Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type

Betsy Wolf and Erica Harbatkin

The Version of Record of this manuscript has been published and is available in the Journal of Research on Educational Effectiveness (2022) https://doi.org/10.1080/19345747.2022.2071364.

This was written in Betsy Wolf's official capacity as part of the national conversation on education, is intended to promote the exchange of ideas among researchers and policymakers and to express views as part of ongoing research, and does not necessarily reflect the official views of the U.S. Department of Education.

Abstract

One challenge in understanding "what works" in education is that effect sizes may not be comparable across studies, raising questions for practitioners and policymakers using research to select interventions. One factor that consistently relates to the magnitude of effect sizes is the type of outcome measure. This paper uses study data from the What Works Clearinghouse to determine average effect sizes by outcome measure type. Outcome measures were categorized by whether the group who developed the measure potentially had a stake in the intervention (non-independent) or not (independent). Using meta-analysis and controlling for study quality and intervention characteristics, we find larger average effect sizes for non-independent measures than for independent measures. Results suggest that larger effect sizes for non-independent measures are not due to differences in implementation fidelity, study quality, or intervention or sample characteristics. Instead, non-independent and independent measures appear to represent partially but minimally overlapping latent constructs. Findings call into question whether policymakers and practitioners should use research based on non-independent measures when they are ultimately responsible for improving outcomes on independent measures.

Introduction

A growing literature is beginning to contend with how to translate effect sizes to make sense of the impacts of educational interventions. As part of this literature, researchers are grappling with how to characterize effect sizes and contextualize them for policymakers and practitioners (Baird & Pane, 2019; Bloom et al, 2008; Kraft, 2020; Lipsey et al, 2012). One challenge in interpreting effect sizes is that they vary along a number of dimensions, including which outcome is being measured, when the outcome is measured relative to the timing of the intervention, and the reliability of the measure (Kraft, 2020). Due to differences in study characteristics, effect sizes are often not comparable across studies (Wilson & Lipsey, 2001).

One study characteristic known to influence the magnitude of effect sizes is whether outcome measures were created either by study authors or researchers involved in the development of the intervention. Specifically, researchers have consistently identified larger average effect sizes on these types of outcome measures relative to outcome measures that were standardized or created by third parties (Cheung & Slavin, 2016; de Boer et al., 2014; Li & Ma, 2010; Lipsey et al., 2012; Lynch et al., 2019; Pellegrini et al., 2019; SWAT Measurement Small Group, 2020; Wilson & Lipsey, 2001). Therefore, differences in outcome measure type may yield effect sizes that are incomparable across studies of different interventions or even across studies of the same intervention.

Less is known about systematic differences in effect sizes for outcome measures covering fewer concepts (“narrow”) versus many concepts (“broad”). A few studies found larger average effect sizes when using narrow versus broad measures (Hill et al, 2008; Lipsey et al., 2012; SWAT Measurement Small Group, 2020). Moreover, measures may be designed for different purposes, such as determining whether students learned one specific skill or whether students are on grade level in a subject area. Using measures for different purposes calls into question whether effect sizes should always be compared across studies.

The purpose of this paper is to examine the extent to which effect sizes systematically vary by outcome measure type, with a particular focus on researcher and developer measures relative to independent measures not created by the same researchers or developers involved in the study. Our findings will help the research community make sense of the magnitude of effect sizes depending on outcome measure type. We ask the following research questions:

1. How often are researcher or developer measures, as opposed to independent measures, the only measures used or reported in studies?
2. How often do researcher and developer measures, as opposed to independent measures, result in positive and statistically significant findings? Relatedly, how often do researcher and developer measures, as opposed to independent measures, result in findings that meet the requirements for the Every Student Succeeds Act (ESSA) evidence tiers?
3. To what extent do effect sizes systematically vary by outcome measure type—defined as independent broad, independent narrow, non-independent developer, and non-independent researcher—controlling for other factors?
4. What mechanisms may explain systematic differences in effect sizes by outcome measure type?

The first two questions provide context on the sample of studies, the outcome measures used, and the extent to which findings by outcome measure type produce favorable results. The third question provides meta-analytic evidence about the extent to which effect sizes vary by measure type. The fourth question is addressed through multiple sensitivity analyses in an effort to explain the mechanisms driving any observed differences in effect sizes.

Literature Review

Researchers have consistently found larger average effect sizes for researcher and developer measures relative to “independent” or “standardized” measures not created by the same authors or developers involved in the study. The average difference in effect sizes for researcher and developer measures ranges from +0.11 to +0.31, most often in terms of Cohen’s d or Hedges’ g (Cheung & Slavin, 2016; de Boer et al., 2014; Gersten et al., 2020; Li & Ma, 2010; Lipsey et al., 2012; Lynch et al., 2019; Pellegrini et al., 2019; SWAT Measurement Small Group, 2020; Williams et al., 2022; Wilson & Lipsey, 2001). Researchers also tend to identify average effect sizes that are larger by +0.12 to +0.16 standard deviations for narrow measures than for broad ones (Lipsey et al., 2012; SWAT Measurement Small Group, 2020). Table 1 summarizes the findings of these studies.

[Table 1]

While the consensus is generally that researcher and developer measures yield larger average effect sizes relative to independent ones, researchers have speculated different hypotheses for *why* there might be a difference in effect sizes based on outcome measure type. One hypothesis for why average effect sizes are systematically larger for researcher and developer measures is that they cover fewer concepts than independent measures; that is, researcher and developer measures capture constructs in a narrow domain, whereas some independent measures capture constructs on a broad domain or across multiple domains. For example, if an independent measure consists of multiple subscales, and a researcher or developer measure consists of only one subscale, the variation in scores will be greater for the independent measure than the researcher or developer measure, which could result in smaller effect sizes for the independent measure. Along those lines, de Boer et al. (2014) hypothesized that researcher measures may focus on whether students can perform specific tasks, whereas broad measures focus on student performance in a content area. However, one study that examined this hypothesis found no relationship between the narrowness of the measure, as determined by a binary indicator, and whether the measure was a researcher or developer measure (SWAT Measurement Small Group, 2020).

A second hypothesis is that researcher and developer measures are more closely aligned with the intervention, and therefore have greater content validity to detect intervention effectiveness than independent measures (Lipsey et al., 2012; Lortie-Forgues & Inglis, 2019; Lynch et al., 2019; SWAT Measurement Small Group, 2020; Wilson & Lipsey, 2001). Moreover, independent measures, such as state standardized tests, may be poorly aligned with the intervention and therefore ill-equipped to detect intervention effectiveness. On the other end of the spectrum, there is concern about possible overalignment of researcher and developer measures with the intervention because overalignment may lead to findings that are an inaccurate indication of the intervention's true effects (WWC Standards Version 4.1, 2020).

A third hypothesis is that use of researcher and developer measures is confounded with greater implementation fidelity because researchers who develop their own measures are more invested in implementation in those studies, which could lead to higher average effect sizes (Li & Ma, 2010; Lipsey, 2009). Yet one study that examined effect sizes *within* studies—therefore holding constant any differences in implementation fidelity across studies—still found larger average effect sizes for researcher and developer measures relative to independent ones (SWAT Measurement Small Group, 2020). However, it is theoretically possible that studies using researcher and developer measures have greater implementation fidelity overall compared with studies using only independent measures.

A final hypothesis is that effect sizes for researcher and developer measures may be larger due to developer involvement in the study for other reasons (Petrosino & Soydan, 2005; Wolf, Morrison, Inns, Slavin, & Risman, 2020). When study authors have a conflict of interest with the intervention, they may "use statistical strategies that skew the changes of a positive result in their program's favor" (Petrosino & Soydan, 2005; p. 443). This idea has also been called "researcher degrees of freedom;" that is, that researchers make numerous decisions in the data collection and analysis processes, and these decisions could be made to yield the most favorable study findings possible (Simmons, Nelson, & Simonsohn, 2011). One study also found evidence of greater publication bias for studies either authored or funded by developers, which may also contribute to higher average effect sizes in studies with developer involvement (Wolf et al., 2020). In summary, research has consistently identified larger average effect sizes on researcher and developer measures compared with independent measures. Researchers have articulated multiple plausible hypotheses for *why* average effect sizes may be larger on researcher or developer measures, but there is not consensus in the field.

Despite agreement in the measurement literature that outcome measures may produce systematically different effect sizes, the literature on "what works" in education does not consistently distinguish between outcome measure types in interpreting effect sizes. It may be the case that researchers do not distinguish between outcome measure types because they assume differences are not relevant. In reality, results from a number of studies show that methodological choices, such as outcome measure type, contribute to effect sizes as much as the characteristics of the interventions themselves (Wilson & Lipsey, 2001). Failing to account for outcome measure type in interpreting effect sizes may lead to inadequate representations of intervention effectiveness.

One question that remains unexamined in the literature is, how do we make sense of findings identified using a researcher or developer measure relative to an independent measure? For example, would we expect studies that find positive effects on a researcher or developer measure to also find positive effects on assessments used by schools and districts for progress monitoring or accountability purposes? In other words, to what extent do researcher and developer measures provide relevant information about student progress and performance to practitioners and policymakers? Another pertinent question is, why do we observe larger average effect sizes on researcher and developer measures relative to independent measures? More research is needed to address these questions.

Methods

Data

This paper draws from study data collected by the What Works Clearinghouse (WWC) at the Institute of Education Sciences (IES) to examine differences in effect sizes by outcome measure type. The advantage of using WWC data is that all studies must demonstrate internal validity in order to be included. The WWC data therefore allow us to examine effect size differences without poor study quality as a potential confounder. In addition, the WWC prohibits poor instrument quality because it requires that outcome measures are collected in the same manner for treatment and comparison groups; have face validity evidenced by a description of the outcome measure; and exhibit a minimum reliability of a Cronbach’s alpha of at least .50, a test-retest reliability of at least .40, or an inter-rater reliability of at least .50 (WWC Standards Version 4.1, 2020).

We use WWC study data in three topic areas: literacy; science, technology, engineering, and mathematics (STEM); and behavior, which largely includes outcomes related to specific student behaviors or competencies. We restrict the sample of studies to these topic areas because they are the most likely to contain proximal intervention outcomes, such as achievement or behavior, as opposed to more distal outcomes such as school retention, dropout, graduation, and post-secondary outcomes. The technical appendix (Appendix A) contains more information about how we compiled and modified the data.

The final dataset includes 1,553 findings from 373 studies that meet WWC standards. Table 2 summarizes the included findings and studies, with findings in the first two columns and studies in the next two columns. The left panel includes the full sample of studies, and the right includes only studies with at least one independent and non-independent measure (i.e., the within-study sample).

[Table 2]

In the full sample, about three-quarters (76%) of studies were randomized controlled trials (RCTs), and 24% of the studies had quasi-experimental designs (QEDs). Fewer than 1% were regression discontinuity designs (RDD). All of the studies meet WWC's research standards, either *without* or *with* reservations. About half (52%) of studies were reviewed under the WWC Standards Version 2.1 or higher, which is more rigorous than the preceding versions of the standards (WWC Standards Handbook Version 4.1, 2020). The WWC reviews studies for different purposes, and about half (52%) of the studies were reviewed by the WWC as part of systematic reviews on particular interventions. The remaining 48% of studies were reviewed because the study was prioritized for review by the WWC for various reasons (e.g., evidence cited for federal grant competitions, research funded by the U.S. Department of Education, individual studies deemed of interest). Studies spanned grade levels from early childhood to high school, with 29% of findings in the upper elementary grades, 27% of findings in the early elementary grades, 23% of findings in the middle school grades, 13% of findings in the high school grades, and 8% of findings in early childhood. Just over half (55%) of the findings related to literacy, 23% to STEM, and 22% to behavior. The studies examined interventions of different types, including supplemental (36%), curricula (26%), practices (9%), teacher-directed (3%), and schoolwide (2%), that were delivered in a variety of ways, including to individuals (25%), small groups (25%), classes (46%), and schools (10%).

The within-study sample is similar to the full sample, except has a higher percentage of supplemental interventions (55%) than the full sample (36%). The within-study sample also has a higher rate of non-independent measures by design, as the within-study sample requires studies to have at least one non-independent measure.

Measures

The WWC study database contains several different types of outcome measures. We conceptualized the outcome measures into two overarching categories: independent measures—when the group who developed the measure was unrelated to the particular intervention—and non-independent measures—when the group who developed the measure was related to the particular intervention and potentially had a stake in whether the intervention was identified as being effective. Within the category of independent measures, we also coded whether the measure intended to capture student achievement in a topic area (e.g., literacy), or whether the measure intended to capture achievement at a more granular level within a topic area (e.g., word fluency). Therefore, we coded each outcome measure in each study as one of the following mutually exclusive categories[1]:

- **Independent broad**: Measures intended to capture student achievement in a topic area, schoolwide climate, or general educational outcomes. This category includes state and district assessments, national surveys and assessments[2], grade point average, graduation rates, and school disciplinary data.
- **Independent narrow**: Measures intended to capture student achievement at a more granular level than a topic area or specific student behaviors. This category includes commercial assessments, measures developed by researchers not involved in the study, and outcomes associated with a specific class or subset of classes (e.g., credits earned, grades, etc.).
- **Non-independent developer**: Measures that were developed for a commercially available intervention and typically only used when the intervention is also being implemented. Commercially available interventions include curricula, online learning products, teacher professional development programs, and others.
- **Non-independent researcher**: Measures developed by study authors, including measures that were created by selecting specific items from preexisting scales.

For brevity, we will refer to the four outcome measure types as broad, narrow, developer, and researcher.

[1] Instruments for researcher and developer measures are often not included in the original studies, making it difficult to determine whether these instruments cover narrow or broad domains. Therefore, researcher and developer measures were coded mutually exclusively from narrow and broad measures.

[2] National assessments are commercial or government assessments used by school districts or post-secondary institutions across the country to assess competency in a topic area.

As we show in Table 2, the most common measure type is narrow, which constituted 43% of the findings, followed by researcher measures (30%), broad measures (22%), and developer measures (5%).[3] We classified each measure by reviewing WWC resources and the original studies. In some cases, we conducted internet searches to learn more about a measure, its formal name, and who created it. When the outcome measure type was ambiguous, we contacted the author, and in very few cases, dropped the finding from the dataset when the outcome measure type could not be resolved. One limitation of this paper is that these classifications are inherently subjective, and other researchers might have chosen different ways of classifying outcome measures.

There was some variation in outcome measure types by intervention type and delivery method. The vast majority (nearly 95%) of school-level interventions use independent measures, whereas teacher-level and supplemental interventions use the greatest share of non-independent measures (50% and 40%, respectively) relative to other intervention types. Additionally, about two-thirds of studies of school and whole-class interventions use broad measures compared with 15% of studies of interventions targeting individual students and small groups. Breakdowns of intervention type and delivery method by outcome measure type are presented in Appendix B.

Meta-analytic approach

We used meta-regression to identify statistically significant differences in effect sizes by outcome measure type, controlling for outcome domain,[4] grade level bands, intervention type, delivery method, study design, WWC study rating, WWC handbook version (2.1+ or higher), and purpose of study review. Researchers have found effect size differences by study design

[3] Given the nature of most behavioral outcomes, very few measures in the behavior area were classified as “broad;” only schoolwide measures of school climate were classified as “broad” in behavior.

[4] The technical appendix contains information about how the WWC’s outcome domains were revised for this paper.

(Baye, Lake, Inns, & Slavin, 2018; Cheung & Slavin, 2016; Wilson, Gottfredson, & Najaka, 2001), intervention types and delivery methods (Lipsey et al., 2012; Slavin and Lake, 2008), grade levels (Slavin, Lake, & Groff, 2009), and outcome domains (Dietrichson, Bøg, Filges, & Jorgensen, 2017; Fryer, 2017). In addition, the WWC standards versions may relate to effect sizes as the WWC standards became substantially more rigorous after version 2.1; the WWC study rating relates to whether the study was a high-quality RCT or QED; and the WWC purpose of review may relate to the degree of publication bias.

We estimated two meta-regression models. The first model estimates effect size differences by outcome measure type in the full sample of studies. The estimates therefore represent differences in effect sizes due to both within- and across-study variation. This model has the benefit of leveraging variation in measure types across all studies in the database to provide estimated differences in effect sizes by outcome measure type. However, these estimates may be biased by unobserved differences across studies that lead researchers to use a particular outcome measure type. For example, if researchers who have greater control of implementation fidelity are more likely to use developer measures, the developer measure estimate would be confounded with greater implementation fidelity, which we cannot observe. We therefore estimated a second model focused on *within-study* effect size differences by outcome measure type by restricting the sample to studies that had both a non-independent measure (i.e., researcher and/or developer) and independent measure (i.e., broad and/or narrow) and adding study fixed effects to the model—which estimates effect sizes differences *within* studies and therefore controls for differences in study design, implementation, and any other study-specific factors.[5] This within-study sample includes 67 studies (18% of the full sample) with both measure types. In both

[5] We dropped any covariates that were redundant with the study fixed effects.

models, we grand-mean center all covariates—with the exception of the dummy indicators for outcome measure types—to facilitate interpretation of the results. The model takes the form:

$$T_{ij} = \beta_0 + \beta_1 Researcher_{ij} + \beta_2 Developer_{ij} + \beta_3 Narrow_{ij} + \beta \boldsymbol{X_{ij}} + \eta_j + \varphi_{\mathrm{ij}} + \varepsilon_{ij}$$

$$\eta_j \sim N(0, \tau^2)$$

$$\varphi_{ij} \sim N(0, \omega^2)$$

$$\varepsilon_{ij} \sim N(0, v_{ij})$$

where $\mathrm{T_{ij}}$ is the WWC-calculated effect size estimate i in study j, β_0 is the estimated effect size for broad measures (the reference category), β_1 is the deviation from the broad effect size for researcher measures, β_2 is the deviation for developer measures, and β_3 is the deviation for narrow measures. β is a vector of regression coefficients for the covariates. $\mathbf{X}_{ij}$ is a vector of covariates that includes outcome domain; grade level band; intervention type; delivery method; study design; and WWC study rating, handbook version, and purpose of study review. η_j is the study random effect, φ_{ij} is the study-by-effect-size random effect, and ε_{ij} is the effect size random effect. τ^2 and ω^2 are estimated by the model, and v_{ij} is the observed sampling variance of $\mathrm{T_{ij}}$. The model also assumes that η_j, φ_{ij}, and ε_{ij} are mutually independent of one another. To account for the dependency of multiple findings within the same study, we conducted multivariate meta-analysis with robust variance estimation using the R packages *metafor* and *clubsandwich* (Olkin & Gleser, 2009; Pustejovsky, 2019; R Core Team, 2018; Viechtbauer, 2010).[6]

[6] We assumed effect sizes within studies to be dependent and correlated at ρ=.80, although we do not know the true covariance structure. Results were not sensitive to changes in the assumed covariance structure.

T_{ij} is the WWC-calculated effect size, which is Hedges' g (WWC Standards Handbook Version 4.1, 2020). The WWC uses the standard deviation, $\widehat{\sigma_T}$, which includes both within- and between-cluster variation for cluster studies (Hedges, 2007). We calculated inverse variance weights for each WWC-calculated effect size using the Hedges' (2007) formula when the clusters are of unequal size (see formula 20). However, robust variance estimation uses these weights for efficiency purposes only in estimating the model (Hedges, Tipton, & Johnson, 2010).

We further explored the results in several ways. First, using the full sample of studies, we examined whether results were robust to inclusion of interaction terms between outcome measure types and covariates. These sensitivity analyses informed whether effect size differences are driven by concurrences with particular topic areas, grade level bands, intervention types, delivery methods, or study designs. While the models were largely underpowered to detect the statistical significance of the interactions, we examined patterns in the average effect sizes for studies in various categories.

Second, using the within-study sample containing both independent and non-independent measures, we re-estimated the models separately for each outcome measure type and calculated the 95% prediction interval for the effect size by outcome measure type.[7] These prediction intervals inform whether the variation in the latent true effects is similar across the different outcome measure types, and therefore, whether the measures are capturing the same underlying constructs within the same outcome domains. Third, to unpack whether non-independent measures might be capturing a subset of the constructs represented by independent measures, we

[7] The 95% prediction interval contains 95% of the values of the effect sizes in the study population and is calculated by $(u - 1.96\sqrt{\tau^2 + \omega^2}, u + 1.96\sqrt{\tau^2 + \omega^2})$ where u is the average effect size, τ^2 is the between-study variance in the effect sizes, and ω^2 is the within-study variance in the effect sizes. While robust variance estimation does not require a normality assumption, estimates of τ^2 and ω^2 are accurately estimated when the normality assumption is met; if the normality assumption is not met, these estimates are approximations.

estimated the correlation coefficient between the effect sizes for independent and non-independent measures within the same studies and outcome domains after correcting the observed correlation coefficients for measurement error.[8] To calculate these correlation coefficients, we paired each effect size for non-independent measures with each effect size for independent measures within the same outcome domain and study. The magnitude of the corrected correlation coefficients informs the extent to which scores on non-independent measures explain variation in scores on independent measures.

Finally, we explored publication bias in effect sizes for each outcome measure type. We applied the Vevea and Hedges (1995) weight-function model to estimate the average effect size adjusted for publication bias for each outcome measure type (Coburn & Vevea, 2019).[9] Because this method can only be applied to study-level data, we first calculated study-level average effect sizes by outcome measure type, and then estimated the model separately for each outcome measure type.[10]

Descriptive Findings

How often are researcher or developer measures, as opposed to independent measures, the only measures used or reported in studies?

About one-fifth of studies in the WWC dataset of high-quality research include only a researcher or developer measure. This finding means that for about one-fifth of studies reviewed

[8] We corrected the observed correlation for measurement error using the following formula $r_{corrected} = \frac{r_{observed}}{\sqrt{r_{xx\prime}}\sqrt{r_{yy\prime}}}$ where $r_{corrected}$ is the correlation corrected for measurement error, $r_{observed}$ is the observed correlation, $r_{xx\prime}$ and $r_{yy\prime}$ are the reliabilities of non-independent and independent measures, respectively (Wiernik & Dahlke, 2020). We could not observe reliability information for the majority of outcome measures in our study data. Therefore, we used the average reliability (.845) for all outcome measures in WWC data. For the subsample of our study data where we did observe the reliability of outcome measures, the reliability for researcher and narrow measures was each .86.

[9] We conducted this analysis using the Vevea & Coburn Shiny application available at https://vevealab.shinyapps.io/WeightFunctionModel/.

[10] These models also include the covariates previously listed; when the covariates varied within studies, we applied the average values by study and outcome measure type.

by the WWC, it is not possible to ascertain whether an intervention's effects were observed on a measure that was independent of both the study authors and the developers involved in the intervention. Figure 1 provides frequency of outcome measure types by topic area. Overall, 38% of studies include at least one broad measure, 41% of studies include no broad but at least one narrow measure, and 21% of studies include only researcher or developer measures. There is variation across the topic areas, however, with higher percentages of studies using only researcher or developer measures in STEM (28%) and behavior (37%) than in literacy (10%).

[Figure 1]

How often do researcher and developer measures, as opposed to independent measures, result in positive and statistically significant findings? Relatedly, how often do researcher and developer measures, as opposed to independent measures, result in findings that meet the requirements for the Every Student Succeeds Act (ESSA) evidence tiers?

Even though researcher and developer measures comprise only about one-third of outcome measures in our sample, 50% of the statistically significant (p<.05) and positive findings across the topic areas are based on researcher or developer measures, while the remaining statistically significant and positive findings are based on narrow (36%) and broad (15%) measures. Put another way, 63% of developer measures and 49% of researcher measures are associated with statistically significant and positive findings compared with narrow (29%) and broad (24%) measures. This finding shows that evidence about "what works" in education is largely based on researcher and developer measures, as opposed to independent measures.

Figure 2 provides the distribution of effect sizes on the y-axis by p-values on the x-axis. The four columns represent the four different measure types, and the two rows provide effect size estimates and p-values for randomized controlled trial (RCT) and regression discontinuity design (RDD) studies (first row) and quasi-experimental design (QED) studies (second row).[11] The vertical line at p=.05 delineates the conventional cutoff for statistical significance. Visually, these figures point to three findings. First, the clustering of markers within one standard deviation in the broad panel compared with wider variation across the other three panels shows smaller effect sizes for broad measures. Second, the high proportion of markers above one standard deviation on the y-axis that are to the left of the p=.05 line suggests that the greater shares of statistically significant and positive findings on researcher and developer measures (and to a lesser extent, on narrow measures) are due to larger effect sizes on these measures. Finally, effect sizes appear to be descriptively larger in studies with RCT or RDD designs compared with studies using QED designs, not controlling for other factors.

[Figure 2]

We next examine differences in statistical significance of findings within the same study and outcome domain to explore whether one would come to the same conclusion about the effectiveness of a particular intervention when using researcher or developer measures or independent measures. This exploration assumes a frequentist approach and the typical convention of p<.05 for the Type 1 error rate. Of the 50 studies that include at least one non-independent (researcher or developer) measure **and** at least one independent (broad or narrow) measure in the same outcome domain[12]:

[11] Along with RCTs, RDDs are eligible for the highest WWC study rating of "Meets without reservations." Similar to RCTs, the treatment assignment procedure in RDDs is part of the research design and fully known (Rossi et al, 2019).

[12] There were 67 studies that included at least one non-independent (researcher or developer) measure and at least one independent (broad or narrow) measure, but only 50 studies contained both an independent and non-independent measure in the same outcome domain as determined by the WWC. We restricted to the same outcome domain for this descriptive analysis because it is plausible that an intervention may affect achievement in one outcome domain (e.g., literacy) but not another (e.g., mathematics). The meta-analytic models control for the outcome domains.

- 25 studies (50%) identify a positive and statistically significant finding on both an independent measure AND on a researcher or developer measure;
- 16 studies (32%) identify a positive and statistically significant finding on a researcher or developer measure but did not find a positive and statistically significant finding on an independent measure;
- 9 studies (18%) do not identify a positive and statistically significant finding on any type of measure; and
- 0 studies (0%) identify a positive and statistically significant finding on an independent measure but do not also find a positive and statistically significant finding on a researcher or developer measure.

Therefore, about one-third of studies would have come to a different conclusion about the effectiveness of the intervention if using only outcome measures that were independent of the researchers and developers involved with the study.

Low statistical power likely contributes to the lack of statistically significant results on independent measures and the mixed results depending on the outcome measure type. Additionally, the typical convention of p<.05 has long been criticized as inadequate for basing decisions about intervention effectiveness (Wasserstein, Schirm, & Lazar, 2019). However, these descriptive findings show that when using typical research conventions, in 32% of studies, researchers would have come to a different conclusion about the effectiveness of interventions based on outcome measure type.

For favorable findings that meet WWC standards, the WWC also assigns evidence tiers that align with the Every Student Succeeds Act (ESSA) (WWC, 2020). ESSA serves as a policy lever because some federal grant programs require schools and districts seeking school improvement funding to select educational interventions that meet the evidence levels outlined in the legislation. The WWC's Tier 1 classifies strong evidence of intervention effectiveness, and Tier 2 classifies moderate evidence of intervention effectiveness. To receive a Tier 1 badge, study findings must meet WWC standards *without* reservations, find favorable results, and include more than one site and a sample size of 350+ individuals (WWC, 2020). To receive a Tier 2 badge, study findings must meet WWC standards *with* reservations, find favorable results, and include more than one site and a sample size of 350+ individuals (WWC, 2020).

Overall, 12% of findings reviewed by the WWC in these topic areas receive a Tier 1 or 2 badge. Figure 3 shows that a disproportionate number of findings receiving a Tier 1 or 2 badge are based on researcher or developer measures, but there is variation across the topic areas. In literacy and STEM, the percent of findings with a Tier 1 or 2 badge based on researcher or developer measures is approximately 8-10 percentage points higher than the percent of findings with a Tier 1 or 2 badge based on independent (broad or narrow) measures. In the behavior topic area, however, fewer than 5% of findings receive a Tier 1 or 2 badge regardless of the outcome measure type. In sum, a disproportionate number of findings earning a Tier 1 or 2 badge are based on non-independent (researcher or developer) measures.

[Figure 3]

Meta-Analytic Findings

To what extent do effect sizes systematically vary by outcome measure type, controlling for other factors?

Across- and within-study findings

We begin with the full-sample model, which includes all studies reviewed by the WWC in the literacy, STEM, and behavior topic areas, and controls for study design and WWC rating, intervention type and delivery method, grade level band, WWC purpose of review and handbook version, and outcome domain. Patterns in effect sizes by outcome measure type may be explained by differences both across and within studies. This model has the benefit of leveraging data from the full sample, though differences in average effect sizes by outcome measure type may be conflated with unobserved factors that vary across studies.

Table 3 provides results from the full-sample model. The adjusted average effect size, which can be calculated as a linear combination of the intercept and the coefficient estimate on outcome measure type, is +0.10 for broad measures (intercept), +0.17 for narrow measures, +0.38 for researcher measures, and +0.41 for developer measures. Therefore, effect sizes using researcher measures are larger than broad measures by an average of +0.28, and larger than narrow measures by an average of +0.21. Similarly, effect sizes using developer measures are larger than broad measures by an average of +0.31, and larger than narrow measures by an average of +0.24. There is no statistically significant difference in the average effect sizes for researcher versus developer measures.[13] Finally, consistent with the visual depiction in Figure 2, narrow measures show statistically significant larger effect sizes than broad measures by an average of +0.07.

[Table 3]

Within-study findings

We turn next to our model examining effect size differences by outcome measure type *within* the same study and outcome domain. This model includes only those studies that contain at least one non-independent (researcher or developer) measure AND at least one independent (broad or narrow) measure. The model includes fixed effects for each study and outcome

[13] We conducted post-hoc Wald tests using the *metafor* R package.

domain, which allows for the examination of effect size differences by outcome measure type within the same study and outcome domain. This model also controls for covariates that vary within studies, such as study design or WWC rating. This model arguably provides the strongest evidence of whether effect sizes vary by outcome measure type because any differences across studies—such as study quality, sample characteristics, or implementation fidelity—are held constant.

Table 4 provides results from the within-study model. Within studies and outcome domains, the adjusted average effect size is +0.19 for broad measures, +0.28 for narrow measures, +0.43 for researcher measures, and +0.51 for developer measures. Therefore, effect sizes using researcher measures are larger by an average of +0.24 relative to broad measures, and by an average of +0.15 relative to narrow measures. Effect sizes using developer measures are larger by an average of +0.32 relative to broad measures, and by an average of +0.23 relative to narrow measures. Put another way, researcher and developer measures show average effect sizes that are 2.3 to 2.7 times larger than effect sizes from broad measures, and about 1.5 to 1.8 times larger than effect sizes from narrow measures within the same study and outcome domain. There is no statistically significant difference in the average effect sizes for researcher versus developer measures, nor is there a statistically significant difference in average effect sizes for broad versus narrow measures. The latter finding implies that effect sizes may not systematically vary across narrow versus broad measures after study quality, implementation fidelity, and sample characteristics are held constant. However, the 95% confidence intervals were (0.03, 0.35) for broad measures and (0.16, 0.40) for narrow measures, suggesting slightly higher average effect sizes on narrow measures than on broad measures.

[Table 4]

Results in Tables 3 and 4 indicate that while there are few differences in average effect sizes by outcome domains and study design, these differences are relatively small in magnitude compared with differences in effect sizes due to outcome measure type. As shown in Figure 4, looking *within* studies and outcome domains and using model estimates from findings in Table 4, the distributions for researcher and developer measures show larger average effect sizes, and this appears to be true across the three topic areas.

[Figure 4]

What mechanisms may explain systematic differences in effect sizes by outcome measure type?

We conducted a number of sensitivity analyses to explore plausible reasons for why effect sizes are systematically larger for non-independent measures. First, using the full study sample, we separately added multiple different sets of interaction terms to the model. These models were largely underpowered to detect the statistical significance of the interactions, but we examined patterns in the average effect sizes for studies in various categories. We present findings from these models in Table 5. We first added interactions between the three topic areas (i.e., literacy, STEM, behavior) and each of three outcome measure types—broad, narrow, and a collapsed type we called "non-independent," which combined researcher and developer due to small counts (Panel A). While the average effect sizes are larger for non-independent measures relative to broad measures in each topic area, the average effect size difference was smaller in the behavioral topic area (+0.15) than in the literacy (+0.25) or STEM (+0.35) topic areas.

[Table 5]

We then estimated a triple-interaction model with study design (i.e., RCT/RDD or QED), grade level band, and outcome measure type (Panel B). The coefficient on the interaction term between QED studies and narrow measures is positive and approaches statistical significance (p<.10), indicating that use of narrow measures may yield even larger effect sizes in QED studies. Moreover, we find that patterns in effect sizes for broad and narrow measures follow conventional wisdom in that effect sizes are generally higher in QED studies and in younger versus older grades (Hill et al., 2008; Wilson, Gottfredson, & Najaka, 2001). Patterns in effect sizes for non-independent measures are less clear across study designs and grade levels but are larger on average across each study type and grade level band relative to broad and narrow measures.

Next, we added interactions between intervention type (e.g., curriculum, policy, practice, teacher-level, supplemental, school-level) and outcome measure type, and in a separate model, between delivery method (e.g., individual, small group, whole class, schoolwide) and outcome measure type. The results are also presented in Table 5 (panels C and D, respectively). While effect sizes are typically larger for supplemental interventions and for interventions delivered individually or in small groups, effect sizes for non-independent measures are consistently larger on average for every intervention type and delivery method. Results of these sensitivity analyses should be interpreted with caution due to the small cell sizes and limited power, though they provide some evidence that the main findings are robust to different types of studies and interventions.

Second, we calculated the 95% prediction intervals of effect sizes by outcome measure type by re-estimating the meta-regression model separately for each outcome measure type and including only the studies that contained both an independent and a non-independent measure. As shown in Figure 5, the latent true effects for developer and researcher measures are much more scattered than the latent true effects for broad and narrow measures. The 95% prediction intervals range from -0.43 to 1.60 for developer measures, from -0.29 to 1.37 for researcher measures, from -0.11 to 0.46 for broad measures, and from -0.32 to 0.69 for narrow measures. One interpretation of these results is that independent and non-independent measures are not capturing the same underlying constructs, although it could still be the case that non-independent measures are capturing subscales of the independent measures.

[Figure 5]

Third, we explored to what extent scores on non-independent measures explain variation in scores on independent measures by calculating the correlation coefficient between each effect size for non-independent and independent measures within the same study and outcome domain. These results are presented in Table 6. The observed correlation coefficient was .338 across all outcome domains, which means that we would expect a true correlation of around .40 after removing measurement error. Therefore, scores on non-independent measures explain 16% of the variation in scores on independent measures in the same outcome domain and study. However, some outcome domains have more overlap between non-independent and independent measures than others. For example, scores on non-independent measures in general literacy explain 17% of the variation in scores on independent measures, whereas scores on non-independent measures in reading comprehension explain 0% of the variation in scores on independent measures. Thus, it is possible that non-independent measures cover subscales of independent measures in some outcome domains. While there may be some overlap in the constructs covered by non-independent and independent measures in some outcome domains, it is also not clear to what extent non-independent measures are capturing subscales of independent measures as opposed to different underlying constructs moderately correlated with constructs in independent measures.

Publication bias

Finally, we explored publication bias using the Vevea & Hedges (1995) weight-function model. As we note in the methods section, these findings derive from the full sample of studies and are based on study-level means of effect sizes by outcome measure type. Because this analysis does not account for multiple findings within each study, these findings should therefore be interpreted with caution.

Evidence of publication bias is presented in Table 7 for each outcome measure type. The first column shows the estimated study-average effect size, while the second column shows the effect size corrected for publication bias, with asterisks to note whether the estimates are statistically significantly different from one another. Significant differences between the two columns provide evidence of publication bias. To the extent that the estimate in first column is larger, studies in the sample have *larger* effect sizes than expected in the study population. To the extent that the second column estimate is larger, studies in the sample have *smaller* effect sizes than expected. Typically, analyses of publication bias relying on peer-reviewed journal articles find the former due to the file drawer problem. Because our sample relies on studies selected by the WWC for review and that are not necessarily peer-reviewed, publication bias could go in either direction. We find evidence of publication bias for developer measures but no evidence of publication bias for researcher measures. We also find fewer than expected statistically significant findings for independent (broad and narrow) measures.

[Table 7]

At a minimum, these findings suggest that a portion of the average effect size difference between developer and independent measures may be explained by publication bias. That is, studies that include developer measures may suffer from the file drawer problem to a greater extent than studies that do not include developer measures (Polanin, Tanner-Smith, & Hennessy, 2016; Wolf et al., 2020). For researcher measures, however, publication bias does not appear to be a key driver of systematically higher average effect sizes relative to independent measures.

The underrepresentation of statistically significant findings on independent measures is curious and may be related to WWC processes for selecting studies for review or WWC procedures when reviewing studies.[14] The WWC reviews both peer-reviewed and non-peer-reviewed studies, which may limit publication bias (John, Loewenstein, & Prelec, 2012; McBee, Makel, Peters, & Matthews, 2017). However, it is unclear why WWC processes would result in an underrepresentation of statistically significant findings only for independent measures, and not for developer and researcher ones as well.

Discussion

Conducing a meta-analysis using WWC study data, we show that effect sizes are systematically larger on researcher and developer measures than on independent (broad and narrow) measures, even when holding constant study quality, implementation fidelity, and sample and intervention characteristics. On average, effect sizes on researcher and developer measures are about 2.3 to 2.7 times larger than effect sizes on broad measures, and about 1.5 to 1.8 times larger than effect sizes on narrow measures within the same study and outcome domain.

We find larger average effect sizes for researcher and developer measures for studies in the literacy, STEM, and behavior topic areas. We also find more similar effect sizes across outcome measure types in the behavior topic area than in the literacy or STEM topic areas. More research

[14] The WWC selects studies to review for a variety of purposes, which may not result in a representative study sample.

is needed to understand why effect sizes may be more similar in magnitude across different outcome types in behavioral domains than in achievement domains.

Results of this paper provide evidence against some hypotheses for why effect sizes are systematically larger on researcher and developer measures than on independent measures. First, this paper provides countervailing evidence for the hypothesis that larger effect sizes on researcher and developer measures are caused by greater implementation fidelity or developer involvement in studies that use these types of measures. The meta-analytic model that includes fixed effects for each study, therefore holding constant implementation fidelity and study characteristics, finds larger average effects for researcher and developer measures than independent measures *within the same study*. Relatedly, meta-analytic models estimated as sensitivity tests show larger effect sizes on researcher and developer measures cannot be explained by concurrences with specific study designs, intervention types, or sample characteristics. On the other hand, this paper finds evidence of publication bias for studies using developer measures, which may contribute to effect size differences observed *across* studies. Publication bias did not appear to explain the effect size difference for researcher measures.

Results of this paper provide some evidence that researcher and developer measures may be capturing subscales of independent measures, yet the overlap is minimal in most outcome domains. The WWC organizes outcome measures into outcome domains (e.g., general literacy, general mathematics), which are somewhat related to the scope of the constructs captured in the measures. Across all outcome domains, scores on researcher and developer measures explain 16% of scores on independent measures, but this ranges from 34% of variation explained in interpersonal behavior to 0% of variation explained in reading comprehension. We provide further evidence that researcher and developer measures are not capturing the same underlying constructs by examining the 95% prediction intervals of the latent true effect sizes. We find much more variation in the distribution of the latent true effects for researcher and developer measures than for independent measures. Findings suggest that either researcher and developer measures capture something beyond the constructs covered in independent measures, or alternatively, researcher and developer measures capture a subset of the constructs covered in independent measures.

This paper cannot rule out the hypothesis that researcher and developer measures may be more properly aligned with the intervention and therefore better equipped to detect intervention effectiveness than independent measures, which may be poorly aligned with the intervention (Lipsey et al., 2012; Lynch et al., 2019; SWAT Measurement Small Group, 2020; Wilson & Lipsey, 2001). If the same researchers or developers who created the intervention also created the outcome measure, it is likely that there is tighter alignment between the measure and the intervention than when using a non-related, independent measure. In some cases, there may not even be a sufficiently aligned independent measure to estimate the impact of an intervention. Yet the hypothesis that researcher and developer measures are more properly aligned with the intervention call into question what the purpose of the research is.

If the purpose of the research is to validate the effectiveness of an intervention in a pilot study or efficacy trial, non-independent measures may be warranted. If the purpose is to help practitioners and policymakers—who are accountable for student progress on independent measures—make decisions about which interventions to implement at scale, then use of measures that are tightly aligned with the intervention may lead to inaccurate, misleading, and unrealistic conclusions about the effectiveness of the intervention. Perhaps there is a mismatch between the evidence needed by researchers or developers to validate an intervention and evidence needed by practitioners and policymakers to select interventions to implement at scale in their settings.

Effect sizes on researcher or developer measures are also generally much larger in magnitude than those on independent measures. Therefore, determining the relative effectiveness of interventions by comparing effect sizes of each without accounting for outcome measure type can result in inaccurate conclusions. Stakeholders in the field of education routinely promote specific interventions based on the magnitude of effect sizes. One example is a study that concluded that intelligent tutoring systems are more effective than other forms of tutoring based on the magnitude of the effect size (Kulik & Fletcher, 2016); the authors also noted that the mean effect size of intelligent tutoring was +0.73 on researcher measures and only +0.13 on standardized measures. Slavin (2020) pointed out that researcher-developed measures accounted for a sizeable portion of the gap between intelligent tutoring systems and other forms of tutoring.

Practitioners may be most interested in the qualitative rating of an intervention's effectiveness, meaning an overall summary of whether an intervention "worked" (SWAT, 2020). An open question is whether a favorable finding on researcher or developer measures translates into something meaningful for practitioners. The best-case scenario is that a favorable finding on a researcher or developer measure is a signal that students have learned concepts and skills along the way towards mastering required academic content. Yet another scenario is that a favorable finding on a researcher or developer measure has no bearing on how well students will perform on a formative or summative assessment in the same content area. This paper suggests minimal overlap between constructs covered on a researcher or developer measure and constructs covered on an independent measure. Therefore, there is not sufficient evidence to conclude that favorable findings on researcher and developer measures will translate into meaningful findings on independent measures. In addition, Song and Herman (2010) argue that using a small subset of constructs to claim effectiveness of an intervention on a broad construct is "unwarranted at best and misleading at worst" (p. 360).

One limitation of this paper is that the classification of outcome measure types is inherently subjective, and other researchers may have classified some outcome measures differently. However, given that the average effect sizes of researcher and developer measures identified in this paper are consistent with those previously identified in the literature, it is unlikely that modifications to the categorization of outcome measures would have resulted in a different conclusion. Another limitation is that the analysis is limited to outcomes in the literacy, STEM, and behavioral domains. More research is needed to understand effect size patterns by outcome measure type in other outcome domains.

One implication of this paper is, whenever possible, researchers should include in their studies outcome measures that have practical significance for practitioners and policymakers. If such measures are not available or appropriate, researchers should always aim to include independent measures along with any researcher and developer measures to compare findings and verify that the intervention is moving the needle according to both outcome measure types. This implication is aligned with the Standards for Excellence in Education Research (Institute of Education Sciences, 2021), which call for the use of high-quality and relevant outcome measures.

Researchers could also use one of several existing statistical approaches to account for differences in effect sizes by outcome measure type. When conducting systematic reviews of educational research, statistical approaches, such as meta-regression or Bayesian modeling, can adjust both the statistical significance and magnitude of effect sizes, accounting for larger average effect sizes when using researcher or developer measures. When reporting findings from a single study, researchers could provide context about the study effect size, such as what is the typical distribution of effect sizes depending on outcome measure type, content area, grade levels, and other factors (Hill et al., 2008; Kraft, 2020). Finally, researchers could be more intentional about how they characterize evidence to different audiences. For example, researchers could clearly articulate when the study provides formative feedback on an intervention and when the study provides evidence that supports adoption of the intervention at scale. Given that outcome measure type is by far the most predictive variable explaining the magnitude of effect sizes in studies reviewed by the WWC, researchers should use the tools available to them to help practitioners and policymakers make sense of the evidence to understand which educational interventions might work best in their contexts.

References

Baird, M. D., Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. Educational Researcher, 48(4), 217–228.

Baye, A., Lake, C., Inns, A., & Slavin, R. (2018). A synthesis of quantitative research on reading programs for secondary students. *Reading Research Quarterly.*

Bloom, H. S., Hill, C. J., Black, A. R., Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289–328.

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

Coburn, K., & Vevea, J. (2019). *weightr: Estimating weight-function models for publication bias*. R package version 2.0.2. Retrieved from https://CRAN.R-project.org/package=weightr

de Boer, H., Donker, A., & van der Werf, M. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*(4), 509-545.

Dietrichson, J., Bǿg, M., Filges, T., & Jorgensen, A. K. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research 87*(2), 243-282.

Fryer Jr, R. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95-322). North-Holland.

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness*, *13*(2), 401-427.

Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*, 341–370.

Hedges, L., Tipton, E., & Johnson, M. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39-65.

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. doi:10.1111/j.1750-8606.2008.00061.x

Institute for Education Sciences. (2021). *Standards for excellence in education research.* https://ies.ed.gov/seer/index.asp

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532.

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

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, *86*(1), 42-78.

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, *22*(3), 215-243.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and offenders*, *4*(2), 124-147.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*.

Lortie-Forgues, H., Inglis, M. (2019). Rigorous large-scale RCTs are often uninformative: Should we be concerned? Educational Researcher, 48(3), 158–166.

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

McBee, M., Makel, M., Peters, S., & Matthews, M. (2017). A manifesto for open science in giftedness research. Retrieved from osf.io/qhwg3

Olkin, I., & Gleser, L. (2009). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 357-376.

Pellegrini, M., Inns, A., Lake, C., & Slavin, R. (2019, March). *Effects of researcher-made versus independent measures on outcomes of experiments in education.* Paper presented at the annual meeting of the Society for Research on Educational Effectiveness. Washington, DC.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of experimental criminology*, *1*(4), 435-450.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*(1), 207-236.

Pustejovsky, J. (2019). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections.* R package version 0.3.5. Retrieved from https://CRAN.R-project.org/package=clubSandwich

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019). *Evaluation: A Systematic Approach* (Eighth edition). Sage.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(5), 369-393.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Slavin, R. (2020). Meta-analysis or muddle-analysis? *Robert Slavin's Blog.* https://robertslavinsblog.wordpress.com/2020/10/08/meta-analysis-or-muddle-analysis/

Slavin, R., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Education Research, 78*(3), 427-515.

Slavin, R., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research 79*(2), 839-911.

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, *32*(3), 351-371.

Statistics, Website, and Training (SWAT). (2020). *How education leaders use the What Works Clearinghouse website* (WWC report). American Institutes for Research (AIR). Manuscript awaiting peer review.

Statistics, Website, and Training (SWAT) Measurement Small Group. (2020). *Preliminary analysis of effect sizes associated with researcher-developed measures* (WWC report). American Institutes for Research (AIR). Manuscript awaiting peer review.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. doi:10.1007/BF02294384

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software,* 36(3), 1-48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*(sup1), 1-19.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94-123.

Williams, R., Citkowicz, M., Lindsay, J., Miller, D., & Walters, K. (2022). Heterogeneity in mathematics intervention effects: Results from a meta-analysis of 191 randomized experiments. *Journal of Research on Educational Effectiveness.* https://airshinyapps.shinyapps.io/math_meta_database/

Wilson, D., Gottfredson, D., & Najaka, S. (2001). School-based prevention of problem behaviors: A meta-analysis. *Journal of Quantitative Criminology*, *17*(3), 247-272.

Wilson, D., & Lipsey, M. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, *6*(4), 413.

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness*, *13*(2), 428-447.

WWC Procedures Handbook Version 4.1. (2020). What Works Clearinghouse Producers Handbook, Version 4.1. US Department of Education, Institute of Education Sciences. *National Center for Education Evaluation and Regional Assistance.* https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

WWC Standards Handbook Version 4.1. (2020). What Works Clearinghouse Standards Handbook, Version 4.1. US Department of Education, Institute of Education Sciences. *National Center for Education Evaluation and Regional Assistance.* https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

WWC. (2020). *Using the WWC to find ESSA tiers of evidence.* What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/essa

Tables

Table 1. Literature review

Reference	Topic areas	Contrast of outcome measure type	Average effect size difference
Cheung & Slavin, 2016	Literacy, mathematics, science, technology, early childhood	Researcher v. Independent	+0.16 for researcher
de Boer et al., 2014	Literacy, mathematics, science	Self-developed v. Independent of the intervention	+0.25 for self-developed
Gersten et al., 2020	Literacy	Researcher/developer v. Standardized or pre-existing	+0.11 for researcher/developer[1]
Li & Ma, 2010	Computer technology	Non-standardized v. Standardized	+0.27 for non-standardized
Lipsey et al., 2012	All subjects but mostly literacy and mathematics	Specialized researcher v. Standardized narrow v. Standardized broad	+0.31 for specialized researcher +0.16 for standardized narrow
Lynch et al., 2019	Mathematics, science	Researcher v. Standardized commercial v. State or district standardized	+0.27 for researcher +0.01 for standardized commercial
Pellegrini et al., 2019	Literacy, mathematics using WWC data	Researcher/developer v. Independent	+0.27 for researcher/developer
Ruiz-Primo et al., 2002	Science	Close (same concepts, same assessment) v. Proximal (same concepts, new assessment) v. Distal (large-scale assessment)	+0.31 to 1.00 for close relative to proximal
SWAT Measurement Small Group, 2020	Literacy, mathematics using WWC data	Researcher/developer v. Narrow v. Not narrow	+0.25 for researcher/developer +0.12 for narrow
Williams et al., n.d.	Mathematics	Unstandardized v. Standardized	+0.24 for unstandardized
Wilson & Lipsey, 2001	Mostly education but some behavior and psychology	Researcher v. Standardized or published instrument	+0.13 for researcher

[1] Difference was not statistically significant at $p<0.05$.

Table 2. Study data descriptives

	Full sample				Within-study sample			
	Number of Findings		Number of Studies		Number of Findings		Number of Studies	
	N	%	N	%	N	%	N	%
Research design								
Randomized controlled trial (RCT)	1,212	78%	283	76%	288	73%	51	76%
Regression discontinuity (RDD)	7	<1%	1	<1%	0	0%	0	0%
Quasi-experimental (QED)	334	22%	89	24%	105	27%	16	24%
WWC study rating								
Without reservations	1,087	70%	238	64%	266	68%	46	69%
With reservations	466	30%	135	36%	127	32%	21	31%
WWC standards version								
Version 2.1+	648	42%	195	52%	135	34%	28	42%
Purpose of review								
Department-funded	65	4%	18	5%	7	2%	2	3%
Grant competition	300	19%	82	22%	53	14%	8	12%
IES performance measure	66	4%	15	4%	29	7%	4	6%
Intervention report	755	49%	178	48%	229	58%	40	60%
Practice guide	72	4%	16	4%	0	0%	0	0%
Quick review	116	7%	20	5%	38	10%	5	7%
Individual study review	179	12%	44	12%	37	9%	8	12%
Grade levels[1]								
Grades PK–K	185	12%	31	8%	91	23%	13	20%
Grades K–3	456	30%	98	27%	107	28%	18	28%
Grades 3–6	425	28%	107	29%	84	22%	14	22%
Grades 6–9	300	20%	84	23%	59	15%	12	18%
Grades 9–12	162	11%	49	13%	45	12%	8	12%
Topic areas								
Literacy	858	55%	220	53%	237	60%	38	57%
STEM	356	23%	135	33%	68	17%	16	24%
Behavior	339	22%	57	14%	88	23%	13	19%
Delivery method								
Individual	391	25%	92	25%	136	35%	20	30%

	Full sample				Within-study sample			
	Number of Findings		Number of Studies		Number of Findings		Number of Studies	
	N	%	N	%	N	%	N	%
School	194	12%	37	10%	13	3%	2	3%
Small group	401	26%	92	25%	80	20%	16	24%
Whole class	589	38%	170	46%	160	41%	29	43%
Intervention type								
Curriculum	329	21%	96	26%	54	14%	12	18%
Policy	6	<1%	1	<1%	6	1%	1	2%
Practice	99	6%	33	9%	14	4%	2	3%
School level	39	3%	9	2%	0	0%	0	0%
Supplemental	630	41%	135	36%	220	56%	37	55%
Teacher level	21	1%	12	3%	4	1%	1	2%
Outcome measure type[2]								
Broad	335	22%			42	11%		
Narrow	674	43%			150	38%		
Researcher	471	30%			132	34%		
Developer	73	5%			69	17%		
Total	1,553	100%	373	100%	393	100%	67	100%

[1] Grade-level bands were determined based on the closest fit to the grade levels included in the study. Grade-level information was missing for a few studies.

[2] The outcome measure types vary at the finding, not the study, level.

Notes. Within each subheading (e.g., research design, purpose of review, etc.), percentages may not sum to 100% due to rounding. The counts by topic area and delivery method may include the same study more than once if the study related to more than one topic area or delivery method. In these cases, the percentages can sum to greater than 100%. Program type counts sum to less than 100% because there are studies that do not fit into any of the program types.

Table 3. Meta-regression results looking both within and across studies

		Estimate	**Standard error**	**t**	**Df**
	Intercept	0.10^{***}	0.02	5.03	128.28
Outcome type	Broad	*Reference*			
	Narrow	0.07^{**}	0.03	2.73	146.94
	Researcher	0.28^{***}	0.03	8.28	119.97
	Developer	0.31^{***}	0.06	5.48	29.64
Study design & rating	QED	0.11^{**}	0.04	2.96	77.93
	With reservations	-0.06^{+}	0.03	-1.84	61.11
Standards version	Version 2.1+	0.02	0.03	0.56	140.95
Purpose of study review	Department-funded	0.04	0.05	0.78	25.63
	Grant competition	0.05	0.03	1.57	111.40
	IES performance	-0.05	0.05	-0.99	23.17
	Intervention report	*Reference*			
	Practice guide	-0.05	0.09	-0.59	14.77
	Quick review	-0.03	0.05	-0.66	20.61
	Single study review	0.10^{+}	0.05	1.97	76.65
Outcome domain	Alphabetics	0.06^{+}	0.04	1.72	78.57
	Comprehension	-0.03	0.03	-1.10	108.37
	Reading fluency	-0.02	0.05	-0.40	60.92
	Inter. behavior	-0.06	0.05	-1.24	51.84
	Intra. behavior	-0.12^{**}	0.04	-3.25	34.51
	Literacy	*Reference*			
	Math	0.04	0.03	1.59	77.29
	Progress in school	-0.15	0.13	-1.13	2.32
	Science	0.01	0.07	0.20	21.74
	Writing	0.11	0.08	1.34	9.18
Model info	Finding N	1,553			
	Study N	373			
	τ^2	0.01			
	ω^2	0.07			

Notes: ***p<.001, **p<.01, *p<.05, $^{+}$p<.10. Do not trust estimates when the degrees of freedom are less than four. The full-sample model also controlled for program types, delivery methods, and grade-level bands.

Table 4. Meta-regression results looking within studies and outcome domains

		Estimate	**Standard error**	**t**	**Df**
	Intercept	0.19[*]	0.08	2.22	24.31
Outcome type	Broad	*Reference*			
	Narrow	0.09	0.06	1.35	29.28
	Researcher	0.24[**]	0.07	3.51	20.57
	Developer	0.32[***]	0.07	4.35	21.81
Study design & rating	QED	0.48[**]	0.09	5.10	4.78
	With reservations	-0.14	0.16	-0.87	7.32
Outcome domain	Alphabetics	0.04	0.07	0.56	10.15
	Comprehension	-0.04	0.07	-0.58	11.43
	Reading fluency	0.05	0.14	0.35	12.95
	Interpersonal behavior	-0.04	0.14	-0.29	4.42
	Intrapersonal behavior	-0.20	0.03	-6.56	1.40
	Literacy	*Reference*			
	Math	-0.16	0.04	-3.81	1.55
	Progress in school	0.18	0.14	1.31	4.48
	Science	0.55+	0.24	2.31	5.53
	Writing	-0.02	0.03	-0.47	20.52
Model info	Finding N	393			
	Study N	67			
	τ^2	0.05			
	ω^2	0.10			

Notes: [***]p<.001, [**]p<.01, [*]p<.05, [+]p<.10. Do not trust estimates when the degrees of freedom are less than four. The model also included fixed effects for each study. All other covariates were redundant with the study fixed effects.

Table 5. Meta-analytic averages from sensitivity analyses

	Broad	Narrow	Non-independent
Panel A			
Topic area (rows) x outcome measure type (columns)			
Literacy	0.09	0.14	0.34
STEM	0.10	0.16	0.46
Behavior	0.24	0.23	0.41
Panel B			
Grade level x outcome measure type x study design (subheads)			
	Randomized controlled trials and regression-discontinuity designs		
Grades PK–K	[a]	0.31	0.60
Grades K–3	0.12	0.12	0.34
Grades 3–6	0.10	0.11	0.36
Grades 6–9	0.07	0.17	0.32
Grades 9–12	0.06	0.16	0.40
	Quasi-experimental designs		
Grades PK–K	-0.24	0.28	0.50
Grades K–3	0.14	0.24	0.39
Grades 3–6	0.12	0.23	0.40
Grades 6–9	0.12	0.32	0.40
Grades 9–12	0.09	0.29	0.47
Panel C			
Intervention type x outcome measure type			
Curriculum	0.02	0.14	0.28
Policy	[a]	0.08	0.17
Practice	0.06	0.12	0.42
School level	0.14	0.15	[a]
Supplemental	0.19	0.19	0.40
Teacher level	0.16	0.20	0.88
Panel D			
Delivery method x outcome measure type			
Individual	0.15	0.23	0.40
School	0.08	0.13	0.25
Small group	0.15	0.10	0.46
Whole class	0.07	0.19	0.36

Notes. Cells provide linear combinations of all relevant interactions and main effects. The models controlled for all other covariates included in Table 3, with the exception of outcome domains for Panel A since they related to the topic areas. When the covariates were not pertinent to calculating the meta-analytic averages for each category, they were grand-mean centered.

[a] No studies in these cells.

Table 6. Correlations between scores on non-independent and independent measures within the same studies and outcome domains

Outcome domain	$r_{corrected}$	Pairwise N	Study N
Alphabetics	0.249	162	9
General literacy	0.417	22	6
General math	0.408	68	14
Interpersonal	0.582	45	9
Intrapersonal	0.190	51	4
Reading comprehension	-0.003	34	6

Notes. The observed correlation coefficients are corrected for measurement error using the observed reliability of outcome measures in WWC data. Pairwise N represents the number of pairwise correlations, and study N represents the number of studies in the analysis. Correlation coefficients are only presented for outcome domains that have both non-independent and independent measures in the same outcome domain and study.

Table 7. Potential for publication bias

	Study-average effect size	With Vevea-Hedges correction
Broad	0.107	0.172***
Narrow	0.192	0.225*
Researcher	0.433	0.398
Developer	0.439	0.282*

Notes. ***p<.001, *p<.05 indicates statistical significance from the likelihood ratio test that indicates whether the model that adjusted for publication bias was a better fit for the data. For the Vevea and Hedges (1995) weight-function model, a two-tailed p-value cutoff of .025 was selected. The model with all of the covariates did not converge for the developer effect sizes. The estimates provided for the developer effect sizes are therefore based on a model without any covariates, though including subsets of covariates did not substantially change the results.

Figures

Figure 1. Number of studies meeting WWC standards by outcome measure type and topic area

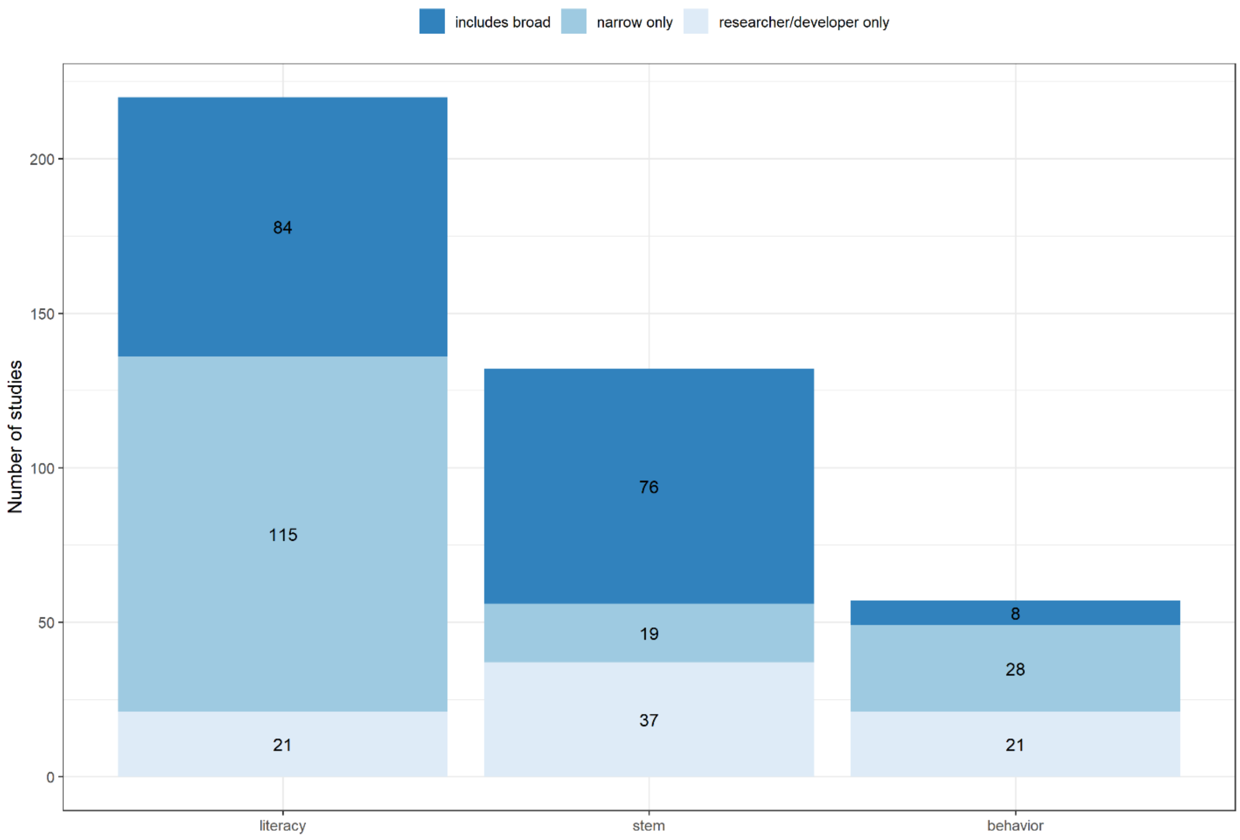


Figure 2. Distributions of effect size by p-values and outcome measure type

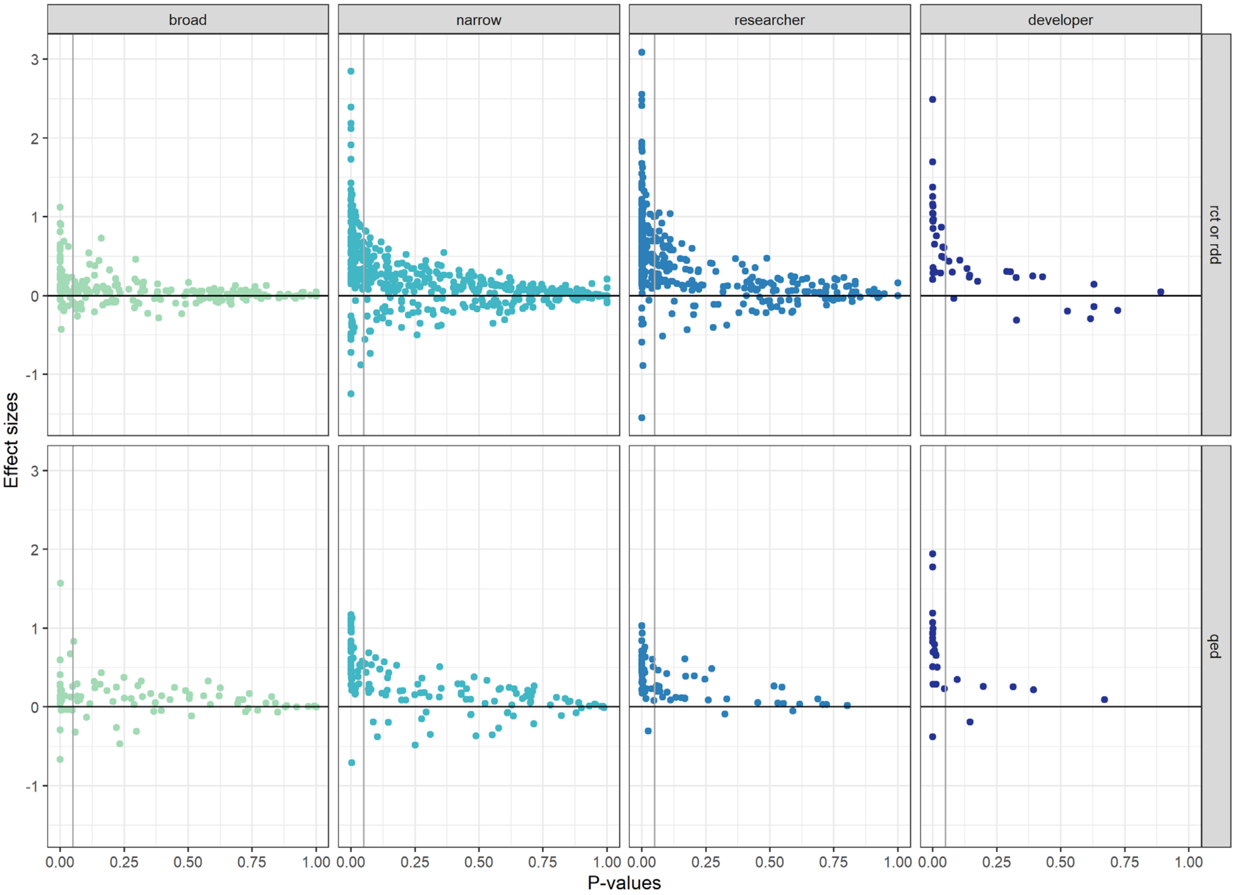
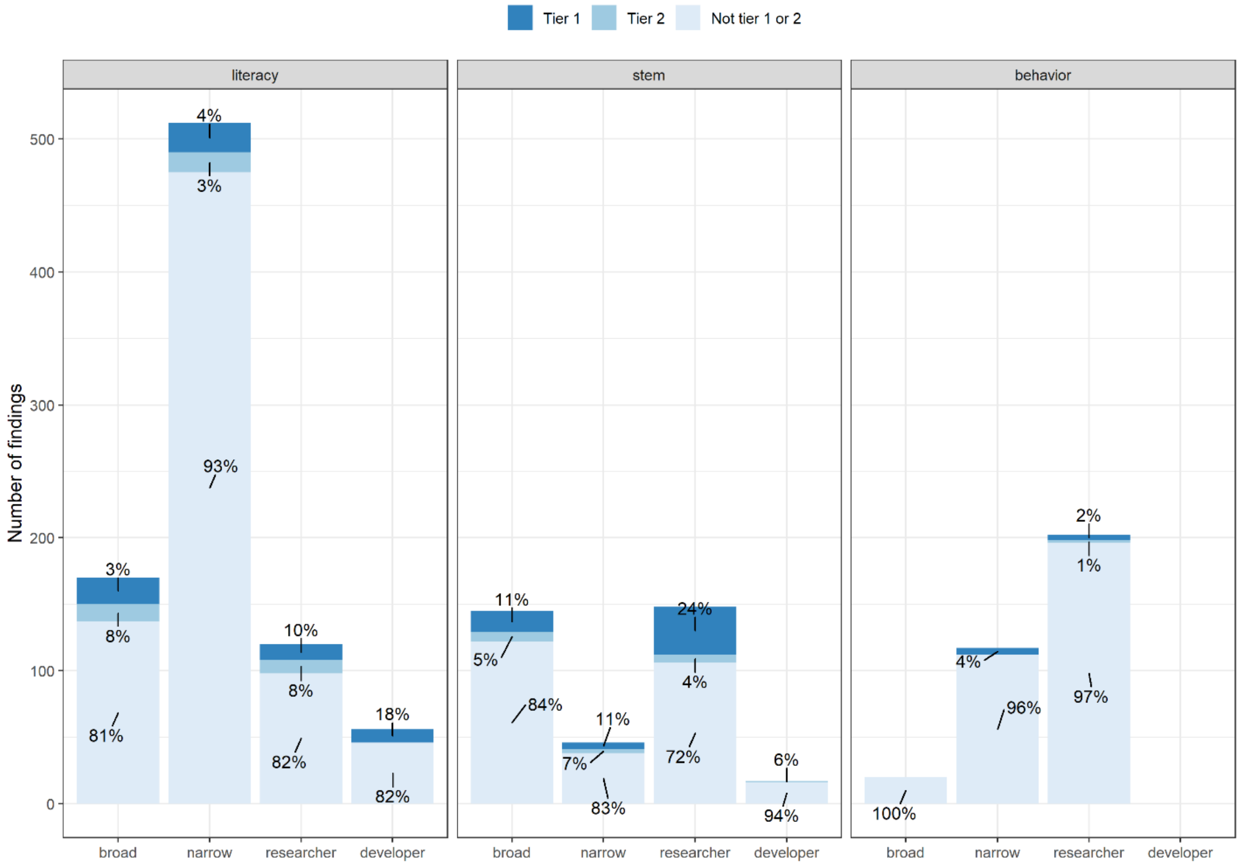
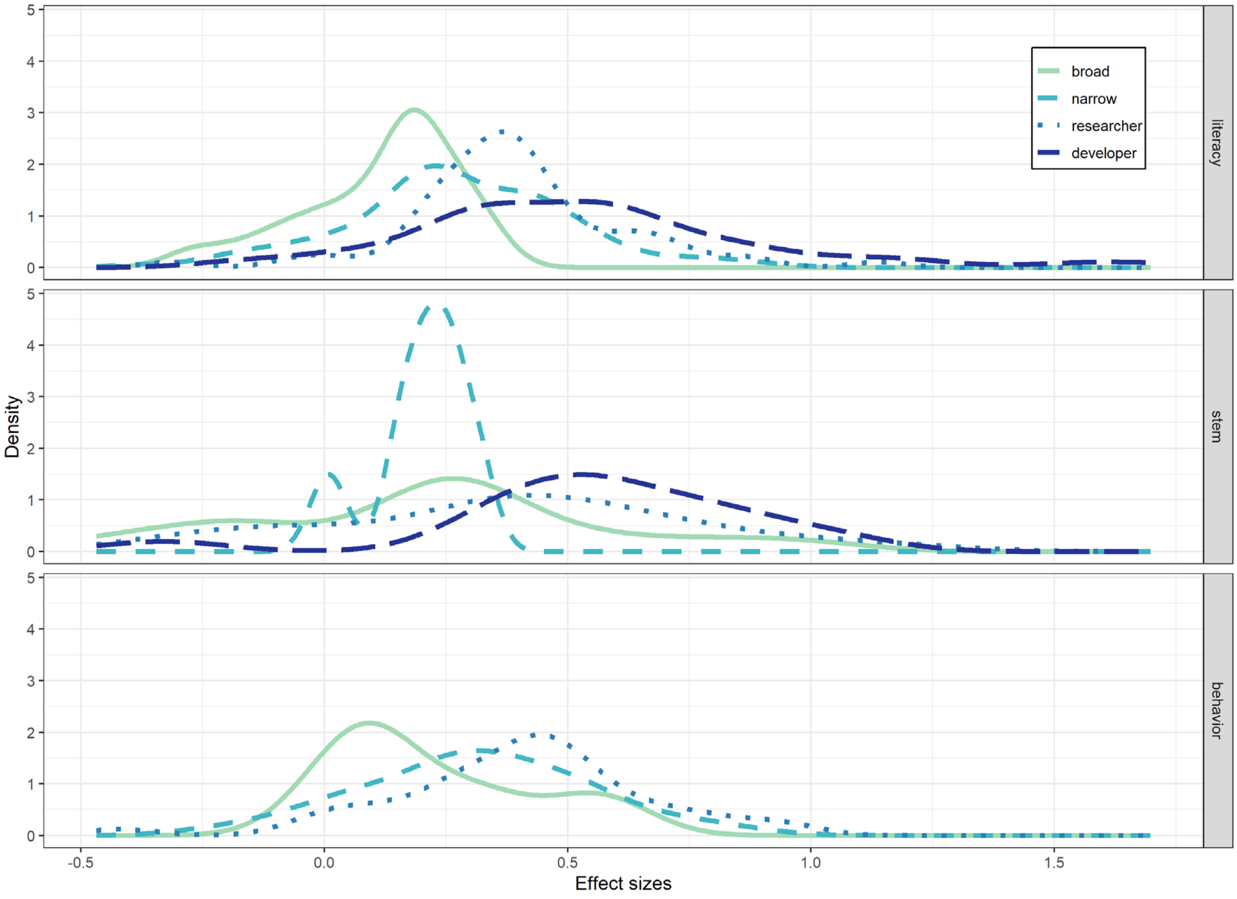


Figure 3. WWC evidence tier badges by outcome measure type and topic area



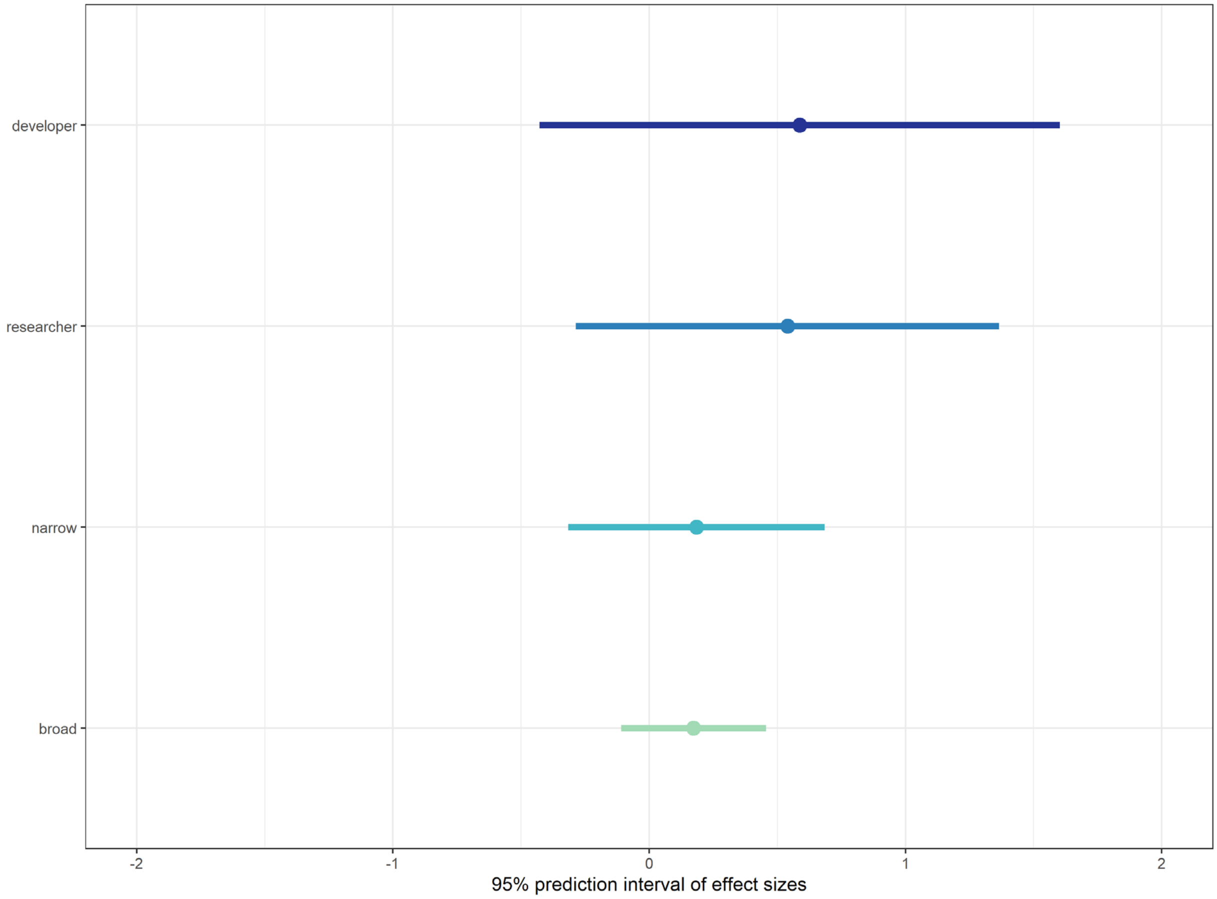
Note: Counts duplicate findings that relate to more than one topic area.

Figure 4. Distributions of the empirical Bayes effect size predictions by topic area



Note: This figure is based on predictions from the within-study meta-regression model.

Figure 5. 95% prediction intervals of effect sizes by outcome measure type



Notes: This figure is based on meta-regression models run separately by outcome measure type for the subset of studies that contain both an independent and non-independent measure. The dots represent the average effect size by outcome measure type, and the line segments represents the 95% prediction interval of effect sizes.