Title: Effects of Researcher-Made vs. Independent Measures on Outcomes of Experiments in Education

Authors and Affiliations: Marta Pellegrini, University of Florence, Italy
(marta.pellegrini@unifi.it)

Amanda Inns, Johns Hopkins University
(ainns1@jhu.edu)

Cynthia Lake, Johns Hopkins University
(clake5@jhu.edu)

Robert E. Slavin, Johns Hopkins University
(rslavin@jhu.edu)

Background

In 2015, the U.S. Congress passed the Every Student Succeeds Act (ESSA), which for the first time defines in law what it means for educational programs to have evidence of effectiveness and requires that very low-achieving schools seeking school improvement funding adopt programs meeting evidence standards. The new law is a demonstration of the growing confidence of educational policy makers in the use of evidence in decision-making and it creates a new urgency for educators to know which programs are proven and which are not. This policy change makes it particularly important that researchers discover methodological factors that may affect outcomes of experimental evaluations.

Cheung & Slavin (2016), Slavin & Madden (2011), de Boer, Donker, & van der Werf (2014), Wolf & Slavin (2018) and other researchers have reported evidence on the effects of design issues on conclusions of meta-analyses. One important design element evaluated in these studies is type of outcome measures. Slavin & Madden (2011) found that measures inherent to experimental treatments compared to independent measures made a substantial difference in effect sizes in What Works Clearinghouse (WWC) reviews, favoring inherent measures. Cheung & Slavin (2016) found that among studies included in the Best Evidence Encyclopedia, measures made by researchers or developers had mean effect sizes twice those of independent measures. The WWC (2017) excludes "over-aligned" measures, so the question is whether there is still an impact among the broader category of measures made by researchers or developers.

Purpose

The purpose of this study is to examine all studies on reading and math (K-12) included in the What Works Clearinghouse database to determine the degree to which studies that used researcher/developer-made measures have different effect sizes from studies that used independent measures.

Method

Data Collection

All accepted studies in WWC reviews in the areas of elementary and secondary reading/literacy and mathematics (K-12) were included in the analysis. The data were obtained from the WWC Individual Study Database (ISD) file, a database that contains information on all single studies included in WWC reviews. The sample consisted of 671 outcomes from 171 studies.

Outcome measures used in each study were coded as follows:

- Independent measures that are primarily state and district tests or other tests that do not specifically measure content or skills only taught in the treatment group, such as the Scholastic Assessment Test (SAT), or Woodcock Johnson Tests;
- Researcher-made measures, created by the researcher or developer for the particular program or study. This category included also measures that are inherent to the treatment, which assess skills or knowledge taught in the experimental group but not in the control group.

A total of 122 outcomes were coded as researcher-made measures and 549 outcomes as independent measures.

Data analysis

We used the meta-analysis approach of subgroup analysis to estimate the weighted mean effect sizes of researcher-made measures and independent measures. Mean effect sizes across studies were calculated after assigning each study a weight based on inverse variance (Lipsey & Wilson, 2001), adjusted as suggested by Hedges (2007), which inflates the variances from school- and class-assigned studies. In combining across studies, we used a random-effects model that takes into account two sources of variance (within-study and between-study variance) (Borenstein et al., 2009).

Findings

The mean effect size of researcher-made measures was more than two times the effect size of independent measures. As indicated in Table 1, the effect sizes for researcher-made measures and independent measures were +0.48 and +0.20, respectively ($Q = 54.79$, $p < .0001$).

Similar outcomes were found when we examined only studies that used both types of measures (Table 2). The difference in effect sizes between researcher-made measures (ES = +0.49) and independent measures (ES = +0.30) was smaller than the one found for all studies, but it is still significant ($Q = 8.89$, $p < 0.01$).

There are several likely explanations for the inflation of effect sizes by the use of researcher- or developer-made measures. One is that researcher/developer-made measures are over-aligned with the content used in the experimental but not the control group. Often, researchers make a strong statement in favor of a particular curricular approach, and then emphasize that content in their measure. For example, a math researcher might advocate for more use of non-routine problem solving, which is the emphasis of his or her program. However, the control group might receive little or no instruction or practice in non-routine problem solving. If the researcher-made test strongly emphasizes non-routine problem solving, it is not fair to the control students. The experimental students may do better on non-routine problem solving, but perhaps worse on many other topics in mathematics that were, however, minimally measured in the researcher's test.

Another reason independent measures are essential is that if researcher-made outcomes are used, participants in the experimental group may align their teaching with the content on the tests, while control teachers may not. A common example of this is in vocabulary interventions, where teachers are fully aware of the vocabulary emphasized in the program and on the test. Measuring knowledge of those specific words of course gives the experimental group a considerable advantage.

Conclusion

This study replicates earlier comparisons, providing clear evidence that measures made by researchers or developers inflate study effect sizes. These findings suggest that meta-analyses should discount or disregard such measures. The findings also suggest that it is not enough to exclude measures *aligned* with treatments. For example, the WWC Standards 3.0 and 4.0 exclude "over-aligned" measures, yet the present analysis was done with current WWC-accepted studies. Measures made by developers and researchers still inflate effect sizes, even after over-alignment has been accounted for.

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

Donker, A. S., de Boer, H., Kostons, D., Dignath-van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic perfor- mance: A meta-analysis. *Educational Research Review*, *11*, 1–26. doi:10.1016/ j.edurev.2013.11.002

de Boer, H., Donker, A. S., & van der Werf, M. P. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509-545. doi:10.3102/0034654314540006

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. doi:10.3102/1076998606298043

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oakes, CA: Sage

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370-380.

What Works Clearinghouse. (2017). *What Works Clearinghouse standards and procedures 4.0*. Institute of Education Sciences, U. S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/

Wolf, R., & Slavin, R. E. (2018). *Do developer-involved evaluations inflate effect sizes among studies accepted by the What Works Clearinghouse*? Manuscript submitted for publication.

Tables

Table 1. Analysis of outcome measures in WWC reviews (all studies, n = 171).

Types of measures	**n outcomes**	**Average ES**	***Q*-Value**	***df*(Q)**	***p*-Value**
Researcher-made	122	+0.48			
Independent	549	+0.20	54.79	1.00	<.0001

Table 2. Analysis of outcome measures in WWC reviews (studies with both types of measures, n = 35).

Types of measures	**n outcomes**	**Average ES**	***Q*-Value**	***df*(Q)**	***p*-Value**
Researcher-made	103	+0.49			
Independent	104	+0.30	8.89	1.00	<.01