Abstract Title Page
Not included in page count.

Title: Estimating cross-site impact variation in the presence of heteroscedasticity

Authors and Affiliations: Howard S. Bloom (MDRC), Kristin E. Porter (MDRC), Michael J. Weiss (MDRC) and Stephen Raudenbush (University of Chicago)

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

To date, evaluation research and policy analysis have focused mainly on average program impacts and paid little systematic attention to their variation. Exceptions to this rule are subgroup analyses, which typically are idiosyncratic[1], and quantile regression analyses, which have limited interpretations[2]. Scholars have argued for some time, however, that to fully understand the implications of programs one must understand the variation in their impacts (e.g. Bryk and Raudenbush, 1988; Heckman, Smith and Clements, 1997; Friedlander and Robins, 1997; Heckman, 2001; Raudenbush and Liu, 2000: Abadie, Angrist and Imbens, 2002; Bitler and Hoynes, 2006; and Habiba and Smith, 2008). Recently, the growing number of multi-site randomized trials that are being planned and conducted make it increasingly feasible to study *cross-site* variation in impacts.

Important technical questions arise when one uses data from a multi-site randomized trial to estimate the magnitude of cross-site impact variation and assess the statistical significance of these estimates. Some of these questions arise from the likelihood that variation in *individual* outcomes (which in the context of multi-level models is often referred to as a "level-one residual variance" or "residual variance" for short) might be different for the treatment group and control group members and/or might be different for different sites. For example, to the extent that program impacts vary across individuals, the variance of individual outcomes (i.e. the residual variance) in the treatment group will differ from that in the control group (Bloom, Raudenbush and Weiss, under review). In addition, to the extent that different populations and local conditions are represented by different sites, the variance of individual "counterfactual" outcomes (i.e. potential outcomes under no treatment) and therefore the variance of control group members' outcomes will differ across sites. This situation can be exacerbated if *the extent of* individual variation in program impacts also varies across sites because of their differences client populations and local conditions. These *differences in individual outcome variation* represent forms of heteroscedasticity which we refer to as: (1) T/C heteroscedasticity[3] and (2) cross-site heteroscedasticity. If not properly accounted for, this heteroscedasticity can result in bias and incorrect statistical inference when estimating cross-site impact variation. These forms of heteroscedasticity are particularly problematic if not accounted for when the sample sizes within sites are small, which is the typical case in practice.

[1] Bloom and Michalopoulos (2011) provide guidance for improving such analyses.

[2] The use of quantile regression analysis to study program impacts (e.g. Friedlander and Robins, 1997; Abadie, Angrist and Imbens, 2002 and Bitler and Hoynes, 2006) is limited by the fact that such analyses can identify impacts on the distribution of individual outcomes but without strong assumptions, they cannot identify the distribution of program impacts.

[3] It is possible to eliminate bias from T/C heteroscedasticity through a balanced design that has the same number of treatment group members and control group members overall (if there is no cross-site heteroscedasticity) or for each site (if there is cross-site heteroscedasticity).

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

Standard practice in multi-site trials in evaluation research and applied social science is to estimate a pooled impact model across sites, with a single pooled individual variance for all treatment and control groups from all sites. A random-effects version of such models is often used to estimate cross-site impact variation (Bloom, Raudenbush and Weiss, under review); however a mixed-effects version of such models (with fixed intercepts and random slopes) resolves some of the problems we highlight. The typical estimators of the cross-site variance (and other parameters) in both of these models are restricted maximum likelihood (RML) estimators.

A straight-forward extension of the single, pooled residual variance random-effects or mixed-effects models, which can be readily implemented by most statistical software packages, is to estimate two residual variances – one for treatment group members and another for control group members. Doing so can properly account for T/C heteroscedasticity. In addition, for most studies that randomize individuals within sites, the number of degrees of freedom from which to estimate these two residual variances is quite large. Hence, they are subject to little estimation error and are unlikely to produce further problems in the analysis.

It is also possible to estimate a separate residual variance for each site in a multi-site study, which in theory could account for cross-site heteroscedasticity. However doing so can dramatically reduce the number of degrees of freedom available for estimating each variance and thereby introduce problematic estimation error into the analysis. Likewise it is possible to estimate a separate residual variance for the treatment group and control group from each site (depending on how many sites there are), which in theory could account for both T/C and cross-site heteroscedasticity. However, doing so further reduces the degrees of freedom that are available for estimating each variance and introduces even more problematic estimation error into the analysis.

Therefore, we address the following methodological questions:

- To what extent are estimators of cross-site impact variation biased and tests of statistical significance incorrect when one does not take into account how the variance of individual outcomes differs between treatment and control group members and/or across sites? In other words, in practice, what are the *risks of estimating too few* residual variances and how does this risk vary under a wide range of conditions?

- Among different estimators that take heteroscedasticity into account – estimators based on either a random-effects model, on a mixed-effects model or on no model (a method of moments estimator) — how do they compare?

- What recommendations can be made for practice?

Relevant statistical theory for the estimators of focus provides some guidance, but is mainly limited to asymptotic or large-sample properties and does not help us fully answer the preceding questions. Consequently, we address these questions through a series of simulations.

Setting:
Description of the research location.
(May not be applicable for Methods submissions)

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

To our knowledge, there is no guidance on how to correctly specify a model for estimating cross-site impact variation or on how to specify the number of distinct level-one residual variances. Analysts who are aware of the risks of misspecification may find themselves between a rock and a hard place. On the one hand, an analyst may be inclined to pool across research groups and/or sites (estimating 1 or 2 level-one residual variances), realizing that with small site sample sizes (typical in most studies), the residual variances will be estimated poorly without pooling. However, on the other hand, relying on a pooled estimate of residual variances, an analyst cannot account for all the various types of heteroscedasticity. Both extremes could potentially lead to nontrivial bias and/or incorrect inference. We demonstrate this and provide preliminary guidance.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

Our study provides guidance to analysts for specifying commonly-used models for estimating cross-site variation in impacts.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Research Design:
Description of the research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).
(May not be applicable for Methods submissions)

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

Findings / Results:
Description of the main findings with specific details.
(May not be applicable for Methods submissions)

Our simulations suggest that the RML estimator of the cross-site variance in a mixed-effects model with two level-one residual variance is robust to T/C and cross-site heteroscedasticity. The consequences for statistical inference are still being investigated.

Although not immediately intuitive, in the case that there is both T/C and cross-site heteroscedasticity, it is sufficient to specify just 2, rather than 2*J* (where *J*=the number of sites), variances. Theory predicts that an estimator of the mean error variance for site-level mean program estimates, which assumes 2 level-one residual variances—a pooled treatment group variance and a pooled control group variance—is unbiased.[4] This implies and our simulations confirm that the RML estimator of the cross-site variance in a mixed-effects model assuming two level-one residual variances is robust when there is both T/C and cross-site heteroscedasticity.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

Continuing research includes the investigation of the properties of the estimators of cross-site variance in additional settings (e.g. different assumptions about the distributions of impact across individuals and across sites). Our research will lead to recommendations for practice, which we expect to include recommended data checks, cautions about model specification and suggestions for conducting tailored simulation studies based on one's actual dataset before conducting any analysis.

[4] Personal memo from Stephen Raudenbush.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Abadie, A., Joshua Angrist and Guido Imbens (2002) Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings, Econometrica, 70: 91 – 117.

Bitler, Marianne P. and Hillary W. Hoynes (2006) What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments, American Economic Review, 96: 988 – 1012.

Bloom, Howard S., Stephen W. Raudenbush and Michael Weiss (under review) "Estimating Variation in Program Impacts: Theory Practice and Applications."

Bloom, Howard S. and Charles Michalopoulos (2011) When is the Story in the Subgroups? Strategies for Interpreting and Reporting Intervention Effects on Subgroups, Prevention Science.

Bryk, Anthony S. and Stephen W. Raudenbush (1988) Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations, Psychological Bulletin, 104 (3): 396 – 404.

Friedlander, Daniel and Philip K. Robins (1997) The Distributional Impacts of Social Programs, Evaluation Review, 21(5): 531 – 553.

Habiba, Djebbari and Jeffrey Smith (2008) Heterogeneous Impacts in PROGRESA, Journal of Econometrics, 145: 64 – 80.

Heckman, James J., Jeffrey Smith and Nancy Clements (1997) Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts, Review of Economic Studies, 64: 487 – 535.

Heckman, James J. (2001) Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, Journal of Political Economy, 109 (4): 673 – 748.

Raudenbush, Stephen W. and Xiaofeng Liu (2000) Statistical Power and Optimal Design for Multisite Randomized Trials, Psychological Methods 5 (2): 199 -213.