Abstract Title Page
Not included in page count.

Title: A General Framework for Effect Sizes in Cluster Randomized Experiments

Authors and Affiliations: Nathan VanHoudnos, Northwestern University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Cluster randomized experiments are ubiquitous in modern education research. Although a variety of modeling approaches are used to analyze these data, perhaps the most common methodology is a normal mixed effects model where some effects, such as the treatment effect, are regarded as fixed, and others, such as the effect of group random assignment or the neighborhoods where the students live are regarded as random. For these models, the standard reference used by education researchers is Raudenbush and Bryk (2002).

Although mixed effects models enjoy wide use in estimating parameters from and testing hypotheses about these experiments, the development of standardized mean difference effect size indices for them is relatively recent. For example, effect sizes were recently defined by Hedges (2007) for a two-level random intercept model, by Hedges (2011) for a three-level random intercept model, and Lai and Kwok (2014) for two-level cross-classified and partially cross-classified models.

Purpose / Objective / Research Question / Focus of Study:

This paper unifies the currently published effect sizes in a general mixed effects modeling framework. We then apply this framework to suggest an effect size for a model that currently lack a published effect size, a random slope model with heterogeneous treatment effects.

Significance / Novelty of study:

We make three contributions:

1. We propose a framework that unifies the definition and estimation of effect sizes in mixed-effects models. Prior work studied only special cases of the general model.

2. We show that the general framework effect sizes have desirable properties: namely that these effect sizes (a) either recover past effect sizes or are comparable with them, (b) are substantively interpretable, (c) have estimators that are easily computable by both primary and meta-analysts, and (d) these estimators have attractive technical properties such as consistency.

3. We use this framework to suggest an effect size for random-slope models, a type of model for which there exists no previously published effect size. We note that this new, random slope effect size has the same metric as prior work on random intercept effect sizes, and further note that the new effect size estimator has substantially increased precision on random slope data.

Statistical, Measurement, or Econometric Model:

Consider the following the normal mixed effects model

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

where $\mathbf{X}$ and $\mathbf{Z}$ are fixed design matrices, $\boldsymbol{\beta}$ is a vector of regression coefficients for the fixed effects, $\mathbf{u}$ is a vector of random effects, and $\mathbf{e}$ is a vector of the residual errors. Let $\boldsymbol{\tau}$ be a vector of variance components and let the functions $\mathbf{D}[\boldsymbol{\tau}]$ and $\mathbf{R}[\boldsymbol{\tau}]$ map the vector of variance

components into variance covariance matrices for the random effect vector $\mathbf{u}$ and error vector $\mathbf{e}$ respectively. If we assume that $\mathbf{u}$ and $\mathbf{e}$ are independent, then

$$\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{bmatrix}\mathbf{D}[\boldsymbol{\tau}] & \mathbf{0}\\ \mathbf{0} & \mathbf{R}[\boldsymbol{\tau}]\end{bmatrix}\right)$$

and it follows that the covariance of $\mathbf{Y}$ is

$$\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{Y}] = \mathbf{Z}\mathbf{D}[\boldsymbol{\tau}]\mathbf{Z}^{\top} + \mathbf{R}[\boldsymbol{\tau}] \quad .$$

For example, a special case of this model is a two-level Hierarchical Linear Model with random slopes and intercepts such as might be used to model heterogeneous treatment effects in a cluster randomized trial. Following the notation of Raudenbush and Bryk (2002), let Y_{ij} be the response of individual $i = 1 \ldots N$ in group (or cluster) $j = 1 \ldots J$ so that

$$Y_{ij} = \gamma_{00} + \gamma_{01} \cdot \mathrm{TREAT}_j + u_{0j} + u_{1j} \cdot \mathrm{TREAT}_j + r_{ij} \qquad (2)$$

where γ_{00} is the intercept, γ_{01} is the average fixed effect of treatment, $\mathrm{TREAT}_j = 1$ for units in the treatment condition, $\mathrm{TREAT}_j = 0$ for units in the control condition, the $u_{0j} \sim \mathrm{N}(0, \tau_{00}^2)$ are the random intercepts, the $u_{1j} \sim \mathrm{N}(0, \tau_{11}^2)$ are the random slopes, the $r_{ij} \sim \mathrm{N}(0, \sigma^2)$ are the individual errors, and the only non-zero covariance terms are $\mathrm{Cov}(u_{0j}, u_{1j}) = \tau_{10}^2$. In our notation then,

$$\mathbf{X} = \begin{bmatrix}1 & \mathrm{TREAT}_1\\ 1 & \mathrm{TREAT}_1\\ \vdots & \vdots\\ 1 & \mathrm{TREAT}_2\\ 1 & \mathrm{TREAT}_2\\ \vdots & \vdots\\ 1 & \mathrm{TREAT}_J\end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix}1 & \mathrm{TREAT}_1 & & & & & & \\ 1 & \mathrm{TREAT}_1 & & & & & & \\ \vdots & \vdots & & & & & & \\ & & 1 & \mathrm{TREAT}_2 & & & & \\ & & 1 & \mathrm{TREAT}_2 & & & & \\ & & \vdots & \vdots & \ddots & & & \\ & & & & & 1 & \mathrm{TREAT}_J\end{bmatrix} \quad \mathbf{u} = \begin{pmatrix}u_{01}\\ u_{11}\\ u_{02}\\ u_{12}\\ \vdots\\ u_{0J}\\ u_{1J}\end{pmatrix}$$

where $\mathbf{X}$ is of order $N \times 2$, $\mathbf{Z}$ is $N \times 2J$, and $\mathbf{u}$ is $2J \times 1$. Furthermore,

$$\boldsymbol{\tau} = \begin{pmatrix}\sigma^2\\ \tau_{00}^2\\ \tau_{10}^2\\ \tau_{11}^2\end{pmatrix} \quad \mathbf{D}[\boldsymbol{\tau}] = \begin{bmatrix}\tau_{00}^2 & \tau_{10}^2 & & & \\ \tau_{10}^2 & \tau_{11}^2 & & & \\ & & \tau_{00}^2 & \tau_{10}^2 & \\ & & \tau_{10}^2 & \tau_{11}^2 & \\ & & & & \ddots\end{bmatrix} \quad \mathbf{R}[\boldsymbol{\tau}] = \sigma^2\mathbf{I}$$

where $\mathbf{D}[\boldsymbol{\tau}]$ is of order $2J \times 2J$ and $\mathbf{I}$ is an $N \times N$ identity matrix.

Definition of an effect size

In this paper, we define effect sizes in relation to hypothesis tests. We only consider hypothesis tests that can be expressed as a linear combination of regression coefficients being equal to zero,

$$\mathrm{H}_0: \boldsymbol{\ell}^{\top}\boldsymbol{\beta} = 0$$

where $\boldsymbol{\ell}^{\top}$ is a vector. For example, $\mathrm{H}_0: \gamma_{01} = 0$, tested with a Wald t test.

This t statistic based approach is also followed by Hedges (2007), Hedges (2011), and Lai and Kwok (2014). These papers, however, only explicitly discussed the definition of effect sizes in the context of models where the diagonal of $\boldsymbol{\Sigma}$ is constant, i.e. where the diagonal of $\boldsymbol{\Sigma}$ makes a natural scale for an effect size. We extend their work to the more general mixed effects model of Equation (1) by defining a scale parameter for the case where the diagonal of $\boldsymbol{\Sigma}$ is non-constant.

One interpretable scale parameter is the average variance of the observed units in the sample. We define the average variance as

$$\overline{\sigma^2} = \frac{1}{N}\text{trace}\left[\mathbf{ZD}[\boldsymbol{\tau}]\mathbf{Z}^\top + \mathbf{R}[\boldsymbol{\tau}]\right]$$

which for the model of Equation 2 would be

$$\overline{\sigma^2} = \sigma^2 + \tau_{00}^2 + \tau_{11}^2\left(\frac{1}{N}\sum_{ij}(\text{TREAT}_j)^2\right) + 2\tau_{10}^2\left(\frac{1}{N}\sum_{ij}\text{TREAT}_j\right) \quad .$$

Note that in models where the diagonal of $\boldsymbol{\Sigma}$ is constant, $\overline{\sigma^2}$ will be equal to the diagonal. This measure, however, can depend on the observed values of the matrix $\mathbf{Z}$, e.g. the TREAT_j values. To address this, we define the expected average variance as

$$\mathbb{E}\left[\overline{\sigma^2}\right] = \frac{1}{N}\text{trace}\left[\mathbb{E}\left[\mathbf{ZD}[\boldsymbol{\tau}]\mathbf{Z}^\top + \mathbf{R}[\boldsymbol{\tau}]\right]\right]$$

which, for the model of Equation 2 would be

$$\mathbb{E}\left[\overline{\sigma^2}\right] = \sigma^2 + \tau_{00}^2 + \tau_{11}^2(\text{Var}[\text{TREAT\_j}] + (\mathbb{E}[\text{TREAT\_j}])^2) + 2\tau_{10}^2\mathbb{E}\left[\text{TREAT}_j\right] \quad .$$

If we further assume that the probability of assignment to treatment is p, i.e. $\text{TREAT}_j \sim \text{Bernoulli}(p)$, then

$$\mathbb{E}\left[\overline{\sigma^2}\right] = \sigma^2 + \tau_{00}^2 + \tau_{11}^2 p + 2\tau_{10}^2 p \quad .$$

We define an effect size of the linear combination $\boldsymbol{\ell}^\top\boldsymbol{\beta}$ as

$$\delta = \frac{\boldsymbol{\ell}^\top\boldsymbol{\beta}}{\sqrt{\mathbb{E}\left[\overline{\sigma^2}\right]}} \quad ,$$

which, for the model in Equation 2 would be

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma^2 + \tau_{00}^2 + \tau_{11}^2 p + 2\tau_{10}^2 p}} \qquad (3)$$

The inclusion of p in the expression for δ is somewhat unsettling: we generally desire that effects size indices not depend on the particular details of a given experiment. Note, however, that the interpretation of δ is as the mean difference (or, more generally, linear combination) scaled by the expected variation of a replication of the experiment. That is, p is a value from the future, not from the current experiment, per se. If, for example, we wish to consider the situation where the intervention is universally implemented, then the appropriate effect size has $p = 1$. If, however, we wish to consider the effect of a very small pilot implementation of the intervention, then we should set $p = 0$. The choice of p will depend on how we wish to interpret δ.

Estimation of δ

Estimation of δ depends on the structure of $\boldsymbol{\Sigma}$. Let

$$\boldsymbol{\Sigma} = \overline{\sigma^2} \cdot \mathbf{V}_U$$

so that $\text{trace}[\mathbf{V}_U] = N$. For example, in a simple two-level random intercept model, $\overline{\sigma^2} = \sigma_w^2 + \sigma_b^2$ where σ_w^2 is the variation within schools, σ_b^2 is the variation between schools. In this case, $\mathbf{V}_U$ is the block diagonal matrix of Intra-Class Correlations (ICCs), where $\rho = \sigma_b^2/(\sigma_w^2 + \sigma_b^2)$ is the ICC.

If $\mathbf{V_U}$ is known a priori, e.g. ρ is known a priori, then Generalized Least Squares (GLS) provides optimal estimates of $\boldsymbol{\beta}$. In the paper, we derive a Uniformly Minimum Variance Unbiased Estimator (UMVUE) of δ and an estimator of its variance. Our approach is a direct extension of the arguments in Hedges (1981), where we note that, under mild regularity conditions, (i) the GLS t statistic is a function of complete and sufficient statistics, (ii) the non-centrality parameter of the GLS t statistic is a simple function of the effect size of interest, and (iii) therefore, an appropriately scaled GLS t statistic is an unbiased estimator composed of complete and sufficient statistics, i.e. via the Lehmann-Scheffe theorem, it is the UMVUE of δ.

If $\mathbf{V_U}$ is not known a priori, then Restricted Maximum Likelihood (REML) has attractive properties. For example, the REML estimate of $\boldsymbol{\tau}$ is a consistent estimator, so a plug in estimator where $\hat{\boldsymbol{\tau}}$ is substituted for $\boldsymbol{\tau}$ in the GLS estimator of δ, is also a consistent estimator under mild regularity conditions. In the paper, we give expressions for both the point estimate and an estimator of its variance. We also show that these estimators can be calculated by meta-analysts.

Usefulness / Applicability of Method:
The new method leads to effect sizes and estimators for a wide variety of mixed effects models. We show in the paper that the definition of δ is useful because it has the following properties:

1. δ is interpretable: it is the observed mean difference (or linear combination) scaled by a measure of the expected variability of a future replication of the experiment.

2. δ can recover prior effect sizes such as those defined by Hedges (2007), Hedges (2011), and Lai and Kwok (2014). For example, Hedges (2007) considered a random intercept model for a group randomized experiment. He defined three effect sizes δ_t, δ_w, and δ_b. We show that the new δ recovers δ_t, δ_w, and δ_b depending on how the replication of the experiment is defined. If the experiment is replicated in new schools with new students, then $\delta = \delta_t$; if the same schools, but new students, then $\delta = \delta_w$; and if new schools, but the same students, then $\delta = \delta_b$.

3. δ is on the same scale as prior effect sizes. For example, the Hedges (2007) $\hat{\delta}_t$ random intercept effect size estimator is on the same scale as the new random slope estimator $\hat{\delta}$. Figure 1 displays the results of a simulation where many datasets where generated from Equation 2 and both the random intercept $\hat{\delta}_t$ estimator and the slope estimator $\hat{\delta}$ are calculated on each dataset. Note that the means of both estimators are similar -- each is estimating the same quantity. Note also, however, that the new estimator is far more efficient on random slope data, i.e. $\hat{\delta}_t$ and $\hat{\delta}$ are comparable, but $\hat{\delta}$ is more efficient.

Conclusions:
The general framework we propose generates interpretable effect sizes for a wide class of mixed effects models. Furthermore, we illustrate the value of this framework by both recovering the definitions of prior effect sizes published in the literature and by deriving a new effect size for a cluster randomized controlled trial with heterogeneous treatment effects.

There are, however, several open questions. Can unconditional and conditional effect sizes be converted between each other in mixed effects models? Are similar effect size indices defined from F and χ^2 tests of hypothesis in the general model of Equation 1 similarly interpretable?

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Hedges, Larry V. 1981. “Distribution Theory for Glass’s Estimator of Effect Size and Related Estimators.” *Journal of Education Statistics* 6 (2): 107–28.

———. 2007. “Effect Sizes in Cluster-Randomized Designs.” *Journal of Educational and Behavioral Statistics* 32 (4): 341–70.

———. 2011. “Effect Sizes in Three-Level Cluster-Randomized Experiments.” *Journal of Educational and Behavioral Statistics* 36 (3): 346–80.

Lai, Mark H C, and Oi-Man Kwok. 2014. “Standardized Mean Differences in Two-Level Cross-Classified Random Effects Models.” *Journal of Educational and Behavioral Statistics* 39 (4): 282–302.

Raudenbush, Stephen W, and Anthony S Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA, USA: Sage Publications.

Appendix B. Tables and Figures
Not included in page count.

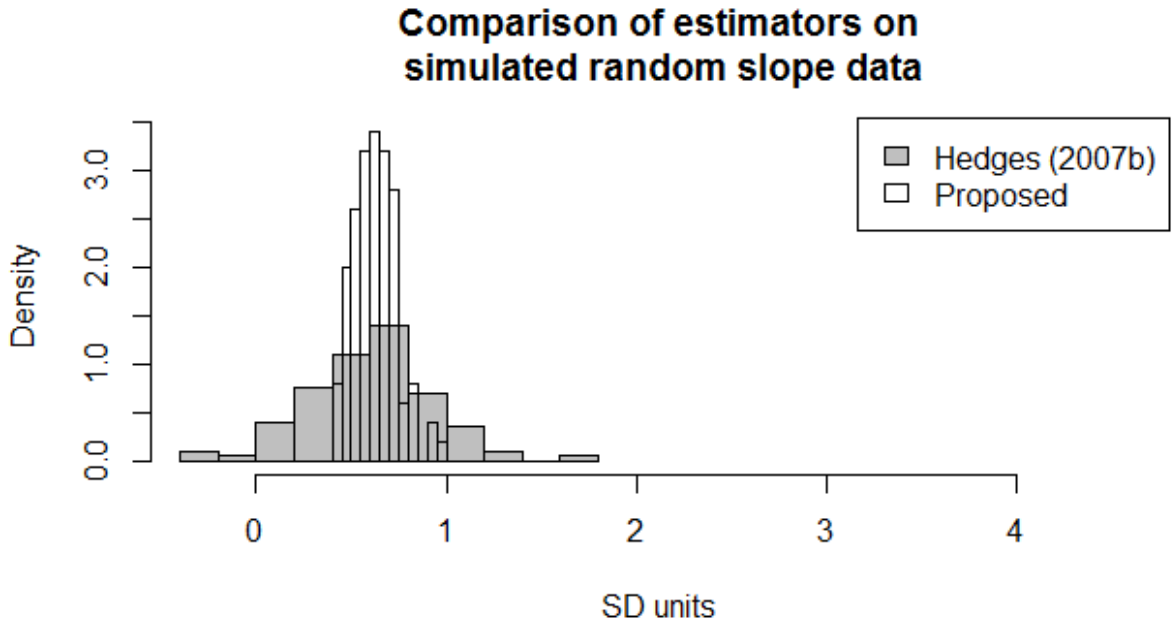

Figure 1: The Hedges (2007b) δ_t random intercept effect size estimator has a wider sampling distribution on random slope data than the new, proposed random slope estimator δ. Both sampling distributions, however, have very similar means.