertainty should be incorporated into inferences. There are multiple possible approaches, including a fully Bayesian analysis in which the attrition mechanism in each treatment arm is explicitly modelled. For education researchers pursuing this strategy, the results we report here represent a strong starting point for priors. However, in keeping with a frequentist framework, we propose augmenting conventional standard errors by the expected magnitude of attrition. The result is an inflated “rule of thumb standard error” for the SATE. This adjustment can be done on any study, given an initial standard error and attrition rates for treated and control units.

The total error in our estimate is the sum of estimation error and any attrition bias induced by attrition and our estimation process:

$$\hat{\tilde{\tau}}_R = \tau_{FULL} + \beta + e$$

In the above, e is our estimation error when we take $\hat{\tilde{\tau}}_R$ as an estimate of $\tilde{\tau}_R$; we assume it is unbiased with $E[\,\hat{\tilde{\tau}}_R] = \tilde{\tau}_R$. To include uncertainty from attrition, we further assume that attrition bias is an unknown random variable independent from estimation error (i.e., we make the simplifying assumption that $cov(e, \beta) = 0$). This gives:

$$E\left[\left(\hat{\hat{\tau}}_R - \tau_{FULL}\right)^2\right] = E[\beta^2] + var(e)$$

We write E[$\beta^2$] because attrition is not necessarily 0 centered. The above is a decomposition of the overall error. The second term is the typical standard error (squared):

$$e \sim N(0, SE^2).$$

We can obtain a working value for $E[\beta^2]$, tailored to a new study $s$, based on observed levels of attrition in the treatment and control arms of $s$. Using (6) we have the following expression for $\eta^2$, the variance of attrition bias across outcomes:

$$\eta^2 = E[\beta^2] = E[(P_T\Delta_T - P_C\Delta_C)^2]$$

$$= P_T^2 E[\Delta_\mathrm{T}^2] + P_C^2 E[\Delta_\mathrm{C}^2] - 2P_T P_C E[\Delta_\mathrm{T}\Delta_\mathrm{C}]$$

To get the above, we treat $P_T$ and $P_C$ as constants, as they are directly observed. We have relatively little empirical information about $E[\Delta_T\Delta_C]$. The information we do have suggests that this term is positive, because estimates of both $\Delta_T$ and $\Delta_C$ tend to be less than zero (see Appendix E for discussion). That said, we argue for dropping this cross term. This is conservative if $\Delta_T$ and $\Delta_C$ have the same sign. For study $s$ this gives us:

$$\eta_s^2 = P_{\mathrm{T},s}^2 \mathrm{E}[\Delta_\mathrm{T}^2] + P_{\mathrm{C},s}^2 \mathrm{E}[\Delta_\mathrm{C}^2]$$

$\eta_s^2$ can then be estimated based on observed attrition ($P_{T,s}, P_{C,s}$) and estimates of the squared magnitude of $\Delta_C$ and $\Delta_T$:

$$\hat{\eta}_s^2 = P_{T,s}^2 \overline{\Delta_T^2} + P_{C,s}^2 \overline{\Delta_C^2}$$

Where $\overline{\Delta_a^2} = \hat{\theta}_a^2 + \hat{\phi}_a^2$ for arm $a$. Estimates for these parameters – with and without conditioning on covariates – are presented in Table 3.

This, in turn, leads to a revised uncertainty estimate, tailored to the observed level of attrition:

$$\widehat{SE}_{revised,s} = \sqrt{\hat{\eta}_s^2 + \widehat{SE}_s^2}$$

Table 3 – Summary of attrition parameters

| | | $\overline{\Delta^2}$ $(magnitude^2)$ | $\widehat{\theta^2}$ $(variance)$ | $\hat{\phi}$ $(mean)$ |
|---|---|---|---|---|
| No covariates | Control | 0.021 | 0.007 | -0.122 |
| | Treatment | 0.065 | 0.021 | -0.210 |
| Covariate adjusted | Control | 0.008 | 0.006 | -0.040 |
| | Treatment | 0.023 | 0.011 | -0.107 |

Notes: this table represents estimates of important attrition bias parameters from our sample of 10 RCTs and 22 outcomes. "No covariates" indicates that attrition bias parameters have been estimated without controlling for any observed characteristics. "Covariate adjusted" estimates condition on observed characteristics. For model equations and parameter definitions, see Section 5.1.

Our attrition-adjusted errors will be strictly larger than the unadjusted standard errors; we have forced any systematic bias into a variance term which protects us from making a priori assumptions about the direction of the bias. Our standard error should now be interpreted as an estimate of the magnitude of our overall error (it is in fact closer to an estimate of the expected RMSE); corresponding confidence intervals take the attrition bias as an unknown, zero-centered variable with a magnitude that depends on proportion of units attrited and evidence from prior studies about the link between attrition and outcomes (for an argument to interpret SEs as an estimate of error in this way, see Sundberg (2003)). Studies with lower rates of attrition will have smaller adjustments. This adjusted standard error does not take into account any "cancelling out" effect we might see due to similar attrition in both arms of a study; we view not doing so as a more cautious, conservative approach.

## 6.2 Sensitivity analyses, based on observed attrition mechanisms

Randomized experiments frequently include sensitivity analyses to assess whether results are robust to attrition. This is often done in terms of missing-data imputation. However, given the possibility that MAR is violated, there is a strong argument for going further, and exploring the potential influence of attrition bias due to unobserved characteristics.

This in itself is not a novel idea and various approaches to bounding unobserved biases have been proposed (e.g. Manski, 1990). The difficulty is that sensitivity analyses often yield very wide ranges that far exceed typical effect sizes in the field and are inconsistent with informed prior expectations. To help resolve this issue, we argue that researchers should ground their sensitivity analyses in estimates of how bias-inducing attrition mechanisms tend to be in practice.

To do this, consider $\hat{\tau}^*$, the estimated SATE, conditional on values for $\Delta_C = \Delta_C^*$ and $\Delta_T = \Delta_T^*$:

$$\hat{\tau}^* = \hat{\tau}_R + P_T\Delta_T^* + P_C\Delta_C^* \quad (7)$$

$$= \hat{\tau}_R + P_T(\Delta_T^* - \Delta_C^*) + \Delta_C^*(P_T - P_C) \quad (8)$$

Nearly all randomized experiments report both impact estimates $(\hat{\tau}_R)$ and attrition rates ($P_T$ and $P_C$). This means that (7) can be widely used to assess how sensitive findings are to attrition. Meanwhile, (8) shows that the sensitivity of an impact estimate depends on two factors: differential attrition *mechanisms* $(\Delta_T - \Delta_C)$ and differential attrition *rates* $(P_T - P_C)$. For any given study, the differential attrition rate is known. Consequently, we recommend that researchers take $P_T$ and $P_C$ as given and focus their sensitivity analyses on the potential influence of differential attrition mechanisms $(\Delta_T - \Delta_C)$.

One way to do this is to generate a simple sensitivity plot with $\tau^*$ on the y-axis and the differential attrition mechanism on the x-axis (either $\Delta_T^* - \Delta_C^*$ or $\Delta_C^* - \Delta_T^*$, depending on whether $\hat{\tau}$ is positive or negative). We suggest that the x-axis span the range from "no differential attrition" to the "worst observed case".

Under "no differential attrition", $\Delta_T^* = \Delta_C^*$. The worst-case estimates of $\Delta^*$ will depend on whether the finding is positive or negative. Consider the case of a positive impact estimate: $\hat{\tau}_R > 0$. This finding will be undermined by positive values of $\Delta_C^* - \Delta_T^*$. A sufficiently large value of $\Delta_C^* - \Delta_T^*$ will send $\tau^*$ zero. To find the "worst observed case" we refer researchers to Table 4, which summarizes our 22 estimates of $\widetilde{\Delta}_C, \widetilde{\Delta}_T, \widetilde{\Delta}_C^X$ and $\widetilde{\Delta}_T^X$. Across our set of 22 outcomes, the worst observed case of attrition for a positive finding is max$(\Delta_C - \Delta_T)$ = 0.350. In the case when we control for covariates this value is max$(\widetilde{\Delta}_C^X - \widetilde{\Delta}_C^X)$ =0.284. Last, to generate the sensitivity analysis, researchers need to select a value of $\Delta_C^*$ (or ${\Delta^*}_C^X$). As a default, we recommend using the median observed value (-0.033). The choice of $\Delta_C^*$ is unlikely to be consequential relative to the impact of differential attrition mechanism $(\Delta_C - \Delta_T)$. However, researchers who are interested in a more comprehensive sensitivity analysis can simply calculate $\tau^*$across a two dimensional grid of $\Delta_T^*$ and $\Delta_C^*$ values. Of course, researchers could choose any parameters they believe are appropriate to include in this grid. But, in keeping with our broader recommendations, we suggest that these values be grounded by attrition mechanisms that have been observed in practice. For example, we suggest that the grid be bounded by the maxima and minima of the relevant $\Delta$ parameters reported in Table 4.

## 6.3 Example: REACH reading intervention

To make these ideas more concrete, we present example analyses using the evaluation of the REACH reading intervention (Sibieta, 2016). This is a randomized experiment in the EEF archive for which the threat of attrition bias is unknown. In this case, the control units were given the treatment straight after the trial concluded as part of a 'wait-list' design. This makes it impossible to use our approach to estimate attrition bias.

Table 4 – Summary values of Δ parameters

| | | | No covariates | | | Covariates | | |
|---|---|---|---|---|---|---|---|---|
| Outcome | Project | Grade | $\widetilde{\Delta}_C$ | $\widetilde{\Delta}_T$ | $\widetilde{\Delta}_C - \widetilde{\Delta}_T$ | $\widetilde{\Delta}_C^X$ | $\widetilde{\Delta}_T^X$ | $\widetilde{\Delta}_C^X - \widetilde{\Delta}_T^X$ |
| Reading | asp | 2 | -0.091 | -0.183 | 0.092 | -0.048 | -0.102 | 0.054 |
| Maths | asp | 2 | -0.027 | -0.110 | 0.083 | 0.049 | -0.045 | 0.093 |
| Reading | cmi | 6 | -0.101 | -0.290 | 0.189 | 0.009 | -0.140 | 0.149 |
| Writing | cmi | 6 | -0.195 | -0.327 | 0.131 | -0.066 | -0.146 | 0.079 |
| Maths | cmi | 6 | -0.051 | -0.153 | 0.102 | 0.079 | -0.024 | 0.103 |
| Reading | cmp | 6 | -0.125 | -0.086 | -0.038 | -0.011 | -0.031 | 0.020 |
| Writing | cmp | 6 | -0.157 | -0.062 | -0.095 | -0.069 | 0.002 | -0.071 |
| Maths | cmp | 6 | -0.145 | -0.049 | -0.096 | -0.045 | 0.001 | -0.047 |
| Reading | dt | 6 | -0.106 | -0.094 | -0.013 | 0.007 | -0.008 | 0.014 |
| Maths | dt | 6 | -0.266 | -0.189 | -0.077 | -0.125 | -0.076 | -0.049 |
| Maths | mtg | 6 | -0.074 | -0.277 | 0.204 | 0.046 | -0.238 | 0.284 |
| Reading | mtg | 6 | -0.070 | -0.118 | 0.048 | 0.034 | -0.091 | 0.125 |
| Maths | ref | 6 | -0.120 | -0.286 | 0.166 | -0.068 | -0.183 | 0.115 |
| Reading | ref | 6 | -0.010 | -0.160 | 0.151 | 0.081 | -0.005 | 0.086 |
| Maths | sm | 6 | -0.139 | -0.467 | 0.328 | -0.029 | -0.230 | 0.202 |
| Reading | sm | 6 | -0.128 | -0.347 | 0.219 | -0.028 | -0.090 | 0.063 |
| Reading | tott | 6 | -0.102 | -0.267 | 0.165 | -0.037 | -0.181 | 0.145 |
| Maths | tott | 6 | -0.089 | -0.104 | 0.015 | -0.024 | 0.003 | -0.027 |
| English | lit | 11 | 0.023 | -0.213 | 0.236 | -0.112 | -0.175 | 0.064 |
| Science | tp | 11 | -0.263 | -0.613 | 0.350 | -0.175 | -0.434 | 0.259 |
| Maths | tp | 11 | -0.140 | -0.227 | 0.086 | -0.146 | -0.115 | -0.031 |
| English | tp | 11 | -0.302 | 0.000 | -0.301 | -0.209 | -0.046 | -0.163 |
| | | Min | -0.302 | -0.613 | -0.301 | -0.209 | -0.434 | -0.163 |
| | | Median | -0.113 | -0.186 | 0.097 | -0.033 | -0.091 | 0.072 |
| | | Max | 0.023 | 0.000 | 0.350 | 0.081 | 0.003 | 0.284 |

Notes: table presents all Δ values, along with their minimums, medians, and maximums, for attrition bias parameters across 10 RCTs and 22 outcomes. "No covariates" indicates that attrition bias parameters have been estimated without controlling for any observed characteristics. "Covariate adjusted" estimates condition on observed characteristics. In all cells we present constrained empirical Bayes estimates of Δ parameters, as described in section 4.3 and Appendix A. The project acronyms are as follows: asp = Act Sing Play; cmi = Changing Mindsets INSET; dt = Dialogic Teaching; lit = LIT programme; mtg = Mind The Gap; ref = ReflectEd; sm = Shared Maths; tott = Talk of The Town; tp = Texting Parents.

The evaluation reports an estimated treatment effect of $\hat{\hat{\tau}}_R = 0.329$ with a standard error of $\widehat{SE} = 0.099$. Attrition in the two arms of the trial was above average: $P_T = 0.278$ and $P_C = 0.323$. The researchers had access to a pre-test, along with demographic covariates, and conditioned on these variables in estimating the treatment effect. As such, we focus on $\widehat{\overline{|\Delta|}}_T^{X^2}$ and $\widehat{\overline{|\Delta|}}_C^{X^2}$ which describe attrition mechanisms after conditioning on covariates.

First, we extend our uncertainty estimates as follows:

$$\hat{\eta}^2 = P_{T,s}^2 \overline{\Delta_T^2} + P_{C,s}^2 \overline{\Delta_C^2}$$

$$\hat{\eta}^2 = 0.278^2 \cdot 0.023^2 + 0.323^2 \cdot 0.008^2$$

$$\hat{\eta}^2 = 0.00256$$

Then, our adjusted error estimate is:

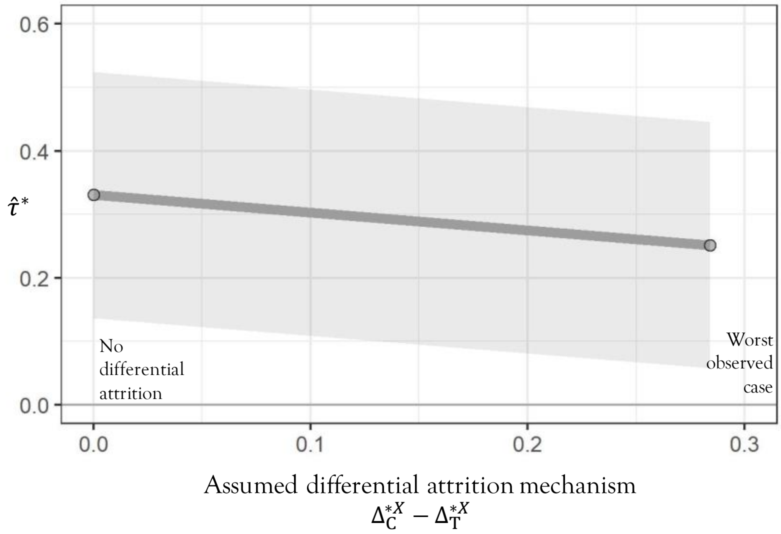
$$\widehat{SE}_{revised} = \sqrt{\widehat{SE}^2 + \hat{\eta}^2}$$

$$= 0.111$$

Generally, adjusting for attrition uncertainty will result in quite modest changes. In this case, the width of the uncertainty interval extended from $0.388\sigma$ to $0.436\sigma$. This is in keeping with empirical findings from section 4.4 that the magnitude of attrition bias in practice is limited, particularly after controlling for predictive covariates.

Turning to sensitivity analysis, note that as $\hat{\hat{\tau}}_R$ is positive, the finding will be overturned for sufficiently large values of $\Delta_C^{*X} - \Delta_T^{*X}$. Using (8), we produce the following plot:

Figure 8 – Example sensitivity analysis for REACH



Note: this figure uses reported attrition rates from the REACH trial and presents sensitivity analyses according to equation (8), setting $\widetilde{\Delta}_C^{X*} = -0.033$.

Under the "worst-observed case" (from the 22 in our sample) the estimated average treatment effect declines from $0.329\sigma$ to $0.251\sigma$. While this is a marked reduction, the point estimate remains positive and the 95% confidence interval does not include zero. Overall, these sensitivity results suggest that the core finding of the REACH evaluation – that the intervention had a positive average treatment effect – is quite robust to attrition bias. This is despite the fact that the RCT suffered from a rate of attrition that, according to the EEF rating scale, would have limited the evaluation to receiving a maximum quality rating of 3 out of 5.

## Section 7: Limitations and conclusions

### Limitations

Our study of attrition bias has two main limitations. First, the census data are themselves subject to some missingness and difficulties with matching. For each RCT in the archive, we searched the National Pupil Database for students who were present at randomization. Across the 10 RCTs, an average of 1.2% of randomized students couldn't be matched to the NPD. A further 2.6% of pupils were missing an outcome measure, which meant they were excluded from our analysis. In total, our analysis included an average of 96.2% of the students who were recorded as being present at randomization. We note that this small level of missingness could bias our estimates of attrition bias $(\tilde{\beta}, \tilde{\beta}^X)$ and the nature of bias mechanisms $(\widetilde{\Delta}, \widetilde{\Delta}^X)$. That said, we found no association between the level of missingness in the National Pupil Database and any of the four aforementioned sets of parameters. See Appendix F for details.

Second, our analyses may not generalize easily to other settings or to more radical educational interventions. The set of interventions examined here are quite diverse and fairly typical of educational programs that are evaluated with RCTs. However, these interventions typically affected a small percentage of total instruction time and were relatively short-lived, generally lasting less than a year.

This may be particularly relevant given the evidence we present suggesting that the relationship between attrition and outcomes is stronger on the treatment side than the control. One interpretation of this finding is that the perniciousness of attrition mechanisms could be a function of intervention intensity. This conjecture is something we hope to test formally in future work. More generally, we note that the nature of attrition bias fundamentally depends on the nature of attrition mechanisms, which may differ from setting to setting – e.g. from the UK to the USA, or from school to university contexts.

## Conclusion

Overall, the analyses presented here suggest that the threat of attrition bias is limited in our context. While attrition is a highly salient risk, other threats – for example, external validity bias due to non-random sampling – may be substantially more problematic in terms of generating practical, useable knowledge from education evaluations.

This is not to say that attrition mechanisms can safely be treated as "missing at random" or "missing completely at random". We find evidence that students who leave studies tend to perform worse than those who remain. This pattern is particularly pronounced for treated students. Moreover, this tendency persists even after conditioning on baseline achievement. That said, we re-emphasize that these associations do not appear to be strong enough to induce large-scale bias.

We suggest that researchers respond to this evidence in two main ways: incorporating 'attrition bias uncertainty' into their inferences and completing sensitivity analyses using empirically-grounded estimates of attrition mechanisms that have been observed in practice. As more studies are added to the RCT archive, we intend to present a more detailed picture of attrition mechanisms, including an exploration of why some evaluations seem to suffer from pernicious attrition. In the meantime, it seems sensible to presume that students who are missing may differ in unobserved ways from those who remain.

## References

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of External Validity Bias When Impact Evaluations Select Sites Nonrandomly. Educational evaluation and policy analysis.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. Educational Research, 60(3), 276-291.

Crawford, C., & Skipp, A. (2014). LIT Programme. IFS and NatCen.

Deke, J., & Chiang, H. (2017). The WWC attrition standard: sensitivity to assumptions and opportunities for refining and adapting to new contexts. Evaluation review, 41(2), 130-154.

Demack, S., Maxwell, B., Coldwell, M., Stevens, A., Wolstenholme, C., Reaney-Wood, S., Stiell, B., & Lortie-Forgues, H. (forthcoming). Review of EEF Reports. Education Endowment Foundation.

DfE. (2015). Statistical First Release: Schools, pupils and their characteristics: January 2015. UK Department for Education, Retrieved on April 22 2020 from www.gov.uk/government/uploads/system/uploads/attachment_data/file/433680/SFR16_2015_Main_Text.pdf.

Dong, N., & Lipsey, M. W. (2011). Biases in Estimating Treatment Effects Due to Attrition in Randomized Controlled Trials and Cluster Randomized Controlled Trials: A Simulation Study. Society for Research on Educational Effectiveness.

Dorsett, R., Rienzo, C., Rolfe, H., Burns, H., Robertson, B., Thorpe, B., & Wall, K. (2014). Mind the Gap. National Institute of Economic and Social Research.

EEF. (2014). Classification of the security of findings from EEF evaluations. Education Endowment Foundation, Retrieved on April 22 2020 from https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Classifying_the_security_of_EEF_findings_FINAL.pdf.

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. Evaluation review, 38(5), 359-387.

Haywood, S., Griggs, J., Lloyd, C., Morris, S., Kiss, Z., & Skipp, A. (2015). Creative Futures: Act, Sing, Play. NatCen.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(1), 137-159.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75(4), 800-802.

Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., & Stevens, A. (2017). Dialogic Teaching. Sheffield Hallam University.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Annals of family medicine, 2(3), 204-208.

Lewis, M. S. (2013). A Series of Sensitivity Analyses Examining the What Works Clearinghouse's Guidelines on Attrition Bias. Ohio University,

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data: John Wiley & Sons.

Lloyd, C., Edovald, T., Morris, S., Skipp, A., KIss, Z., & Haywood, S. (2015). Durham Shared Maths Project. NatCen Social Research.

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

Manski, C. F. (1990). Nonparametric bounds on treatment effects. American Economic Review, 80(2), 319-323.

Miller, S., Davison, J., Yohanis, J., Sloan, S., Gildea, A., & Thurston, A. (2016). Texting Parents. Queen's University Belfast.

Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, A. (2016). ReflectED. Manchester Institute of Education.

NPD. (2015). National Pupil Database User Guide. UK Department for Education, Retrieved on April 22 2020 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/472700/NPD_user_guide.pdf.

Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). Changing Mindsets. National Institute for Economic and Social Research.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R., & Wong, S. (1998). A Method for Exploring the Effects of Attrition in Randomized Experiments With Dichotomous Outcomes. Psychological Methods, 3(1), 3-22.

Sibieta, L. (2016). REACH Evaluation Report. Institute of Fiscal Studies.

Sundberg, R. (2003). Conditional statistical inference and quantification of relevance. Journal of the Royal Statistical Society: Series B, 65(1), 299-315.

Thurston, A., Roseth, C., O'Hare, L., Davison, J., & Stark, P. (2016). Talk of the Town. Queens University Belfast.

Weidmann, B., & Miratrix, L., (Forthcoming). Lurking inferential monsters: selection bias in non-experimental evaluations of school programs. Journal of Policy Analysis and Management

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, 1-34.

WWC. (2014). Assessing Attrition Bias. What Works Clearinghouse.

WWC. (2017). Standards Handbook 4.0. What Works Clearninghouse, Retrieved on April 20 2020 from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

## Appendix A: Simulation-based uncertainty estimates

To generate uncertainty estimates for attrition bias, we conduct a simulation-based procedure. We use a simulated procedure because of the nested nature of our data. Bias estimates from outcomes in the same study are highly correlated, and we wanted to capture these dependencies in our inference.

We perform two related procedures. The first generates uncertainty estimates for $\hat{\beta}_{kw}$. Here we simulate a world in which attrition is completely random, i.e. a world in which the MCAR assumption is true by design. We condition on several dimensions: the number of pupils at randomization (N), observed attrition rate ($P$), the observed treatment assignment ($T^{obs}$), and observed outcomes ($Y^{obs}$). For each study-outcome pair, we complete the following two-step process 1000 times:

(a) Permute the observed binary attrition indicator. We define the 'null responder' sample as all units for whom this permuted indicator is equal to zero.

(b) Generate an estimate of $\hat{\beta}_{kw}^{Null}$ by estimating Model 1 twice: once for the full sample, once for the 'null responder' sample.

The second procedure generates uncertainty estimates for $\hat{\beta}_{kw}^{X}$. Here we simulate a world in which attrition is determined by observed covariates $X$. We again condition on several dimensions: N, $P_T, P_C$, $T^{obs}$ and $Y^{obs}$. For each study-outcome pair, we complete the following three-step process 1000 times:

(a) Fit two propensity score models for attrition:
   (i) $P_i\left(A_i^{obs} = 1 \middle| T_i^{obs} = 1\right) = logit^{-1}(\boldsymbol{\gamma_t X_i})$
   (ii) $P_i\left(A_i^{obs} = 1 \middle| T_i^{obs} = 0\right) = logit^{-1}(\boldsymbol{\gamma_c X_i})$

(b) Define a set of responders, based on each student's observed covariate profile $\boldsymbol{X_i}$.
   (i) For treated units, draw from $Bern\left(1 - \hat{P}_i\left(A_i^{obs} = 1 \middle| T_i^{obs} = 1\right)\right)$. If this equals one then the unit is a 'null treatment responder'.
   (ii) For control units, draw from $Bern\left(1 - \hat{P}_i\left(A_i^{obs} = 1 \middle| T_i^{obs} = 0\right)\right)$. If this equals one then the unit is a 'null control responder'.

(c) Generate an estimate of $\hat{\beta}_{kw}^{X,null}$ by estimating Model 2 twice: once for the full sample, once for the 'null responder' sample.

Models are defined in section 4.

## Appendix B: Meta-analysis Details

This appendix presents our approach to generating constrained empirical Bayes' estimates of bias. We use a meta-analysis framework to model 'attrition sampling variation' (Weidmann and Mirtarix, forthcoming). Our observed bias estimates $\hat{\beta}_{kw}$ are assumed to be made up of several components:

$$\hat{\beta}_{kw} | \beta_{kw} \sim N(\beta_{kw}, \sigma_{kw}^2)$$
$$\beta_{kw} \sim N(\nu, \eta^2)$$

Where:

(a) $\nu =$ the mean attrition bias across all interventions and outcomes

(b) $\beta_{kw} =$ the true attrition bias for outcome $k$ in intervention $w$. This has a variance of $\eta^2$ reflecting the fact that attrition bias may vary due to context, the nature of the program and so on.

(c) Observed bias $\hat{\beta}_{kw}$ deviates from underlying bias $\beta_{kw}$ with a variance of $\sigma_{kw}^2$. This sampling variation largely depends on how many schools participated in intervention $w$.

To estimate the variance of bias $\widehat{var}(\beta_{kw}) = \hat{\eta}^2$, we use the method-of-moments approach from Higgins et al. (2009):

$$\hat{\eta}^2 = \max\left\{0, \frac{Q-(K-1)}{\sum\hat{\sigma}_{kw}^{-2} - \frac{\sum\hat{\sigma}_{kw}^{-4}}{\sum\hat{\sigma}_{kw}^{-2}}}\right\}$$

Where:

$$Q = \sum\left(\hat{\beta}_{kw} - \bar{\beta}\right)^2 \hat{\sigma}_{kw}^{-2}$$

$$\bar{\beta} = \frac{\sum\hat{\beta}_{kw} \cdot \hat{\sigma}_{kw}^{-2}}{\sum\hat{\sigma}_{kw}^{-2}}$$

Estimates of $\hat{\sigma}_{kw}^{-2}$ come from the simulations under the null, described in Appendix A: $\hat{\sigma}_{kw}^{-2} = var\left(\hat{\beta}_{kw}\right)$. K is the effective sample size and is a function of the total number of outcomes, N, the mean number of outcomes per study (k) and the estimated intra-class correlation of $\hat{\beta}$ ($\hat{\rho}$), as per Killip et al. (2004):

$$K = \frac{N}{1-(k-1)\cdot\hat{\rho}} = \frac{22}{1-(2.2-1)\cdot 0.76} = 11.5$$

The estimate of $\hat{\rho}$ comes from a multilevel model in which $\hat{\beta}_{kw} \sim N(\alpha_w, \sigma_e^2)$, $\alpha_w \sim N(\gamma_0, \sigma_a^2)$, and $\hat{\rho} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}$. Letting $\widehat{\omega}_{kw} = \left(\hat{\sigma}_{kw}^2 + \hat{\eta}^2\right)^{-1}$, we estimate the mean attrition bias:

$$\hat{\nu} = \frac{\sum\hat{\beta}_{kw}\widehat{\omega}_{kw}}{\sum\widehat{\omega}_{kw}}$$

Next, we generate simple, parametric empirical Bayes estimates of the attrition bias for intervention $w$ and outcome $k$:

$$\beta_{kw}^* = \hat{\lambda}_{kw}\hat{\nu} + \left(1 - \hat{\lambda}_{kw}\right)\hat{\beta}_{kw}$$

where $\hat{\lambda}_{kw} = \frac{\hat{\sigma}_{kw}^2}{\hat{\sigma}_{kw}^2 + \hat{\eta}^2}$.

While individual estimates of $\beta_{kw}^*$ minimize RMSE, an empirical distribution based on these estimates will underestimate the variability in bias estimates across studies and outcomes (Weiss et al., 2017). As such, we follow the procedure of Weiss et al. (2017, p.13) and scale our shrunken estimates so that their variance is equal to the estimated value of $\hat{\eta}^2$.

We follow the same procedure for other parameters of interest: $\beta^X, \Delta_T, \Delta_C, \Delta_T^{\mathrm{X}}, \Delta_C^{\mathrm{X}}$.

## Appendix C: tests of MAR and MCAR at the study-outcome level

This section presents the results for study-outcome level tests for two assumptions: "Missing At Random" (MAR) and "Missing Completely At Random" (MCAR). In both cases, we perform null-hypothesis tests.

For MAR, we test the null $H_0\colon \delta_2 = \gamma_2 = 0$ using a Likelihood Ratio test applied to Model 4:

$$Y_{ij} = \alpha_j + \gamma_2 A_{ij} + \lambda_2 T_{ij} + \delta_2 A_{ij} T_{ij} + \boldsymbol{\nu_C X_{ij}} + \boldsymbol{\nu_T X_{ij}} T_j + e_{ij} \qquad (Model\ 4)$$
$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$
$$e_{ij} \sim N(0, \sigma^2)$$

For MCAR, we examine the null $H_0: \delta = \gamma = 0$, using a Likelihood Ratio test applied to Model 3:

$$Y_{ij} = \alpha_j + \gamma A_{ij} + \lambda_1 T_{ij} + \delta A_{ij} T_{ij} + e_{ij} \qquad (Model\ 3)$$
$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$
$$e_{ij} \sim N(0, \sigma^2)$$

Variables are defined in Section 2. Table 5 provides the p-values for the tests.

Table 5 – p-values for MCAR and MAR for individual studies

| | Missing Completely At Random ($H_0: \delta = \gamma = 0$) | | | | | Missing At Random ($H_0: \delta_2 = \gamma_2 = 0$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Project* | Maths | Reading | Writing | English | Science | Maths | Reading | Writing | English | Science |
| asp | 0.394 | 0.373 | - | - | - | 0.294 | 0.457 | - | - | - |
| cmi | 0.686 | 0.074 | *0.004* | - | - | 0.314 | 0.326 | 0.168 | - | - |
| cmp | 0.069 | 0.514 | 0.062 | - | - | 0.102 | 0.337 | 0.166 | - | - |
| dt | **<0.001** | 0.052 | - | - | - | *0.035* | 0.851 | - | - | - |
| lit | - | - | - | **<0.001** | - | - | - | - | **<0.001** | - |
| mtg | *0.024* | 0.337 | - | - | - | *0.018* | 0.441 | - | - | - |
| ref | 0.058 | 0.396 | - | - | - | 0.137 | 0.468 | - | - | - |
| sm | **<0.001** | **<0.001** | - | - | - | **0.001** | 0.309 | - | - | - |
| tott | 0.539 | *0.046* | - | - | - | 0.968 | 0.112 | - | - | - |
| tp | **<0.001** | - | - | **<0.001** | **<0.001** | **<0.001** | - | - | **<0.001** | **<0.001** |

Notes: this table presents the p-values for hypothesis tests examining the Missing Completely At Random (MCAR; left panel) and Missing At Random (MAR; right panel) assumptions. The project acronyms are as follows: asp = Act Sing Play; cmi = Changing Mindsets INSET; dt = Dialogic Teaching; lit = LIT programme; mtg = Mind The Gap; ref = ReflectEd; sm = Shared Maths; tott = Talk of The Town; tp = Texting Parents. Cells with '-' represent a domain that wasn't tested. **Bold font** indicates that individual null hypothesis tests were rejected at $\alpha$ =0.05, after a Hochberg (1988) multiple-comparison correction.

## Appendix D: predictors of pernicious attrition mechanisms

### Is there evidence that larger studies tend to suffer from more problematic attrition mechanisms?

To examine the relationship between attrition mechanisms and study size we fit the following models:

$$\left|\widetilde{\Delta}_{akw}^{X}\right| = \delta_{0\mathrm{C}} + \delta_{0\mathrm{T}} T_a + \delta_{1\mathrm{C}} \mathrm{n}_{\mathrm{aw}} + \delta_{1\mathrm{T}} \mathrm{n}_{\mathrm{aw}} T_a + \epsilon_{akw} \quad \text{(H1)}$$

$\widetilde{\Delta}^X$ is is the estimated difference in mean outcomes for 'attriters' and 'responders', after conditioning on covariates with a linear model. $a$ is a treatment arm (Control or Treatment); $T_a$=1 if the arm is treatment, rather than control; $k$ is outcome domain (maths, English, science, reading, writing), w is the intervention. $n_{aw}$ is the number of pupils at randomization in arm a of intervention w.

To avoid strong distributional assumptions about the residuals, and to account for clustering of outcomes within studies, we perform a permutation test of the null hypothesis that the number of pupils has no effect on the magnitude of bias. We begin by fitting model H1 and estimating the t-statistics for $\delta_{1C}$ and $\delta_{1T}$. Then we complete the following procedure 10,000 times:

(a) Permute the identity of the interventions (w). Based on this permutation, generate a new list of sample sizes, $n'_{aw}$.

(b) Fit model H1 again, using $n'_{aw}$ as a predictor in place of $n_{aw}$

(c) Calculate the t-statistics for $\delta'_{1C}$ and $\delta'_{1T}$

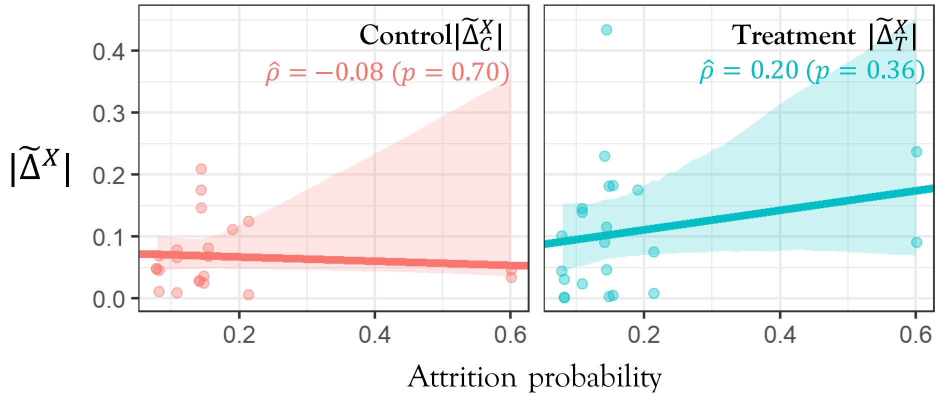
The p-value for the null hypothesis that $\delta_{1C} = 0$ is p=0.0004. This is defined as the proportion of estimated t-statistics that are smaller in magnitude than the observed t-stat for $\delta_{1T}$ (2.22). This is evidence to reject the null of no association.

We then repeat this process on the treatment side. Here the p-value for the null hypothesis of no association is p=0.102. We fail to reject the null at $\alpha = 0.05$. As can be seen in the right-hand panel of Figure 7, this is because the relationship on the treatment side depends on an outlier, which introduces extra uncertainty relative to the control side.

### Is there evidence that studies with more attrition tend to suffer from more problematic attrition mechanisms?

Here we examine the possibility that 'problematic' attrition is associated with the probability of attrition. A simple graphical analysis suggests that there is no association. We confirm this by estimating the correlations (reported in Figure 9). Even without correcting for the clustering of outcomes within studies, neither of the estimated correlations is significantly different from zero.

Figure 9: Relationship between attrition probability and attrition mechanism

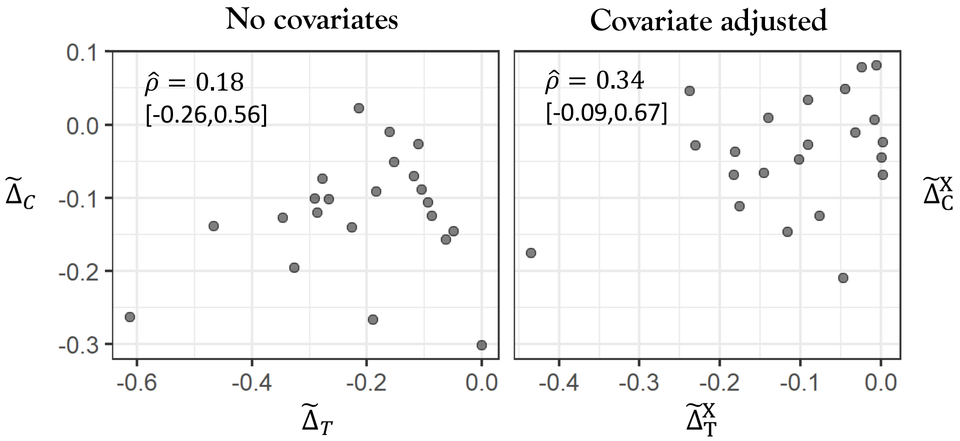


## Appendix E: assessing the relationship between $\Delta_T$ and $\Delta_C$

In section 6, we suggest the simplifying assumption that the cross term $E[\Delta_T \Delta_C]$ be set to zero. This assumption influences $\eta^2$ and consequently the extent to which accounting for attrition bias will widen uncertainty estimates. For example, if $\Delta_T$ and $\Delta_C$ have the same sign, then the threat of attrition bias decreases as bias in each arm is offsetting. Equally, if $\Delta_T$ and $\Delta_C$ have different signs, then threat of bias increases along with the uncertainty due to attrition bias.

The evidence we have about the relationship between $\Delta_T$ and $\Delta_C$ is presented in Figure 10 (left panel), along with an analogous plot for $\Delta_C^X$ and $\Delta_T^X$ (right panel). It suggests the lack of a clear association. While the point estimates of correlation are positive in both plots, the data fail to reject the null that $\rho = 0$ in both cases. Moreover, in both cases the estimated positive correlations, reported on the plots, are substantially shaped by an outlier. Clearly, as discussed in the main text, the estimated mean of both $\Delta_T$ and $\Delta_C$ is negative.

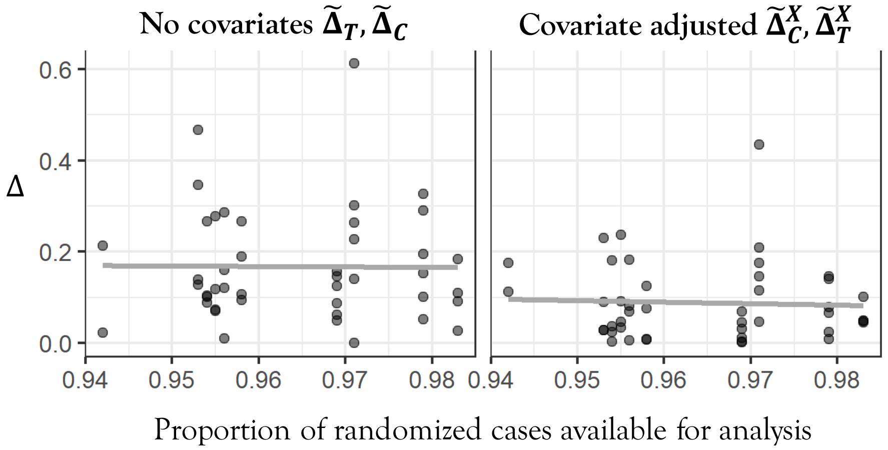
Figure 10: Relationship between treatment attrition and control attrition



## Appendix F: Testing the association between missingness in the national pupil database, and bias parameters

Figure 11 presents the association between the proportion of randomized cases available for the analyses we present in this paper, and attrition parameters. We find no association.

Figure 11: association between attrition mechanisms and missingness from the census


No covariates $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{T}}, \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{C}}$
Covariate adjusted $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{C}}^{\boldsymbol{X}}, \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{T}}^{\boldsymbol{X}}$
Δ
Proportion of randomized cases available for analysis