What Works Clearinghouse Working Paper

Average differences in effect sizes by outcome measure type

Betsy Wolf

February 2021

About the WWC

The What Works Clearinghouse (WWC) reviews existing research on programs, products, practices, and policies in education. The goal is to provide educators with the information they need to make evidence-based decisions. The WWC focuses on the results from high-quality research to answer the question, "What works in education?"

Acknowledgements

The author thanks colleagues at the Institute of Education Sciences (IES) for providing technical direction and logistical support for this working paper: Matthew Soldner, Commissioner of the National Center for Education Evaluation and Regional Assistance (NCEE), Elizabeth Eisner, the Associate Commissioner of the Knowledge Utilization Division at NCEE, Jonathan Jacobson, Senior Research Scientist at NCEE, and Erin Pollard, Education Research Analyst at NCEE. The author also thanks members of the WWC Statistics, Website, and Training (SWAT) Team and the Statistical, Technical, and Analysis Team (STAT) Measurement Small Working Group for providing comments and engaging in thoughtful conversations about the implications of this working paper.

What Works Clearinghouse Working Paper

February 2021

Average differences in effect sizes by outcome measure type

Wolf, R.

The What Works Clearinghouse (WWC) seeks to provide practitioners information about "what works in education." One challenge in understanding "what works" is that effect sizes may not be comparable across studies, which limits the ability to compare the relative effectiveness of multiple interventions. One factor that consistently relates to the magnitude of effect sizes is the type of outcome measure. This paper uses WWC study data to examine differences in average effect sizes by outcome measure type. Controlling for other factors and using advanced meta-analysis, effect sizes found on researcher and developer measures are substantially larger on average than those found on independent measures not related to the intervention under study or the study authors. One implication of this finding is that the WWC should consider whether findings based on researcher and developer measures should be differentiated from those based on independent measures to meet the evidence needs of all WWC stakeholders.

Introduction

The What Works Clearinghouse (WWC) reviews rigorous research on educational practices, policies, programs, and products with a goal of identifying "what works" and making that information accessible to practitioners. One complication in communicating "what works" to practitioners is that effect sizes—the degree to which an intervention produces positive (or negative) outcomes—are not comparable across different interventions, in large part due to differences in study characteristics (Wilson & Lipsey, 2001). Therefore, one challenge is that the WWC may not currently provide practitioners with enough information to make the most informed decisions in selecting educational interventions.

One study characteristic that has been shown to significantly relate to effect sizes is the type of outcome measure used in the study. Therefore, ignoring outcome measure type may yield effect sizes that are incomparable across studies of different interventions. Researchers have consistently identified larger average effect sizes when outcome measures were created either by study authors or researchers involved with the development of the intervention than when outcome measures were standardized or created by third parties (SWAT Measurement Small Group, 2020; Cheung & Slavin, 2016; de Boer et al., 2014; Li & Ma, 2010; Lipsey et al., 2012; Lynch et al., 2019; Pellegrini et al., 2019; Wilson & Lipsey, 2001). Less is known about differences in effect sizes for outcome measures covering fewer concepts ("narrow") versus measures covering many concepts ("broad"), but a few studies have found larger average effect sizes when using narrow versus broad measures (SWAT Measurement Small Group, 2020; Lipsey et al., 2012). Moreover, measures may be designed for different purposes, such as to gauge implementation fidelity or determine intervention effectiveness, which calls into question whether effect sizes should be compared across outcome measures with different purposes.

The WWC study database contains several different types of outcome measures, and one question that has not yet been thoroughly addressed is to what extent effect sizes in the WWC

systematically differ by outcome measure type. For the purpose of this paper, the outcome measure types are defined by the following mutually exclusive categories:

- **Broad**: Measures intended to capture student achievement in a content area, schoolwide climate, or general educational outcomes. This category includes state and district assessments, national surveys and assessments[1], grade point average, graduation rates, and school disciplinary data.
- **Narrow**: Measures intended to capture student achievement at a more granular level than a content 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 (credit, grades, etc.).
- **Developer**: Measures that were developed for a particular intervention and typically only used when the intervention is also being implemented.
- **Researcher**: Measures developed by study authors, including measures that were created by selecting specific items from preexisting scales.

Research questions

The purpose of this paper is to examine to what extent effect sizes in WWC study data systematically vary by outcome measure type, with a particular focus on researcher and developer measures. Findings may be used to inform whether the WWC should adjust the way it reports findings to account for systematic differences in effect sizes, which may hinder "apples-to-apples" comparisons of intervention effectiveness across studies using different types of outcome measures. This paper addresses the following research questions:

1) To what extent do effect sizes systematically vary by outcome measure type, controlling for other factors?
2) How often are researcher, developer, or narrow measures the only measures used in studies reviewed by the WWC?
3) What percent of the positive and statistically significant findings in the WWC are based on researcher, developer, or narrow measures?
4) If the WWC's ESSA evidence tier badges were restricted to non-researcher and non-developer outcome measures only, how many studies reviewed by the WWC would lose an ESSA Tier 1 or 2 badge?

Literature review

Researchers have consistently found larger average effect sizes for researcher and developer measures relative to "independent" or "standardized" measures, ranging from +0.11 to +0.31, typically in terms of Cohen's d or Hedges' g (SWAT Measurement Small Group, 2020; 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; Williams et al., n.d.; Wilson & Lipsey, 2001). In addition, researchers have identified average effect sizes that were +0.12 to +0.16 larger for narrow measures than for broad measures (SWAT Measurement Small Group, 2020; Lipsey et al., 2012). Table 1 summarizes the results of the literature review.

[1] National assessments are commercial or government assessments used by school districts or post-secondary institutions across the country to assess competency in a content area. To be classified as broad, the measure's composite or subscale score had to be used.

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 (e.g., letter-word identification) v. Not narrow (e.g., reading comprehension)	+0.25 for researcher/developer +0.12 for very 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<.05.

While there is generally consensus among researchers that researcher and developer measures yield larger average effect sizes relative to independent measures, researchers have speculated different reasons for *why* we might see 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 these measures are narrower compared with independent measures; that is, researcher and developer measures capture constructs on a small domain, whereas standardized tests capture constructs on a broad domain or multiple domains. For example, 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 explicitly 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 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). This hypothesis has been countered by one study that examined effect sizes *within* studies—therefore holding constant any differences in implementation fidelity across studies—and found larger average effect sizes for researcher and developer measures relative to independent ones (SWAT Measurement Small Group, 2020).

A third hypothesis is that researcher and developer measures are more properly aligned with the intervention, and therefore have greater content validity to detect intervention effectiveness than independent measures (Lipsey et al., 2012; Lynch et al., 2019; SWAT Measurement Small Group, 2020; Wilson & Lipsey, 2001). Moreover, broad, independent measures may be poorly aligned with the intervention and therefore ill-equipped to detect intervention effectiveness. On the other end of the spectrum, the WWC prohibits the "over-alignment" of measures with the intervention; for studies to meet WWC standards, students in the treatment condition cannot be exposed to the tested concepts when comparison students are not (WWC Standards Version 4.1, 2020).

A fourth hypothesis is that effect sizes for researcher and developer measures may be larger due to developer involvement in the study rather than bias in the outcome measures themselves (Petrosino & Soydan, 2005; Wolf et al., 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 process, and these decisions could be made to yield the most favorable study findings possible (Simmons, Nelson, & Simonsohn, 2011; p. 1359). Another 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).

A final hypothesis is that researcher or developer measures may be less valid and reliable than standardized measures with well-documented psychometric properties, which could lead to inflated effect sizes (Li & Ma, 2010). To meet WWC standards, the WWC 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 content within; and exhibit a minimum reliability of a Cronbach 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). However, these criteria regarding validity and reliability may be less stringent than those required of widely administered standardized tests.

In summary, researchers have consistently identified larger average effect sizes on researcher and developer measures than on independent measures. Researchers have also articulated multiple plausible explanations for *why* average effect sizes may be larger on researcher or developer measures, but there is not consensus in the field. Wilson and Lipsey (2001) noted that "issues related to the quality and appropriateness of outcome measurement are not extensively discussed in the literature," and therefore, "the operationalization for the dependent variable is generally not discussed or explained in any depth in reports" (p. 425). They further noted that studies' lack of attention to outcome measure type is based on the "assumption that this matter is not especially problematic" (p. 425) when results from meta-analyses indicate that "methodological choices made by the researcher have nearly as much influence on observed effect sizes as the features of the intervention phenomena under study" (p. 423). They conclude by calling for more attention to the operationalization of outcome measures.

One question that remains unexamined in the literature is, for a particular intervention, to what extent do positive effects on a researcher or developer measure translate into 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 meaningful information about student progress and performance to practitioners and policymakers? More research is needed to address this question.

Data

This paper uses WWC study data in the literacy; science, technology, engineering, and mathematics (STEM); and behavior topic areas as a starting point to explore the differences in magnitude and statistical significance of effect sizes by outcome measure type. The technical appendix contains more information about the compilation of the data. The data analyzed in this paper represent about half of the findings available in the WWC study data across all topic areas.

This paper analyzes 1,553 findings from 373 studies that meet WWC standards across the literacy, STEM, and behavior topic areas. About three-quarters (76%) of the study data were from randomized controlled trials (RCTs), most of which met WWC standards without reservations. About half (52%) of studies were reviewed under the WWC Standards Version 2.1 or higher. About half (48%) of the study data came from intervention reports, and about one-fifth (22%) of the study data came from evidence cited for U.S. Department of Education grant competitions. Findings spanned grade levels from early childhood to high school, with the most findings in the upper elementary grades, and with fewer findings on both ends of the grade level spectrum. Just over half (55%) of the findings related to literacy, 20% to mathematics, 22% to behavior, and only 3% to science.

Outcome measure types varied within studies and were coded as broad, narrow, researcher, or developer measures.[2] Narrow measures constituted 43% of the findings in the WWC study data in these topic areas, followed by researcher measures (30%), broad measures (22%), and developer measures (5%). Given the nature of most behavioral outcomes, very few measures in the behavior topic area were

[2] The author classified each measure by reviewing WWC resources and the original studies. The categories of outcome measure types were mutually exclusive. 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. The classification of each outcome measure for the purpose of this paper can be found here.

classified as “broad”; only schoolwide measures that captured school climate were classified as “broad.” One limitation of this paper is that these classifications are inherently subjective. These descriptive statistics are shown in Table 2.

Table 2. Study data descriptives

	Number of Findings		Number of Studies	
	N	%	N	%
Research Design				
Randomized controlled trial (RCT)	1,212	78%	283	76%
Regression discontinuity (RDD)	7	<1%	1	<1%
Quasi-experimental (QED)	334	22%	89	24%
Study Rating				
Without reservations	1,087	70%	238	64%
With reservations	466	30%	135	36%
Standards Version				
Version 2.1+	648	42%	195	52%
Purpose of Review				
Department-funded	65	4%	18	5%
Grant competition	300	19%	82	22%
IES performance measure	66	4%	15	4%
Intervention report	755	49%	178	48%
Practice guide	72	4%	16	4%
Quick review	116	7%	20	5%
Single study review	179	12%	44	12%
Grade Levels[1]				
Grades PK–K	185	12%	31	8%
Grades K–3	456	30%	98	27%
Grades 3–6	425	28%	107	29%
Grades 6–9	300	20%	84	23%
Grades 9–12	162	11%	49	13%
Content areas				
Literacy	858	55%	220	53%
Mathematics	306	20%	117	28%
Science	50	3%	18	4%
Behavior	339	22%	57	14%
Outcome measure type[2]				
Broad	335	22%		
Narrow	673	43%		
Developer	73	5%		
Researcher	472	30%		
Total	1,553	100%	373	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. The percentages may not sum to 100% due to rounding error. The counts by content areas may include the same study more than once if it related to more than one content area.

Methods

Meta-regression was used to identify statistically significant differences in effect sizes by outcome measure type, controlling for the following covariates:

- Outcome domain[3]
- Grade level bands
- Program type
- Program delivery method
- Study design
- WWC study rating
- Version of handbook (2.1+ or higher)
- Purpose of study review

Multivariate meta-analysis with robust variance estimation and the R packages *metafor* and *clubsandwich* were used to account for the dependency of multiple findings within the same study.[4] Effect sizes were calculated by the WWC, and the WWC uses Hedges' g as the effect size metric (WWC Standards Handbook Version 4.1, 2020).

Two meta-regression models were estimated. The first model estimates *within-study* effect size differences by outcome measure type by narrowing to studies that had both a researcher or developer measure as well as a narrow and/or broad measure, and by adding study-level fixed effects to the model.[5] The second model estimates effect size differences by outcome measure type without the study-level fixed effects, meaning that it examines differences in effect sizes due to both within- and across-study differences. All covariates were grand-mean centered to facilitate interpretation of the results.

Descriptive Findings

How often are researcher, developer, or narrow measures the only measures included in studies reviewed by the WWC?

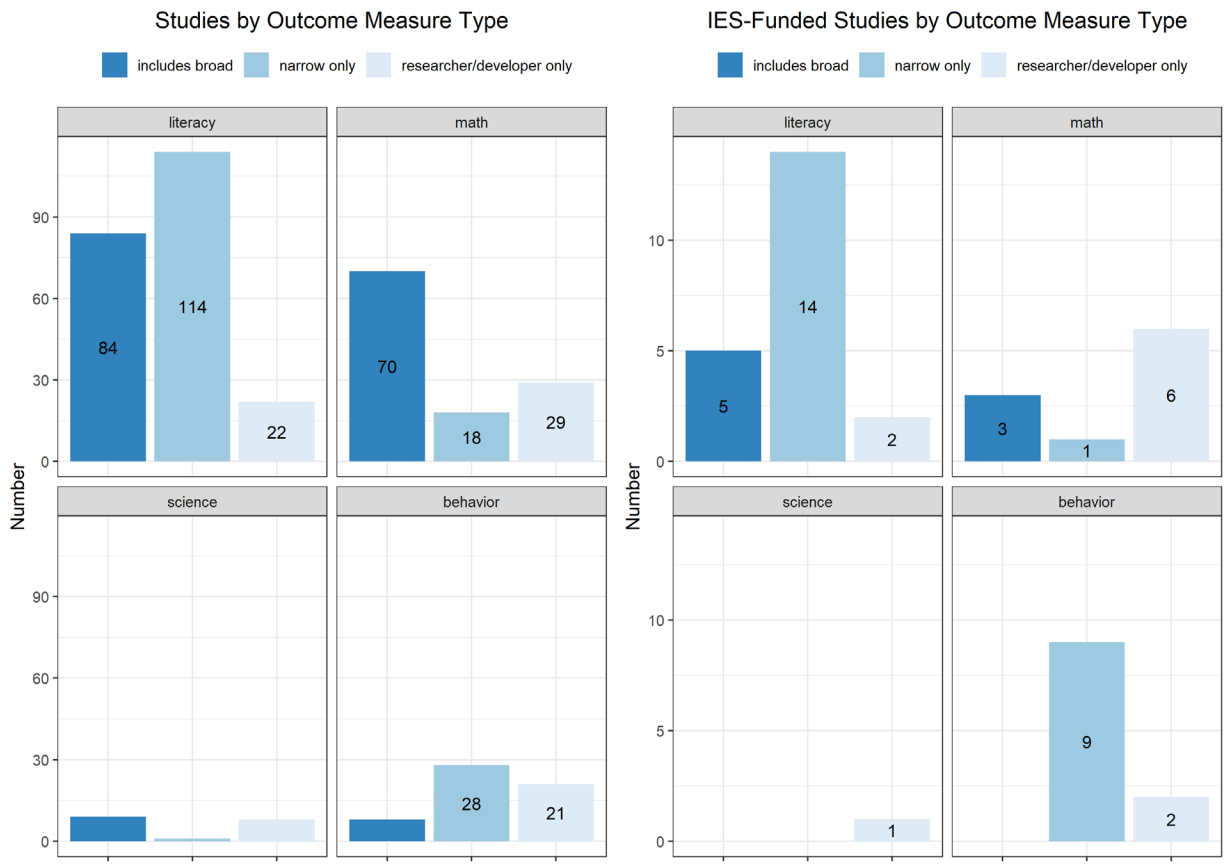
Across all studies reviewed by the WWC in the topic areas of literacy, STEM, and behavior, 38% of studies included at least one broad measure, 41% of studies included no broad but at least one narrow measure, and 21% of studies included only researcher or developer measures. The percentage of studies that included at least one broad measure varied across content areas, with 60% of studies in mathematics, 50% in science, 38% in literacy, and 14% in behavior including at least one broad measure, as shown in Figure 1.

[3] Appendix 1 contains information about how the WWC's outcome domains were revised for this paper.

[4] Effect sizes within studies were assumed to be dependent and correlated at ρ=.80, although the covariance structure was unknown. Results were not sensitive to changes in the assumed correlation.

[5] Some covariates were redundant with the study fixed effects.

Figure 1. Number of studies reviewed by the WWC by outcome measure type and content area



Note. The counts duplicate studies that relate to more than one content area.

When restricting to studies that were funded by IES[6], only 20% included a broad measure, 53% did not include a broad measure but included a narrow measure, and 28% included only researcher or developer measures. Therefore, the evidence suggests that IES-funded research may not fare better in terms of analyzing educational impacts on broad measures, and in fact, may fare worse than other studies. However, this finding should be interpreted with caution because not all studies funded by IES could be identified in the WWC study data. Moreover, IES-funded studies may be more likely to focus on new concepts or skills that may not be well captured by existing broad measures.

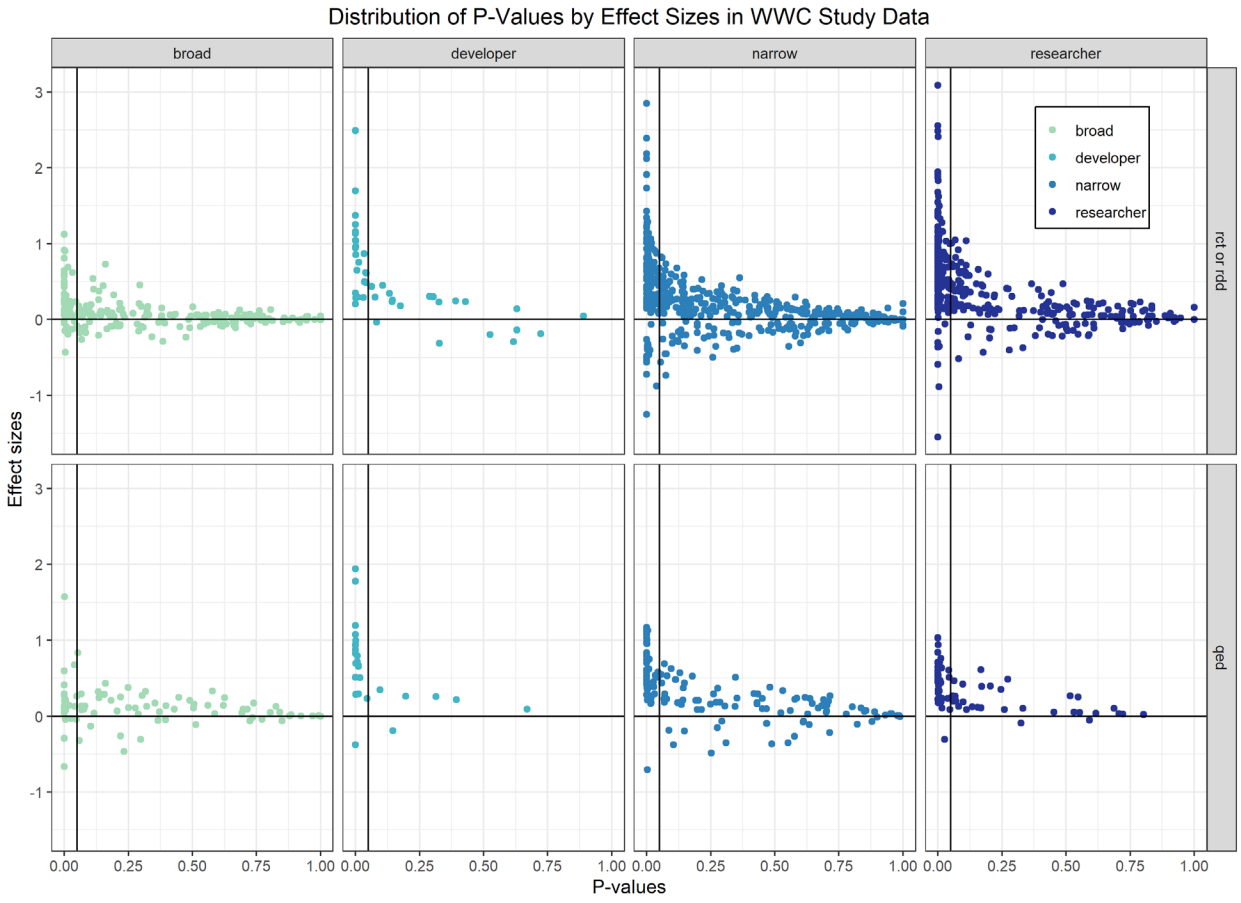
What percent of the positive and statistically significant findings in the WWC are based on researcher, developer, or narrow measures?

Fifty-one percent of the statistically significant (p<.05) and positive findings in the WWC across the content areas were based on researcher or developer measures, and the remaining statistically significant and positive findings were based on broad (15%) and narrow (35%) measures. Put another way, 62% of developer measures and 47% of researcher measures were associated with statistically significant and positive findings compared with 28% of narrow measures and 20% of broad measures.

[6] In this paper, IES-funded refers to studies that were funded by either NCER or NCSER that were both included in ERIC and flagged as being funded by IES in the ERIC API as of September 2020.

When descriptively examining the relationship between effect sizes and p-values, it appears that the greater shares of statistically significant and positive findings on researcher and developer measures are due to larger effect sizes on these measures relative to broad and narrow measures, as shown in Figure 2. It also appears that effect sizes are systematically larger in studies with QED designs compared with studies with RCT or RDD designs.

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



To explore whether one would come to the same conclusion about the effectiveness of a particular intervention using researcher or developer measures versus broad or narrow ones, results from different types of outcome measures were descriptively examined within the same study and outcome domain. Out of 50 studies that included at least one researcher or developer measure **and** at least one broad or narrow measure in the same outcome domain:

- 25 (50%) studies identified a positive and statistically significant finding on both a broad or narrow measure AND on a researcher or developer measure;
- 16 (32%) studies identified a positive and statistically significant finding on a researcher or developer measure but did not find a positive and statistically significant finding on a broad or narrow measure;

- 9 (18%) studies did not identify a positive and statistically significant finding on any type of measure; and
- 0 (0%) of studies identified a positive and statistically significant finding on a narrow or broad measure but did not find a positive and statistically significant finding on a researcher or developer measure.

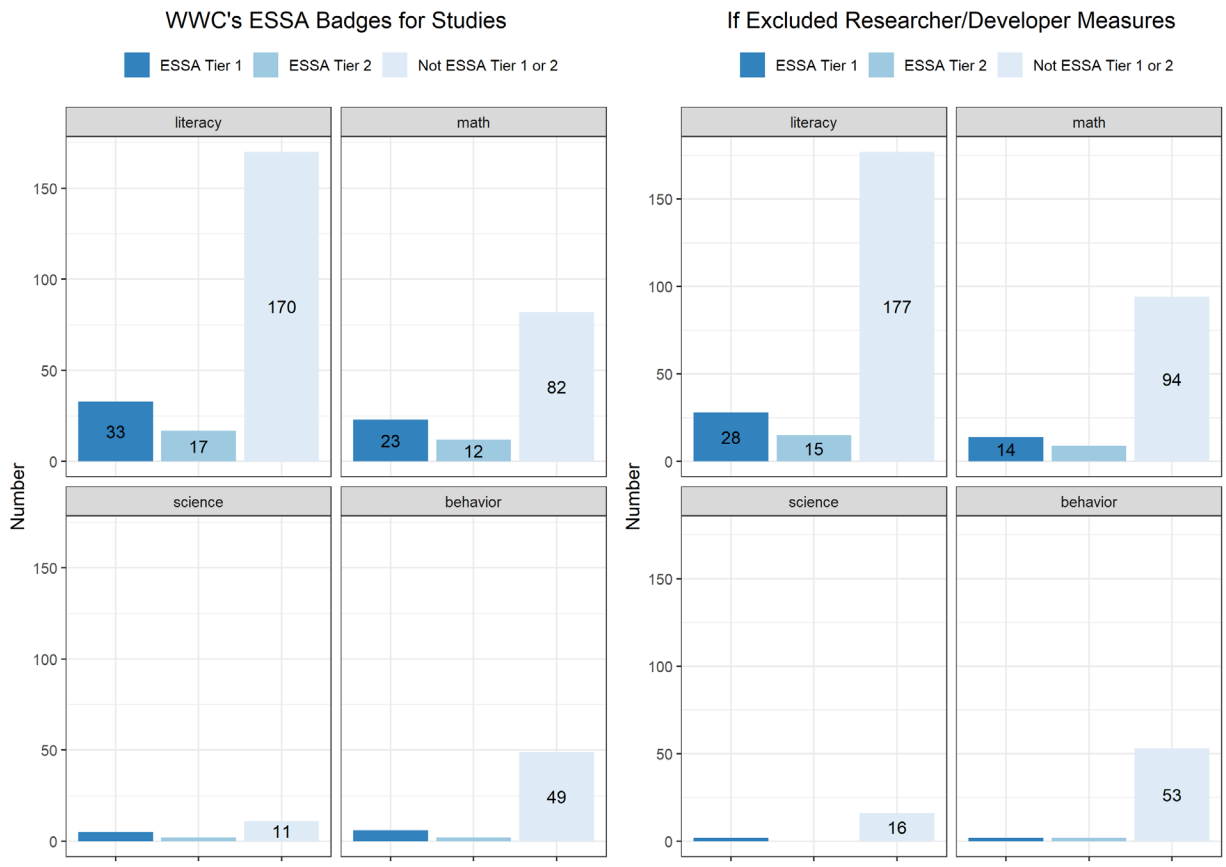
Therefore, about one-third of studies would have come to a different conclusion about the effectiveness of the intervention if using only outcome measures that were independent of the researchers and developers involved with the study. This descriptive analysis was limited to only a subset of studies that included both types of outcome measures (researcher or developer and broad or narrow). Although low statistical power may contribute to some of the studies that showed mixed effects depending on the outcome measure type, this finding illustrates that a substantial proportion of studies that identified positive findings on researcher or developer measures also identified null findings when using broad or narrow measures.

If the WWC's ESSA evidence tier badges were restricted to non-researcher and non-developer outcome measures only, how many studies reviewed by the WWC would lose an ESSA Tier 1 or 2 badge?

For studies that meet WWC standards, the WWC also assigns badges for ESSA evidence tier designations (WWC, 2020). The WWC's ESSA Tier 1 badge refers to strong evidence of intervention effectiveness from well-designed and well-implemented experimental studies, while the WWC's ESSA Tier 2 badge refers to moderate evidence of intervention effectiveness from well-designed and well-implemented quasi-experimental studies.

About 23% of studies reviewed by the WWC in these topic areas were labeled with an ESSA Tier 1 or 2 badge. Of studies not receiving an ESSA Tier 1 or 2 badge, 61% did not have a positive and statistically significant finding, and 39% of studies did not receive an ESSA badge for other reasons, such as not meeting the multisite or sample size requirements. If the WWC reserved ESSA Tier 1 and 2 badges for findings using non-researcher and non-developer measures only, the percentage of studies earning an ESSA Tier 1 or 2 badge would fall by eight percentage points to 15%, which is approximately a one-third reduction. The biggest differences would occur in mathematics (from 30% to 20% of studies) and in science (from 39% to 11% of studies), although there are few science studies, as shown in Figure 3. In addition, if ESSA Tier 1 and 2 badges were reserved only for findings using broad (as opposed to narrow) measures, only 9% of studies reviewed by the WWC would receive an ESSA Tier 1 or 2 badge in these content areas. The percentage of studies that would receive an ESSA Tier 1 or 2 badge using broad measures only ranges from 0% in behavior, to 6% in science, to 11% in literacy, and to 17% in mathematics.

Figure 3. Change in WWC's ESSA Tier 1 and 2 badges if exclude researcher and developer measures



Note. The counts duplicate studies that relate to more than one content area.

Meta-Analytic Findings

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

We first examine differences in effect sizes by outcome measure type by looking *within* studies (by including study fixed effects). Studies were included in this meta-analysis if they contained at least one researcher or developer measure **and** at least one broad or narrow measure. This analysis examines whether effect sizes vary by outcome measure type within the same study and outcome domain. This analysis arguably provides the strongest evidence of whether effect sizes vary by outcome measure type because any differences across studies—such as study quality or implementation fidelity—are held constant.

Within studies and outcome domains, effect sizes using researcher measures were larger by an average of +0.24 (in standardized units) relative to broad measures, and by an average of +0.15 relative to narrow measures. Effect sizes using developer measures were 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 showed average effect sizes that were about 1.75 to 2 times larger than effect sizes from broad measures, and about 1.4 to 1.6 times larger than effect sizes from narrow

measures within the same study and outcome domain. There was no statistically significant difference in the average effect sizes for researcher versus developer measures, nor was there a statistically significant difference in average effect sizes for broad versus narrow measures.[7] The latter finding implies that effect sizes may not systematically vary across narrow versus broad measures, once study quality, implementation fidelity, and other study characteristics are held constant. These findings are presented in Table 3.

Table 3. Meta-regression results looking within studies and outcome domains

		Estimate	Standard error	t	Df	p-value
Within-study model						
	Intercept	0.31	0.07	4.35	10.26	**
Outcome type	Researcher	0.24	0.07	3.51	20.57	**
	Developer	0.32	0.07	4.35	21.81	***
	Narrow	0.09	0.06	1.35	29.28	
	Broad	*Reference*				
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	<.10
	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. 1. ***p<.001, **p<.01, *p<.05. 2. Do not trust estimates when the degrees of freedom are less than four. 3. The model also included fixed effects for each study. All other covariates were redundant with the study fixed effects.

The next, full-sample model includes all of the studies reviewed by the WWC in the literacy, STEM, and behavior topic areas, and patterns in effect sizes by outcome measure type may be explained by differences both across and within studies. While the previous, within-study model inherently controls for factors such as study quality and implementation fidelity, this analysis does not. Therefore, differences in average effect sizes by outcome measure type may be conflated with factors that vary across studies.

Using the full-sample model, effect sizes using researcher measures were larger than broad measures by an average of +0.28, and larger than narrow measures by an average of +0.21, as shown in

[7] Post-hoc Wald tests were conducted using the *metafor* R package.

Table 4. Effect sizes using developer measures were larger than broad measures by an average of +0.31, and larger than narrow measures by an average of +0.24. Consistent with the results from the previous model, there was no statistically significant difference in the average effect sizes for researcher versus developer measures.

Unlike the results from the previous model, narrow measures showed statistically significant larger effect sizes than broad measures (by an average of +0.07). While the regression coefficients for narrow measures were similar across the two models, it could have been the case that the within-study model was under-powered to detect a statistically significant difference. Alternatively, use of narrow measures may be confounded with unrelated and unobserved factors that varied across studies.

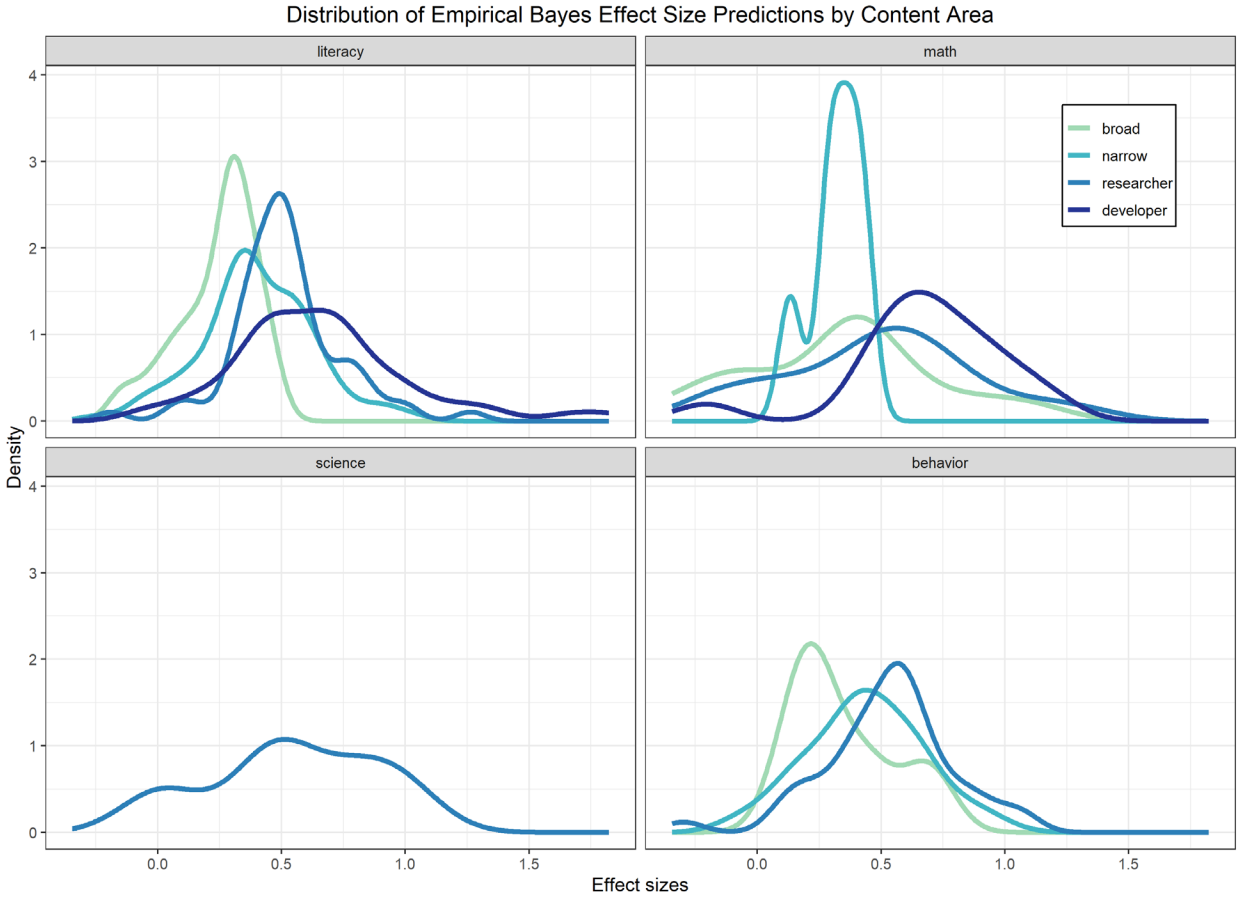
Table 4. Meta-regression results looking both within and across studies

		Estimate	Standard error	t	Df	p-value
Full-sample model						
	Intercept	0.23	0.01	16.94	200.77	***
Outcome type	Researcher	0.28	0.03	8.24	121.08	***
	Developer	0.31	0.06	5.47	29.65	***
	Narrow	0.07	0.03	2.68	146.38	**
	Broad	*Reference*				
Study design & rating	QED	0.11	0.04	3.01	77.97	**
	With reservations	-0.06	0.03	-1.82	61.12	<.10
Standards version	Version 2.1+	0.02	0.03	0.51	141.22	
Purpose of study review	Department-funded	0.05	0.05	0.90	25.72	
	Grant competition	0.06	0.03	1.67	111.17	<.10
	IES performance	-0.05	0.05	-0.91	23.21	
	Intervention report	*Reference*				
	Practice guide	-0.05	0.09	-0.54	14.74	
	Quick review	-0.03	0.05	-0.66	20.49	
	Single study review	0.10	0.05	2.00	76.72	*
Outcome domain	Alphabetics	0.06	0.04	1.75	78.44	<.10
	Comprehension	-0.03	0.03	-1.09	108.32	
	Reading fluency	-0.02	0.05	-0.38	60.90	
	Interpersonal behavior	-0.06	0.05	-1.19	51.91	
	Intrapersonal behavior	-0.12	0.04	-3.29	34.53	**
	Literacy	*Reference*				
	Math	0.05	0.03	1.63	77.28	
	Progress in school	-0.15	0.13	-1.13	2.32	
	Science	0.02	0.07	0.21	21.72	
	Writing	0.11	0.08	1.33	9.18	
Model info	Finding N	1,553				
	Study N	373				
	τ^2	0.02				
	ω^2	0.07				

Notes: 1. ***p<.001, **p<.01, *p<.05. 2. Do not trust estimates when the degrees of freedom are less than four. 3. The full-sample model also controlled for program types, delivery methods, and grade-level bands.

Results in Tables 3 and 4 further indicate that while there are few differences in average effect sizes by outcome domains, study design, and purpose of study review, 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 3, the distributions for researcher and developer measures show larger average effect sizes, and this appears to be true across the four content areas.

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



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

Sensitivity Analyses

Several sensitivity analyses were conducted. First, using the full study sample, we tested whether interaction effects were statistically significant to determine whether the overall effects of outcome measure types were larger in either a particular content area or in QED studies. Using the full study sample, there was a statistically significant and positive interaction term between researcher measures and the mathematics content area, which indicates that effect sizes based on researcher measures in mathematics were systematically larger. Additionally, the interaction term between QED studies and narrow measures was positive and approached statistical significance ($p<.10$), indicating that use of narrow measures may yield even larger effect sizes in QED studies, but more research is needed to confirm this finding.

Second, we re-estimated the full study model for each topic area separately to confirm that results were consistent across areas. The direction and magnitude of the regression coefficients for the researcher and developer measures were similar to those in Table 4 in the topic areas of literacy and STEM. However, researcher measures were not statistically significantly related to effect sizes for studies within the behavior topic area, though the direction of the regression coefficient for researcher measures was positive (+0.06).[8] This analysis indicates that effect sizes for researcher measures within the behavior topic area were more similar to those from independent measures.

Discussion

This paper uses advanced meta-analysis and demonstrates that effect sizes, which represent the degree to which educational interventions are effective in improving student outcomes, relate to outcome measure type. On average, researcher and developer measures showed effect sizes that were about 1.75 to 2 times larger than broad measures, and about 1.4 to 1.6 times larger than narrow measures within the same study and outcome domain. Looking both across and within studies, findings based on researcher and developer measures also showed larger average effect sizes when separately considering studies in the literacy and STEM topic areas. However, average differences in effect sizes by outcome measure type in the behavior topic area were smaller and not statistically significant.

Researcher and developer measures may be useful to validate the effectiveness of an intervention in a pilot study or efficacy trial. Yet practitioners and policymakers, who are held accountable for student progress on independent measures, may not find this evidence sufficient to inform their decisions. Perhaps there is a mismatch between the evidence needed by researchers or developers to validate an intervention versus evidence needed by practitioners and policymakers to select interventions to implement at scale in their settings.

Practitioners are most interested in the qualitative ratings of an intervention's effectiveness (SWAT, 2020). One open question is whether positive and statistically significant findings on researcher or developer measures translate into something meaningful for practitioners. Descriptive findings suggest that in 32% of studies, positive and statistically significant effects were identified on a researcher or developer measure, yet null effects were identified on all independent (broad or narrow) measures in the same study and outcome domain. There are several plausible explanations for this discrepancy. First, findings using independent measures may lack statistical significance because they were under-powered, given that effect sizes for independent measures are lower on average and sample size requirements increase as effect sizes decrease. Second, 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). Third, researcher and developer measures may measure a smaller subset of constructs than independent measures, making it easier to detect intervention effectiveness using researcher and developer measures (Song & Herman, 2010). Yet Song and Herman (2010) argue that using a narrow measure to claim effectiveness of an intervention on a broad construct is "unwarranted at best and misleading at worst" (p. 360).

The best-case scenario is that statistical significance 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 statistical significance on a researcher or developer

[8] No outcome measures were coded as developer measures in the behavior topic area.

measure has no bearing on how well students will perform on a formative or summative assessment in the same content area. The results of this paper suggest that there is not sufficient evidence to conclude that positive and statistically significant findings on researcher and developer measures will translate into positive and statistically significant findings on independent measures.

Aside from statistical significance, a related issue is that effect sizes on researcher or developer measures can be 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. This issue is a concern only to the extent that WWC users compare the relative effectiveness of different educational interventions based on the magnitude of their effect sizes. Yet 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 intelligent tutoring systems are not more effective than other forms of tutoring when outcome measure type is taken into account.

Results of this paper call into question whether using researcher and developer measures leads to inaccurate and misleading conclusions about the effectiveness of educational interventions. There are several existing statistical approaches that could be used to account for differences in outcome measure types. Statistical approaches, such as meta-regression or Bayesian modeling, could be used to adjust both the statistical significance and magnitude of effect sizes, accounting for larger average effect sizes when using researcher or developer measures. Alternatively, additional study qualifications could be made to differentiate findings that were assessed using a researcher or developer measure versus an independent one. Different characterizations of evidence could be applied for different WWC stakeholder groups to avoid confusion and align the evidence with stakeholders' needs. Given that outcome measure type is by far the most predictive variable explaining the magnitude of effect sizes in studies reviewed by the WWC in some topic areas, 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.

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. A second limitation is that this paper does not account for unobserved factors, such as the alignment between the intervention and outcome measure, which may also relate to effect sizes and confound these results. A final 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.

References

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

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.

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.

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*.

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.

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.

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/.

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/

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.* American Institutes for Research (AIR). Internal WWC report: unpublished.

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

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/

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

Williams, R., Citkowicz, M., Linsay, J., Miller, D., & Walters, K. (n.d., under review). *Heterogeneity in mathematics intervention effects: Results from a meta-analysis of 190 randomized experiments.* https://airshinyapps.shinyapps.io/math_meta_database/

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 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

Technical Appendix

This technical appendix provides more information about the extraction of the WWC study data and construction of the database for this paper. WWC study data were extracted in August 2020 and included all publicly available data in all topic areas. Because studies may have been reviewed by the WWC more than once, the data were deduplicated prioritizing the most recent study review.[9] Subgroup findings were also excluded during this process. The data analyzed in this paper are made publicly available here: https://github.com/betsyjwolf/Average-Effect-Sizes-by-Outcome-Measure-Type.

Studies were initially flagged as being in a particular topic area based on the review protocol. However, given that a large number of studies were reviewed under generic review protocols, additional studies were flagged as being relevant to the topic area based on outcome domains, outcome measures, and intervention names. If any of these related to a topic area, the study was flagged for inclusion in the topic area. As a result, some studies were flagged for more than one topic area.

Outcome measures were categorized by reviewing original studies as well as information provided by the WWC on the website, in the WWC study database, or in a WWC publication, such as an intervention report. In many cases, the outcome measure name in the WWC study data was ambiguous, and it was necessary to review the original study to determine what the outcome measure was. The data were also changed in the following ways:

- A small number of implausible effect sizes (greater than +200) were recoded to missing.
- Some effect sizes in behavioral domains that reflected reductions in negative behaviors were negative numbers when they should have been positive numbers, and the absolute value of the effect size was applied when the improvement indexes were positive.
- In some cases, the WWC-calculated effect size was missing, but the finding received an ESSA badge. This discrepancy meant that a slightly different set of studies would be used across the different descriptive analyses. To ensure that the study samples were consistent across all analyses, missing WWC-calculated effect sizes were replaced with non-missing study-reported effect sizes when the WWC-calculated p-value was also non-missing; this resulted in the replacement of 226 missing WWC-calculated effect sizes out of 1,673 total findings. In addition, another 120 findings were excluded from the analyses because the findings were missing both WWC-calculated effect sizes and WWC-calculated p-values. Additional sensitivity analyses was conducted by re-estimating both meta-analytic models with only the WWC-calculated effect sizes to determine whether these replacements changed the results, and the results were consistent with those presented in the paper.
- Findings in which the outcome measure type was unknown were excluded. There were few cases in which the outcome measure could not be determined because the original study was not available or because the information that the reviewer recorded could not be reconciled with the information in the original study.

In addition, outcome domains were recoded to standardize the outcome domains across different review protocols and versions of review protocols, as each new review protocol could redefine and rename outcome domains. In general, sub-domains that were not used consistently over time were combined in more encompassing, related domains. This crosswalk is provided in Table 5.

[9] The review with the most recent review protocol or standards version was retained, unless the most recent review was conducted for a practice guide, quick review, or grant competition, in which case a previous review was retained, if the previous review was conducted for a different purpose.

Table 5. Crosswalk of original and revised outcome domains

Original Outcome Domains	Revised Outcome Domain
• Alphabetics • Print knowledge	Alphabetics
• Academic achievement • Communication/language • English language arts achievement • English language development • English language proficiency • Language arts • Language development • Literacy achievement • Oral language • Reading achievement	General literacy
• Academic achievement • Algebra • Conceptual knowledge • General mathematics achievement • Geometry • Geometry and measurement • Number and operations • Procedural flexibility • Procedural knowledge	General math
• Academic achievement	General other subject
• Academic achievement • Science achievement	General science
• Behavior • External behavior • Problem behavior • Social outcomes • Student behavior • Student social interaction	Interpersonal behavior
• Academic self-efficacy • Emotional/internal behavior • Executive functioning • Knowledge, attitudes, & values • Organization • Self-concept • Self-determination • Social outcomes • Social-emotional competence • Social-emotional development • Student emotional status	Intrapersonal behavior
• Academic achievement	Progress in school
• Comprehension	Reading comprehension

• Reading comprehension • Vocabulary development	
• Reading fluency	Reading fluency
• Audience • Genre elements • Overall writing quality • Writing achievement • Writing output • Writing processes • Writing quality	Writing