Abstract Title Page

Title:
Estimating the Variance of Design Parameters

Authors and Affiliations:

E. C. Hedberg
Arizona State University

L. V. Hedges
Northwestern University

A.M. Kuyper
Northwestern University

Background / Context:
Description of prior research and its intellectual context.

Randomized experiments are generally considered to provide the strongest basis for causal inferences about cause and effect. Consequently randomized field trials have been increasingly used to evaluate the effects of education interventions, products, and services. Populations of interest in education are often hierarchically structured (such as students nested within classrooms, nested within schools, nested within school districts). The sampling designs used in educational experiments often exploit this hierarchical structure by sampling entire intact units (such as schools or classrooms). The two most frequently used designs in education field experiments are variants of two designs: The hierarchical design or the (generalized) randomized block design (see Spybrook and Raudenbush, 2009).

The hierarchical design assigns entire intact groups (such as schools) to treatments, so that every individual (or lower level hierarchical unit) in the group receives the same treatment. Because the intact groups assigned to treatments are statistical clusters (in the sense of sampling theory) the hierarchical design is often called the cluster-randomized design. For example, an experiment using the hierarchical design might assign entire schools (and all the classrooms and students within them) to receive the same treatment.

In the randomized block design, individuals (or lower level units) are assigned to treatments *within* intact groups such as schools. Because the intact groups within which treatment assignment takes place are often geographical site, this design is often called the multisite design. It is also sometimes called the matched design because the units assigned to different treatments occur within the same higher level intact unit and are therefore matched by virtue of being in that same higher level unit. For example, an experiment using the randomized block design might involve several schools, but assign different individuals within each school to different treatments. Alternatively, it might assign entire different classrooms (and all of the individuals with those classrooms) to different treatments within each school.

A crucial aspect of designing field experiments is the assessment of the statistical power of the test for treatment effects so that the investigator can be sure that the design has adequate sensitivity to detect the smallest treatment effects that are judged to be important. The statistical power of hierarchical designs depends on the effect size, the significance level, the sample size at each level of the design (e.g., the number of clusters and the number of individuals in each cluster), the way that the outcome variance is distributed across levels of the design (usually summarized by the intraclass correlation structure), and if covariates are used, the effectiveness of the covariates in explaining variation at each level of the design (usually summarized by variance explained or R^2 statistics at each level of the design).

While the significance level and the sample sizes are under the control of the investigator, the intraclass correlation structure and the effectiveness of the covariates are not. Moreover these parameters are often difficult to know precisely before the experiment has been carried out. Because of their importance in planning experiments a literature has emerged that provides an empirical grounding for intraclass correlations and covariate R^2 values. These values come from experiments that have already been conducted (e.g., Schochet, 2008), large urban school districts (e.g., Bloom, et al.,2007), sample surveys (e.g., Hedges and Hedberg, 2007), or state longitudinal data systems (e.g., Hedges and Hedberg, 2014). Such reports generally provide estimates of the design parameters and an uncertainty (standard error) of those estimates.

Although intraclass correlation was originally introduced in relation to two level sampling models, the concept extends naturally to sampling models with three or more levels. The intraclass correlation concept in cases of three and four level sampling models is of great interest in the design and analysis of experiments in education (Hedges and Rhoads, 2010; Konstantopoulos, 2008ab, 2009).

Consequently there has been considerable interest in the estimation of intraclass correlations from sample surveys using multistage samples to estimate intraclass correlations (e.g., Hedges & Hedberg, 2007). Such studies typically fit unconditional multilevel models to the survey data to estimate variance components at each level of the sampling design. Other studies use datasets that were assembled in the course of carrying out randomized experiments (Bloom et al., 2007). While the surveys often have large total sample sizes, the number of sampled units at some levels (typically the higher levels) may not be large enough make negligible the sampling uncertainty of estimates of variance components and functions of variance components (such as intraclass correlations). Even experiments that are normatively large (that is large for experiments) typically involve much smaller sets of schools than national surveys (typically less than 100 schools). Consequently, for either type of study designed to provide reference values of intraclass correlations, it is important to provide some assessment of the uncertainty of the estimates. However because the sample sizes in surveys are actually large, estimates of sampling uncertainty based on large sample methods should be accurate enough to give this guidance.

In the case of experiments with randomized block designs, statistical power depends on an additional parameter: the variation of treatment effects across higher-level units. Different investigators have used several different variants of this parameter, but there has been little research on the estimation of these treatment effect heterogeneity parameters.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this paper is to consider estimators of the treatment effect heterogeneity parameters and their sampling distributions. In addition, the estimation and sampling distributions of intraclass correlations are also reviewed. This research should be of interest to investigators who wish to report empirical estimates of treatment effect heterogeneity parameters (and their uncertainties) from individual experiments or larger datasets.

Significance / Novelty of study:

The literature on estimation of intraclass correlations, however, has largely been restricted to the case of two level models. We review this literature. However, little work as been done to outline the variance of heterogeneity parameters. This paper brings to the literature new formulas for the estimation of the variance of heterogeneity parameters.

Statistical, Measurement, or Econometric Model:

Variance of ICC for Two Level Model

In a two level model there are two variance components σ_1^2 and σ_2^2, which are estimated by s_1^2 and s_2^2, respectively, where $s_1^2 \approx \sigma_1^2$. Let v_1 and v_2 be the variances of s_1^2 and s_2^2. The condition $s_1^2 \approx \sigma_1^2$ implies that $v_1 = 0$. Let m denote the number of clusters (level 2 units) and n_i

denote the number of level 2 units in the i^{th} level 2 unit. When the design is balanced, $n_1 = \cdots = n_m = n$. The interclass correlation in the two level model is

$$\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

which is estimated by

$$r = \frac{s_2^2}{s_1^2 + s_2^2}.$$

Then the large sample variance of r, the estimate of ρ in a balanced design is

$$\frac{(1-\rho)^2 v_2}{\sigma_T^4} \qquad (1)$$

where v_2 is the variance of s_2^2, the sample estimate of the level 2 variance component and $\sigma_T^2 = \sigma_1^2 + \sigma_2^2$ is the total variance. The sample estimate of the variance of r is obtained by replacing all of the parameters in (1) by their samples estimates, that is the estimate of the variance of r is

$$\frac{(1-r)^2 v_2}{\left(s_1^2 + s_2^2\right)^2}.$$

The standard error of r is just the square root of its estimated variance.

Variance of Heterogeneity Parameters for Two Level Block Randomized Design

Suppose that there are m level 2 units (blocks or clusters such as schools) and there are n_{ij} level 1 units (individuals) in each level 2 unit. Let Y_{ij} be the outcome score for the jth level 1 unit in the ith level 2 unit. The level 1 model is

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_{ij} + \varepsilon_{ij},\ i = 1, \ldots, m; j = 1, \ldots n_i, \qquad (2)$$

Where β_{0i} is the mean and β_{1i} is the treatment effect in the ith level 1 unit, T_{ij} is a treatment indicator, and ε_{ij} is a normally distributed level 1 residual with mean zero and variance σ_1^2. The level 2 model is

$$\beta_{0i} = \gamma_0 + \eta_{0i},\ i = 1, \ldots, m, \qquad (3)$$
$$\beta_{1i} = \gamma_1 + \eta_{1i},\ i = 1, \ldots, m,$$

where γ_0 is the grand mean, γ_1 is the mean treatment effect, η_{0i} is a normally distributed level 2 residual with mean zero and variance σ_2^2, and η_{1i} is a normally distributed level 2 residual with mean zero and variance τ^2.

Statistical power in randomized block designs depends on the significance level and the sample sizes at each level, but also on the effect size, the intraclass correlation, and the treatment effect heterogeneity, although the precise relation depends on the way the treatment effect size and treatment effect heterogeneity are expressed.

The effect size can be defined as

$$\delta_T = \gamma_1 \Big/ \sqrt{\sigma_1^2 + \sigma_2^2} \qquad (4)$$

or as

$$\delta_W = \gamma_1/\sigma_1, \qquad (5)$$

(see Hedges, 2007).

Note that the effect size and the intraclass correlation are scale free, that is, δ and ρ are defined as a proportion of the total variance of an untreated population and therefore (like the effect size) it does not depend on the scale of the outcome variable.

The most direct parameter used to express the treatment effect heterogeneity is the treatment effect variance (the treatment by level 2 unit interaction variance component) τ^2. For the purposes of providing reference values that might apply across experiments, τ^2 has the disadvantage that is not scale free—it depends on the scale of the outcome variable. Consequently, discussions of statistical power in randomized block experiments have usually relied on transformations of τ^2 that are scale free. One of these has been called the effect size variance

$$ESV = \tau^2/\sigma_1^2, \qquad (6)$$

which corresponds to the variance of the effect sizes that would be computed if each level 2 unit were a separate experiment, that is the variance of the $\delta_i = \beta_{1i}/\sigma_1$ values (see Spybrook, et al, 2006). Another parameter that has been used is

$$\omega = \tau^2/\sigma_2^2, \qquad (7)$$

which is the ratio of treatment effect variance to the variance of level 2 units (Hedges and Rhoads, 2010; Hedges and Borenstein, 2014).

If the number of units receiving each treatment is identical within every level 2 unit, the design is balanced. In balanced designs, there are simple presentations of the treatment heterogeneity indices in terms of *F*-statistics from the treatment by level 2 unit analysis of variance. The analysis of variance *F*-statistic for the treatment by level 2 unit interaction (F_{AB}) has a sampling distribution that is $[(n\tau^2/2 + \sigma_1^2)/\sigma_1^2]$ times a central *F* random variable with $(m - 1)$ and $(mn - 2m)$ degrees of freedom (see, e.g., Searle, 1971). Because the expected value of this central *F* is $(mn - 2m)/(mn - 2m - 2)$, it follows that the expected value of F_{AB} is

$$\left(\frac{n}{2}\frac{\tau^2}{\sigma_1^2}+1\right)\left(\frac{mn-2m}{mn-2m-2}\right),$$

it follows that

$$\left(\frac{2}{n}\right)\left[\left(\frac{mn-2m-2}{mn-2m}\right)F_{AB}-1\right] \qquad (8)$$

is an unbiased estimator of *ESV*. Because the variance of the relevant central *F* is

$$\frac{2(mn-2m)^2(mn-m-3)}{(m-1)(mn-2m-2)^2(mn-2m-4)}$$

it follows that the variance of (8) is

$$\frac{\left(n^2\tau^4+4n\tau^2\sigma_1^2+4\sigma_1^4\right)\left[2(mn-m-3)\right]}{n^2\sigma_1^4(m-1)(mn-2m-4)}$$

$$=\frac{(nESV+2)^2\left[2m(n-1)-6\right]}{n^2(m-1)(mn-2m-4)}. \qquad (9)$$

The estimator of ω is obtained from the variance component estimates

$$\frac{2(MSAB-MSE)}{MSB-MSE}=\frac{2(F_{AB}-1)}{F_B-1}, \qquad (10)$$

where *MSAB, MSB, MSE*, F_{AB}, and F_B are the treatment by level 2 interaction, level 2 main effect, and within cell or error mean squares and F_{AB} and F_B are the treatment by level 2 interaction and level 2 main effect *F*-test statistics from the two factor analysis of variance. Although the numerator and denominator of (10) are unbiased estimates of τ^2 and σ_2^2, respectively, the ratio is not an unbiased estimator of τ^2/σ_2^2, but generally estimates a quantity that is larger than ω. The variance of (10) is approximately

$$\frac{8a\tilde{\rho}^2}{n^2} + \left(\frac{8}{n(m-1)} - \frac{8\tilde{\rho}}{n^2 m(n-2)}\right)\tilde{\rho}\omega + \left(\frac{4}{m-1} + \frac{4\tilde{\rho}}{n(m-1)} + \frac{2a\tilde{\rho}^2}{n^2}\right)\omega^2, \qquad (11)$$

where

$$a = \frac{mn-m-1}{m(n-2)(m-1)}$$

and $\tilde{\rho} = (1-\rho)/\rho$. Obviously if we are estimating ω, neither ω nor ρ will be known so that we will have to substitute estimates of ω and ρ for the parameter values in (11), which results in a consistent estimator of the variance.

Usefulness / Applicability of Method:

Stata programs to estimate the variance of ICCs are available in the package ICCVAR (to install, type "ssc install iccvar" into the stata command prompt). Code for estimating the variance of heterogeneity parameters is also being developed.

Conclusions:

This work allows the growing community of design scientists to estimate variances of important parameters so that users of compendiums are aware of the sampling variability of their recommendations.

One limitation of the methods presented here is that they assume a balanced design. We are working on derivations that allow for unbalanced designs and hope to have them ready for the conference.

Appendices

Appendix A. References

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30-59.

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

Hedges, L. V., & Borenstein, M. (2014). Conditional Optimal Design in Three-and Four-Level Experiments. *Journal of Educational and Behavioral Statistics*, 1076998614534897.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass Correlations and Covariate Outcome Correlations for Planning Two-and Three-Level Cluster-Randomized Experiments in Education. *Evaluation review*, 0193841X14529126.

Hedges, L. V., & Rhoads, C. (2010). Statistical Power Analysis in Education Research. NCSER 2010-3006. *National Center for Special Education Research*.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66-88.

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265-288.

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation review*, *33*(4), 335-357.

Raykov, T. (2011). Intraclass correlation coefficients in hierarchical designs: Evaluation using latent variable modeling. *Structural Equation Modeling*, *18*(1), 73-90.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62-87.

Searle, S. R. (1971). A Biometrics Invited Paper. Topics in Variance Component Estimation. *Biometrics*, 1-76.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298-318.

Spybrook, J., Raudenbush, S. W., Liu, X. F., Congdon, R., & Martínez, A. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. *Survey Research Center of the Institute of Social Research at University of Michigan*.