Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1082

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

Joshua B. Gilbert
Harvard University

James 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 45 effect sizes from 30 studies shows that both greater standard deviations of item-specific treatment effects and lower correlations between item-specific effects and item easiness predict larger effect sizes and reduce the observed difference between RD and ID measures from .24 SDs to .15 SDs. The findings advance our understanding of how item properties predict educational intervention outcomes and underscore the affordances of analyzing item-level data for building theory in education research.

VERSION: November 2024

Suggested citation: Gilbert, Joshua B., and James Soland. (2024). Mechanisms of Effect Size Differences Between Researcher Developed and Independently Developed Outcomes: A Meta-Analysis of Item-Level Data. (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: A Meta-Analysis of Item-Level Data

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

[1]Harvard University Graduate School of Education
[2]University of Virginia School of Education and Human Development

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 45 effect sizes from 30 studies shows that both greater standard deviations of item-specific treatment effects and lower correlations between item-specific effects and item easiness predict larger effect sizes and reduce the observed difference between RD and ID measures from .24 SDs to .15 SDs. The findings advance our understanding of how item properties predict educational intervention outcomes and underscore the affordances of analyzing 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

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 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 files to include participants' item responses (Domingue et al., 2023). Consequently, examining item characteristics and the outcome's psychometric properties as potential moderators is rare.

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.

In this study, we leverage item-level outcome data from 45 datasets from 30 randomized controlled trials (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 the correlation between item-specific effects and item easiness negatively moderates effect sizes, and controlling for these moderators explains about 40% of the observed outcome type moderation. Other characteristics, such as treatment effects on item discrimination, the internal consistency of the outcome, or the SD of item easiness,

do not significantly moderate effect sizes in this population 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 overaligned) 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, is a key consideration. 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), which also moderates effect sizes.

Given the replicability 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, the properties of the measure used to produce the dependent variable often receive little intention, 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 measure is some sort of summary score, such as a sum or IRT-based scaled score. 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. (2024) 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%. 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 location can create bias in treatment by covariate interaction effects (Gilbert, 2024a; Gilbert, Hieronymus, et al., 2024; Gilbert, Himmelsbach, et al., 2024; Gilbert, Kim, & Miratrix, 2024; Gilbert, Miratrix, et al., 2024).

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 a prior study of item-level data from a large set of RCTs (Gilbert, Himmelsbach, et al., 2024). The authors examine item-level heterogeneous treatment effects—a statistical modeling approach that allows for unique treatment effects on each item—in 75 datasets from 48 RCTs from education, economics, political science, and health. Here, we limit our analysis to 45 datasets from 30 RCTs examining educational interventions and outcomes 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, and age groups, are summarized in Table 1.

Table 1: Descriptive statistics for the data sets in our analysis

ID	Study	Location	Subjects	Items	Population	Outcome	Type
1	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
10	J. S. Kim et al., 2021	USA	4834	20	G1-G2	Reading Self Concept	ID
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
22	Berry et al., 2018	Ghana	5290	10	Adolescents	Saving Attitudes	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
25	Schreinemachers et al., 2020	Nepal	775	15	Children 8-12	Food Preferences	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
33	Aladysheva et al., 2017	Kyrgyzstan	1242	18	Adolescents	Social Trust	RD
35	Persson et al., 2020	Sweden	1152	12	High School	Democratic Values	RD
36	Persson et al., 2020	Sweden	1108	7	High School	Political Knowledge	RD
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., 2023	South Africa	3068	134	Early Elementary	Oral Reading Fluency	ID
46	Glatz et al., 2023	Netherlands	120	42	G1	Language	RD
48	Cárdenas et al., 2023	Mexico	1150	30	Early Childhood	Child Development	ID
55	Wang et al., 2024	Bangladesh	1704	15	Elementary	Academic Achievement	RD

56	Sebele et al., 2023	Liberia	2307	4	Preschool	Literacy	RD
64	Zhao et al., 2023	Jordan	4041	9	Preschool	Social Emotional Learning	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	Gilbert, Kim, and Miratrix, 2024	USA	1225	12	G2	Vocabulary	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 original study. 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 equation).

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 1SD would be extremely unlikely (Kraft, 2020). Note that, while some studies examine the effects of individual item characteristics on patterns of treatment effects within a single study (e.g., Gilbert, 2024a, 2024c; Gilbert, Hieronymus, et al., 2024; Gilbert, Kim, and Miratrix, 2024), 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 (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 *Me and My Reading Profile* 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.

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. The extent to which these item-level moderators explain moderation by outcome type will depend on the extent to which these moderators differ by outcome type. 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 then, 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 that ranges from 0 to 1. α therefore provides a rough proxy for the psychometric quality of the outcome measure. Note that low reliability will attenuate effect sizes derived from observed scores, because measurement error inflates the standard deviation of the observed scores (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 (Gilbert, 2024b; Soland, 2022; Soland et al., 2022). 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, if a researcher designs a measure more tightly linked to a narrow intervention), while 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 remains unclear.

2. **The standard deviation of item-level heterogeneous treatment effects (IL-HTE).** A body of work—including the 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., 2024; Gilbert, Himmelsbach, et al., 2024; 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 treatment effect size on item discrimination.** Item discrimination (denoted a_i in item response theory) captures how a change in the latent construct translates to a change in the probability of a correct item response. More discriminating items therefore better distinguish among participants along the latent construct and are more reliable indicators of the construct. Research shows that ignoring item discrimination can yield biased treatment effect estimates (Soland, 2022; Soland et al., 2022). In the RCT context, if a treatment makes participants more familiar with the format of the test items, we might expect average item discrimination be higher in the treatment group due to a reduction in construct-irrelevant variance, and also higher accuracy emerging from increased familiarity. Similarly, increases to item discrimination could act as an effect multiplier because the treatment effect on an item is equal to the true treatment effect on the latent construct multiplied by a_i (assuming the changes to a_i are not incorporated into the estimation model, which would theoretically account for this). Thus, positive treatment effects on item discrimination may be associated with larger effect sizes. We model treatment effects on a_i on a log scale (Cho et al., 2014; Gilbert, Zhang, et al., 2024).

4. **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 (Gilbert, Miratrix, et al., 2024). 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 on 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.

5. **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 size on item discrimination	Relationship between the latent construct and the probability of a correct response.	Increased familiarity with item format due to treatment; effect multiplier.
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

Table 3 shows descriptive statistics of each moderator by outcome type. We see that RD measures tend to show lower α, a larger SD of item-specific effects, fewer items, fewer subjects, and a larger SD of item easiness.

Table 3: Descriptive statistics of moderators by outcome type

Moderator	ID	RD
α	0.94 (0.06)	0.82 (0.16)
TE on a_i	-0.01 (0.06)	-0.04 (0.07)
SD(Item TEs)	0.06 (0.04)	0.22 (0.22)
N Items	31.21 (31.31)	21.52 (16.59)
N Subjects	5617.57 (4648.59)	4442.87 (5201.12)
r(Item, TE)	-0.24 (0.66)	-0.25 (0.61)
SD(Item Easiness)	0.97 (0.34)	1.24 (0.64)

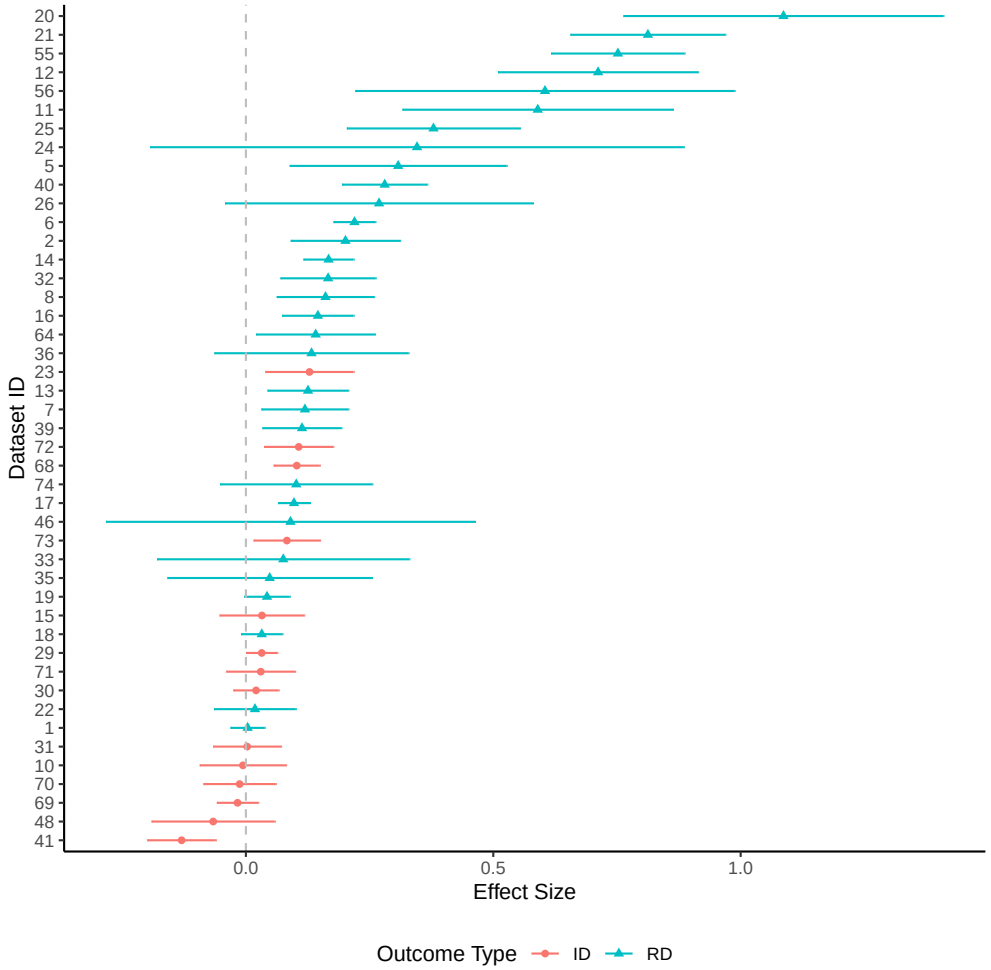
Notes: Cells contain means and SDs in parentheses. ID = independently developed (N = 14), RD = researcher developed (N = 31). We report the raw correlations here and use the absolute value in our models. The treatment effect on a_i is on a log scale, and the SD of item treatment effects in on an SD scale.

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 show larger effect sizes for RD measures.

Figure 1: Forest plot of intervention effect sizes



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

4.3 Meta-regression Models

Table 4 shows the estimates from our meta-regression models. All coefficients on item-level moderators are divided 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 0.02 provides the estimated mean effect size for ID measures (suggesting an effect that is null, on average, across ID outcomes), and RD measures show effect sizes 0.24 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 ($\beta = -0.08$ per doubling of the participant sample), but 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 ($\beta = 0.07$), and when controlling for this moderator, the coefficient on outcome type is reduced substantially, but still significant ($\beta = 0.15$). 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 (D. 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 (Gilbert, Himmelsbach, et al., 2024, Figure 1). (Interestingly, the outlying word is *settle*, which was reported as not directly taught through the intervention.)

Table 4: Results of meta-regression models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Intercept	0.02	0.04	0.03	0.24	−0.03	−0.02	0.13*	−0.05	0.16	0.05
	[−0.06; 0.10]	[−0.05; 0.12]	[−0.05; 0.12]	[−0.04; 0.49]	[−0.10; 0.04]	[−0.10; 0.08]	[0.03; 0.22]	[−0.15; 0.05]	[−0.42; 0.70]	[−0.05; 0.15]
RD Outcome	0.24*	0.22*	0.22*	0.27*	0.15*	0.26*	0.21*	0.24*	0.16*	0.15*
	[0.17; 0.31]	[0.15; 0.28]	[0.14; 0.29]	[0.20; 0.35]	[0.08; 0.22]	[0.19; 0.32]	[0.14; 0.28]	[0.18; 0.31]	[0.08; 0.26]	[0.07; 0.21]
Subjects (log)		−0.08*							−0.05	
		[−0.13; −0.04]							[−0.10; 0.02]	
Items (log)			0.02						0.02	
			[−0.01; 0.05]						[−0.03; 0.08]	
α				−0.03					−0.01	
				[−0.06; 0.00]					[−0.07; 0.05]	
SD(Item TEs)					0.07*				0.03	0.06*
					[0.04; 0.10]				[−0.01; 0.07]	[0.03; 0.08]
TE on a_i						−0.09*			−0.05	
						[−0.13; −0.04]			[−0.10; 0.01]	
r(Item, TE)							−0.02*		−0.01*	−0.01*
							[−0.02; −0.01]		[−0.02; −0.00]	[−0.02; −0.00]
SD(Item Easiness)								0.01*	0.00	
								[0.00; 0.01]	[−0.00; 0.01]	
SD: study_id	0.21	0.22	0.22	0.23	0.17	0.24	0.21	0.22	0.22	0.18
R^2	0.75	0.79	0.75	0.77	0.80	0.78	0.77	0.77	0.81	0.80
Num. obs.	45	45	45	45	45	45	45	45	45	45

* Null hypothesis value outside the credible interval.

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.

The correlation between treatment effect size and item easiness is a significant and negative moderator ($\beta = -0.02$), in line with one of the mechanisms we discussed. However, the point estimate is close to zero and the practical significance of this finding is unclear. Contrary to our initial hypothesis, the SD of item easiness in the control group is a *positive* and significant moderator, but with a very small point estimate ($\beta = 0.01$). The treatment effect on item discrimination is a *negative* moderator of effect size ($\beta = -0.09$). However, this moderator does not seem to explain the RD versus ID effect. 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 much of the difference between estimated treatment effects using RD versus ID measures.

Because these moderators may be correlated with each other, we fit a saturated model including all moderators (Model 9). Our final model (Model 10) contains only the moderators whose 95% credible intervals exclude 0. The coefficients from Model 10 are interpreted as follows: holding constant the SD of item-specific treatment effects and the correlation between item-specific treatment effects and item easiness, RD outcomes show effect sizes 0.15 SDs higher than ID effect sizes. This difference is substantially smaller than that of the unadjusted model. Holding constant the other variables in the model, a 0.1 unit difference in the SD of item-specific treatment effects predicts effect sizes that are 0.06 SDs higher, and a 0.1 unit difference in the absolute value of the correlation between treatment effect size and item easiness predicts effect sizes 0.01 units lower, on average.

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 RCTs 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 significant differences between RD and ID measures is 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 40% from our unconditional model, with a coefficient of 0.15 SDs compared to 0.24 SDs in the unadjusted model. Furthermore, moderators such as participant sample size and the treatment effect on item discrimination were significant when considered in isolation, but not when we controlled for other moderators. Thus, while item characteristics explain a large portion of moderation by outcome type, 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 reduce the magnitude of outcome type as an effect size moderator substantially.

While we show that several item characteristics were significant predictors of treatment effect size across our models, only one led to a large 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 only other significant moderator in our final model was the correlation between item-specific treatment effect size and item easiness. In line with one of our proposed mechanisms, this correlation was negatively associated with effect size. The magnitude of this effect is small but precisely estimated. This result suggests that interventions that produce effects concentrated on the easier or harder end of the content spectrum predict slightly smaller effect sizes overall.

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; D. M. Koretz, 2008). The use of measures like MAP, which are computer-adaptive, also raise 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. Our results provide evidence that item characteristics moderate treatment effect sizes in educational interventions, and that these characteristics at least partially explain the well known moderation by outcome type. One implication of our study is that researchers should heed calls to share item level data as part of their replication materials (Domingue et al., 2023). A second implication is that program evaluators should not expect effect sizes of equal magnitudes across all types of measures, and that examining how consistent item-specific treatment effects are within an impact evaluation can be a powerful tool to explore the sensitivity of results to the items assessed (Ahmed et al., 2024; Gilbert, Himmelsbach, et al., 2024; Halpin & Gilbert, 2024). Thus, by better understanding how item 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 original datasets are publicly available at the following URL: https://doi.org/10.7910/DVN/C4TJCA. The datasets are also available in the Item Response Warehouse (IRW, Domingue et al., 2023), with the prefix `gilbert_meta`: https://redivis.com/datasets/as2e-cv7jb41fd. Our data, code, results, and supplemental materials will be made available at the following URL upon publication: https://researchbox.org/3579.

6.3 Author Contributions

Conceptualization: Author 1, Author 2

Methodology: Author 1

Software: Author 1

Formal Analysis: Author 1

Writing—original draft preparation: Author 1, Author 2

Writing—review and editing: Author 1, Author 2

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., & Benjamin, W. (2024). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*. https://doi.org/https://doi.org/10.1080/19345747.2024.2361337

Aladysheva, A., Asylbek Kyzy, G., Brück, T., Esenaliev, D., Karabaeva, J., Leung, W., & Nillesen, E. (2017). Impact evaluation of the Livingsidebyside peacebuilding educational programme in Kyrgyzstan.

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 [Publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418]. *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.

Berry, J., Karlan, D., & Pradhan, M. (2018). The impact of financial education for youth in Ghana. *World Development*, *102*, 71–89.

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.

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.

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

Cárdenas, S., Evans, D. K., & Holland, P. (2023). Parent Training and Child Development at Low Cost? Evidence From a Randomized Field Experiment in Mexico. *Journal of Research in Childhood Education*, *38*(sup1), S130–S160.

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

Cho, S.-J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*, 84–104.

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., Cho, S.-J., & Wilson, M. (2016). Explanatory item response models [Publisher: Wiley Online Library]. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 247–266.

Domingue, B., Kanopka, K., Braginsky, M., Zhang, L., Caffrey-Maffei, L., Kapoor, R., Liu, Y., Zhang, S., & Frank, M. (2023). The Item Response Warehouse (IRW).

Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental Evidence on Four Policies to Increase Learning at Scale. *The Economic Journal*, 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*.

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. B. (2024a). Estimating treatment effects with the explanatory item response model. *Journal of Research on Educational Effectiveness*, 1–19. https://doi.org/https://doi.org/10.1080/19345747.2023.2287601

Gilbert, J. B. (2024b). How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights. *Edworkingpapers.com*.

Gilbert, J. B. (2024c). 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.

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*(1). https://doi.org/doi:10.1515/em-2024-0006

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2024). Estimating Heterogeneous Treatment Effects with Item-Level Outcome Data: Insights from Item Response Theory [Version Number: 3]. https://doi.org/10.48550/ARXIV.2405.00161

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.

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/https://doi.org/10.1080/08957347.2024.2386934

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2024). 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*. https://doi.org/https://doi.org/10.3102/10769986241240085

Gilbert, J. B., Zhang, L., Ulitzsch, E., & Domingue, B. W. (2024). Polytomous Explanatory Item Response Models for Item Discrimination: Assessing Negative-Framing Effects in Social-Emotional Learning Surveys. *arXiv preprint arXiv:2406.05304*.

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.

Halpin, P., & Gilbert, J. (2024). Testing Whether Reported Treatment Effects are Unduly Dependent on the Specific Outcome Measure Used [Version Number: 2]. https://doi.org/10.48550/ARXIV.2409.03502

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

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.

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.

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.

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.

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

Koretz, D. M. (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. (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/ncser/pubs/20133000/

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. (2023). 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*, 1–24. https://doi.org/https://doi.org/10.1080/19345747.2023.2279123

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.

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

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., Jeffery, M., Mwanza, M., Mwai, L., & Rudasingwa, M. (2023). Improving reading proficiency in early childhood education classrooms: Evidence from Liberia. https://doi.org/https://riseprogramme.org/sites/default/files/inline-files/Rudasingwa_Improving_Reading_Proficiency_ECE_Liberia.pdf

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. (2022). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*.

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

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

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.

Zhao, V. Y., Hilgendorf, D., Yoshikawa, H., & Michael, D. (2023). Impacts of Ahlan Simsim TV Program in Pre-Primary Classrooms in Jordan on Children's Emotional Development: A Randomized Controlled Trial.

Appendices

A Treatment Effect Estimator in the Source Studies

Treatment effects for each study are derived from the following Explanatory Item Response Model (De Boeck et al., 2016; Gilbert, Himmelsbach, et al., 2024). All datasets include a pretest variable, either a lagged outcome or a similar metric to the outcome, which is included to improve the precision of the estimates.

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

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

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

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

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
- 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.