Kaitlyn G. Fitzgerald
PhD Candidate
Department of Statistics
Northwestern University
kgfitzgerald@u.northwestern.edu

Adjusting standard deviation estimates to account for homogeneity of samples

Research Methods, Work-in-progress poster

Background

The IES RFA for Education Research Grants encourages study designs to use ideal conditions that include "a more homogeneous sample of students, teachers, schools, and/or districts" (U.S. Department of Education, 2018, p. iv). This is considered good research design for precision of estimation and power considerations. Therefore, education RCTs often have study samples that are more homogeneous than policy-relevant inference populations. Much work has been done on generalizability methods for how to adjust sample average treatment effect estimates to better estimate population average treatment effects (Olsen, Orr, Dell, & Stuart, 2013; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013; Tipton & Olsen, 2018). However, most of these methods focus on adjustments for the mean, and little work has been done to explore whether a similar adjustment is needed for the standard deviation.

There is evidence that samples in NCER-funded studies only include a fraction of the variation in covariates as compared to the total variation that exists in the population of US schools (Tipton & Hedges, 2017). If education RCTs do not observe the total variation in covariates, they also likely do not observe the total variation in outcomes. This matters because variation in outcomes is used directly in calculating common effect sizes, such as the standardized mean difference $\left(\frac{\bar{Y}_T - \bar{Y}_C}{s_Y}\right)$, which are in turn used in meta-analyses. When individual studies have more homogeneous samples, it is likely that the observed s_Y will be smaller than σ_Y^* in the desired inference population. This can lead to systematically smaller standard deviations on the denominator and therefore systematically inflated effect sizes. Better understanding is needed regarding how much this matters in practice and what type of adjustments would be appropriate.

Purpose

The goal of this study is to develop a method for adjusting standard deviations and to understand the analytic properties of an adjusted effect size estimate. We will illustrate empirically how much standard deviation estimates on the same outcome vary in practice and the implications this has on effect size estimates.

Methods

Creating an adjusted meta-analytic estimate

Assume that there are *m* studies, indexed by $j = 1, 2, \ldots m$, that use the same outcome measure. The usual estimate of the standardized mean difference is given by

$$\widehat{\delta_j} = \frac{\bar{Y}_{T_j} - \bar{Y}_{C_j}}{s_{Yj}},$$

where $\bar{Y}_{T_j}$ and $\bar{Y}_{C_j}$ are the treatment and control group means, respectively, and s_{Yj} is the pooled standard deviation of the outcomes in study j. The usual meta-analytic estimate of the true treatment effect, δ , is given by $\hat{\delta} = \frac{\sum w_j \widehat{\delta_j}}{\sum w_j}$, where the $w_j s$ are the usual inverse variance weights for a fixed or random effects meta-analytic model.

In order to adjust for the bias in s_{Yj} due to homogeneity of samples, we propose meta-

analyzing the numerator and denominator of the standardized mean difference separately. The estimand of interest here is

$$\frac{\mu_T - \mu_C}{\sigma_{Y^*}},$$

where $\mu_T - \mu_C$ is the population (unstandardized) mean difference, and σ_Y^* is the total variation in Y_{ij} across all individuals (*i)* and studies (*j*) in the population. Let

$$\widehat{\theta_j} = \bar{Y}_{T_j} - \bar{Y}_{C_j}$$

be the estimate of the (unstandardized) mean difference for outcomes Y_{ij} in study *j*. For the numerator then, the *m* estimates $\widehat{\theta_j}$ can be meta-analyzed in the usual way, with $\hat{\theta} = \frac{\sum w_j \widehat{\theta_j}}{\sum w_j}$ and $E(\hat{\theta}) = \theta = \mu_T - \mu_C$.

The denominator, meant to capture the total variation in outcomes, can be partitioned into within and between study variance using the law of total variance, such that

$$\sigma_Y^* = \sqrt{\sigma^2 + \tau^2},$$

where σ^2 is the average within-study variation (pooled across the study-specific variances ${\sigma_j}^2$), and τ^2 is the between-study variance. An estimate of σ_Y^* will be used, similar to that used in Tipton & Shuster (2017), in which a random effects estimate is used to pool the study-specific standard deviations on the log-scale. The proposed adjusted meta-analytic effect size then is $\widehat{\delta^*} = \frac{\hat{\theta}}{\hat{\sigma}_Y^*}$, and its sampling distribution will be derived. Properties of $\widehat{\delta^*}$and $\hat{\delta}$ will be compared via simulations.

Example data collection and comparisons

A merged dataset with information on 964 studies that meet What Works Clearinghouse (WWC) standards was extracted directly from the WWC website. Case studies were chosen for outcome measures that were used in multiple studies and for which data on sample sizes, means, and standard deviations were available for both the treatment and control groups. This enables us to have multiple estimates of the standard deviation of an outcome (s_Y), and thus to explore how much these estimates vary in practice and implications that adjustments will have on effect size estimates.

Additionally, in some cases, there are known national norms for the standard deviation of the outcome measure, such as with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). In this case, adjusted effect sizes are created, standardizing by the population standard deviation rather than the sample standard deviations. Both sets of effect sizes (those using the original standard deviation and those using the population standard deviation) are meta-analyzed, and the difference in the two meta-analytic estimates is compared.

(Preliminary) Results

For the purposes of this proposal, we focus only on the example data, and within this, on a single outcome measure for which there is a known national norm: DIBELS. There are 12 studies in WWC that measure DIBELS on Kindergarteners. In Figure 1, we provide comparisons between effect sizes computed using the original study standard deviations with those standardizing by the national norm standard deviation. In each of the 12 studies, standardizing by the national SD

reduces the effect sizes, and the percent reductions range from 3% (Study 2) to 38% (Study 11). At SREE, results related to simulations and analytic work will be provided.

Conclusions

Preliminary results suggest that there is substantial variation in standard deviation estimates of the same outcomes across studies, which has implications for how we calculate and interpret effect sizes and for the meta-analysis of such estimates. In order to more fully understand this problem and address it moving forward, improved reporting of standard deviations in both treatment and control groups is necessary in individual studies. Work on the analytic properties of the proposed adjusted effect size as well as other methods of adjustment are ongoing.

References

Olsen, R. B., Orr, L. L., Dell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32*(1), 107–121.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 174*(2), 369–386. Retrieved from JSTOR.

Tipton, E. (2013). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266. https://doi.org/10.3102/1076998612441947

Tipton, E., & Hedges, L. V. (2017). The Role of the Sample in Estimating and Explaining Treatment Effect Heterogeneity. *Journal of Research on Educational Effectiveness*, *10*(4), 903–906. https://doi.org/10.1080/19345747.2017.1364563

Tipton, E., & Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

Tipton, E., & Shuster, J. (2017). A Framework for the Meta-Analysis of Bland-Altman Studies Based on a Limits of Agreement Approach. *Statistics in Medicine*, *36*(23), 3621–3635. https://doi.org/10.1002/sim.7352

10/1/2019 10:15:00 PMU.S. Department of Education. Institute of Education Sciences. (2018). *Request for applications:*

Education Research Grants, CFDA Number: 84.305A. Retrieved from https://ies.ed.gov/funding/pdf/2019_84305A.pdf

Appendix

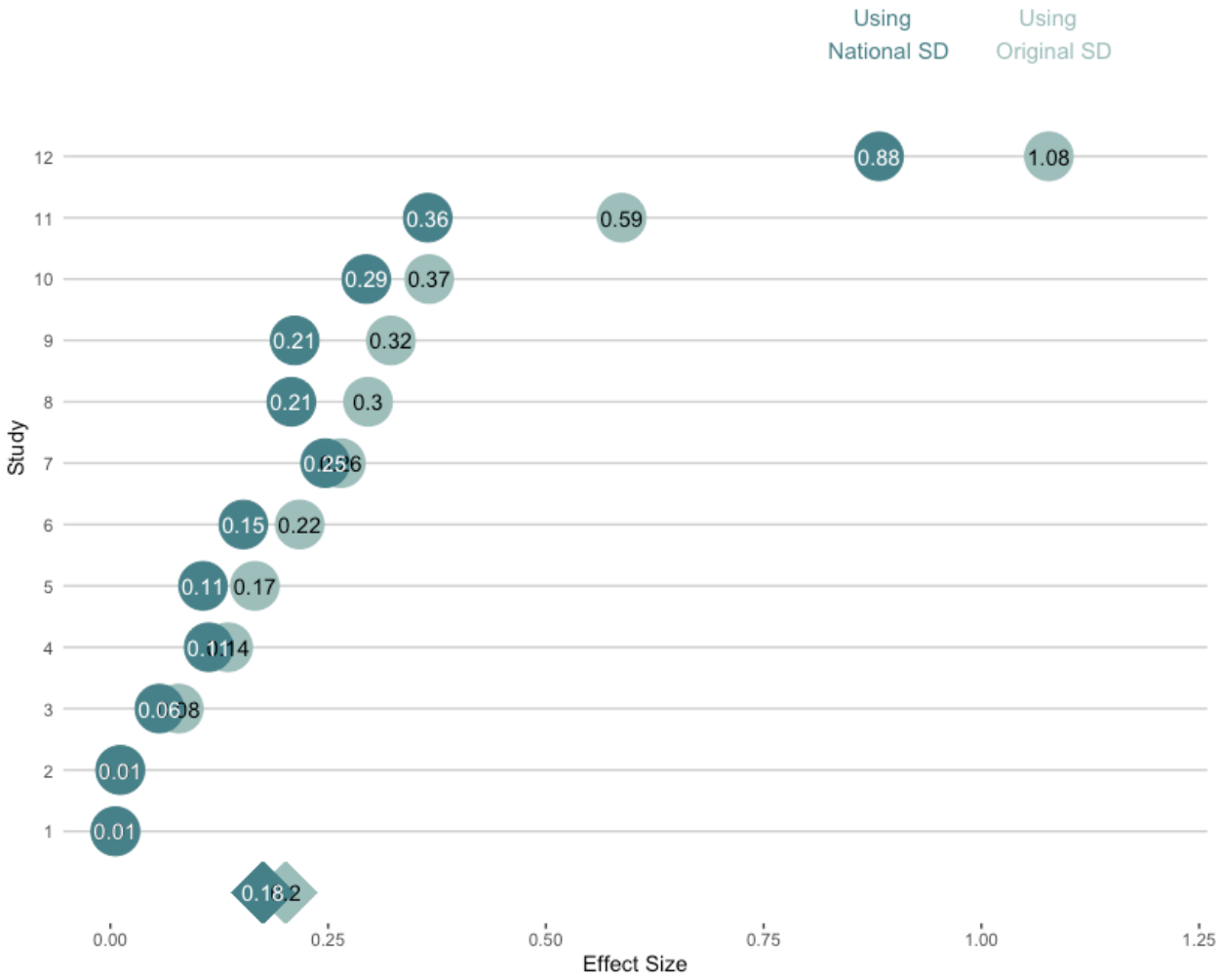

Figure 1: *Effect sizes for 12 WWC studies using DIBELS, using original study standard deviations and the national norm standard deviation*