ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-754

Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials

Ishita Ahmed
Stanford University

Masha Bertling
Harvard University

Lijin Zhang
Stanford University

Andrew D. Ho
Harvard University

Prashant Loyalka
Stanford University

Hao Xue
Stanford University

Scott Rozelle
Stanford University

Benjamin W. Domingue
Stanford University

Researchers use test outcomes to evaluate the effectiveness of education interventions across numerous randomized controlled trials (RCTs). Aggregate test data—for example, simple measures like the sum of correct responses—are compared across treatment and control groups to determine whether an intervention has had a positive impact on student achievement. We show that item-level data and psychometric analyses can provide information about treatment heterogeneity and improve design of future experiments. We apply techniques typically used in the study of Differential Item Functioning (DIF) to examine variation in the degree to which items show treatment effects. That is, are observed treatment effects due to generalized gains on the aggregate achievement measures or are they due to targeted gains on specific items? Based on our analysis of 7,244,566 item responses (265,732 students responding to 2,119 items) taken from 15 RCTs in low-and-middle-income countries, we find clear evidence for variation in gains across items. DIF analyses identify items that are highly sensitive to the interventions—in one extreme case, a single item drives nearly 40% of the observed treatment effect—as well as items that are insensitive. We also show that the variation of item-level sensitivity can have implications for the precision of effect estimates. Of the RCTs that have significant effect estimates, 41% have patterns of item-level sensitivity to treatment that allow for the possibility of a null effect when this source of uncertainty is considered. Our findings demonstrate how researchers can gain more insight regarding the effects of interventions via additional analysis of item-level test data.

VERSION: April 2023

Suggested citation: Ahmed, Ishita, Masha Bertling, Lijin Zhang, Andrew D. Ho, Prashant Loyalka, Hao Xue, Scott Rozelle, and Benjamin W. Domingue. (2023). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. (EdWorkingPaper: 23-754). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/1nw4-na96

Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials

Ishita Ahmed[a,1], Masha Bertling[2], Lijin Zhang[1], Andrew D. Ho[2], Prashant Loyalka[1,3], Hao Xue[3], Scott Rozelle[3], and Benjamin W. Domingue[a,1]

[1]Graduate School of Education, Stanford University
[2]Harvard Graduate School of Education
[3]Stanford Center on China's Economy and Institutions, Freeman Spogli Institute for International Studies, Stanford University
[a]Correspondence concerning this article should be addressed to Ishita Ahmed (iahmed2@stanford.edu) & Ben Domingue (bdomingue@stanford.edu).

Acknowledgements

This article reflects contributions from many organizations and individuals. We would like to thank the research teams for sharing the item-level outcome data from their randomized controlled trials (RCTs). An early analytic design of this study is pre-registered at `https://osf.io/mjrx3/`. This work is partially funded by the Jacobs Foundation.

Abstract

Researchers use test outcomes to evaluate the effectiveness of education interventions across numerous randomized controlled trials (RCTs). Aggregate test data—for example, simple measures like the sum of correct responses—are compared across treatment and control groups to determine whether an intervention has had a positive impact on student achievement. We show that item-level data and psychometric analyses can provide information about treatment heterogeneity and improve design of future experiments. We apply techniques typically used in the study of Differential Item Functioning (DIF) to examine variation in the degree to which items show treatment effects. That is, are observed treatment effects due to generalized gains on the aggregate achievement measures or are they due to targeted gains on specific items? Based on our analysis of 7,244,566 item responses (265,732 students responding to 2,119 items) taken from 15 RCTs in low-and-middle-income countries, we find clear evidence for variation in gains across items. DIF analyses identify items that are highly sensitive to the interventions—in one extreme case, a single item drives nearly 40% of the observed treatment effect—as well as items that are insensitive. We also show that the variation of item-level sensitivity can have implications for the precision of effect estimates. Of the RCTs that have significant effect estimates, 41% have patterns of item-level sensitivity to treatment that allow for the possibility of a null effect when this source of uncertainty is considered. Our findings demonstrate how researchers can gain more insight regarding the effects of interventions via additional analysis of item-level test data.

1 Introduction

Recent decades have witnessed a large growth in the number of educational interventions evaluated via randomized controlled trials (RCTs [1, 2]). Much of this work has been done in low-and-middle-income countries (LMICs) and evidence from this work offers insights into how interventions can improve children's educational achievement in the context of LMICs [3, 4]. RCTs assess the impact of an intervention by comparing outcomes—typically aggregate test scores—between treatment and control groups. The tests used in RCTs most often measure math or language ability. Information about whether interventions are effective in increasing academic skills as measured by the tests used as outcomes is an important question for improving educational practice thus motivating the outlay of time and resources required to conduct RCTs.

Effect estimates are premised on the choice of specific outcome measures. Any information about the intervention—whether the intervention works, for whom, in which settings—assumes that stakeholders already agree upon and understand the outcome. Using a high-quality outcome—an outcome that has strong psychometric properties and is appropriately aligned to the intervention—is thus crucial. However, a systematic review of RCTs in education in LMICs [5] demonstrates room for improvement in transparency and quality about the outcome measures that many RCTs use. Studies rarely provide information about tested subdomains and reliability (e.g., only 4% of studies reported reliability) [5]. Although aggregate test scores are supposed to be representative of broad academic abilities in math or language, there is a large variation in how many subdomains are represented and the number of items within each subdomain [5]. The nature of the outcome measure used to determine the effectiveness of education interventions varies greatly across studies.

We build on this recent work [5] focusing on the overall psychometric properties of outcomes used in RCTs by focusing on a related issue: the degree to which treatment effects may vary across items. Suppose a math intervention focuses on basic arithmetic operations with fractions. An item on a test that involves adding two fractions with different denominators would be directly aligned with the skills targeted by the intervention. In contrast, another item may involve manipulation of fractions as just one part of a larger sequence of operations; in this sense, the item is relatively distal from the treatment and may show relatively less sensitivity to treatment (i.e., utiliziation of such an item may show a smaller, possibly null, treatment effect). The current project is designed to study such variation. One related issue has to do with the misalignment of an intervention and outcome measure. An outcome that is fairly distal in terms of its connect to the intervention or otherwise inappropriate for evaluating a specific intervention may lead to the erroneous evaluation that an intervention is ineffective. Selection of an appropriate outcome is thus imperative (see guidance given in [6]); psychometric analyses may help ensure outcomes have reasonable technical properties and, as used here, may offer additional information about the effectiveness of interventions.

The study of such heterogeneity in item-level sensitivity to treatment that we undertake here allows us to ask, for example, whether effects of interventions manifest as general increases in the targeted domain.

Or, rather, are they suggestive of more narrow gains in subsets of items (i.e., that may be associated with specific skills).[1] We consider these issues using item-level response data—data from over 200,000 students responding to over 2,000 items—from 15 RCTs from the REAP and JPAL consortia [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. We also consider how variation in sensitivity across items within a test may be a facet of uncertainty that may have implications for our attempts to understand the degree to which the magnitude of estimated effects is driven by random variation. Given the centrality of the outcome measure in subsequent inferences about the efficacy of educational interventions, we view the insights shown here as highly relevant for the design of future RCTs. In the following section, we offer technical details on the problem and our proposed approach for analyzing it.

2 The problem

Our approach is motivated by ideas from the literature on "Differential Item Functioning" (DIF; [20]). DIF analysis is typically used to examine whether certain items used on educational or psychological measures show DIF as a function of group membership; such analyses do so by asking whether item response behavior varies across groups once we have controlled for underlying ability. Here, we anticipate findings of DIF if the effects of an RCT intervention on achievement outcomes are specific to certain items (i.e., if there is variation in the level of sensitivity to treatment across the test's items). We would not interpret such findings as related to fairness but as potential evidence for the item's enhanced sensitivity to the treatment effect. For example, if an intervention leads to increased math test scores, DIF analysis will demonstrate if the positive effect is generally uniform across all of the items or if there are effects on specific items. Alternatively, null effects may mask the existence of item-specific effects (e.g., [21]).

We use techniques from item response theory (IRT; [22]) to analyze item responses; specifically, we consider the 1PL model [23].[2] This approach should generalize to a variety of settings given the common usage of multiple choice items that are easily scored as correct/incorrect in such scenarios. For a generic item in an outcome measure of the kind we consider here, this model for a dichotomous outcome is built upon the supposition that if $\Pr(Z_{ij} = 1) \equiv p_{ij}$ is the probability of a correct response Z_{ij} from person i to item j then the log-odds of the outcomes are equivalent to the distance between person ability θ_i and item difficulty b_j,

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i - b_j. \tag{1}$$

Algebraic manipulation leads to something akin to the traditional logistic regression model

$$\Pr(Z_{ij} = 1|\theta_i) = \frac{1}{1+\exp(-(\theta_i - b_j))}. \tag{2}$$

We now extend this basic model in a few key ways to allow for the study of item-level sensitivity to treatment effects.

To study the effect of treatment, we introduce item-treatment interactions β_j and update Eqn 2 so that the item response function—defined as $\mathbb{E}(Z_{ij}|\theta_i)$, which is equivalent to $\Pr(Z_{ij} = 1|\theta_i)$ here as Z is a Bernoulli random variable—varies as a function of treatment status. We assume that

$$\Pr(Z_{ij} = 1|\theta_i) = \frac{1}{1+\exp(-(a\theta_i + \delta_i\beta_j - b_j))} \tag{3}$$

where $\delta_i = 0$ if person i is in the control group and 1 otherwise. Note that, if the treatment is effective, we would anticipate item j being easier had the respondent been in the treatment group and we would have $\mathbb{E}(\beta_j) > 0$. If the effect of treatment is homogeneous across items (ie. all items are equally sensitive to the treatment effect), then we would have $\beta_j = \beta_{j'}$ for all j'. The a coefficient is a test-level loading that we estimate given the identification constraints we now discuss.

[1] One motivation for this project stemmed from an observation regarding the specificity of a exogenous shock in a different setting. Spousal loss is known to lead to an increase in depressive symptoms, but it is not well-characterized as a general increase in the liability to report all symptoms but a highly specific increase of the probability of reporting, for example, loneliness [7].

[2] The 1PL is effectively a reparametrization of the Rasch [24] model.

The 1PL requires additional constraints to be properly identified (i.e., note that Eqn 2 would be invariant under the transformation $\theta \to \theta_i + \epsilon$ or $b_j \to b_j + \epsilon$ for some constant ϵ). For identification, we assume that $\theta_i \sim \text{Normal}(0, 1)$. Note in particular that we assume that there is no difference in mean for treatment and control groups (i.e., the distribution of θ_i does not depend on δ_i). This assumptions results in treatment effects being estimated as item-level offsets; i.e., in the values of β_j, consistent with our above comments regarding $\mathbb{E}(\beta_j)$. The key assumptions being made here are that (i) the ability variances of the two groups are identical and that (ii) the a value is equivalent across groups. These are required for identification and consistent with the basic notion of a constant treatment effect (assuming identical distributions of ability pre-randomization). We view it as a relatively weak assumption but comment more on this point in the discussion.

Our utilization of Eqn 3 is critical so we pause to emphasize its relevance with a few examples. Suppose first that $\beta_j \approx c$ for some constant c across all j items. This suggests that the intervention had a generic effect (potentially a null effect if c is near 0); the value c is integrally linked to the global sensitivity of outcome to the intervention. Our primary focus here is on variation in the β_j parameters. Variation in the β_j parameter suggests that group membership is uniquely associated with item response behavior for the j-th item. Let us suppose that the treatment had some effect on the outcome; i.e., that $\beta_{j'} \approx c > 0$ for $j' \neq j$ (i.e., the RCT had a constant effect on the j' items) but that $\beta_j \neq c$. Focusing on item j, if $\beta_j > c$, item j is relatively easier for members of the treatment group than for members of the control group. We generally consider this to be evidence that the treatment effect was larger on item j—perhaps due to synergies between the nature of the intervention and the kinds of skills tapped by this particular item (note that we will generally be agnostic as to the specific skills in question)—than items j'. If $\beta_j < c$, we interpret this as evidence that the treatment effect was smaller for item j, perhaps due to that item focusing on content that is not especially aligned to the intervention in contrast to the other items included in the assessment.

3 Methods

3.1 Data

We obtained item-level response data on academic outcomes from multiple RCTs through our collaborations with the Education Measurement Initiative at the Abdul Latif Jameel Poverty Action Lab (JPAL; [8, 9, 10, 11, 12, 13]) and the Rural Education Action Program (REAP; [14, 15, 16, 17, 18, 19] plus additional unpublished reports and follow-up studies) at Stanford's Freeman Spogli Institute for International Studies. We separately describe data from JPAL and REAP.

- JPAL: The JPAL data includes 6 different RCTs from countries in Sub-Saharan Africa and South Asia with item-level variables for math and literacy outcomes across all grade levels.[3] Five of the RCTs target children in primary school grades and one RCT includes children in junior high school [9]. There are interventions in school infrastructure and management, teacher incentives and training, educational technology, and financial education. In general, the researchers for these studies typically focused on broader skills in math and literacy rather than specific items for their intervention and many items were taken from publicly available questions released by existing assessments such as national exams [5]. After preprocessing, we use data from 6 RCTs with 64 separate grade-subject combinations.

- REAP: The REAP data focuses on 9 RCTs conducted in China with item-level variables on math outcomes. The interventions include teacher training and performance incentives, education technology, and health (e.g., providing eyeglasses). Seven of the interventions were in primary schools and two of the interventions were in junior high schools. The items for the math tests were drawn from standardized math curricula for primary and junior high school students and experts checked the content validity of these tests. Pilot data during test development was used to analyze the psychometric properties of the tests. After preprocessing, we use data from 9 interventions that we classify as 20 separate tests (after accounting for different grade levels).

[3] Note that item-level data from 16 JPAL RCTs are described in [5]. Here we focus on those that contained information on treatment status.

We refer generically to a 'test' as the outcome/data used for a particular intervention in either JPAL or REAP in a specific grade. We also combine all treatment arms into a single "treatment" that we contrast with the control condition. We also focus on simple post-treatment comparisons and do not adjust for other factors (e.g., we do not control for pretest scores and do not account for any clustering or stratification).

Data were preprocessed in a common way. In particular, items with mean correct responses below 0.02 or above 0.98 were excluded and we required tests to have at least 50 respondents and 4 items that met exclusion criteria.[4] Across the 84 tests, we use data from 265,732 respondents (162,359 in JPAL and 103,373 in REAP) on 2119 items (1466 in JPAL and 653 in REAP; we emphasize that the same item may occur in many tests). In total, we analyze over seven million item responses.

3.2 Analysis

To estimate Eqn 3, we use the multiple group approach [25] estimated via the EM algorithm [26] as implemented in `mirt` [27].[5] We focus on estimates of β_j; power analyses suggest we should be generally well-powered in a test with the mean number of respondents to detect $|\beta_j| > 0.3$ (this value is on the logit scale, see Appendix A for guidance on interpretation). Estimates of β_j, denoted $\widehat{\beta}_j$, are informative about the role of items in a specific test context.[6] We use estimates from Eqn 3 in a variety of subsequent analyses. We first consider the direction and magnitude of the $\widehat{\beta}_j$ estimates. We also examine associations between these quantities and item characteristics emphasized in classical test theory [28]: difficulty (the mean proportion of correct responses to an item, $\sum_i Z_{ij}/N_i$) and discrimination (the correlation between the item responses and the total score $r(Z_{ij}, \sum_j Z_{ij})$). We also study the association between $\widehat{\beta}_j$ and estimates of treatment effects that result from removing the focal item from the observed matrix of item responses.

To index the overall homogeneity of item-level treatment sensitivity, we consider the ratio, denoted R, of variation in the treatment-specific offsets to the variation in difficulty parameters,

$$R = \frac{\sigma(\widehat{\beta}_j)}{\sigma(\widehat{b}_j)} \tag{4}$$

(where σ is the SD operator). Relatively large values of R would suggest tests with significant item-treatment interactions as compared to tests with smaller R values. Note that R is not a measure of DIF but rather an index about the degree to which items are varying in treatment-related sensitivity relative to the overall variation in difficulty of the items. A uniform treatment effect across items will yield $R = 0$. Usage of R presumes variation in the b_j parameters; that is, we are indexing the variation in item-treatment interactions as a proportion of the variation in item difficulty across the test. In practice, items tend to vary in difficulty as a matter of good item writing practice and the lack of such variation would be problematic for other reasons (i.e., this would lead to the test providing limited information about respondents whose abilities are not near the common difficulty across items).

We also consider a simulation study to examine the imprecision of treatment effect estimates associated with variation in item-level sensitivity to treatment. We focus on the difference in standardized mean sum score between treatment and control groups, denote this difference as d. For values $\widehat{b}_j$ and $\widehat{\beta}_j$ for a single test of J items from N respondents, we simulate a new set of outcomes as follows:

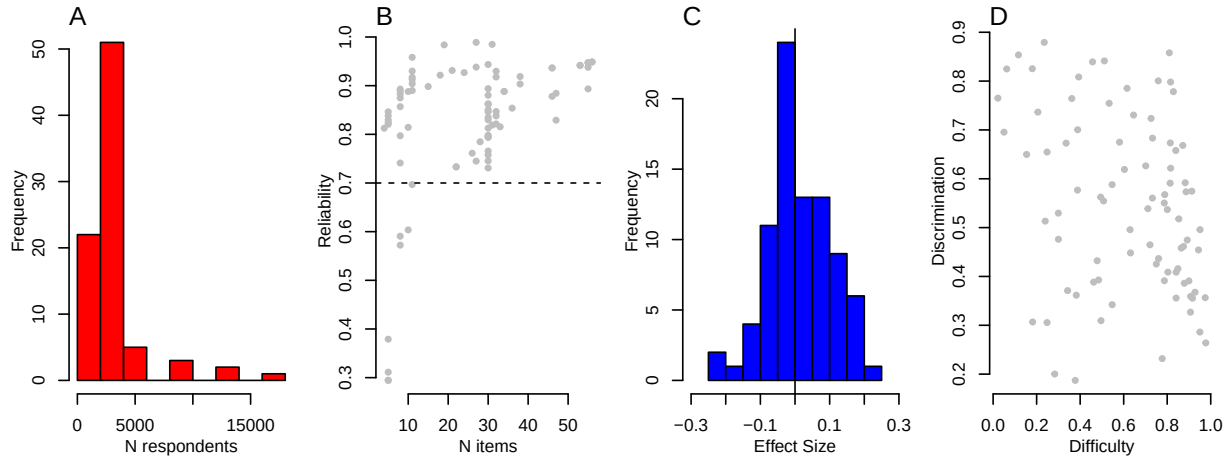
1. We generate J values from Unif(0,1).
2. Treating these as probabilities, we sample β'_j (with replacement) values based on the empirical cumulative density function (ECDF) of the $\widehat{\beta}_j$ values. These β'_j values are meant to indicate alternative arrangements of sensitivity as captured by the distribution of $\widehat{\beta}_j$ values.

[4] These criteria were used to ensure reasonable sample sizes; in particular, items that are especially difficult (with mean below 0.02) or easy (with mean above 0.98) would have very little variation in responses. Whether such items are useful for discerning treatment effects is a separate question (we would, in general, argue that they are not).

[5] Our use of multigroup analysis departs from conventional DIF studies. Conventional DIF studies analyze each item separately whereas the approach used here allows for simultaneous estimation of all β_j parameters.

[6] Note that [27] actually estimates a model parametrized slightly differently than Eqn 3; in particular it estimates a difficulty parameter separately for each group. If we denote these as b_c and b_t (omitting the j subscript), we let $b = b_c$ and then allow $\beta_j = b_c - b_t$.

Figure 1: Summary of item response data across 84 RCTs. A: Number of respondents per RCT. B: Scatterplot of number of items by reliability. C: Histogram of effect sizes. D: Item-level difficulties and discriminations.



3. We simulate N abilities from a standard normal and randomly assign those N respondents to treatment or control based on the overall proportion of respondents in the treatment group.

4. We use empirical estimates $\widehat{b_j}$ and $\widehat{a}$ and Eqn 3 with values of θ_i and β'_j from above to simulate new item responses.

5. We then construct the estimated treatment effect for this simulated set of item responses.

We simulate 5000 such item response datasets for each test to examine variation in the effect estimate driven by variation in the possible configuration of the $\widehat{\beta_j}$ parameters.

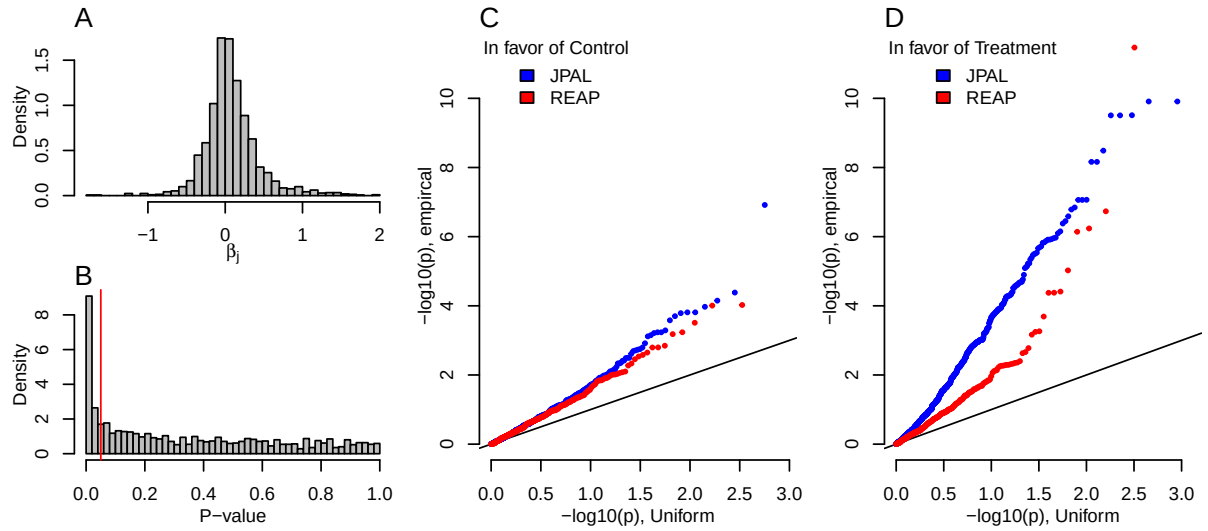
3.3 Preliminary Description of Data

We summarize data from the 84 tests considered here in Figure 1 (brief description of the intervention and target population of each RCT shown in Appendix B). In Panel A, we first consider the relative sample sizes. Most RCTs involve fewer than 5000 grade-level respondents (mean=3163) although a few involve much larger samples. In total, we analyze 7,244,566 item responses from 265,732 respondents (although, note that respondents are typically nested across multiple tests and, for the REAP studies, the studies utilize a common test across multiple studies). In Panel B, we examine the number of items and associated reliability [29]. Test outcomes with more items tended to be more reliable, but there was still variation in reliability even for a fixed number of items. Of the 84 tests, 76 had reliabilities above 0.7. Panel C shows effect sizes. We focus on effect sizes derived from the sum of correct responses. This is consistent with common treatment of these outcomes but note that other approaches may be superior in some cases [30].[7] Intervention effectiveness varied widely; the average intervention had an effect size of 0.015.[8] Finally, we also show the difficulties and discriminations for the 2,199 items we consider here in Panel D. The smaller variance of a Bernoulli random variable with expected value far from 0.5 leads to the curved shape observed here such that very hard or easy items tend to be somewhat less discriminating. Items with a given difficulty clearly exhibit a wide range of discriminations; this is relevant for understanding the overall reliability of the test as less discriminating items will tend to lead to less reliable tests.

[7]Note that this approach may not allow us to precisely cover treatment effects from original studies given that those might be based on, for example, multiple treatment arms. We view our approach as straightforward for the purpose of illustrating our key point regarding item-level sensitivity to treatment.

[8]When our goal is to discuss overall treatment effect, we rely on the difference in standardized sum scores as this is fairly conventional. However, the sample average of the $\widehat{\beta_j}$ offsets ($\sum_j \widehat{\beta_j}/N_j$) could also be considered as an estimate of treatment effect. These two metrics for treatment effect are strongly correlated ($r = 0.77$) but not identical due to the fact that the sum score and θ scales are monotonically but non-linearly related.

Figure 2: Summary of item-level DIF across all items. A: Histogram of $\widehat{\beta}_j$ values for both JPAL and REAP data. B: Histogram of P-values across both JPAL and REAP data (vertical red line represents 0.05 threshold). C & D: QQ plots for P-values separately for items in favor of control and treatment.



4 Results

4.1 Overall variation in $\widehat{\beta}_j$

We first consider estimates of β_j for all items; a histogram of these $\widehat{\beta}_j$ values is shown in Panel A of Figure 2. Estimates of β_j range from -1.73 to 1.98 (IQR=-0.10 to 0.22). A right skew is apparent in these values (skew=0.85); this is consistent with the notion that items tend to show be marginally easier for treatment respondents rather than marginally easier for control respondents; this is plausibly a consequence of the choice of measures that are relatively well-targeted for a given intervention. Values of β_j are challenging to interpret as they depend on both b_j and a in Eqn 3. Thus, to faciliate intuition, we also consider linear probability model-based estimates of a classic DIF model.[9] These estimates describe the expected change in a corect response due to group membership (here, treatment status) net of ability (as proxied by the sum score). Expected changes in the probability of a correct response associated with being in the treatment rather than the control groups range from -0.12 to 0.10 (IQR=-0.012, 0.012).

Across the REAP and JPAL datasets, 25% of items (527 of 2119) were flagged as exhibiting DIF at the $\alpha = 0.05$ level. DIF was identified more frequently in the JPAL data (28%) as compared to REAP (17%). The enrichment of small p-values is apparent in Panel B of Figure 2 (under the null, these p-values should be uniformly distributed). In Panels C and D, we consider QQ plots which compare p-values to those expected under the null model. We separately consider p-values from items that show DIF in favor of control ($\widehat{\beta}_j < c$) and those that show DIF in favor of the treatment (i.e., $\widehat{\beta}_j > c$); items favoring the treatment exhibit somewhat more extreme levels of DIF. We also consider the proportion of items flagged for DIF within a test after Bonferroni correction for the number of items on each test. As compared to the 527 of 2119 DIF based on the conventional $\alpha = 0.05$ level, only 178 were significant after performing a Bonferroni correction.[10] After Bonferonni correction, 45 tests have items exhibiting DIF.

We further consider the patterning of heterogeneity in sensitivity across item-level difficulty and discrimination estimates [28]. Item difficulty was negatively correlated ($r = -0.16$) with estimates of $\widehat{\beta}_j$ while

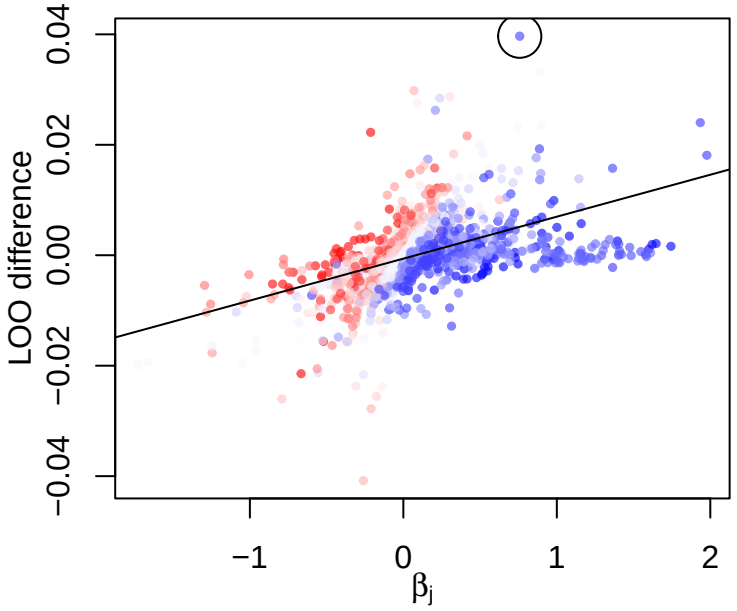
[9]Specifically we estimate

$$\mathbb{E}(Z_{ij}) = \alpha_{0,j} + \alpha_{1,j} \sum_j Z_{ij} + \alpha_{2,j}\delta_i \tag{5}$$

via OLS where $\alpha_{2,j}$ now describes the change in probability associated with group membership net of the sum score.

[10]The Bonferonni approach is fairly conservative compared to alternatives such as the Romano-Wolf [31] approach. However, we view it as sufficient for our purposes which is to indicate that there is still DIF after controlling for multiple testing.

Figure 3: LOO differences as a function of β_j. Dots are colored as a function of the intervention's effect size (dark red indicates interventions with more negative effects, darker blue indicates interventions with more positive effects). OLS regression line is shown.



item discrimination was positively correlated ($r = 0.29$) with $\widehat{\beta}_j$, see Table 1 (which also includes Spearman rank-order correlations). Alongside the correlations in the full sample, we also recomputed item difficulties and discriminations amongst the control-only respondents ($\widehat{\beta}_j$ are still computed as before in the whole sample). This analysis was motivated by an interest in observing whether relevant patterns are contingent on item responses observed following treatment. Correlations are somewhat attenuated in the control-only sample, but in general the consistency of the associations suggest that the pattern holds even in difficulty and discrimination estimates derived from non-experimental data.

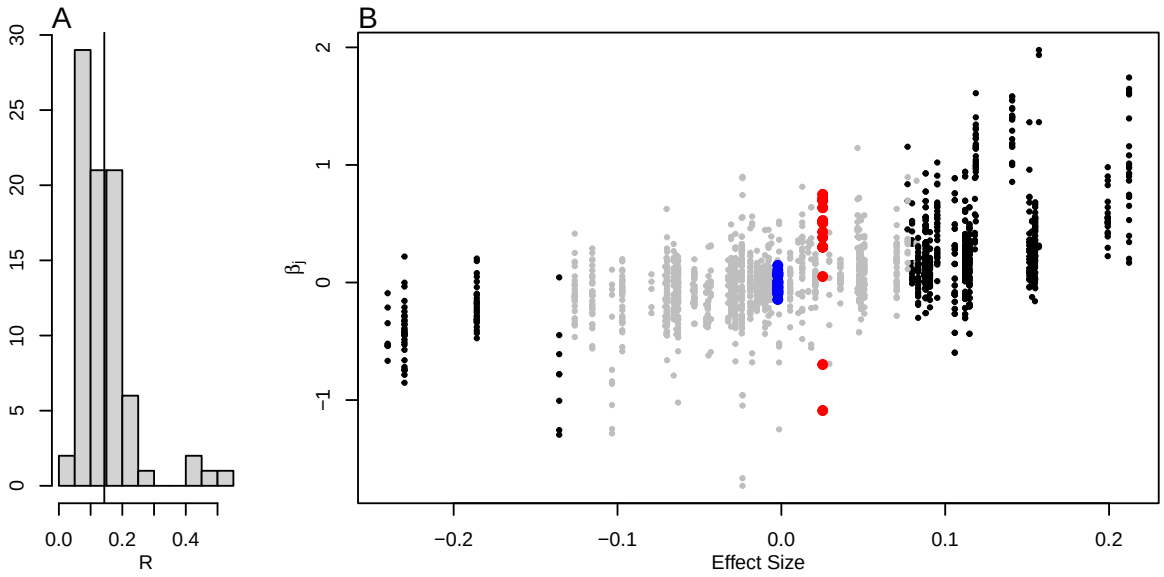
Table 1: Correlations between β_j and CTT-based item parameters

	All respondents		Controls only	
	Difficulty	Discrimination	Difficulty	Discrimination
Pearson	-0.16	0.29	-0.13	0.28
Spearman	-0.12	0.21	-0.09	0.19

Findings thus far suggest clear evidence of variation in sensitivity, as identified via $\widehat{\beta}_j$, across items. In particular, some items show substantial sensitivity to treatment (e.g., the skew in Figure 2 Panel A). To emphasize why this matters, we now consider an analysis focused on the role of $\widehat{\beta}_j$ in the estimated treatment effect. For each item, we first re-estimate the effect of treatment in the study after removing that item from the test: a "leave-one-out" (LOO) estimate. We then subtract this updated effect size from the original estimate of the treatment effect; a value of 0 would indicate that removing the item led to no difference in the estimated treatment effect of the resultant item response matrix. We consider these LOO differences as a function of β_j.

Findings are shown in Figure 3. There is a positive correlation between β_j and the LOO difference. This correlation is mechanical in the sense that inclusion of items with larger $\widehat{\beta}_j$ will tend to increase estimated treatment effects. But, as a point of emphasis, consider the point highlighted in the black circle in Figure 3. This item comes from an intervention with an effect size of 0.11 (p=0.036). However, after we remove this item, the adjusted effect size falls to 0.066. If we standardize the LOO difference by the original effect size, this single item accounts for 37% of the original effect. Without this item, a different inference about

Figure 4: A: Histogram of R values for all RCTs. B: $\widehat{\beta_j}$ values as a function of effect size (black dots represent RCTs with significant effect estimate).



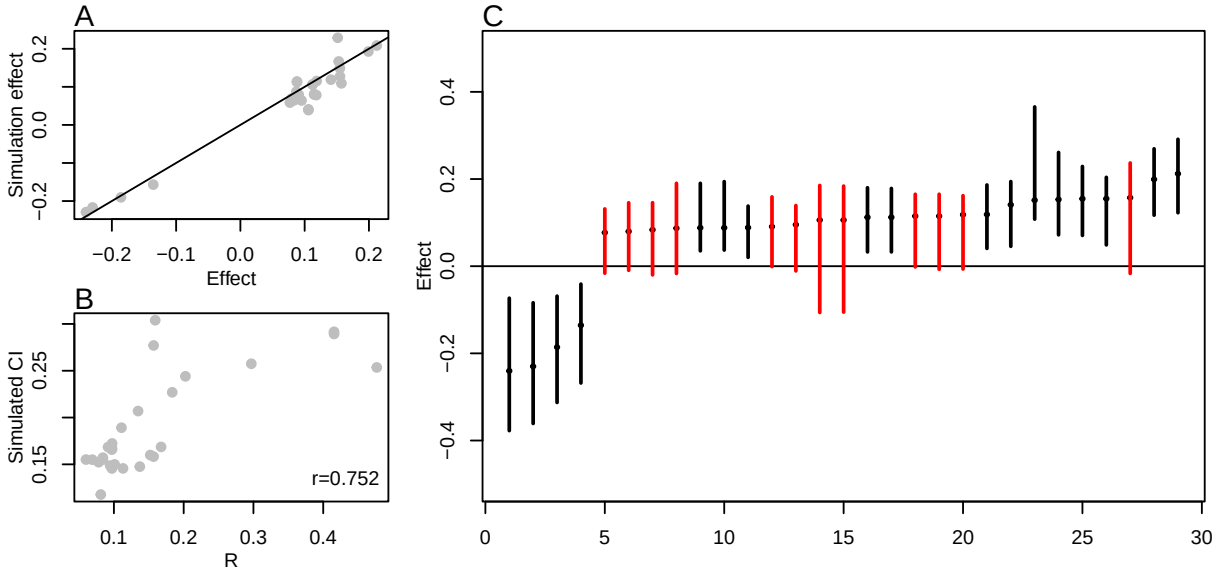
the relative efficacy of this intervention would result. While this might be a relatively extreme example (considering the value of this point on the y-axis), we argue that, given the time and money associated with both running the RCTs and subsequent operational decisions associated with their findings, this is a type of sensitivity that should be documented and scrutinized.

4.2 Test-specific variation in $\widehat{\beta_j}$

We now begin to focus on within-test variation in the $\widehat{\beta_j}$ estimates via the R quantities (Eqn 4). Recall that a perfectly homogeneous test would have $R = 0$; values of $R >> 0$ indicate that a test is highly heterogeneous such that there is wide item-level variation in the degree to which items indicate the existence of treatment effects. Note that a homogeneous test might be entirely insensitive or sensitive to treatment. Adjudication of that question would require substantive analysis of the content of the intervention and outcome measure; our emphasis is on the fact that the items all deliver identical information about the effect of the treatment. The average test has $R = 0.14$ but there is clear variation (see the histogram of R values in Panel A of Figure 4). The IQR spans from 0.09 to 0.17. For example, consider an intervention involving the delivery of eyeglasses to children that would benefit from them [18]. This intervention has, as we may anticipate, relatively small R values (0.066 and 0.076 for grades 1 and 2 respectively). We next utilize the variation in R to ask about the implication of variation in $\widehat{\beta_j}$ for our inferences related to treatment effects.

Figure 4 Panel B is a scatterplot of effect sizes for RCTs versus $\widehat{\beta_j}$ values (the x-axis values are common for a test thus leading to the vertical bands). RCTs with significant effects are shown in black. We emphasize our point via a focus on two RCTs with non-significant effects which we designate by colors (blue and red). Both RCTs had effect estimates that are near zero and insignificant; conventional analysis might stop there but we argue that more information can be ascertained through the kind of analysis performed here. The RCT shown in blue—the aforementioned eyeglass-related intervention [18] for second graders—has a set of items that are highly homogeneous in terms of sensitivity. In contrast, the RCT shown in red—a program that offered schools a mixture of unconditional grants and teacher incentives based on student performance [13]—is highly heterogeneous. We can similarly observe the difference in terms of R values which also account for possible differences in the variation in difficulties across the tests: $R = 0.076$ for blue and $R = 0.21$ for red. Whether the precise outcome measure used in the blue RCT is the optimal one would require detailed content analysis. However, this plot shows that we can be reasonably confident that a similar set of items would yield a similar inference. In contrast, inferences about the red RCT seem highly sensitive to the

Figure 5: Simulation analysis associated with significant empirical results. A: Empirical effect estimates versus average effect estimates across simulations. B: R values (Eqn 4) compared to width of simulated CIs. C: Treatment effects (points) along with 95% confidence intervals based on alternative outcomes created via resampling with replacement of items on the test.



specific configuration of items; i.e., if the two items with the lowest $\widehat{\beta}_j$ values had not been included on the assessment, it may have resulted in a significant effect estimate.

We build on this notion with a simulation exercise but focusing on those RCTs that had a significant treatment effect (again based on our calculation of the grade-specific effect). Conventional assessment of significance effectively treats the items as fixed but we can probe the degree to which variation in the architecture of the $\widehat{\beta}_j$ parameters may need to be accounted for if we aim to fully understand the uncertainty of the resulting effect estimate. Specifically, we ask about the degree to which heterogeneously sensitivity items can lead to uncertainty in estimates of treatment effectiveness using the simulation described in Section 3.2. For a given outcome's configuration of item-treatment interactions, we consider an experiment that asks about potential variation in outcomes for different configurations of items (obtained via resampling with replacement) in the outcome measure. For outcomes that display relatively homogeneous sensitivity across items, we expect relatively little additional uncertainty in effect estimates that might arise from considering somewhat different outcomes. For outcomes with more heterogeneous sensitivity across items, we anticipate relatively more uncertainty being introduced when we consider alternative possible forms.

Results of this simulation study are shown in Figure 5 which focuses on the 29 tests showing significant effects. Note first that simulated datasets have similar average effects as the empirical data (Panel A). As anticipated, larger R values translated into simulated CIs of larger width (Panel B).[11] We now re-examine effects after accounting for uncertainty related to variation across items of treatment effects (Panel C). Point estimates are shown as dots with associated uncertainty (i.e., the variation over the simulated datasets) shown as vertical bars. The uncertainty due to variation in item-level sensitivity was 19% larger on average than the uncertainty based on the standard error for the difference. This additional uncertainty has implications: in (12/29=) 41% of those cases, the CI related to item-level sensivitity suggests that these may be false positives (i.e., this CI included zero whereas the conventional CI did not).

5 Discussion

We argue that researchers and policymakers can move beyond using aggregate test scores to evaluate the effectiveness of RCT interventions by incorporating DIF analysis of item-level test data. Our work is similar

[11] Note that we compute standard errors here that do not account for potential clustering.

in spirit to other recent efforts suggesting that more insights can be gained from the results of RCTs by considering item-level variation rather than just overall test-level variation [21]. Using item-level data from a range of RCTs in education in LMICs, we find significant presence of DIF across test forms, which indicates that the effects of an intervention are not uniform across all of the items. Certain items are more sensitive to the intervention (those with DIF favoring the treatment group) and certain items are not as sensitive to the intervention (those with DIF favoring the control group). Analyses of the kind considered here can thus be used to inform policymakers and researchers about whether there are specific underlying academic skills that are more affected by certain interventions and what skills could be targeted in the future.

To illustrate, suppose we find DIF in specific math items with the treatment resulting in better performance on fraction items and not other skills. Subsequent research can then focus on whether this finding is likely due to the nature of the intervention and, if so, how it could be altered to lead to more generalized skill improvement in the future. Further, if the aim is to increase math ability overall, highlighting the presence of DIF can demonstrate the skill areas that future interventions may need to target (i.e., those not affected by the intervention). Examining DIF across multiple RCTs will provide a better understanding of whether certain types of interventions are more linked to improvements of specific subsets of academic skills.

As the use of RCTs in education gains in popularity, it is important to improve the outcome measures that are the crux of understanding the effectiveness of the intervention. Specifically, outcome measures that are more aligned to the intervention can better answer whether and how learning outcomes were improved. DIF analysis is one tool that can be used to explicate which items are more sensitive to an intervention and thus to demonstrate which skills were improved. Cost-effectiveness analysis of RCTs could use DIF analysis to show the effects across specific subdomains tested per dollar spent to determine how to target future education research and policy resources. Estimates of treatment effects may also need to more fully account for variation that may exist as a function of outcome-related sampling; our observation that all facets of variability may need to be considered is reminiscent of the insights from generalizability theory [32, 33]. The considerations raised here don't uniquely apply to RCTs. We have chosen to focus on RCTs given the unique opportunity in those settings for prospective choice of outcome. However, in many quasi-experimental studies, similar analysis of item-level sensitivities to specificities of the intervention could similarly be considered.

We acknowledge limitations. First, our results are based on an assumption that the variation in abilities is constant across groups. This assumption is required due to the fundamental identification problem associated with these kinds of analyses (see further discussion in [34, 35]); future work could probe the degree to which findings may be sensitive to violations of this assumption. Second, we focus on uniform DIF as opposed to non-uniform DIF (see Figure 7.6 in [20]). Addition of Eqn 5 to include a $X_i G_i$ interaction term would be one means of probing for non-Uniform DIF and other approaches exist as well [36]. However, we view Uniform DIF as both appropriate for this first analysis of this question and also focused on the central issue of whether items are targeted to treatment or control (rather than being more or less discriminating for treatment or control). Third, as we noted above, we do not attempt to perfectly replicate previous analysis meant to estimate treatment effects instead relying on a rough approach (e.g., group differences) that is easily applicable to the range of different RCTs that we consider. We have ignored issues associated with, for example, multiple treatment arms and randomization within blocks; our view is that these issues are largely orthogonal to our main point regarding between-item differences in sensitivity to treatment but it does suggest potential opportunities for additional refinement by further analyzing such sensitivities as a function of these complex experimental designs.

We note two additional caveats related to the simulation study. First, as just noted, we do not adjust for either pre-test scores or stratification. Such approaches would lead to improved precision of estimates. On the one hand, this might result in more studies being significant in Figure 5. On the other hand, our point is that conventional analysis overlooks a critical source of uncertainty and so from this perspective the improved precision that comes with other refinements (i.e., adjustment for pre-treatment performance) does not address our concern. Second, we are assuming that the reshuffling of items leads to plausible alternative outcomes. In our view this is consistent with the desire for "parallel forms" [28] and it is generally not implausible to imagine psychometric outcomes that utilize slightly different sets of items. But, in many cases, the outcomes were chosen due to content specifications that the reshuffled forms might not meet.

The choice of outcome for use in a given RCT is a difficult problem that will need to be addressed

holistically.[12] While the issues raised here cannot substitute for expert judgement by subject matter experts, we make some notes that might be considered in future attempts to choose outcome measures for education-related RCTs. First, if IRT approaches are used in future work to scale outcomes they may need to carefully consider whether previously acquired item parameters (i.e., parameters that come from a pre-existing item bank [37] or are otherwise based on data collected in other settings) apply to the current scenario. Such parameters would be similar to those estimated using only control respondents here and may clearly not generalize. Analyses of the kind considered here might clarify whether they can be safely be used. Second, how should items be selected? In an ideal case, we would suggest items that are aligned with the intervention and that themselves are predictive of future learning outcomes. Targeted interventions might be able to more nimbly select appropriate items emphasizing targeted skills whereas more generic interventions (e.g., distributing eyeglasses) might consider a larger number of items. The latter approach might end up using many items that are not well-aligned but this seems a small cost as compared to not having any aligned items if more targeted measures are used. Finally, what might considerations of possible variation across items in sensitivity to treatment suggest for experimental design and pre-registration? Analysts should discuss whether the outcome is fixed or random (i.e., [38]). If the items in the outcome measures are treated as random—that is, the items are viewed as one sample of items taken from a larger universe of possible items—consideration will need to be made regarding potential variation in uncertainty of the kind illustrated in Figure 5.

RCTs are a critical tool for understanding the efficacy of different ideas for improving educational outcomes. We encourage heightened attention be placed on outcome measures used in RCTs. The outcome measures are the lens through we are attempting to observe any effects of RCTs. Given both the need for better understanding of "what works" and the existence of data that would allow for more refined understanding, these outcomes deserve closer attention than they have historically received. This research is one step in that direction.

References

[1] Paul Connolly, Ciara Keenan, and Karolina Urbanska. 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, 2018.

[2] Esther Duflo and Abhijit Banerjee. *Poor economics*, volume 619. PublicAffairs New York, NY, USA, 2011.

[3] Paul Glewwe and Karthik Muralidharan. Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, volume 5, pages 653–743. Elsevier, 2016.

[4] Alejandro J Ganimian and Richard J Murnane. Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*, 86(3):719–755, 2016.

[5] Masha Bertling. Improving measures of student achievement in rcts in lmics. Manuscript in preparation, 2022.

[6] Institute of Education Sciences. What works clearinghouse standards handbook, version 4.1.

[7] Benjamin W Domingue, Laramie Duncan, Amal Harrati, and Daniel W Belsky. Short-term mental health sequelae of bereavement predict long-term physical health decline in older adults: Us health and retirement study analysis. *The Journals of Gerontology: Series B*, 76(6):1231–1240, 2021.

[8] Rukmini Banerji, James Berry, and Marc Shotland. The impact of mother literacy and participation programs on child learning: Evidence from a randomized evaluation in india. *Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL)*, 2013.

[12] See, for example, the considerations raised in the "Overalignment" discussion on page 84 of [6].

[9] James Berry, Dean Karlan, and Menno Pradhan. The impact of financial education for youth in ghana. *World Development*, 102:71–89, 2018.

[10] Moussa P Blimpo, David K Evans, and Nathalie Lahire. School-based management and educational outcomes: Lessons from a randomized field experiment. *Unpublished manuscript*, 7, 2011.

[11] Anca Dumitrescu, Dan Levy, Cara Orfield, and Matt Sloan. Impact evaluation of niger's imagine program final report september 13, 2011. 2011.

[12] Paul Glewwe, Nauman Ilias, and Michael Kremer. Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–227, 2010.

[13] Isaac Mbiti, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. Inputs, incentives, and complementarities in education: Experimental evidence from tanzania. *The Quarterly Journal of Economics*, 134(3):1627–1673, 2019.

[14] Fang Chang, Huan Wang, Yaqiong Qu, Qiang Zheng, Prashant Loyalka, Sean Sylvia, Yaojiang Shi, Sarah-Eve Dill, and Scott Rozelle. The impact of pay-for-percentile incentive on low-achieving students in rural china. *Economics of Education Review*, 75:101954, 2020.

[15] Prashant Loyalka, Anna Popova, Guirong Li, and Zhaolei Shi. Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, 11(3):128–154, 2019.

[16] Prashant Loyalka, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi. Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*, 37(3):621–662, 2019.

[17] Yue Ma, Robert W Fairlie, Prashant Loyalka, and Scott Rozelle. Isolating the "tech" from edtech: experimental evidence on computer assisted learning in china. Technical report, National Bureau of Economic Research, 2020.

[18] Yaojiang Shi, Wei Nie, Ming Mu, Shuyi Song, Lanxi Peng, Lifang Zhang, Jie Yang, Hongyu Guan, Yiqi Zhu, Qiufeng Gao, et al. What can children learn from a free trial of eyeglasses use? evidence from a cluster-randomized controlled trial in rural china. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 57:0046958020968776, 2020.

[19] Hongmei Yi, Di Mo, Huan Wang, Qiufeng Gao, Yaojiang Shi, Paiou Wu, Cody Abbey, and Scott Rozelle. Do resources matter? effects of an in-class library project on student independent reading habits in primary schools in rural china. *Reading Research Quarterly*, 54(3):383–411, 2019.

[20] Gregory Camilli. Test fairness. *Educational measurement*, 4:221–256, 2006.

[21] Luke W. Miratrix Joshua B. Gilbert, James S. Kim. Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging online formative assessments to pinpoint the impact of educational interventions.

[22] Wim J Van der Linden and RK Hambleton. Handbook of item response theory. *Taylor & Francis Group. Citado na pág*, 1(7):8, 1997.

[23] Item response theory. In RL Brennan, editor, *Educational measurement*.

[24] Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[25] Bengt Muthen and James Lehman. Multiple group irt modeling: Applications to item bias analysis. *Journal of Educational Statistics*, 10(2):133–142, 1985.

[26] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

[27] R Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29, 2012.

[28] Linda Crocker and James Algina. *Introduction to classical and modern test theory*. ERIC, 1986.

[29] G Frederic Kuder and Marion W Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[30] Joshua B. Gilbert. Estimating treatment effects with the explanatory item response model, November 2022.

[31] Joseph P Romano and Michael Wolf. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40, 2016.

[32] Robert L Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 1992.

[33] Richard J Shavelson, Noreen M Webb, and Glenn L Rowley. Generalizability theory. *American Psychologist*, 44(6):922, 1989.

[34] Peter F Halpin. Differential item functioning via robust scaling. *arXiv preprint arXiv:2207.04598*, 2022.

[35] Ben Stenhaug, Michael C Frank, and Ben Domingue. Treading carefully: Agnostic identification as the first step of detecting differential item functioning. 2021.

[36] R Philip Chalmers. Improving the crossing-sibtest statistic for detecting non-uniform dif. *Psychometrika*, 83(2):376–386, 2018.

[37] Timothy J Muckle. Web-based item development and banking. *Handbook of test development*, pages 257–274, 2015.

[38] Paul De Boeck. Random item irt models. *Psychometrika*, 73:533–559, 2008.

A Power Analysis

We conducted a power analysis to ensure that we are well-powered to detect variation in β_j given the sample sizes considered here. We simulate data as per our model. We use Eqn 3 with $\theta_i \sim \text{N}(0,1)$ for $i \in \{1, ..., N\}$, $b_j \sim \text{N}(0,1)$ for $j \in \{1, ..., 20\}$, $\delta_i \sim$ Bernoulli(0.5), and allow for variation in β_1 while all other $\beta_j = 0$ (we simulate 100 datasets for each set of simulation conditions). We then use the same approach to estimation and conduct inference on $\widehat{\beta_1}$.

Results are shown in Figure A.1. We obtain power above 0.8 for β_1 values bigger than roughly 0.2 with sample sizes of 5000 and around 0.5 for sample sizes of 1000 (these values of N can be compared to sample size numbers from Figure 2). For a sample of 2500—which is below the mean sample of 3163 here—we are powered to detect β_1 of roughly 0.3. To assess the magnitude of this value, consider a scenario where $b_1 = 0$, $\beta_1 = 0.3$, and $\theta \sim \text{N}(0,1)$. In that scenario, we would anticipate a marginal accuracy of 0.5 for one group versus 0.56 for the second.

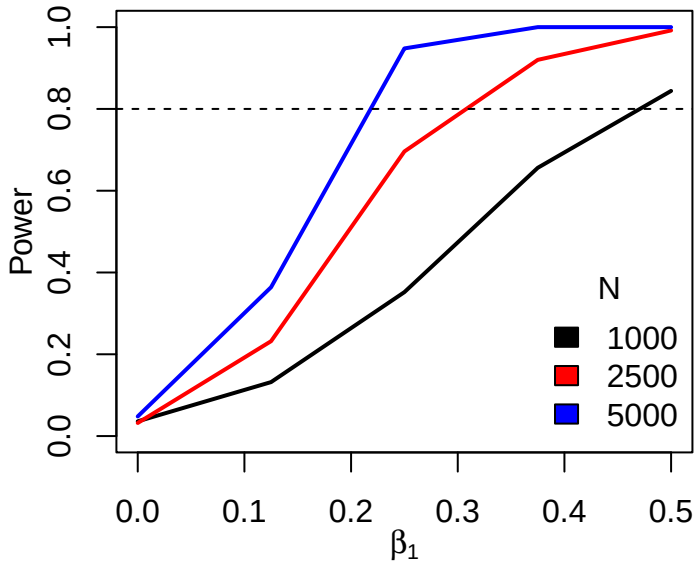


Figure A.1: Power analysis for $\widehat{\beta_1}$ for different values of N.

B Information on RCTs

Table B.1: JPAL Data

Reference	Intervention	Country
-Banerji, R., Berry, J., & Shotland, M. (2013). The impact of mother literacy and participation programs on child learning: Evidence from a randomized evaluation in India. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL).	Maternal Literacy	India
-Berry, J., Karlan, D., & Pradhan, M. (2018). The impact of financial education for youth in Ghana. World Development, 102, 71-89.	Financial Education	Ghana
-Blimpo, M., & Evans, D. (2013). School-based management and educational outcomes: lessons from a randomized field experiment (No. 81457, pp. 1-2). The World Bank.	School Management	Gambia
-Dumitrescu, A., Levy, D., Orfield, C., & Sloan, M. (2011). Impact evaluation of Niger's IMAGINE program (No. 9910b41833e34355bf9b2e2d7dc1e9c2). Mathematica Policy Research.	School Infrastructure	Niger
-Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. American Economic Journal: Applied Economics, 2(3), 205-27.	Teacher incentives	Kenya
-Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627-1673.	School Grants & Teacher Incentives	Tanzania

Table B.2: REAP Data

Reference	Intervention	School Level
-Chang, F., Wang, H., Qu, Y., Zheng, Q., Loyalka, P., Sylvia, S., ... & Rozelle, S. (2020). The impact of pay-for-percentile incentive on low-achieving students in rural China. Economics of Education Review, 75, 101954. -Liu, C., Liu, C., Loyalka, P., Shi, Y., & Sylvia, S. (2017) The distributional impacts of pay for percentile: Evidence from a randomized trial in rural China. Unpublished Report.	Teacher incentives	Primary
-Loyalka, P., Popova, A., Li, G., & Shi, Z. (2019a). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. American Economic Journal: Applied Economics, 11(3), 128-154.	Teacher training	Junior High
-Loyalka, P., Sylvia, S., Liu, C., Chu, J., & Shi, Y. (2019b). Pay by design: Teacher performance pay design and the distribution of student achievement. Journal of Labor Economics, 37(3), 621-662.	Teacher incentives	Primary
-Ma, Y., Fairlie, R. W., Loyalka, P., & Rozelle, S. (2020). Isolating the "Tech" from Edtech: experimental evidence on computer assisted learning in China (No. w26953). National Bureau of Economic Research.	EdTech	Primary
-Shi, Y., Nie, W., Mu, M., Song, S., Peng, L., Zhang, L., ... & Nie, J. (2020). What Can Children Learn from a Free Trial of Eyeglasses Use? Evidence from a Cluster-Randomized Controlled Trial in Rural China. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 57, 0046958020968776.	Eyeglasses	Primary
-Yi, H., Mo, D., Wang, H., Gao, Q., Shi, Y., Wu, P., ... & Rozelle, S. (2019). Do resources matter? Effects of an in-class library project on student independent reading habits in primary schools in rural China. Reading Research Quarterly, 54(3), 383-411.	Academic	Primary
-Follow-up of Yi et al. (2019)	Academic	Primary
-Follow-up of Ma et al. (2020)	EdTech	Primary

Table B.3: Summary of RCT data.

RCT	Domain	Grade	N item	N people
Mbiti et al. (2019)	literacy	1	15	2904
Mbiti et al. (2019)	literacy	2	24	2916
Mbiti et al. (2019)	literacy	3	30	2914
Mbiti et al. (2019)	literacy	1	19	3023
Mbiti et al. (2019)	literacy	2	27	2979
Mbiti et al. (2019)	literacy	3	31	2967
Mbiti et al. (2019)	math	1	18	3089
Mbiti et al. (2019)	math	2	21	3010
Mbiti et al. (2019)	math	3	27	2992
Berry et al. (2018)	literacy	unknown	5	2759
Berry et al. (2018)	math	unknown	5	2759
Berry et al. (2018)	literacy	unknown	5	2532
Berry et al. (2018)	math	unknown	5	2532
Blimpo & Evans (2013)	literacy	4	55	776
Blimpo & Evans (2013)	literacy	6	55	861
Blimpo & Evans (2013)	literacy_oral	4	8	814
Blimpo & Evans (2013)	literacy_oral	6	8	890
Blimpo & Evans (2013)	math	4	32	781
Blimpo & Evans (2013)	math	6	32	861
Banerji et al. (2013)	literacy	1	10	3747
Banerji et al. (2013)	literacy	2	11	2979
Banerji et al. (2013)	literacy	3	10	2791
Banerji et al. (2013)	literacy	4	8	2215
Banerji et al. (2013)	literacy	5	8	2010
Banerji et al. (2013)	math	1	4	3747
Banerji et al. (2013)	math	2	5	2979
Banerji et al. (2013)	math	3	5	2791
Banerji et al. (2013)	math	4	5	2215
Banerji et al. (2013)	math	5	5	2010
Glewwe et al. (2010)	literacy	3	53	4376
Glewwe et al. (2010)	literacy	4	56	3906
Glewwe et al. (2010)	literacy	5	55	3127
Glewwe et al. (2010)	literacy	6	46	2742
Glewwe et al. (2010)	literacy	7	34	2748
Glewwe et al. (2010)	literacy	8	22	1549
Glewwe et al. (2010)	literacy	3	53	4376
Glewwe et al. (2010)	literacy	4	56	3906
Glewwe et al. (2010)	literacy	5	55	3127
Glewwe et al. (2010)	literacy	6	46	2742
Glewwe et al. (2010)	literacy	7	34	2748
Glewwe et al. (2010)	literacy	8	22	1549
Glewwe et al. (2010)	math	3	31	4500
Glewwe et al. (2010)	math	4	32	3907
Glewwe et al. (2010)	math	5	32	3127
Glewwe et al. (2010)	math	6	32	2814
Glewwe et al. (2010)	math	7	28	2731
Glewwe et al. (2010)	math	8	27	1599
Glewwe et al. (2010)	math	5	26	3119
Glewwe et al. (2010)	math	6	36	2764
Glewwe et al. (2010)	math	7	38	2730
Glewwe et al. (2010)	math	8	38	1599
Dumitrescu et al. (2011)	literacy	1	5	2009
Dumitrescu et al. (2011)	literacy	2	8	2769
Dumitrescu et al. (2011)	literacy	3	8	1989
Dumitrescu et al. (2011)	literacy	4	8	1530
Dumitrescu et al. (2011)	literacy	5	8	1158

Dumitrescu et al. (2011)	literacy	6	8	1297
Dumitrescu et al. (2011)	math	0	11	5233
Dumitrescu et al. (2011)	math	1	11	2009
Dumitrescu et al. (2011)	math	2	11	2766
Dumitrescu et al. (2011)	math	3	11	1989
Dumitrescu et al. (2011)	math	4	11	1529
Dumitrescu et al. (2011)	math	5	11	1157
Dumitrescu et al. (2011)	math	6	10	1295
Loyalka et al. (2019a)	math	7	30	16746
Loyalka et al. (2019a)	math	8	30	12740
Loyalka et al. (2019a)	math	9	30	4094
Ma et al. (2020)	math	4	30	3030
Ma et al. (2020)	math	5	30	3588
Ma et al. (2020)	math	6	30	3965
Follow-up of Ma et al. (2020)	math	4	47	2693
Follow-up of Ma et al. (2020)	math	5	46	2891
Follow-up of Ma et al. (2020)	math	6	47	3063
Yi et al. (2019)	math	3	30	713
Yi et al. (2019)	math	4	30	671
Yi et al. (2019)	math	5	30	686
Yi et al. (2019)	math	6	30	1428
Follow-up of Yi et al. (2019)	math	4	30	2552
Follow-up of Yi et al. (2019)	math	5	30	2665
Shi et al. (2020)	math	4	30	8424
Shi et al. (2020)	math	5	30	8882
Liu et al. (2017)	math	7	33	12139
Loyalka et al. (2019b)	math	6	30	8559
Chang et al. (2020)	math	5	30	3844