Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1082

Mechanisms of Effect Size Differences Between Researcher Developed and Independently Developed Outcomes: An Item-Level Meta-Analysis

Joshua B. Gilbert
Massachusetts General Hospital

James G. Soland
University of Virginia

Differences in effect sizes between researcher developed (RD) and independently developed (ID) outcome measures are widely documented but poorly understood in education research. We conduct a meta-analysis using item-level outcome data to test potential mechanisms that explain differences in effects by RD or ID outcome type. Our analysis of 54 effect sizes from 32 studies shows that greater standard deviations of item-specific treatment effects predict larger effect sizes and reduce the observed difference between RD and ID measures from .14 SDs to .12 SDs, leaving the remainder attributable to other factors. The findings advance our understanding of how item properties predict educational intervention outcomes and underscore the importance of considering item-level data for building theory in education research.

VERSION: July 2026

Suggested citation: Gilbert, Joshua B., and James Soland. (2026). Mechanisms of Effect Size Differences Between Researcher Developed and Independently Developed Outcomes: An Item-Level Meta-Analysis. (EdWorkingPaper: 24-1082). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/8axs-y713

Mechanisms of Effect Size Differences Between Researcher Developed and Independently Developed Outcomes: An Item-Level Meta-Analysis

Joshua B. Gilbert [1] and James Soland [2]

[1]Center for School Behavioral Health, Massachusetts General Hospital
[2]University of Virginia School of Education and Human Development

July 20, 2026

Abstract

Differences in effect sizes between researcher developed (RD) and independently developed (ID) outcome measures are widely documented but poorly understood in education research. We conduct a meta-analysis using item-level outcome data to test potential mechanisms that explain differences in effects by RD or ID outcome type. Our analysis of 54 effect sizes from 32 studies shows that greater standard deviations of item-specific treatment effects predict larger effect sizes and reduce the observed difference between RD and ID measures from .14 SDs to .12 SDs, leaving the remainder attributable to other factors. The findings advance our understanding of how item properties predict educational intervention outcomes and underscore the importance of considering item-level data for building theory in education research.

Keywords: meta-analysis, moderation, psychometrics, item analysis, randomized controlled trials

Corresponding author: joshua_gilbert@g.harvard.edu

Forthcoming in *Multivariate Behavioral Research*, DOI: 10.1080/00273171.2026.2707650

1 Introduction

When evaluating the efficacy of an educational intervention, researchers must choose an outcome measure as a metric for success. An important decision point in program evaluation is whether the researchers design their own outcome measure or use an existing, independently developed measure. The inherent trade-off arising from this choice is that while a researcher developed (RD) measure may be more aligned with the substantive content and goals of an intervention, researchers also risk constructing measures that are over-aligned with the intervention, producing large effects that fail to generalize more broadly to outcome measures another researcher may have chosen or to independently developed (ID) measures. Conversely, ID measures may better reflect broader content domains, but may be under-aligned with the intervention and fail to capture effects of substantive interest (Francis et al., 2022). Thus, understanding the extent to which the effects of educational interventions are dependent on the outcome measure has posed a long-standing challenge in interpreting intervention research in education.

Empirically, differences in intervention effect sizes between RD and ID outcome measures are widely documented in education research, with RD measures consistently showing substantially larger effect sizes across a wide range of contexts (Wolf & Harbatkin, 2023). The central question motivating the present study is what explains this large difference in intervention effect sizes. Several meta-analyses have explored this question by examining study characteristics that also moderate effect sizes, such as implementation fidelity, study quality, or participant characteristics, given that RD versus ID outcome type may be correlated with these study features. However, these meta-analyses show that outcome type remains a significant moderator even after conditioning on these study design and implementation factors (Cheung & Slavin, 2016; J. Kim et al., 2021; Kraft, 2020; Wolf & Harbatkin, 2023). Thus, our understanding of the mechanisms underlying the moderation of effect sizes by RD versus ID outcome type remains incomplete.

While valuable, the standard approach of using study-level characteristics as additional moderator variables in meta-analyses overlooks a potentially informative data source: the individual

participants' item response data from each outcome measure. Typically, meta-analysts collect aggregate data on effect sizes from a range of studies, and the effect sizes then serve as outcome variables in meta-regression models as functions of other study characteristics. While analyzing item response data from outcome measures could enable additional tests of specific mechanisms that may explain moderation by outcome type, it is relatively uncommon for study replication materials to include participants' item responses (Domingue et al., 2025). Consequently, examining item characteristics and the outcome's psychometric properties as potential moderators is rare (with some exceptions, see J. B. Gilbert, Young, et al., 2026), as empirical item-level meta-analyses tend to focus on the psychometric properties of single assessments across a range of studies and populations (e.g., Anglim et al., 2022; Scherer and Campos, 2022).

There are many reasons to suspect that item-level analyses could shed new light on differences in effect sizes between RD and ID measures. For example, the content of RD measures may be more closely aligned with the intervention than when ID measures are used. If only some items from ID measures parallel the intervention, then one might observe treatment effects on specific items that are aligned, but not on others. At the same time, if researchers leading the intervention focus particularly on a few key concepts (for example, using the specific language that shows up in an item they in fact wrote), those items might see very large effects, but other items might see smaller effects, even for an RD measure. Such hypotheses have not been fully tested, in part due to the relative inaccessibility of item-level outcome data from randomized controlled trials (RCTs).

In this study, we leverage item-level data from 54 outcomes from 32 RCTs in education to propose and test novel hypotheses about mechanisms explaining effect size moderation by outcome type. By analyzing the item-level outcome data directly, we can gain new insight into potential mechanisms underlying outcome type moderation because such mechanisms predict distinct empirical patterns that are observable in the item responses, but would be masked by a point estimate of the treatment effect that is the typical unit of analysis.

In short, our results show that the standard deviation (SD) of item-specific treatment effects around the average positively moderates effect sizes, and controlling for this moderator explains

about 15% of the observed outcome type moderation. Other characteristics, such as the internal consistency of the outcome, or the SD of item easiness parameters, do not significantly moderate effect sizes in this sample of studies. Our results underscore the potential contributions of item-level data analysis to meta-analysis of educational interventions more broadly.

2 Background

2.1 Prior Research on Outcome Type Moderation

A growing body of literature shows that treatment effect sizes in educational interventions tend to be larger, on average, for RD versus ID measures. For example, effect sizes for RD measures, which are presumably more aligned (or potentially over-aligned) with the treatment, are often twice as large as effect sizes derived from ID measures (Cheung & Slavin, 2016; Hill et al., 2008; Lipsey et al., 2012; Lynch et al., 2019). In a comprehensive study, Wolf and Harbatkin (2023) use data from the What Works Clearinghouse to examine average effect sizes by outcome type. Controlling for study quality and characteristics, they find larger mean effect sizes for RD measures than for ID measures. Their results suggest that larger effect sizes for RD measures are not driven by differences in implementation fidelity or study quality, nor by characteristics of the intervention or sample.[1]

Kraft (2020) examines these issues in the context of empirical benchmarks for interpreting effect sizes in education research. He points out that when producing effect sizes, empirical benchmarks, alignment between the construct being measured and the intervention, or both, are key considerations. While he reviews research suggesting that RD measures can yield tighter alignment, he also argues that larger effects on RD measures can be misleading when they capture impacts on narrow knowledge and skills that are not easily generalizable. J. Kim et al. (2021) draw similar conclusions in a meta-analysis of educational apps for children in pre-K to grade 3. In particular, they show that effects are larger for RD versus ID measures, and that this result holds even when

[1]Note that prior studies use varying terminology, such as "narrow" vs. "broad" measures, "independent" vs. "non-independent", RD vs. "standardized", or RD vs. ID. While the definitions vary, all studies target similar general trends. We use the terms RD and ID throughout.

controlling for whether the outcome tested a constrained or unconstrained skill (e.g., letter naming vs. vocabulary, respectively), which also moderates effect sizes.

Given the consistency of moderation of effect sizes by outcome type across a wide range of studies, this body of work raises important questions for understanding what works in education. Without understanding the moderating effects of outcome type on intervention results, major challenges arise when trying to compare the effectiveness of studies that use RD versus ID measures. Further, when trying to use empirical benchmarks to articulate the practical significance of a given finding, these comparability issues mean researchers probably should not use the same benchmarks for all outcome types. In short, for all the attention paid to best practices in evaluating studies, including those contained in resources like the What Works Clearinghouse (What Works Clearinghouse, 2022), the properties of the measure used to produce the dependent variable often receive little attention, including how the measure was developed and by whom.

2.2 Leveraging item-level outcome data in RCTs

In RCTs of educational interventions, the typical outcome variable is some sort of summary score, such as a sum score or IRT-based scaled score (J. Gilbert et al., 2026). Recent scholarship argues for the affordances of item-level analysis of RCT outcome measures to provide more fine-grained insight into the nature of intervention impacts. For example, Ahmed et al. (2025) examine item-level outcome data from 15 RCTs in education and demonstrate that impact estimates can be highly sensitive to the included items, including one case in which exclusion of a single item reduced the treatment effect size by 40%. Similarly, Student (2026) uses a moderated non-linear factor analysis approach to identify and adjust for items that show treatment effects that differ from what would be expected from an impact on the target construct. Related work demonstrates that when treatment effects vary across the items of the outcome measure, standard errors become inflated because of the added uncertainty of which items were selected for test administration and that correlations between item easiness parameters and item-specific treatment effects can create bias in treatment by

covariate interaction effects (J. B. Gilbert, Himmelsbach, et al., 2025; J. B. Gilbert, Miratrix, et al., 2025).

Most similar to the present study, Halpin and Gilbert (2024) propose a statistical method to distinguish treatment effects directly on the target construct that in principle will generalize to other measures of the same construct from item-specific effects that will not generalize. The authors then show how their approach partially explains moderation by RD versus ID outcome type in an empirical application to a set of RCTs. In particular, the authors argue that items that are more sensitive to treatment are more likely to be selected (deliberately or incidentally) by researchers implementing an intervention, which could produce a subset of items with large but idiosyncratic effects. However, the authors only examine their proposed test statistic as a single potential moderator of effect sizes. In the present study, we build on this work to examine how other item characteristics may provide new insights into effect moderation in educational interventions more broadly.

3 Methods

3.1 Data Source

We draw our sample from prior studies of item-level data from a large set of RCTs (Domingue et al., 2025; J. B. Gilbert, Himmelsbach, et al., 2025). These data come from two sources. First, J. B. Gilbert, Himmelsbach, et al. (2025) examine item-level heterogeneous treatment effects—a statistical modeling approach that allows for unique treatment effects on each item—in 75 outcomes from 48 RCTs in education, economics, political science, and health. Second, further RCTs with item-level data are available in the Item Response Warehouse (IRW) (Domingue et al., 2025). Here, we draw data from both sources to analyze 54 outcomes from 32 RCTs examining educational interventions and outcomes. We focus on education in large part because RD versus ID outcome type moderation is most documented and of theoretical interest in education. The datasets, covering

a wide range of geographic regions, assessed outcomes, disciplinary traditions, and age groups, are summarized in Table 1.[2]

Table 1: Descriptive statistics for the datasets in our analysis

ID	Study	Location	Subjects	Items	Population	Outcome	Type
1	J. B. Gilbert et al., 2023	USA	7797	30	G3	Reading Comprehension	RD
2	J. S. Kim et al., 2023	USA	2174	20	G2	Reading Comprehension	RD
5	Woods-Townsend et al., 2021	UK	2486	7	Adolescents	Health Literacy	RD
6	Bruhn et al., 2016	Brazil	15395	10	Adolescents	Financial Literacy	RD
7	J. S. Kim et al., 2024	USA	1352	36	G3	Vocabulary	RD
8	J. S. Kim et al., 2024	USA	1303	29	G3	Reading Comprehension	RD
11	J. S. Kim et al., 2021	USA	2565	24	G1	Vocabulary	RD
12	J. S. Kim et al., 2021	USA	2580	24	G2	Vocabulary	RD
13	Romero et al., 2020	Liberia	3381	20	Elementary	Literacy	RD
14	Romero et al., 2020	Liberia	3381	44	Elementary	Math	RD
15	Romero et al., 2020	Liberia	3381	10	Elementary	Raven's Progressive Matrices	ID
16	de Barros et al., 2024	India	3202	32	G4	Math	RD
17	A. Duflo et al., 2024	Ghana	17344	21	G1-G3	Math	RD
18	A. Duflo et al., 2024	Ghana	17344	21	G1-G3	English	RD
19	A. Duflo et al., 2024	Ghana	17331	21	G1-G3	Local Language	RD
20	Jayanthi et al., 2021	USA	186	93	G5	Math	RD
21	Davenport et al., 2023	USA	3671	13	G5	Math	RD
23	Bang et al., 2023	USA	886	38	K-G1	Math	ID
24	Llauradó et al., 2014	Spain	495	13	Elementary	Dietary Behavior	RD
26	Schreinemachers et al., 2020	Nepal	775	15	Children 8-12	Food Knowledge	RD
29	Banerji et al., 2017	India	14576	15	Children	Language	ID
30	Banerji et al., 2017	India	14576	10	Children	Math	ID
31	E. Duflo et al., 2015	India	11893	6	Elementary	Academic Achievement	ID
32	Maruyama, 2022	El Salvador	3619	20	G7	Math	RD
36	Persson et al., 2020	Sweden	1108	7	High School	Political Knowledge	RD

[2]While some of these studies come from disciplines outside of education narrowly construed, the methods used in these disciplines are often employed in education, and related outcomes are of substantive interest to educators. For example, principles of economics are often used to understand phenomena like teacher turnover, and health measures are used to quantify the psychological and social-emotional needs of students. More generally, our research is meant to shed light on how RD versus ID measures impact intervention studies, a topic of demonstrated interest to educational evaluators given how sensitive results in the What Works Clearinghouse are to the type of measure used (Wolf & Harbatkin, 2023).

39	Berry et al., 2022	Malawi	6196	10	G5-G8	Cognitive Test	RD
40	Berry et al., 2022	Malawi	6188	20	G5-G8	Computation	RD
41	Mohohlwane et al., 2024	South Africa	3068	134	Early Elementary	Oral Reading Fluency	ID
46	Glatz et al., 2023	Netherlands	120	42	G1	Language	RD
55	Wang et al., 2024	Bangladesh	1704	15	Elementary	Academic Achievement	RD
56	Sebele et al., 2025	Liberia	2307	4	Preschool	Literacy	RD
68	Banerjee et al., 2017	India	5974	35	G1-G4	Hindi	ID
69	Banerjee et al., 2017	India	5966	30	G1-G4	Math	ID
70	Banerjee et al., 2017	India	3543	24	G1-G2	Hindi	ID
71	Banerjee et al., 2017	India	3448	20	G1-G2	Math	ID
72	Banerjee et al., 2017	India	2669	35	G3-G5	Hindi	ID
73	Banerjee et al., 2017	India	2682	30	G3-G5	Math	ID
74	J. B. Gilbert, Kim, and Miratrix, 2024	USA	1225	12	G2	Vocabulary	RD
76	Thai et al., 2022	USA	428	78	K	Math	ID
77	Cabell et al., 2025	USA	1081	26	K	Language Fundamentals	ID
78	Cabell et al., 2025	USA	1082	186	K	Vocabulary	ID
79	Cabell et al., 2025	USA	1107	30	K	Narrative Language	ID
80	Cabell et al., 2025	USA	1080	35	K	Vocabulary	ID
81	Cabell et al., 2025	USA	1080	18	K	Science	ID
82	Cabell et al., 2025	USA	1080	18	K	Social Studies	ID
100	J. B. Gilbert, Domingue, and Kim, 2026	USA	2118	12	G2	Vocabulary	RD
102	Relyea et al., 2025	USA	1743	36	G3	Vocabulary	RD
103	Relyea et al., 2025	USA	1743	29	G3	Reading Comprehension	RD
104	Relyea et al., 2025	USA	1743	9	G3	Background Knowledge	RD
109	Mosher and Kim, 2025	USA	936	9	G3	Recall	RD
110	Mosher and Kim, 2025	USA	929	9	G3	Reading Comprehension	RD
111	Mosher and Kim, 2025	USA	907	9	G3	Reading Comprehension	RD
112	De Weerdt et al., 2026	Belgium	243	9	G5-G6	Science (Forces)	RD
113	De Weerdt et al., 2026	Belgium	239	9	G5-G6	Science (DNA)	RD

Notes: We exclude dataset 47 from our analysis because it examines a math outcome in an intervention focused on literacy. Dataset 46 represents the literacy outcome from this study. We use the dataset identifiers from the IRW (Domingue et al., 2025) to facilitate replicability. G = grade.

3.2 Meta-regression Models

We use Bayesian meta-regression to model predictors of intervention effect size using the R software `brms` (Bürkner, 2021). The outcome is the covariate-adjusted Cohen's *d* derived from a latent

variable model fit directly to the item responses. (See Appendix A for the mathematical model and the `lme4` code for estimating treatment effects from the individual participant item response data and Appendix B for example `brms` meta-regression code.)

We estimate meta-regression models of the following general form:

$$\delta_{ij} = \mathbf{X}_{ij}\beta + u_j + e_{ij} \tag{1}$$

$$u_j \sim N(0, \tau^2) \tag{2}$$

$$e_{ij} \sim N(0, s_{ij}^2), \tag{3}$$

where δ_{ij} is the true treatment effect size in dataset i in study j. δ_{ij} is in turn a function of a matrix of moderator variables $\mathbf{X}_{ij}$ multiplied by a vector of regression coefficients β, a random effect for study u_j, and the residual e_{ij}. τ^2 represents the between-study variance and s_{ij}^2 is the (known) variance of each effect size. This approach allows us to account for both studies that contribute a single effect size to the analysis and studies that contribute multiple effect sizes. We use a Bayesian approach because of the small sample size and the ability to incorporate priors into our analysis. We use moderately informative priors of $N(0, 1)$ for the intercept, $N(0, .5)$ for regression coefficients, and half-normal$(0, .5)$ for the variance parameter because effect sizes in education research tend to be small and moderator effects greater than 1 SD would be unlikely (Kraft, 2020). Note that, while some research examines the effects of individual item characteristics on patterns of treatment effects within a single study (e.g., J. B. Gilbert, 2024; J. B. Gilbert, Hieronymus, et al., 2024; J. B. Gilbert, Kim, and Miratrix, 2024; J. B. Gilbert et al., 2023), here we are modeling outcomes at the dataset level. The model is structured in this way because our item characteristics (described shortly) are summarized across a dataset. For example, the correlation between item easiness and the item-specific treatment effect is estimated by dataset, not by item.

3.3 Moderators

We examine the following moderator variables. We focus primarily on item characteristics because intervention and participant characteristics have already been studied in prior meta-analyses of educational interventions, including to understand RD versus ID differences (e.g., J. Kim et al., 2021; Kraft, 2020; Wolf and Harbatkin, 2023).

3.3.1 Outcome Characteristics

Our focal moderator is an indicator variable for whether the outcome measure is researcher developed (RD) or independently developed (ID), determined by our review of study materials and/or data replication files. We coded outcomes as ID when the outcome was not developed explicitly for the purposes of the study. Examples of ID measures include the *Annual Status of Education Report*, *Raven's Progressive Matrices*, and the *Peabody Picture Vocabulary Test* assessments. For the small proportion of studies that did not include sufficient empirical information to identify a specific ID measure after an extensive review of study materials and existing assessments, we assumed the measure to be RD. One way we test hypotheses in this study is by examining the coefficient on this ID versus RD moderator in isolation, then including other moderators to see if they reduce the magnitude of the ID versus RD effect. Next, we describe several of those other moderators.

3.3.2 Study Characteristics

While study characteristics are not our primary focus, we nonetheless control for some related facets of the intervention, including ones germane to how the outcome measure was constructed. Specifically, we examine sample size (both number of participants and number of items) as potential moderators. Because of the severe right skew in these variables, we apply a $\log_2$ transformation. Therefore, the coefficients can be interpreted as the predicted difference in effect size for a doubling of the person or item sample size.

3.3.3 Item Characteristics

We examine several item characteristics as potential moderators. For each moderator, we provide an explanation of a potential mechanism that could predict treatment effect sizes. However, as will become clear shortly, these characteristics do not always lend themselves to obvious hypotheses about mechanisms. Almost all of these hypotheses relate in some way to (mis)alignment between what is tested and what the intervention is designed to impact. For example, tighter alignment between the content on the assessment and the intervention will, all else equal, lead to a larger effect size across items. The extent to which this mechanism explains differences between RD versus ID measures would in turn depend on how the degree of alignment differs between RD and ID measures. In what follows, we emphasize how item characteristics might predict the effect size, rather than hypothesize about how these item characteristics are correlated with the type of measure used. Our item characteristic moderators include the following and are summarized in Table 2. We emphasize that the possibilities we explore here are not intended to be exhaustive, but rather illustrative of the affordances of examining item characteristics as potential effect size moderators.

1. **The internal consistency of the outcome (Cronbach's α).** α is an estimate of reliability, the proportion of observed score variance accounted for by true score variance—equivalently, the expected correlation between replications of measurement procedures—that ranges from 0 to 1. α therefore provides a rough proxy for the psychometric quality of the outcome measure. Low reliability will mechanically attenuate effect sizes derived from observed scores, because measurement error inflates the standard deviation of the observed scores in the denominator of the effect size calculation (Hedges, 1981). However, this attenuation is already at least partially accounted for by using effect sizes derived from a latent variable model that adjusts for the unreliability of the outcome (J. B. Gilbert, 2025; Soland, 2022; Soland et al., 2024). Thus, our use of a latent variable model to estimate the average treatment effects in each dataset already corrects for any mechanical attenuation of the effect size due to unreliability (see Appendix A). Even with the mechanical effects accounted for, α could

still moderate effect sizes for more substantive reasons. That is, all else equal, higher values of α could emerge from (a) more items, (b) higher inter-item correlations, (c) measures of narrower content areas, and (d) more heterogeneous populations (because the numerator of the reliability coefficient contains the true score variance). Higher α arising from narrower content areas covered by an assessment may be easier to experimentally manipulate. For example, a higher effect size could result from a researcher designing a measure of a narrow content area—yielding a high α—more tightly linked to a narrow intervention. In contrast, higher α arising from more heterogeneous populations may be harder to experimentally manipulate. Thus, the direction of the relationship between α and disattenuated effect sizes derived from a latent variable model is not straightforward.

2. **The standard deviation of item-level heterogeneous treatment effects (IL-HTE).** A body of work—including the primary source of the data sets for the present study—has examined how individual items comprising the outcome measure may be differentially affected by treatment (Ahmed et al., 2025; J. B. Gilbert, Himmelsbach, et al., 2025; J. B. Gilbert et al., 2023). The degree of item-level heterogeneous treatment effects (IL-HTE) in a dataset is captured by the SD of item-specific effects around the average effect. If the value is 0, it means that all items are equally affected by treatment, as would be predicted from a treatment effect directly on the latent construct (Halpin & Gilbert, 2024). Larger values indicate greater heterogeneity in the item sensitivity to treatment. IL-HTE could be caused by, for example, over- or under-alignment between the intervention and the items of the measure, and therefore may be predictive of larger or smaller effect sizes depending on the direction of the alignment. While high IL-HTE SDs could occur when ID measures are not aligned well with the intervention, they could also occur when researchers select items that are more sensitive to the treatment for their RD measure, but researchers are imperfectly able to select items with larger effects in advance (yielding heterogeneity of item-level effects). Regardless of the specific cause, the likely culprit in the context of our study is variability in alignment between the measure and the intervention.

3. **The correlation between item-specific effect size and item easiness.** If items are differentially sensitive to treatment, it is possible that effects would be concentrated on the easier or harder items of the outcome. This would result in a positive or negative correlation between treatment effect size and item easiness, respectively (J. B. Gilbert, Miratrix, et al., 2025). A large correlation could be induced by, for example, incentive structures that encourage teachers to focus on bringing students above a proficiency threshold, causing teachers to focus on specific content areas or subscales from a broader domain. If the intervention targets lower-performing students, then there may be a correlation between the easiness of the item (with easier items providing the most psychometric information for lower-performing students) and the magnitude of the item-specific effect. Thus, larger correlations may predict smaller effect sizes because treatment effects will be concentrated on a subset of items rather than spread equally across the measure. (One might further suspect that this misalignment between the measure and the population targeted by the intervention is more probable for ID measures because they are not tailored to the RCT, but that hypothesis need not hold.) We use the absolute value of the correlation in our models because this argument does not depend on the sign of the correlation.

4. **The standard deviation of item easiness in the control group.** Relative to the distribution of student abilities, the SD of item easiness provides an index of the range of proficiency captured by the items of the measure. If narrower ranges of item easiness are easier to experimentally improve, then larger values of this SD would predict lower effect sizes. For example, researchers might be able to more easily improve students' ability to add one-digit numbers (small SD) than to improve addition of one- and two-digit numbers (large SD).

Table 2: Summary of item characteristic moderator variables

Item Characteristic	Definition	Potential Mechanism

Cronbach's α	Internal consistency, ranging from 0 to 1.	Narrower measures and heterogeneous populations lead to high α. The former may be easier to experimentally manipulate while the latter may be more difficult to experimentally manipulate.
Item-level heterogeneous treatment effects	The SD of item-specific treatment effects around the average treatment effect.	Larger SDs may result from over- or under-alignment between the intervention and specific items. The former may predict larger effect sizes, the latter may predict smaller effect sizes.
Treatment effect-easiness correlation	Correlation between item-specific effect size and item easiness.	Interventions that focus only on easier or harder content areas while the assessment covers the entire range of difficulty/academic skills may lead to concentrated (and therefore diluted) estimated effects.
SD of item easiness in the control group	The SD of item easiness parameters (relative to the distribution of student ability)	Narrower easiness ranges may be easier to improve through intervention.

4 Results

4.1 Descriptive Statistics and Baseline Balance

Table 3 shows descriptive statistics of each moderator by outcome type. We see that RD measures tend to show lower and more variable α, a larger SD of item-specific effects,[3] fewer items, similar numbers of subjects, a smaller SD of item easiness, and a stronger negative correlation between how easy the items are and the item-specific treatment effects. To provide evidence on the quality of the randomization, we conduct balance tests on the standardized baseline variable for each outcome using simple linear regression. We find that the median baseline treatment-control difference is -.02 SDs, with an IQR of -.06 to .04 and an unweighted mean of -.03. 90% of baseline differences are

[3]These IL-HTE magnitudes are comparable to past analyses of RCTs with item-level outcome data; see J. B. Gilbert, Himmelsbach, et al. (2025, Figure 1). Substantively, the value of .20 in the RD sample means that, on average, item-specific treatment effects have an SD of .20 effect size units around the average effect size.

less than or equal to .25 SDs, the What Works Clearinghouse benchmark for baseline equivalence (What Works Clearinghouse, 2022). Thus, the randomization appears to have been effective at equalizing the groups in terms of baseline scores on average across the datasets. Nevertheless, our calculation of the effect size for each dataset includes the baseline score as a covariate, both to increase precision and to correct for any residual baseline imbalance.

Table 3: Descriptive statistics of moderators by outcome type

Moderator	ID	RD
α	0.87 (0.1)	0.75 (0.17)
IL-HTE	0.1 (0.1)	0.2 (0.22)
N Items	40.95 (45.56)	21.03 (16.37)
N Persons	5123.74 (4749.85)	3880.85 (5043.4)
r(TE, easiness)	-0.14 (0.6)	-0.18 (0.63)
SD(easiness)	1.31 (0.67)	1.08 (0.59)

Notes: Cells contain means and SDs in parentheses. TE = treatment effect, ID = independently developed (N = 19), RD = researcher developed (N = 35). We report the raw correlations here and use the absolute value in our models. The SD of item treatment effects is on an SD scale.

4.2 Forest Plot

Figure 1 shows a forest plot of the average treatment effect sizes from each outcome with 95% CIs, color-coded by outcome type. Visually, we see a range of effect sizes; overall, those derived from ID measures tend to be substantially smaller and less variable than effect sizes derived from RD measures. These results suggest that our datasets are descriptively similar to prior work that examines moderation by outcome type, which consistently shows larger effect sizes for RD measures.

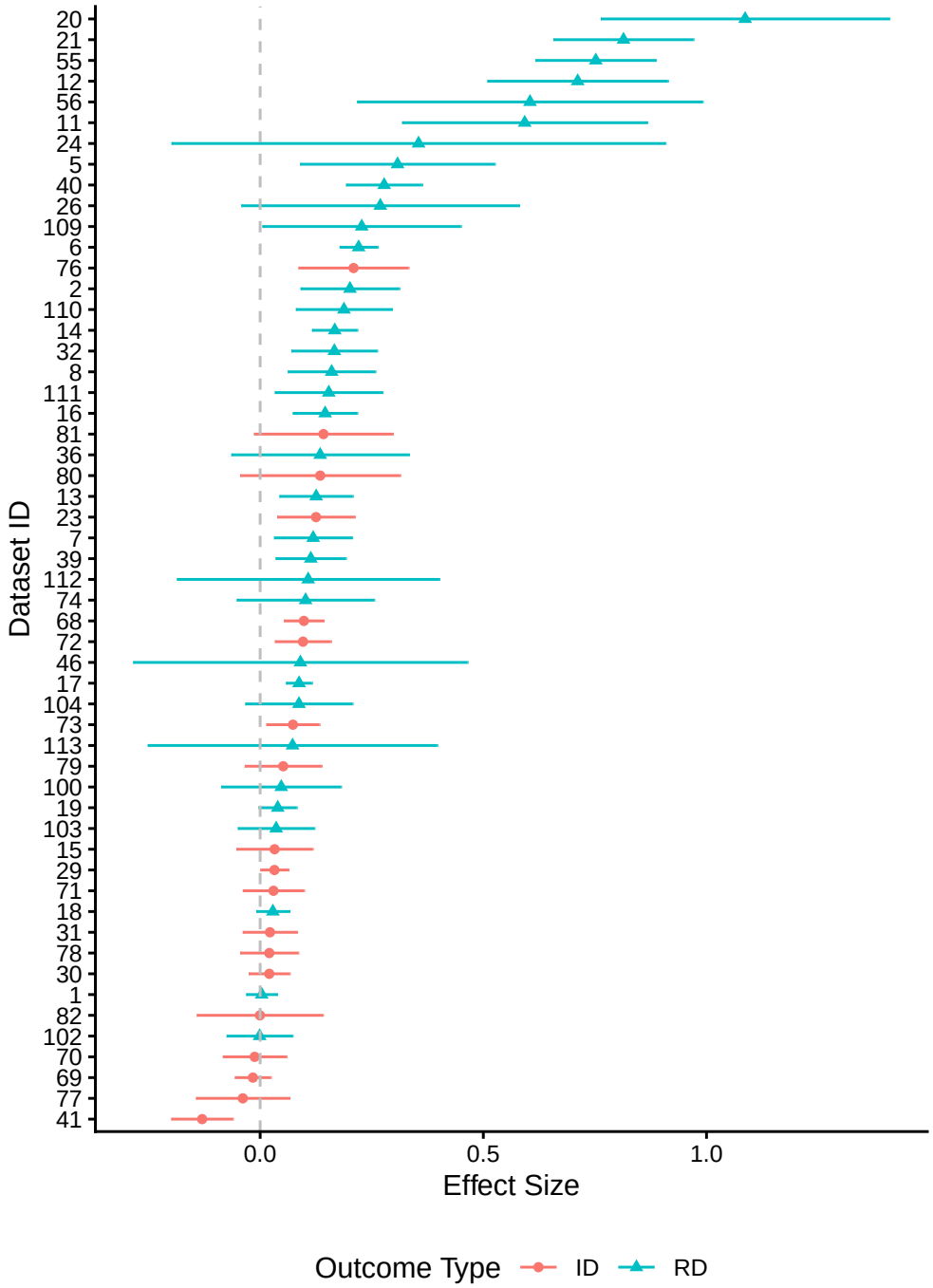
4.3 Meta-regression Models

Table 4 shows the estimates from our meta-regression models. All coefficients on item-level moderators are multiplied by 10 for interpretability so that the regression coefficients represent the

difference in effect size associated with a 0.1 unit difference in the moderator. We begin with a simple bivariate model of the effect size on outcome type, and replicate the widely documented result of larger effects on RD measures. The intercept of 0.10 provides the estimated mean effect size for ID measures (suggesting an effect that is small, on average, across ID outcomes), and RD measures show effect sizes 0.14 SDs larger, on average. We include outcome type as a moderator in all models. In subsequent models, we control for each additional moderator separately. We find that, in these models, participant sample size is a significant negative moderator ($\beta = -0.05$ per doubling of the participant sample), but the number of items in the measure is not. This result could be indicative of potential publication bias if smaller sample studies are more likely to be reported when they show positive results (Thornton, 2000). (Such bias would be at least partially controlled for by including sample size in the model.)

The SD of item-level treatment effects is a positive predictor of effect size ($\beta = 0.05$), and when controlling for this moderator, the coefficient on outcome type is reduced, but still significant ($\beta = 0.12$). While the exact mechanism underlying this result is unclear, there are potential explanations, such as those suggested in Table 2. For example, larger variability in item-specific treatment effects could be due, in part, to particular items showing exceptionally large treatment effects. Under such a scenario, those outlier items could lead to a higher point estimate of the overall treatment impact and potentially reflect differential gaming, coaching, or score inflation more generally (Koretz, 2005). We see at least one example of this phenomenon in our datasets, as dataset 11—a content literacy intervention measured with a vocabulary assessment—shows a single item with a treatment effect of over 3 SDs compared to an average effect size of about 0.5 SDs (J. B. Gilbert, Himmelsbach, et al., 2025, Figure 1). (Interestingly, the outlying word is *settle*, which was reported as not directly taught through the intervention.)

Figure 1: Forest plot of intervention effect sizes



The figure shows the effect size for each outcome with 95% confidence intervals. The points are color-coded by outcome type. ID = independently developed, RD = researcher developed.

Table 4: Results of meta-regression models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Intercept	0.10*	0.11*	0.11*	0.14*	0.04	0.06	0.12*	0.05
	[0.01; 0.18]	[0.03; 0.20]	[0.01; 0.19]	[0.00; 0.28]	[−0.04; 0.11]	[−0.04; 0.16]	[0.03; 0.21]	[−0.04; 0.13]
RD Outcome	0.14*	0.14*	0.14*	0.16*	0.12*	0.15*	0.14*	0.12*
	[0.08; 0.22]	[0.07; 0.21]	[0.06; 0.21]	[0.08; 0.24]	[0.05; 0.19]	[0.08; 0.22]	[0.07; 0.21]	[0.05; 0.19]
Subjects (log)		−0.05*						−0.02
		[−0.08; −0.01]						[−0.06; 0.02]
Items (log)			0.01					
			[−0.01; 0.02]					
α				−0.01				
				[−0.02; 0.01]				
SD(item TEs)					0.05*			0.04*
					[0.03; 0.07]			[0.02; 0.07]
SD(item easiness)						0.00*		
						[0.00; 0.01]		
r(item easiness, TE)							−0.00*	
							[−0.01; −0.00]	
SD: study_id	0.23	0.22	0.23	0.23	0.18	0.23	0.23	0.19
R^2	0.84	0.85	0.84	0.84	0.84	0.85	0.84	0.85
Num. obs.	54	54	54	54	54	54	54	54
LOO-IC	−78.57	−81.89	−79.84	−76.49	−74.86	−74.13	−76.97	−77.70
WAIC	−97.04	−99.26	−96.50	−94.28	−92.82	−96.70	−97.03	−95.83

Notes: The table shows point estimates and 95% credible intervals from Bayesian meta-regression models. The outcome is covariate-adjusted Cohen's *d* derived from a latent variable model. Subject and item sample size moderators are mean centered. LOO-IC and WAIC are leave-one-out information criterion and Watanabe-Akaike information criterion, respectively. Both information criteria are designed to capture out-of-sample prediction error, and lower values are preferred (Vehtari et al., 2017).

The correlation between treatment effect size and item easiness and the SD of item easiness in the control group are not significant moderators. Similarly, Cronbach's α was not a significant moderator, likely due in part to the use of a latent variable model from the outset (i.e., lower α is mechanically related to lower effect sizes when using sum scores, but our analysis accounts for this). In general, the correlation between item easiness and effect size, the SD of item easiness, and Cronbach's α do not appear to explain the difference between estimated treatment effects using RD versus ID measures.

Because the statistically significant moderators may be correlated with each other, we fit an additional model with both the SD of item-level treatment effects and participant sample size (Model 8). In Model 8, the coefficient on participant sample size is no longer significant when controlling for the SD of item-level treatment effects. Model fit indices such as R^2, LOO-IC, and WAIC are essentially indistinguishable across model specifications. For example, we find that the WAIC SE from Model 5 is 17.8, a value much larger than the differences in WAIC point estimates between any pair of models. These results suggest that additional moderators beyond outcome type do not add substantial predictive power. Therefore, Model 5, which includes RD outcome and the SD of item-level treatment effects, is our preferred model. The terms from Model 5 are interpreted as follows. An ID measure with constant item-level treatment effects is predicted to show an effect size of 0.04 SDs. Holding constant the SD of item-specific treatment effects, RD outcomes show effect sizes 0.12 SDs higher than ID effect sizes. This difference is about 15% smaller than that of the unadjusted model. Holding constant RD vs. ID outcome type, a 0.1 unit difference in the SD of item-specific treatment effects predicts effect sizes that are 0.05 SDs higher, on average.

As a sensitivity check, we refit our final model (Model 5) with a moderator representing outcome type, categorized as literacy, math/science, or other. We find that the outcome domain moderators are not substantially different from zero. We also find that the main substantive results are unchanged: the SD of item-level treatment effects is still a significant moderator ($\beta = .05$, SE = .01) and RD measures show larger effect sizes than ID measures, though the estimate is less precise ($\beta = .08$, SE = .05).

5 Discussion

Meta-analyses of educational interventions demonstrate differences in effect size by outcome type, with RD outcomes showing significantly larger effect sizes than ID outcomes. However, the source of this moderation by outcome type remains poorly understood. In this study, we leverage item-level outcome data from a diverse set of RCTs in education to develop and test novel hypotheses about potential moderators of effect size that may explain some of the moderation by outcome type.

In their meta-analysis examining moderation by outcome type, Wolf and Harbatkin (2023, Table 3) show that differences between RD and ID measures are between 0.24 and 0.32 SDs within the same study. In our final model, the coefficient on RD outcome type is reduced by about 15% from our unconditional model, with a coefficient of 0.12 SDs compared to 0.14 SDs in the unadjusted model. Furthermore, the participant sample size moderator was significant when considered in isolation, but not when we controlled for the SD of item-level treatment effects. Thus, while item characteristics explain some portion of moderation by outcome type, the study and item characteristics explored here appear insufficient to fully explain why RD measures produce larger effect sizes in our sample of RCTs. These results suggest that there is still more work to be done to fully understand moderation by outcome type, though the examination of item characteristics in this study provides an important first step in understanding this phenomenon, and does somewhat reduce the magnitude of outcome type as an effect size moderator.

Only one item characteristic led to a meaningful reduction of the magnitude of outcome type as a moderator of the treatment effect. Specifically, we find that the SD of item-level heterogeneous treatment effects (IL-HTE) was a significant moderator of effect size, and that controlling for this SD reduces the magnitude of the outcome type moderator coefficient substantially. A priori, we had two theories for how the IL-HTE SD might be associated with measure type as a moderator. First, imperfect alignment between the assessment content and nature of the intervention could produce more variability in item-specific effects for ID measures. Under such a scenario, IL-HTE could be associated with lower treatment effects if it captures poor alignment. Second, and in

line with results from Halpin and Gilbert (2024), a larger spread of item-specific treatment effects could be associated with larger average effect sizes across items, which could explain some of the moderation by outcome type phenomenon. Conceptually, this relationship could exist because researchers producing RD measures may try to optimally align items with the intervention, but cannot always predict which items will be best aligned in advance. Thus, the overall alignment between the construct and the intervention would be higher for an RD measure, but the variability in the item-specific effects would nonetheless be higher. This second theory is more consistent with our results.

The study is strengthened by the use of individual participant item-level outcome data (Braeken et al., 2026), the pooling of coefficients derived from latent variable models that correct for measurement error (Borenstein et al., 2022), the relatively novel exploration of item features as potential moderators (J. B. Gilbert, Young, et al., 2026), and the use of Bayesian meta-analysis methods to incorporate prior information and guard against overfitting in our relatively small sample (Grant & Di Tanna, 2025). Methodologically, our approach provides a novel framework to test hypotheses about effect size moderation in meta-analysis when individual participant item-level data allows psychometric quantities to be calculated directly for each study. Substantively, our results suggest that item-level data can provide valuable insight into effect moderation that would be masked by traditional methods that consider only study-level features as moderators.

While providing new insights into moderators of treatment effect sizes by leveraging item-level outcome data, several limitations temper the conclusions of our study. First and foremost, our analysis is limited by the availability of item-level outcome data. Our sample is therefore unlikely to be fully representative of the broader pool of educational interventions assessed in previous studies, though the fact that we see average effects in the typical range for educational research and replicate the moderation by outcome type that motivates this study suggests that we are at least partially capturing broader trends in education research. Second, only a handful of our studies contained a mix of both RD and ID measures, such as Romero et al. (2020), who include RD measures of math and language and the *Raven's Progressive Matrices* ID measure. As such, most

of our contrasts are driven by between- rather than within-study comparisons, and therefore may be confounded by omitted variables that differ between studies. In this vein, some studies do in fact report a mix of RD and ID measures, but item responses are available only for the RD measure. For example, various studies by Kim and colleagues report ID outcomes such as the *Measure of Academic Progress* (MAP) or other state standardized test score outcomes, but items are only available for inclusion in our analysis from their RD measures of reading comprehension or vocabulary, limiting informative within-study comparisons that may help to distinguish potential score inflation from genuine improvement on the construct being measured (Halpin & Gilbert, 2024; Koretz, 2008). The use of measures like MAP, which are computer-adaptive, also raises further complexities: while such tests are not designed to match the intervention, they are designed to match items to the student's proficiency, which itself represents a different way to conceptualize alignment. Last, it is possible that the item characteristic moderators function differently depending on outcome type, but given our small sample size, including interactions among moderators would be likely to overfit the data and, as such, we do not explore the possibility here.

With these limitations in mind, we view our contribution as necessarily exploratory and hypothesis-generating and as a foundation for future research. While empirical item-level meta-analyses are rare, the use of item-level features as effect size moderators is a promising area of inquiry that can shed light on fundamental issues in measurement and program evaluation (J. B. Gilbert, Young, et al., 2026; Halpin & Gilbert, 2024). Our results contribute to emerging evidence that measurement characteristics moderate treatment effect sizes in educational interventions (Halpin & Gilbert, 2024; Wolf & Harbatkin, 2023), and that these characteristics at least partially explain the well known moderation by RD vs. ID outcome type.

As such, our results suggest the following implications:

1. Researchers should heed calls to share item-level data as part of their replication materials, so that these issues can be fully investigated in future meta-analyses (Domingue et al., 2025). When item-level data cannot be shared, authors should share item-level summaries from fitted measurement models (e.g., item easiness/location estimates and item-by-treatment effect

estimates, with uncertainty), so future meta-analyses can incorporate item-level moderators without requiring access to raw student responses. At a broader level, program evaluators should routinely report psychometric information for their outcome measures, both for scientific transparency and also so that this information can be used in future meta-analyses. Psychometric reporting quality in RCTs is generally low, as J. Gilbert et al. (2026) report that across 112 RCT outcomes, 72% use a sum score, 31% report reliability, and only 7% test for unidimensionality, a finding similar to explorations of related issues in psychological research (Flake et al., 2017; Flake & Fried, 2020).

2. Meta-analyses of psychometric outcomes such as educational test scores should explicitly adjust for measurement error to avoid spurious heterogeneity induced by unreliability, either through classical corrections (e.g., dividing the effect size by the square root of reliability) or the use of latent variable models as pursued in this study to estimate treatment effects (Borenstein et al., 2022; J. B. Gilbert, 2025).

3. Applied researchers designing RD measures should explicitly consider construct representation and concentration at the item level, for example, by constructing item pools that span the intended construct and examining whether estimated impacts are driven by a small number of highly sensitive items. Similarly, when item response models are used for treatment effect estimation, researchers should consider and report checks for treatment-induced measurement noninvariance (e.g., item-level shifts not attributable to the focal construct), since concentrated item effects may reflect mechanisms distinct from uniform construct improvement. Estimating treatment effects with models that allow for item-level heterogeneity can provide a useful sensitivity check on empirical findings (Ahmed et al., 2025; J. B. Gilbert, Himmelsbach, et al., 2025; Halpin & Gilbert, 2024; Student, 2026).

In sum, by better understanding how measurement properties are related to intervention effectiveness, researchers can test new hypotheses and build better theories for intervention impact in education (Borsboom et al., 2021).

6 Declarations

6.1 Funding

The authors received no funding for this study.

6.2 Data and Code Availability

The datasets are available in the Item Response Warehouse (IRW, Domingue et al., 2025), with the prefix `gilbert_meta`: https://itemresponsewarehouse.org. Our data, code, results, and supplemental materials are available at the following URL: https://researchbox.org/3579.

6.3 Author Contributions

Conceptualization: JG, JS

Methodology: JG

Software: JG

Formal Analysis: JG

Writing—original draft preparation: JG, JS

Writing—review and editing: JG, JS

6.4 Acknowledgments

The authors wish to thank Ben Domingue, Betsy Wolf, and Andrew Ho for their helpful comments on drafts of this manuscript.

References

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Domingue, B. W. (2025). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*, *18*(4), 854–877. https://doi.org/10.1080/19345747.2024.2361337

Anglim, J., Dunlop, P. D., Wee, S., Horwood, S., Wood, J. K., & Marty, A. (2022). Personality and intelligence: A meta-analysis. *Psychological Bulletin*, *148*(5-6), 301–336. https://doi.org/10.1037/bul0000373

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

Banerji, R., Berry, J., & Shotland, M. (2017). The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in India. *American Economic Journal: Applied Economics*, *9*(4), 303–337.

Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. *Early Childhood Education Journal*, *51*(4), 717–732. https://doi.org/10.1007/s10643-022-01332-3

Berry, J., Kim, H. B., & Son, H. H. (2022). When student incentives do not work: Evidence from a field experiment in Malawi. *Journal of Development Economics*, *158*, 102893.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2022). *Introduction to meta-analysis* (2nd ed.). John Wiley & Sons.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Braeken, J., Lehre, A.-C., & Marcq, K. (2026). Pearls and perils when interpreting item-specific results from international large-scale educational assessments: The case of the "Nordic"

fractions. *Studies in Educational Evaluation*, *88*, 101552. https://doi.org/10.1016/j.stueduc.2025.101552

Bruhn, M., de Souza Leão, L., Legovini, A., Marchetti, R., & Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*, *8*(4), 256–295.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cabell, S. Q., Kim, J. S., White, T. G., Gale, C. J., Edwards, A. A., Hwang, H., Petscher, Y., & Raines, R. M. (2025). Impact of a content-rich literacy curriculum on kindergarteners' vocabulary, listening comprehension, and content knowledge. *Journal of Educational Psychology*, *117*(2), 153–175. https://doi.org/10.1037/edu0000916

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Davenport, J. L., Kao, Y. S., Johannes, K. N., Hornburg, C. B., & McNeil, N. M. (2023). Improving children's understanding of mathematical equivalence: An efficacy study. *Journal of Research on Educational Effectiveness*, *16*(4), 615–642.

de Barros, A., Fajardo-Gonzalez, J., Glewwe, P., & Sankar, A. (2024). The limitations of activity-based instruction to improve the productivity of schooling. *The Economic Journal*, *134*(659), 959–984.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12

De Boeck, P., Cho, S.-J., & Wilson, M. (2016). Explanatory item response models. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 247–266). Wiley.

De Weerdt, D., Simons, M., & Struyf, E. (2026). Team teaching or solo teaching? Evidence from a crossover experiment on the effects of team teaching on student achievement. *Teaching and Teacher Education*, *172*, 105403. https://doi.org/10.1016/j.tate.2026.105403

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L. A., Gilbert, J., Kanopka, K., Kapoor, R., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2025). An introduction to the Item Response Warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*, *57*. https://doi.org/10.3758/s13428-025-02796-y

Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental evidence on four policies to increase learning at scale. *The Economic Journal*, ueae003. https://doi.org/10.1093/ej/ueae003

Duflo, E., Berry, J., Mukerji, S., & Shotland, M. (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. *3ie Impact Evaluation Report*, *22*.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Francis, D. J., Kulesz, P. A., Khalaf, S., Walczak, M., & Vaughn, S. R. (2022). Is the treatment weak or the test insensitive: Interrogating item difficulties to elucidate the nature of reading intervention effects. *Learning and Individual Differences*, *97*, 102167.

Gilbert, J., Soland, J., & Young, W. (2026). Is the replication crisis a measurement crisis? Evidence from over 100 randomized trial outcomes. https://doi.org/10.31234/osf.io/9zgj8_v1

Gilbert, J. B. (2024). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. *Behavior Research Methods*, *56*(5), 5055–5067. https://doi.org/10.3758/s13428-023-02245-8

Gilbert, J. B. (2025). How measurement affects causal inference: Attenuation bias is (usually) more important than outcome scoring weights. *Methodology*, *21*(2), 91–122. https://doi.org/10.5964/meth.15773

Gilbert, J. B. (2026). Explanatory item response models for continuous data: A tutorial in R. *Behavior Research Methods*, *58*(5), 124. https://doi.org/10.3758/s13428-026-02986-2

Gilbert, J. B., Domingue, B. W., & Kim, J. S. (2026). Estimating causal effects on psychological networks using item response theory. *Psychological Methods*, *31*(1), 110–135. https://doi.org/10.1037/met0000764

Gilbert, J. B., Hieronymus, F., Eriksson, E., & Domingue, B. W. (2024). Item-level heterogeneous treatment effects of selective serotonin reuptake inhibitors (SSRIs) on depression: Implications for inference, generalizability, and identification. *Epidemiologic Methods*, *13*(S2). https://doi.org/10.1515/em-2024-0006

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*, *44*(4), 1417–1449. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2024). Leveraging item parameter drift to assess transfer effects in vocabulary learning. *Applied Measurement in Education*, *37*(3), 240–257. https://doi.org/10.1080/08957347.2024.2386934

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2025). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*, *50*(1), 72–101. https://doi.org/10.3102/10769986241240085

Gilbert, J. B., Young, W. S., Himmelsbach, Z., Ulitzsch, E., & Domingue, B. W. (2026). Conditional dependencies between response time and item discrimination: An item-level meta-analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644261426972

Glatz, T., Tops, W., Borleffs, E., Richardson, U., Maurits, N., Desoete, A., & Maassen, B. (2023). Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers: A cluster randomised controlled trial. *PeerJ*, *11*, e15499.

Grant, R., & Di Tanna, G. L. (2025). *Bayesian meta-analysis: A practical introduction* (1st ed.). Chapman & Hall/CRC.

Halpin, P., & Gilbert, J. (2024). Testing whether reported treatment effects are unduly influenced by item-level heterogeneity. https://osf.io/preprints/psyarxiv/9ru45_v1

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Jayanthi, M., Gersten, R., Schumacher, R. F., Dimino, J., Smolkowski, K., & Spallone, S. (2021). Improving struggling fifth-grade students' understanding of fractions: A randomized controlled trial of an intervention that stresses both concepts and procedures. *Exceptional Children*, *88*(1), 81–100.

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *AERA Open*, *7*, 23328584211004183. https://doi.org/10.1177/23328584211004183

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3–26. https://psycnet.apa.org/doi/10.1037/edu0000465

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2023). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*, *115*(1), 73–98. https://psycnet.apa.org/record/2022-69392-001

Kim, J. S., Gilbert, J. B., Relyea, J. E., Rich, P., Scherer, E., Burkhauser, M. A., & Tvedt, J. N. (2024). Time to transfer: Long-term effects of a sustained and spiraled content literacy intervention in the elementary grades. *Developmental Psychology*, *60*(7), 1279–1297. https://doi.org/10.1037/dev0001710

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Teachers College Record*, *107*(14), 99–118.

Koretz, D. (2008). *Measuring up*. Harvard University Press.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K., & Busick, M. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms* (tech. rep.). Institute of Education Sciences. https://ies.ed.gov/ies/2025/01/translating-statistical-representation-effects-education-interventions-more-readily-interpretable

Llauradó, E., Tarro, L., Moriña, D., Queral, R., Giralt, M., & Solà, R. (2014). EdAl-2 (Educacio en Alimentacio) programme: Reproducibility of a cluster randomised, interventional, primary-school-based study to induce healthier lifestyle activities in children. *BMJ Open*, *4*(11), e005496.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293. https://doi.org/10.3102/0162373719849044

Maruyama, T. (2022). Strengthening support of teachers for students to improve learning outcomes in mathematics: Empirical evidence on a structured pedagogy program in El Salvador. *International Journal of Educational Research*, *115*, 101977.

Mohohlwane, N., Taylor, S., Cilliers, J., & Fleisch, B. (2024). Reading skills transfer best from home language to a second language: Policy lessons from two field experiments in South Africa. *Journal of Research on Educational Effectiveness*, *17*(4), 687–710. https://doi.org/10.1080/19345747.2023.2279123

Mosher, D. M., & Kim, J. S. (2025). Building a science of teaching reading and vocabulary: Experimental effects of structured supplements for a read aloud lesson on third graders' domain-specific reading comprehension. *Scientific Studies of Reading*, *29*(1), 7–31. https://doi.org/10.1080/10888438.2024.2368145

Persson, M., Andersson, K., Zetterberg, P., Ekman, J., & Lundin, S. (2020). Does deliberative education increase civic competence? Results from a field experiment. *Journal of Experimental Political Science*, *7*(3), 199–208.

Relyea, J. E., Gilbert, J. B., Burkhauser, M., Scherer, E., Mosher, D. M., Wei, Z., Tvedt, J., & Kim, J. S. (2025). Asset-based implementation of structured adaptations in an online third-grade content literacy intervention. *Reading Research Quarterly*, *60*(4), e70048. https://doi.org/10.1002/rrq.70048

Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review*, *110*(2), 364–400.

Scherer, R., & Campos, D. G. (2022). Measuring those who have their minds set: An item-level meta-analysis of the implicit theories of intelligence scale in education. *Educational Research Review*, *37*, 100479. https://doi.org/10.1016/j.edurev.2022.100479

Schreinemachers, P., Baliki, G., Shrestha, R. M., Bhattarai, D. R., Gautam, I. P., Ghimire, P. L., Subedi, B. P., & Brück, T. (2020). Nudging children toward healthier food choices: An experiment combining school and home gardens. *Global Food Security*, *26*, 100454.

Sebele, M., McManus, J., Mwanza, M., Mwai, L., & Rudasingwa, M. (2025). Improving reading proficiency in early childhood education classrooms: Evidence from Liberia. *Journal of Development Effectiveness*, 1–17. https://doi.org/10.1080/19439342.2025.2540954

Soland, J. (2022). Evidence that selecting an appropriate item response theory–based approach to scoring surveys can help avoid biased treatment effect estimates. *Educational and Psychological Measurement*, *82*(2), 376–403. https://doi.org/10.1177/00131644211007551

Soland, J., Kuhfeld, M., & Edwards, K. (2024). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, *29*(5), 1003–1024. https://doi.org/10.1037/met0000506

Sørensen, Ø. (2024). Multilevel semiparametric latent variable modeling in R with "galamm". *Multivariate Behavioral Research*, *59*(5), 1098–1105. https://doi.org/10.1080/00273171.2024.2385336

Student, S. R. (2026). Causal parameter moderation: Applying moderated nonlinear factor analysis to causal inference with latent outcomes. *Journal of Educational and Behavioral Statistics*, 10769986251414869. https://doi.org/10.3102/10769986251414869

Thai, K.-P., Bang, H. J., & Li, L. (2022). Accelerating early math learning with research-based personalized learning games: A cluster randomized controlled trial. *Journal of Research on Educational Effectiveness*, *15*(1), 28–51. https://doi.org/10.1080/19345747.2021.1969710

Thornton, A. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216. https://doi.org/10.1016/S0895-4356(99)00161-4

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wang, L. C., Vlassopoulos, M., Islam, A., & Hassan, H. (2024). Delivering remote learning using a low-tech solution: Evidence from a randomized controlled trial in Bangladesh. *Journal of Political Economy Microeconomics*, *2*(3), 562–601. https://doi.org/10.1086/730456

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0* (tech. rep.). Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE).

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

Woods-Townsend, K., Hardy-Johnson, P., Bagust, L., Barker, M., Davey, H., Griffiths, J., Grace, M., Lawrence, W., Lovelock, D., Hanson, M., et al. (2021). A cluster-randomised controlled trial of the LifeLab education intervention to improve health literacy in adolescents. *PLoS One*, *16*(5), e0250545.

Appendices

A Treatment Effect Estimator in the Source Studies

Treatment effects for each study are derived from the following Explanatory Item Response Model (EIRM, De Boeck et al., 2016; J. B. Gilbert, Himmelsbach, et al., 2025). All datasets include a pretest variable, either a lagged outcome or a similar metric to the outcome, which we include as a covariate to improve the precision of the treatment effect estimates and adjust for any residual baseline differences between groups.

$$\text{logit}(Y_{ij} = 1) = \eta_{ij} = \theta_j + b_i + \zeta_i T_j \tag{4}$$

$$\theta_j = \beta_0 + \beta_1 T_j + \beta_2 X_j + \varepsilon_j \tag{5}$$

$$\begin{bmatrix} b_i \\ \zeta_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_b^2 & \rho\sigma_b\sigma_\zeta \\ \rho\sigma_b\sigma_\zeta & \sigma_\zeta^2 \end{bmatrix} \right) \tag{6}$$

$$\varepsilon_j \sim N(0, \sigma_\theta^2). \tag{7}$$

The model terms are interpreted as follows:

- Y_{ij} is the dichotomous item response for person j to item i
- θ_j is the latent trait under investigation
- b_i is item location or "easiness"
- ζ_i is item-specific sensitivity to treatment
- T_j is a treatment indicator
- β_1 is the average treatment effect
- X_j is a pre-treatment covariate (either a lagged outcome or a similar metric to the outcome)

- ε_j is a residual term with standard deviation σ_θ
- ρ is the correlation between item location and item-specific treatment effect size

The treatment effect coefficient β_1 is standardized by dividing it by the pooled within-group standard deviation σ_θ derived from an analogous model that includes only the treatment indicator (J. B. Gilbert et al., 2023).

We include `lme4` code to fit Equation 4 below, where `resp` is the 0/1 response, `treat` is the 0/1 treatment indicator, `std_baseline` is the standardized baseline covariate, `id` is the subject identifier, and `item` is the item identifier. Please see our references for detailed tutorials for fitting this model and related extensions in R (De Boeck et al., 2011; J. B. Gilbert, 2024, 2026).

```
glmer(resp ~ treat + std_baseline + (1|id) + (treat|item),
              data = df,
              family = binomial)
```

Equation 4 is a One-Parameter Logistic (1PL) EIRM in which each item is assumed to be equally discriminating with respect to θ_j. While 2PL EIRMs are estimable with `brms` or `galamm` (Bürkner, 2021; Sørensen, 2024), we maintain the 1PL approach for the primary analysis of this study for several reasons. First, prior analyses of 10 of these datasets show that average treatment effect estimates and standard errors are quite similar whether using 1PL or 2PL EIRMs (J. B. Gilbert, 2025, Figure 4). Second, when only average treatment effects at a single time point are of interest, the use of equal or variable scoring weights tends to be inconsequential (J. B. Gilbert, 2025). Thus, the use of equal item weights in a 1PL model is unlikely to be driving our empirical results.

B `brms` Code for Meta-regression Models

The code below shows how to fit the relevant Bayesian meta-regression models using `brms`. Note that we include only a random effect for study, rather than study and outcome, because the average study had only 1.7 outcomes in our data. Refitting our final model (Model 5) with an additional random effect for outcome in a three-level multilevel meta-regression gives similar results to our

main specification: the coefficient on RD outcome is .11 (SE = .05) and the coefficient on IL-HTE SD is .06 (SE = .02).

```
# set the priors
prior <- prior(normal(0,1), class = Intercept) +
  prior(normal(0,.5), class = b) +
  prior(normal(0,.5), class = sd)

# declare the model
# es | se(se) tells brms that the variable
# es has an associated SE
# other moderators would be added after outcome_type
f1 <- bf(es | se(se) ~ outcome_type + (1|study_id))

# fit the model
mod <- brm(
     formula = f1,
     data = dat,
     prior = prior,
     iter = 2000,
     cores = 4,
     chains = 4,
     threads = threading(4),
     backend = "cmdstanr",
     silent = 2
  )
```