Do Developer-Commissioned Evaluations Inflate Effect Sizes?

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 that meets ESSA standards. Some have suggested that evidence that meets ESSA standards could be determined according to whether the evidence meets the rigorous standards of the What Works Clearinghouse (WWC) (Lester, 2018).

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? On the other hand, perhaps interventions evaluated in developer-commissioned evaluations are simply more effective than interventions studied by independent parties, so any differences favoring developer-commissioned studies may be due to greater effectiveness, not to bias.

Data. We used data from the WWC database 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. The mean number of findings per study was 4.5.

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). For the purposes of this article, studies that were not commissioned by developers were labeled as "independent studies." In total, there were 300 findings in our database from 73 developer-commissioned studies, and 455 findings from 96 independent studies.

Research Design. We used a meta-regression model with robust variance estimation to conduct the meta-analysis (Hedges, Tipton, & Johnson, 2010). This approach has several advantages. First, our data included multiple effect sizes per study, and robust variance estimation accounts for this dependence without requiring knowledge of the covariance structure (Hedges et al., 2010). Second, this approach allows for moderators to be added to the meta-regression model and produces the statistical significance of each moderator in explaining

variation in the effect sizes (Hedges et al., 2010). We used the R package *robumeta* to conduct the analysis (Fisher, Tipton, & Zhipeng, 2017).

We estimated several meta-regression models. First, we estimated a model with an intercept and the covariates (e.g., grade level and publication year) to estimate the overall mean effect size. We did not include subject (e.g., mathematics or reading/literacy) because differences between these subjects in effect sizes were not statistically significant. Second, we added a developer dummy indicator to the model. Third, we re-estimated the previous model while also controlling for study design features (e.g., quasi-experiment or experiment, researcher- or developer-made or independent measure, and natural logarithm of student sample size), program type (e.g., curriculum, practice/professional development, whole school, or supplemental), delivery method (e.g., individual student, small group, whole class, or whole school), and whether the intervention included educational technology.

While the previous model accounts for differences in study design features and program characteristics for developer and independent studies, it is hypothetically possible that interventions in developer studies are simply more effective than those in independent studies. To further explore this possibility, we narrowed the sample to interventions for which there were both developer and independent studies and estimated a fourth meta-regression model that included fixed effects for each intervention, as well as additional covariates that were not redundant.

Findings. Studies commissioned by developers produced larger average effect sizes than studies by independent parties. Developer-commissioned studies had an average effect size of +0.307 compared with +0.173 for independent studies. In other words, developer-commissioned studies produced average effect sizes that were more than one and a half times those of independent studies.

	Model 1: Baseline + grade level, subject, study year	**Model 2: Model 1 + developer effect**	**Model 3: Model 2 + factors known to influence effect sizes**	**Model 4: Model 3 + intervention fixed effects**
Intercept	0.239*** (0.021)	0.173*** (0.023)	0.212*** (0.025)	0.194*** (0.039)
"Developer" effect		0.134*** (0.036)	0.117** (0.034)	0.156* (0.060)

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

We attempted to determine to what extent this developer effect could be explained by study design features or program characteristics. Developer-commissioned studies were more likely to use quasi-experimental as opposed to experimental designs, researcher- or developer-made measures as opposed to independent ones, and smaller sample sizes, all of which could result in inflated effect sizes (Cheung & Slavin, 2016). Controlling for study design features and

program characteristics, developer-commissioned studies had an average effect size of +0.329 compared with +0.212 for independent studies.

Yet what if interventions in developer-commissioned studies are more effective than those in independent studies? If this were true, then we would expect developer-commissioned and independent studies *of the same program* to produce similar effect sizes. Yet controlling for intervention fixed effects, study design features, and other covariates, developer-commissioned studies had an average effect size of +0.350 compared with +0.194 for independent studies. Therefore, while it did appear that some interventions were more effective than others, when looking within the same intervention, developer-commissioned studies had effect sizes that were systematically greater than those of independent studies, on average, and the developer effect could not be explained by study design features alone.

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 features and program characteristics. Study design features and program characteristics explained only a small portion of the "developer effect," and the developer effect was largely unexplained by observed study and program characteristics available in the WWC data.

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, Rosenbaum, & Weinkam, 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

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.

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.

Lester, P. (2018). *Evidence-based comprehensive school improvement.* Retrieved from http://socialinnovationcenter.org/wp-content/uploads/2018/03/CSI-turnarounds.pdf.

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.