Do Developer-Commissioned Evaluations Inflate Effect Sizes Among Studies Accepted By the What Works Clearinghouse?

Rebecca Wolf
betsywolf@jhu.edu

Jennifer Morrison
JRMorrison@jhu.edu

Robert Slavin
rslavin@jhu.edu

Kelsey Risman
klrisman@jhu.edu

Johns Hopkins University
School of Education
Center for Research and Reform in Education
300 E Joppa Road
Baltimore, MD 21286

Background. Educational decision-making should be based on rigorous evidence. While researchers have advocated for the use of rigorous evidence in decision-making for many years, policymakers have recently mandated the use of evidence in selecting educational programs. The Every Student Succeeds Act (ESSA) of 2015 requires that districts seeking certain types of educational funding from the federal government select programs supported by evidence, and is encouraging use of evidence more broadly.

One challenge practitioners face is identifying educational programs that are supported by evidence. To this end, the federal government funds the What Works Clearinghouse (WWC), a database of educational program evaluations. Given ESSA, it is important to understand the WWC database itself as well as any potential sources of bias in the studies reviewed by the WWC.

Purpose. One issue that has not been previously explored is whether studies carried out or commissioned by developers produce larger effect sizes than studies carried out by independent third parties. The purpose of this article is to determine whether there is a developer effect. If there is a systematic difference in effect sizes for studies commissioned by developers and independent parties, we will attempt to determine why: Are there specific features of developer-commissioned evaluations that account for systematic differences in effect sizes?

Data. This paper uses study data from the WWC database, and other information from individual studies, to understand how developer-commissioned research affects study effect sizes. We used WWC data as of January 2018 in the areas of K–12 mathematics and reading/literacy. Only studies that met WWC standards were retained in the sample, as the necessary study data were populated only for such studies. The data were further restricted to whole-sample analyses, excluding subgroup analyses. The final database of studies consisted of 755 findings in 169 studies.

Table 1: *Study Sample Descriptives*

	All (%)	Developer-Commissioned (%)	Independent (%)	Chi-Square p
Study Design				
Experimental	71	49	85	***
Quasi-experimental	29	51	15	
Outcome Measure Type				
State, district, or other independent	83	71	92	***
Researcher-made	17	29	8	
WWC Study Rating				
Meets standards *without* reservations	63	47	74	***
Meets standards *with* reservations	37	53	26	
Small Sample Size (≤250 students)	71	65	75	**

External Evaluator	70	41	89	***
Study Author				
Developer	23	59	0	***
Research organization	19	12	24	
School district	3	0	5	
University	39	30	46	
Graduate student	16	0	26	
Study Funder				***
Developer	23	58	0	
Federal government	49	35	59	
Foundation	5	4	6	
No funding	20	0	33	
State	2	3	1	
Unknown source of funding	1	0	1	
WWC Study Rating				***
Meets standards *without* reservations	63	47	74	
Meets standards *with* reservations	37	53	26	

Note. **p<.001, ***p<.001

Practice. For the purposes of this study, a developer was defined as the organization responsible for developing the proprietary intervention that was being studied. Each study was coded as being commissioned by a developer either if an employee of the developer was one of the authors of the study or if the developer had funded the study. Each study was individually reviewed to identify author type (e.g., developer, district, graduate student, research firm, university) and funder type (e.g., developer, federal government, foundation, no funding, state, unknown source). In total, there were 300 developer-commissioned findings in 73 studies and 455 study findings in 96 studies by independent parties.

Research Design. We used a meta-regression model to estimate the mean standardized effect sizes of developer-commissioned versus independent studies. A meta-regression model with random effects is similar to a mixed effects model where each study is first weighted by its inverse variance (Borenstein et al., 2009; Tipton, 2015). We used the R package, robumeta, to conduct the meta-analysis (Fisher, Tipton, & Zhipeng, 2017).

We estimated several models. First, we estimated the developer effect by including a dummy variable indicating whether a study was commissioned by a developer. Then, we cumulatively added covariates in the model to attempt to explain the developer effect by controlling for factors known from previous research to affect effect sizes. We controlled for whether the studies had experimental or non-experimental designs, small student sample sizes of less than 250 students, and researcher-made outcome measures (e.g., measures created for the purpose of the study) (Cheung & Slavin, 2016); and whether the study was conducted by an external evaluator.

Findings. Studies commissioned by developers produced larger average effect sizes than studies by independent parties. Developer-commissioned studies had an average effect size of +0.27, compared with +0.13 for independent studies. Hence, developer-commissioned studies produced average effect sizes that were twice as large as those of independent parties.

Table 2: *Meta-Analysis Regression Results*

	Developer effect model (No controls)	***+ Control for quasi-experiment***	***+ Control for researcher-made measure***	***+ Control for small sample size***	***+ Control for external evaluator***
Developer effect	0.143*** (0.036)	0.156*** (0.041)	0.116** (0.036)	0.105** (0.034)	0.089* (0.034)
Quasi-experiment		-0.037 (0.039)	-0.021 (0.036)	-0.010 (0.035)	-0.019 (0.034)
Researcher-made measure			0.332*** (0.064)	0.297*** (0.062)	0.291*** (0.060)
Small sample size (≤250 students)				0.129*** (0.034)	0.122*** (0.035)
External evaluator					-0.060 (0.033)
Intercept	0.131*** (0.021)	0.140*** (0.024)	0.102*** (0.021)	0.066** (0.021)	0.120** (0.036)
I^2	96.107	96.063	95.640	95.576	95.595
τ^2	0.0177	0.0178	0.0161	0.0158	0.0161

Note. *p<.05, **p<.01, ***p<.001

We attempted to determine to what extent this developer effect was explained by study design features. One potential explanation is that developer-commissioned studies were more likely to use quasi-experimental as opposed to experimental designs, which could result in inflated effect sizes (Cheung & Slavin, 2016). However, controlling for quasi-experimental design did not explain the developer effect.

Another potential explanation for the developer effect is that developer-commissioned studies were more likely to make use of researcher-made measures, thus producing inflated effect sizes (Cheung & Slavin, 2016). Controlling for outcome measure type explained 18% of the initial developer effect, but the developer effect persisted (ES=+0.22 for developer compared with +0.10 for independent studies).

Developer studies may have smaller sample sizes, and small-scale studies typically have larger effect sizes than large-scale studies (Cheung & Slavin, 2016). Controlling for small

sample sizes explained 8% of the initial developer effect, and yet the developer effect persisted (ES=+0.17 for developer compared with +0.07 for independent studies).

Finally, perhaps the developer effect was explained by whether an external evaluator conducted the evaluation versus an internal research team; an external evaluator may be less likely to manipulate researcher degrees of freedom to optimize study findings (Simmons et al., 2011). Controlling for external evaluator explained another 11% of the initial developer effect. Yet the developer effect still persisted despite all of these controls (ES=+0.21 for developer compared with +0.12 for independent studies).

Conclusion. Effect sizes for developer-commissioned studies were inflated, relative to effect sizes for studies conducted by independent researchers. The developer effect was partly explained by study design and features. Controlling for these, however, explained only 38% of the developer effect, and the average effect size for developer studies was still more than 1.5 times larger than the average effect size of independent studies.

Our inability to fully account for the developer effect in program evaluations by observable characteristics alone leaves open the possibility that the explanation lies elsewhere. Potential factors likely contributing to the developer effect are the file drawer problem and researcher degrees of freedom (Gelbach & Robinson, 2018; Sterling et al., 1995). A potential solution to both the file drawer effect and researcher degrees of freedom would be to require studies to be pre-registered in order to be listed in the What Works Clearinghouse.

References

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *An Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd.

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

Fisher, Z., Tipton, E., Zhipeng, H., & Fisher, M. (2017). Package 'robumeta'. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

Gehlbach, H., & Robinson, C. (2018). Mitigating illusory results through pre-registration in education. *Journal of Research on Educational Effectiveness*, *11*(2), 296-315.

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

Sterling, T., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108-112.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375.