# Analysis of Small Ensembles of Social Experiments

David R. Judkins
Abt Associates, 4550 Montgomery Ave, Suite 800 North, Bethesda, MD 20814

## Abstract

Many randomized trials of social interventions involve randomization of clients (e.g., students, welfare recipients) at each of a small collection of service facilities (e.g., schools, welfare offices). Only rarely are the facilities randomly selected; most often they volunteer or agree to participate after intensive recruitment efforts. Actors at the local facilities mediate the effects of the intervention -- both by the fidelity of their implementation and by the charisma and energy they bring to their organizations. For this reason, generalization of results to broader implementation phases can be hazardous. Nonetheless, if the intervention is successful in the trial, advocates for the intervention will use the results to urge more widespread implementation. If they succeed, the results may be disappointing if it turns out that the intervention only works in the hands of a few skilled actors. This hazard of can be reduced by placing confidence intervals on estimated effects that reflect the variation in effects across sites (the random slopes). The question then arises of how many sites are required to get valid generalization inferences. This paper reviews the literature, lays out some philosophy, and shares some new simulations.

**Key Words:** Random slopes, external validity, multicenter trials, treatment-by-center interaction

## 1. Local Actors

In many individually randomized trials of social interventions (e.g., education reform, childcare reform, welfare reform, parole reform), multiple sites are used as a means to build sample size. With multiple sites, it often becomes easier to find adequate numbers of subjects to recruit for the trial. However, this easing comes with the challenge of using local workers and facilities to administer the treatment. In this paper, I refer to the staff working to administer the treatment as local actors, or just actors. Although trialists work hard to train these local actors to deliver the intervention uniformly, these actors bring their own skills and attitudes to their work. If the treatment works better when delivered by some actors than others, the question arises of how to combine the results and how to measure the precision of the combined results. From the field of drug development, Senn (2007, p223), lists five estimands that we might try to make inferences about from a multicenter trial. Numbers one and five from his list are:

Q1) The true mean of the effects for all patients in the trial.
Q5) The true effect for any future patient or centre to which we might wish to apply the results.

If there are treatment-by-center interactions, then different methods must be used to answer the two questions. The terminology around these two questions varies by

substantive field. Often, inferences that are valid for Q1 are said to be "internally valid" while those that are valid for Q5 are said to be "externally valid." Study of Q1 is sometimes referred to as concerning as "proof of concept" or "efficacy," while the study of Q5 is referred to as concerning "effectiveness." Answers to Q1 may be referred to as the effect in the "test ground," while answers to Q5 may be labeled as "combined response to treatment" (Dragalin et al., 2001; Fedorov and Jones, 2005).

Among research areas where randomized human trials are conducted, concern about interactions between local actors and treatments is perhaps highest in the field of behavioural research. In this field, the question of whether to aim for external validity by fitting models with random effects for both the actors and the interaction of treatment with local actor (also known as a random-slopes model in the econometric and meta-analysis literature) or to be satisfied with aiming for internal validity by fitting models with fixed effects for local actors and a single additive treatment effect for all actors is a central issue that has been hotly debated (Walwyn and Roberts, 2010; Feaster, Miklulich-Gilbertson, and Brincks, 2011). Walwyn and Roberts paraphrase Martindale (1978) and Crits-Christoph, et al. (2003) as stating that, "there is little, if no, scientific value in treating therapists as fixed."

Hesitations about the use of random slopes models usually run along one of three lines. The first is that an experiment with good power for Q1 may have low power for Q5. The difference in power might be extreme if the number of centers is small relative to the total patient sample size. The second hesitation is that unless the actors are drawn from a probability sample of a well-defined population, it is unclear how much progress along the path from internal validity to external validity is achieved by fitting a random-slopes model. The third involves technical concerns about the adequacies of model-fitting procures for random-slopes models with a very small number of actors. The focus of this paper is on the third argument against fitting random slopes model in multi-center trials, but in section 2, I also discuss the first two arguments.

In section 3, I review the history of research into the third anti-random-slopes argument. In section 4, I describe the framework for a simulation study I conducted to extend the research in this field to smaller samples sizes (in terms of actors) than had previously been studied. In section 5, I present the results of the simulation study. I close with some concluding remarks in section 6.

## 2. Imperfect External Validity as a Goal

As the Institute of Education Science has increased the numbers of experiments being conducted in education research (Angrist, 2004, Cook, 2007), the issue of whether to aim for internal or external validity in the analysis of multi-site trials has also arisen in the field of education research. Schochet (2008) addressed the issue briefly in his guide to power projections for education trials, drawing on the two standard anti-random-slopes arguments mentioned above (low power for Q5 and lack of probability sampling from a well-defined population):

> *Although this issue needs to be addressed for each study, we believe that the fixed effects case is usually more realistic in evaluations of education interventions. Most evaluations are efficacy trials where a relatively small number of purposively-selected sites are included in the study. Thus, in many*

> *instances, it is untenable to assume that the study sites are representative of a broader, well-defined population. Furthermore, inflating the standard errors to incorporate between-site effects will slant the study in favor of finding internally valid impact estimates that are not statistically significant, thereby providing less information to policymakers on potentially promising interventions. Instead, we believe, in general, that it is preferable to treat site effects as fixed, and to assess the "generalizability" of study findings by examining the pattern of the impact estimates across sites (for example, by calculating the percentage of sites with beneficial impacts). This approach is likely to yield credible information on the extent to which specific interventions could be effective, and whether larger-scale studies are warranted to examine whether they are effective.*

Senn's advice is similar. He is firmly opposed to random sampling of centers, and generally against fitting random-slopes models in this context. With respect to the possibility of randomly selecting centers, he notes:

> *In my view this proposal, however theoretically desirable, is so far removed from practical reality as to be infeasible. The task of proving that a treatment works at all is so difficult and faces so many ethical, practical and financial constraints that most trialists struggle to design a trial that is adequate to proving that the treatment works in the patients studied. …, trialists seek not typical patients but suitable patients for their trials. In any case, there is a philosophical problem in that one cannot sample at random from the future. Thus, I maintain that most trialists will continue to be satisfied with the limited aim of proving that the treatment works at all. To the extent that they attempt to answer the more ambitious question of what the effect of treatment will be more generally, they are unlikely to want to move beyond the conventional random-effects [slopes] estimator. Trying to define more complicated combined responses to treatment is unlikely to be attractive and in particular is unlikely to be reflected in trial design …*

Also, on the fitting random slopes models, he notes:

> *I would nearly always propose a fixed-effect analysis of a clinical trial. I might also consider that a random-effect analysis would be useful on occasion; especially if there were rather many centres which had been fairly widely selected.*

The advice of Schochet and Senn appears to be accepted by most researchers using multi-site trials in social science. However, there has been one very important exception. The best example of a social experiment in which external validity was the explicit goal is the Head Start Impact Study (Puma, et al., 2010). In that evaluation, the set of studied Head Start Centers was a large probability sample of all eligible centers across the U.S. that was explicitly powered to yield useful results even analysis methods appropriate for Q5 are used. This study appears to widely admired, but it was very expensive to conduct and did not find important long-term effects of Head Start, so it by itself is not likely to set the pattern for future social science experimental research. Most future experiments are likely to involve haphazard collections of willing sites numbering from a handful to a few dozen at most. So I think it is important to deconstruct Schochet's arguments and consider whether Senn's advice for drug research applies to social science research.

I think there are two core issues. The first is the frequency with which an answer to Q1 is a satisfactory answer to Q5, and the second is what happens after a positive answer to Q1 is published without any attempt to answer Q5.

As discussed by Michael and O'Muircheartaigh (2008), medical research seems mostly content to study Q1. This seems odd to me given current interest in personalized medicine and the ranking of doctors and health care facilities. As James (2010) discussed, there is substantial variation in practice among doctors, and when conscious efforts are made to reduce the variation, average patient outcomes improve. So it is reasonable to suppose that effectiveness of a pharmaceutical or medical device might vary across doctors. Moreover, Kempthorne and Doerfler (1959) discussed the necessity of jumping from answers to something like Q1 to answers to something like Q5, albeit without supplying many clues about how to make the jump:

> *The fact that such inferences [on Q5] are difficult does not mean, of course, that we should not make them. We have to do so. But we should be clear in our thinking about the basis on which we form such opinions. The basis is outside our experiment. No amount of analysis of our own data will totally justify such an extrapolation.*

One study where this possibility was explored (Lingsma, et al., 2011) failed to find significant variation in the effect of a drug treatment across centers but they advocated continued study of the issue for complex interventions:

> *We consider our results to be applicable to drug interventions, which work on physiological mechanisms. Trials investigating a more complex intervention … may be more sensitive to differences in quality of care.*

This issue of intervention complexity is critical. In social science the interventions we study are typically highly multi-dimensional with wide scope for interpretation by local actors. There so much concern about the lack of faithful and consistent implementation during trials that the number of centers is often kept small so as to facilitate the mounting of parallel qualitative studies of "fidelity" (Judkins, 2011). Moreover, Bell et al. (2011) clearly established that schools that volunteer for experiments are systemically different from other schools in important ways and that the effect of a major education initiative was different in the two sets of schools.

The dispute about whether to reflect the extra uncertainty in relationships due to interactions with local conditions dates back at least to Kish and Frankel (1974). On this subject, Hansen, Madow, and Tepping (1983) commented:

> *Failure to recognize such [design] effects may lead to serious understatement of confidence intervals and to serious overstatements of precision in inferences to the causal system. We believe that misinterpretations are especially likely when design effects due to cluster sampling are not included in the models used for inferences.*

So in social science, we must assume that the answer to Q5 could be very different from the answer to Q1. This brings us to the question of what happens following publication of

an answer to Q1. In the field of drug development, Paul Flyer[1] writes, "We approve a new drug under carefully controlled circumstances. Doctors/patients decide if the published data are supportive for a particular application with respect to the extrapolation of risk/benefit seen in the artificial clinical trial to the actual clinical application. Post marketing studies are often relied upon to provide reassurance that the treatment isn't causing adverse reactions that wouldn't be seen without the intervention but this information is often difficult to interpret where the natural history is not well understood." So the responsibility for implementation of drug research findings devolves to doctors and patients. The situation is rather different in social science research.

Teachers and social workers will generally be compelled to follow protocols to the best of their ability for most of their students, parolees, welfare beneficiaries, or unemployment insurance beneficiaries. If an intervention is widely implemented based on a favorable answer to Q1, and the general pool of workers is not able to implement as well as the specially trained workers in the Q1 study, there may be substantial societal costs. To prevent such losses, as indicated by Schochet, there is a concept in public policy research, that research should be staged – a study is first mounted to answer Q1, a favorable answer is followed by a larger study such as the Head Start Impact Study to answer Q5. This is a reasonable idea, but it is unclear how much such staged research has been conducted. Moreover, I think there is also some danger that Institutional Review Boards (IRBs) may decide that effectiveness trials are unethical once efficacy has been established.

The authors of a report on a randomized trial of a social intervention must be cognizant of the fact that their study, even if carefully poised as an answer to Q1, will be acted upon in the same manner as if the study answered Q5. Schochet is clearly aware of this danger and mentions some ad hoc defenses against it. However, I find his advocacy of the examination of the pattern of site-specific effects is too vague to be useful. Furthermore, his suggestion for pooled inference based on the proportion of sites with beneficial site-specific effects is not well defined. What criterion would be used for "beneficial" and what proportion would need to be significant? The best way to make sense of a pattern of site-specific effects is to fit a random-slopes model. Moreover, if one informally examines the pattern and decides that that one site or another is so different from the others that the effect estimates cannot be pooled, then the power loss might be greater than one would incur by just fitting a random-slopes model.

My position is that even one has a volunteer sample of sites, it is better to aim for imperfect external validity by formally using whatever information is available about cross-site effect heterogeneity. As Senn noted, "If we attempt to answer this difficult question [Q5], the random effects [slopes] model will almost certainly produce a better approximation [than the fixed effects model]." Power to answer Q5 will be lower than it would have been to answer Q1, but if one is lucky in the selection of sites (in the sense that heterogeneity of effects in the general population is mirrored in the volunteer sample), then one will have achieved external validity. At the worst, if the selected sites have unnaturally similar effects, then the instability in the variance estimation process one can cause one to end up with smaller standard errors than if one used a fixed effects model, but this problem can be solved by taking the maximum of standard errors from the two approaches. The instability of variance estimates when the number of sites is small

[1]Personal communication. Former Team Leader Biometrics, CDER, FDA, currently heading a biostatistical consulting group, Pacific Northwest Statistical Consulting.

can lead one to wonder how low one can go (Bell, et al, 2014). This is topic of the balance of this paper.

## 3. The Limits of Design-Based and Cluster-Robust Inference

The third standard anti-random slopes argument, as discussed above, concerns the performance of analytic software when the number of clusters is small. This concern is related to a small literature on the limits of design-based survey inference and more general cluster-robust inference as the number of clusters (known in the survey literature as primary sampling units, or PSUs for short) becomes very small. In fact, the primary founder of design-based survey inference, Morris Hansen, in an otherwise fierce defence of the methodology (Hansen, Madow, and Tepping, 1983) expressed uncertainty about the validity of these methods when the survey has a small number of PSUs:

> *… in most practical problems the application of probability-sampling theory is essentially assumption-free only if the sample is acceptably large. When surveys use relatively small samples, the samples may be too small for the application of the theory to be essentially assumption-free. Under such conditions, model-dependent inferences may be preferable. Much research needs to be undertaken on the applicability of asymptotic theory to relatively small samples… No general rules can be given for what is a large enough sample. … Ordinarily, one can reasonably regard samples of less than 25 as small, …*

Following up on this challenge years later, Bell and McCaffrey (2002) found bias in various cluster-adjusted variance estimator for regression coefficients on multi-stage samples when the number of clusters. They also proposed an estimator to reduce this bias. Unfortunately, the estimator is very complex, and, to date, none of the major analysis software systems have added it as an option. However, their Theorem 1 shows that this bias in the variance estimator depends on the intraclass correlation (ICC) of the covariate. If the ICC is low, then the bias should also be low. In fact, if $\mathbf{X}_i'\mathbf{X}_i\left(\mathbf{X}'\mathbf{X}\right)^{-1}\ell$ is constant across the clusters, then the bias in the estimate of $\mathrm{Var}\left(\ell'\hat{\beta}\right)$ is zero, where $\ell'\hat{\beta}$ is an arbitrary linear combination of estimated regression coefficients. If the regressor in question is randomization-to-treatment status, and both the total sample size and the treatment sample vary little across clusters, then this condition should be approximately met. (If no covariates are used, and both the total sample size and the treatment sample size are constant across clusters, then the condition is exactly met.) Supporting this observation, Bell and McCaffrey have simulations with as few as 20 clusters where the bias in the variance estimate is nearly zero for a regressor with ICC near zero.

In a completely independent strand of related research, Bell et al. (2014) found encouraging results for inferences about regressors in random slopes models with a mixture of runs with 10, 20, and 30 clusters. However, they did not report results separately by the number of clusters, so the paper does not really plumb the depths of how low one can go. Based on the encouraging results from these two unrelated papers, I simulated performance of variance estimators for as few as 3 or 5 clusters as is discussed in the next section.

## 4. Simulation Framework

I simulated person level responses as

$$y_{ij} = \pi_{0j} + \pi_{1j} T_{21ij} + \pi_{2j} T_{22ij} + \sum_{m=1}^{4} \gamma_{mj} X_{ijm} + \varepsilon_{ij},\ j = 1,...,10;\ i = 1,...,n_j,$$

$$X_{ijm} = \mu_{mj} + u_{ijm}$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_j^2\right)(iid)$$

$$\sigma_j^2 = \max\left\{0.05, 1 - \sum_{m=1}^{4} \gamma_{mj}^2\right\}$$

$$u_{ijm} \sim N\left(0, 1 - \rho_m\right)(iid)$$

where:

$\pi_{0j}$ = the site-specific component of the mean outcome of beneficiaries in site *j* absent the T21 and the T22 treatments,

$\pi_{1j}$ = mean impact of the T21 treatment on beneficiary outcomes in site *j*,

$\pi_{2j}$ = mean impact of the T22 treatment on beneficiary outcomes in site *j*,

$T_{21ij}$ = an indicator of whether beneficiary *i* in site *j* has been randomized into the T21 group (= 1 if so, = 0 if not),

$T_{22ij}$ = an indicator of whether beneficiary *i* in site *j* has been randomized into the T22 group (= 1 if so, = 0 if not),

$X_{ijm}$ = a measure of baseline characteristic *m* for individual *i* in site *j*,

$\gamma_{mj}$ = regression coefficient of $X_{ijm}$ in site *j*,

$\varepsilon_{ij}$ = person level error independent of all other terms in the model and across sites,

$\sigma_j^2$ is the residual variance of *y* in site *j*,

$\mu_{mj}$ = mean of $X_{ijm}$ in site *j*,

$u_{ijm}$ = person level error in covariate $X_{ijm}$,

$\rho_m$ = intraclass correlation in in covariate $X_{ijm}$,

$n_j$ = sample size at site *j* (across arms).

I simulated site parameters as

$$n_j \sim N\left(100, 100\right)(iid)$$

$$\pi_{0j} = \beta_{00} + \nu_{0j},\ j = 1,...,10$$
$$\pi_{1j} = \beta_{10} + \nu_{1j},\ j = 1,...,10$$
$$\pi_{2j} = \beta_{20} + \nu_{2j},\ j = 1,...,10$$
$$\nu_{0j} = \sqrt{\tau_{00} - |\tau_{01}| - |\tau_{02}|}\,\xi_{0j} + \sqrt{|\tau_{01}|}\,\xi_{01j} + \sqrt{|\tau_{02}|}\,\xi_{02j}$$
$$\nu_{1j} = \sqrt{\frac{\tau_{11} - |\tau_{01}| - |\tau_{12}|}{\ell/(\ell-2)}}\,\xi_{1j} + \frac{|\tau_{01}|}{\tau_{01}}\sqrt{|\tau_{01}|}\,\xi_{01j} + \sqrt{|\tau_{12}|}\,\xi_{12j}$$
$$\nu_{2j} = \sqrt{\frac{\tau_{22} - |\tau_{02}| - |\tau_{12}|}{\ell/(\ell-2)}}\,\xi_{2j} + \frac{|\tau_{02}|}{\tau_{02}}\sqrt{|\tau_{02}|}\,\xi_{02j} + \frac{|\tau_{12}|}{\tau_{12}}\sqrt{|\tau_{12}|}\,\xi_{12j}$$
$$\xi_{0j}, \xi_{01j}, \xi_{02j}, \xi_{12j} \sim \mathrm{N}(0,1)\,(iid)$$
$$\xi_{1j}, \xi_{2j} \sim t(\ell)\,(iid)$$
$$\mu_{mj} \sim N\left(0, \rho_m\right)\,(iid),$$
$$\gamma_{mj} = \theta_{m0} + \sqrt{c_m}\,\psi_{mj},\ j = 1,...,10,$$
$$\psi_{1j}, \psi_{2j}, \psi_{3j}, \psi_{4j} \sim \mathrm{N}(0,1)\,(iid),$$

where:

$\beta_{00}$ = the grand mean of outcome *y* across the 10 sites absent either the T21 or the T22 treatment,

$\beta_{10}$ = the overall impact of the T21 treatment (versus the no treatment of the C2 group),

$\beta_{20}$ = the overall impact of the T22 treatment (versus the no treatment of the C2 group),

$\theta_{m0}$ = average regression coefficient across sites for $X_{ijm}$

$c_m$ = variance of regression coefficient for $X_{ijm}$ across sites,

$\nu_{0j}$ = intercept offset for site *j* under control conditions (this term is often also called the random intercept),

$\nu_{1j}$ = difference between the effect of T21 in site *j* and the average effect of T21 (this term could also be called the random effect or slope of T21),

$\nu_{2j}$ = difference between the effect of T22 in site *j* and the average effect of T22 (this term could also be called the random effect or slope of T22),

$\ell$ = degrees of freedom for *t*-distribution used to generate nonnormal variation in treatment effects across sites, and

$\xi_{0j}, \xi_{01j}, \xi_{02j}, \xi_{12j}, \xi_{1j}, \xi_{2j}, \psi_{1j}, \psi_{2j}, \psi_{3j}, \psi_{4j}$ = latent independent random variables used in the background to create the desired

variance-covariance structure for random intercepts, random treatment effects, and random slopes.

With this setup, the covariance matrix for $\begin{bmatrix} v_{0j} & v_{1j} & v_{2j} \end{bmatrix}'$ is

$$\begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{bmatrix},$$

but the distributions of the random effects of T21 and T22 have heavier tales than the distribution of random intercepts in the control arm.

I developed the formula for $\sigma_j^2$ above so that the unconditional outcome variance in every site would be one. This makes it easier to interpret all other parameters as effect sizes. Another advantage is that this made the conditional level-one errors heteroscedastic. This was useful because some concerns have been raised about results from survey-sensitive regression software when level-one residual errors are heteroscedastic. I achieved this be letting the explanatory power of the covariates vary across sites. Using this general structure, I created several scenarios for simulation. Before discussing them and the differences among them, I first mention all the parameter values that were constant across the scenarios.

## 4.1 Scenarios

I created 6 scenarios corresponding to two different structures of effects across sites and three cluster counts. They are sketched in bullet form below and then closely defined in Tables 1 and 2. I generated 5000 replicates of each scenario.

- Structure 1. Strong local variation in treatment effects with no overall effect of either T21 or T22. Balanced variation in local treatment effects for T21 and T22.
- Structure 2. Strong local variation in treatment effect of T21 but minimal local variation in treatment effect of T22. No overall effect of either T21 or T22.

Table 1. Common Parameter Settings

| Concept | Symbol | Value |
|---|---|---|
| Average sample size per site | | 100 |
| Standard deviation of site-level sample sizes | | 10 |
| Grand mean under control conditions | $\beta_{00}$ | |
| Proportion randomized to T21 at each site | | 0.33 |
| Proportion randomized to T22 at each site | | 0.33 |
| Intraclass correlation in covariate #1 | $\rho_1$ | 0.001 |
| Intraclass correlation in covariate #2 | $\rho_2$ | 0.05 |
| Intraclass correlation in covariate #3 | $\rho_3$ | 0.10 |
| Intraclass correlation in covariate #4 | $\rho_4$ | 0.26 |
| Variance in the site-specific regression coefficients of covariate #1 | $c_1$ | 0.05 |
| Variance in the site-specific regression coefficients of covariate #2 | $c_2$ | 0.05 |
| Variance in the site-specific regression coefficients of covariate #3 | $c_3$ | 0.05 |
| Variance in the site-specific regression coefficients of covariate #4 | $c_4$ | 0.05 |
| Degrees of freedom for *t*-distribution used to generate nonnormal variation in treatment effects across sites | $\ell$ | 5 |
| Average regression coefficient across sites for *y* on covariate #1 | $\theta_{10}$ | 0.20 |
| Average regression coefficient across sites for *y* on covariate #1 | $\theta_{20}$ | 0.20 |
| Average regression coefficient across sites for *y* on covariate #1 | $\theta_{30}$ | 0.20 |
| Average regression coefficient across sites for *y* on covariate #1 | $\theta_{40}$ | 0.20 |

Table 2. Parameter Settings for Specific Scenarios

| | | Value Under Structure # | |
|---|---|---|---|
| Concept | Symbol | 1 | 2 |
| Average effect of T21 | $\beta_{10}$ | 0 | 0 |
| Average effect of T22 | $\beta_{20}$ | 0 | 0 |
| Variance of site intercepts | $\tau_{00}$ | 0.10 | 0.10 |
| Variance of site-level effect of T21 | $\tau_{11}$ | 0.20 | 0.20 |
| Variance of site-level effect of T22 | $\tau_{22}$ | 0.20 | 0.01 |
| Covariance of site intercept with site-level effect of T21 | $\tau_{01}$ | 0.02 | 0.02 |
| Covariance of site intercept with site-level effect of T22 | $\tau_{02}$ | 0.02 | 0 |
| Covariance of site-level effects of T21 and T22 | $\tau_{12}$ | 0.02 | 0 |

## 4.2 Analysis Methods

For each generated population under each scenario, I used two different analytic approaches. Method 1 uses survey-sensitive regression software that corrects for clustering. Pseudocode is as follows:

```
proc surveyreg data=pop;
 cluster Site;
 class Treat;
     model y= Treat x1 x2 x3 x4/ solution;
run;
```

Method 2 involves a random slopes model fit with restricted maximum likelihood and the Kenward-Roger degrees of freedom estimation. Pseudocode is as follows:

```
proc mixed data=pop method=reml;
 class treat site;
 model y=treat x1 x2 x3 x4/solution ddfm=kr;
 random intercept t1 t2/subject=site;
run;
```

## 4.3 Performance Standards

For each generated population under each scenario and each different analytic approach, I calculated several performance statistics:

- Whether estimated 95% confidence interval for the average effect of T21 includes the truth.
- Ditto for the average effect of T22, as well as the average regression coefficients of the four covariates
- Whether the estimated standard error for T21 was only half or less of the true standard error for the estimation process
- Ditto for T22
- Whether the estimated standard error for T21 was half or less of the estimated standard error for T22
- Whether the estimated standard error for T22 was half or less of the estimated standard error for T21

The rationale for the first two standards is clear, but some additional motivation for the others may be helpful. Standard errors being way off occasionally does not invalidate the frequentist properties of the confidence intervals and related hypothesis testing but do create problems for analysts who prefer likelihood-based methods and therefore condition on the observed data. Mismatched estimates of standard errors for T21 and T22 are particularly eye-catching to such analysts.

Joseph Sedransk in his 2007 Hansen Memorial Lecture reminded the audience of a clever toy problem by Buehler (1959). In his toy problem, Buehler posited two players and a referee. Based on a sample, Peter forms a confidence interval for the mean of a normal population with unknown mean and variance, and Paul wagers whether it is true or not in this specific instance. The referee then informs Paul whether he has won. If Paul's strategy is to bet against the confidence interval whenever the sample estimate of the

standard error is smaller than some arbitrary constant *a*, then over a long run of repetitions of the process (draw a sample, Peter forms confidence interval, Paul decides whether to accept it, referee decides the issue), Paul's expected gain is nonnegative. Moreover, if he has any information about the true standard error, then Paul ought to be able to pick a value for *a* such that his expected gain is strictly positive. For the problem at hand of making inferences about an average treatment effect, it seems plausible that knowing that the estimated standard error is either smaller than what one would obtain from a fixed effects model or smaller than the standard error on the average effect of a similar treatment is the sort of information that Paul could use to craft a winning betting strategy.

With 5000 simulations, the standard errors on coverage rates are on the order of 0.31 percentage points, so estimated coverage rates above 94.3 percent indicate good performance.

## 5. Results

Table 3 shows that design-based regression software is valid for testing for experimental effects with as few as 3 clusters if the between-cluster variance in treatment effects is non-negligible, as in structure #1 for treatment variations and in structure #2 for treatment T21. If the between-cluster variance is negligible, as in structure #2 for treatment T22, then design-based regression software is a little liberal with three clusters but delivers valid inferences with as few as five clusters. Inferences are not valid for the covariates with higher intraclass correlation, but since they are nuisance parameters in the analysis of experiments, this is not important.

However, Table 3 also shows that with fewer than 10 clusters, the estimated standard errors are wildly unstable. With 3 clusters, estimated standard errors are too small by more than a factor of about a quarter of the time. This is not surprising with so few degrees of freedom. What I found more surprising is that the hypothesis testing procedure is still valid even though p-values based on estimated standard errors might often give misleading evidence about the strength of the evidence against the null.

The fact that the standard error in particular sample was poorly estimated would be largely invisible if there was a single treatment, but with three-arm experiments, it can be more obvious. With three PSUs one estimated standard error will be more than twice as large as the other estimated standard error about 40 percent of the time if the random effect variances are balanced (as in structure #1) and about 57 percent of the time if the random-effect variances are unbalanced (as in structure #2).

Table 4 shows broadly similar results for a random slopes model fit with REML and Kenward-Roger options in SAS PROC MIXED. The random slopes model appears to need at least 5 clusters to provide valid inference for treatment effects, a little more than design-based regression but not much worse. In exchange, dramatic underestimation of variances is less common with the random slopes model – particularly when the true between-cluster in treatment effect is small, as for treatment T22 under structure #2. Also, dramatic mismatches between estimated standard errors for the two treatment arms in a three-arm trial are much less common with random slopes regression modeling than with design-based regression. I think this might be due to the feature in SAS PROC MIXED that forces variance components with negative likelihood-based estimates to be

slightly positive. Also, the inferences for the other covariates are far too liberal, but this could be largely repaired by adding random slopes for them to the model.

Table 3. Performance of Survey Software with Cluster-Adjusted Standard Errors

| Population Structure | 1 | 1 | 1 | 2 | 2 | 2 |
|---|---|---|---|---|---|---|
| Number of clusters | 3 | 5 | 10 | 3 | 5 | 10 |
| Performance Measure | | | | | | |
| CI covers $\beta_{10}$ (%) | 95.3 | 95.1 | 95.7 | 95.3 | 95.1 | 95.5 |
| CI covers $\beta_{20}$ (%) | 95.4 | 95.6 | 95.0 | **93.7** | 94.5 | 94.9 |
| CI covers $\theta_{10}$ (%) | 94.6 | 94.9 | 94.6 | 94.7 | 94.4 | 94.7 |
| CI covers $\theta_{20}$ (%) | **93.9** | **94.0** | 94.6 | **93.7** | **93.6** | 95.0 |
| CI covers $\theta_{30}$ (%) | **93.9** | **94.1** | **93.8** | **94.1** | **94.0** | 94.4 |
| CI covers $\theta_{40}$ (%) | **92.9** | **92.6** | **92.6** | **92.8** | **92.7** | **93.2** |
| Est. StdErr on $\beta_{10}$ <half true StdErr on $\beta_{10}$ (%) | 26.3 | 12.6 | 2.2 | 26.9 | 13.4 | 2.0 |
| Est. StdErr on $\beta_{20}$ <half true StdErr on $\beta_{20}$ (%) | 26.0 | 12.4 | 2.4 | 24.6 | 11.4 | 1.8 |
| Est. StdErr on $\beta_{10}$ <half est. StdErr on $\beta_{20}$ (%) | 18.9 | 12.0 | 4.8 | 6.5 | 1.2 | 0 |
| Est. StdErr on $\beta_{20}$ <half est. StdErr on $\beta_{10}$ (%) | 20.8 | 11.9 | 4.3 | 50.3 | 49.0 | 47.4 |

Table 4. Performance of Multi-level model with Random Slopes for Treatment Effects

| Population Structure | 1 | 1 | 1 | 2 | 2 | 2 |
|---|---|---|---|---|---|---|
| Number of clusters | 3 | 5 | 10 | 3 | 5 | 10 |
| Performance Measure | | | | | | |
| CI covers $\beta_{10}$ (%) | **92.4** | 95.1 | 95.8 | **92.7** | 95.0 | 95.5 |
| CI covers $\beta_{20}$ (%) | **92.6** | 95.3 | 95.3 | 96.3 | 95.6 | 96.1 |
| CI covers $\theta_{10}$ (%) | **53.8** | **53.6** | **54.0** | **52.5** | **52.9** | **53.6** |
| CI covers $\theta_{20}$ (%) | **54.1** | **53.6** | **54.4** | **52.7** | **54.9** | **54.4** |
| CI covers $\theta_{30}$ (%) | **55.2** | **56.6** | **54.6** | **54.3** | **55.2** | **55.2** |
| CI covers $\theta_{40}$ (%) | **58.7** | **58.9** | **58.0** | **59.2** | **57.8** | **59.1** |
| Est. StdErr on $\beta_{10}$ <half true StdErr on $\beta_{10}$ (%) | 23.0 | 10.7 | 1.6 | 23.0 | 11.1 | 1.1 |
| Est. StdErr on $\beta_{20}$ <half true StdErr on $\beta_{20}$ (%) | 22.9 | 11.4 | 1.7 | 0 | 0 | 0 |
| Est. StdErr on $\beta_{10}$ <half est. StdErr on $\beta_{20}$ (%) | 14.5 | 9.9 | 4.2 | 1.0 | 0.1 | 0 |
| Est. StdErr on $\beta_{20}$ <half est. StdErr on $\beta_{10}$ (%) | 16.1 | 10.1 | 3.5 | 36.4 | 38.8 | 43.6 |

## 6. Discussion

I have demonstrated that valid external inference can be obtained from ensembles of experiments containing as few as three sites. Survey-sensitive software protects against liberal tests of treatment effects better than multi-level modeling software, but the difference is slight, and multi-level modeling is less likely to yield wild standard error estimates. It should be noted that although my simulation did not involve weights, if the

sites have been selected with a probability-sampling procedure, then it is easier to find appropriate survey-sensitive software that will accommodate both clustering and weights than it is to find appropriate multi-level modeling software. The MIXED procedure in SAS specifically does not use weights in an appropriate manner (Carle, 2009).

The other commonly advanced reasons for ignoring clustering when analyzing small ensembles still hold. Namely, power for Q5 is often much lower than power for Q1, and, unless the sites are randomly selected or one is lucky to have a balanced set of sites, external validity will not be fully attained. One will only be a step closer. If there is a good prospect for following up encouraging results (a finding of "efficacy" for a particular set of actors) with a large study with probability sampling of sites, then a good argument can be made for being satisfied with internal validity. However, if a study is likely to be the last formal evaluation of an intervention prior to widespread implementation, then I think the analyst should not be satisfied with internal validity. Instead, the analyst should reflect observed variation in effects in the formal finding about the value of the intervention. In this context, I assert that an imperfect evaluation of effectiveness is better than a perfect evaluation of efficacy. At a minimum, one can present both sets of standard errors and counsel users on which they should use for their personal inferences.

Finally, to forestall criticisms from Bayesians about prima facie plausibility, it might make sense to bound the standard errors from below by the fixed-effect solution. This will make the tests even more conservative unconditionally, but will reduce the number of times that results do not make sense conditioned on other information. Bayesians might want to go farther in the case of a three-arm study to average the standard errors for the two treatments, and then use the average to form confidence intervals for both treatment variations, but I did not explore this option.

## References

Angrist, J. D. (2004), "American education research changes tack," *Oxford Review of Economic Policy*, 20, 198-212.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., and Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 10, 1-11.

Bell, R. M. and McCaffrey, D. F. (2002), "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169-181.

Bell, S. H., Olsen, R. B., Orr, L. L., and Stuart, E. A. (2011), "Estimates of bias when impact evaluations select sites purposively," Presentation at the annual conference of the Association for Public Policy Analysis and Management, November 2011.

Buehler, R. J. (1959), "Some validity criteria for statistical inferences," *Annals of Mathematical Statistics*, 30, 845-863.

Carle, A. C. (2009), "Fitting multilevel models in complex survey data with design weights: Recommendations," *BMC Medical Research Methodology*, 9, 49.

Cook, T. D. (2007), "Randomized experiments in Education: Assessing the objections to doing them," *Economics of Innovation and New Technology*, 16, 331-355.

Crits-Christoph, P., Tu, X., and Gallop, R. (2003), "Therapists as fixed versus random effects – Some statistical and conceptual issues: A comment on Siemer and Joorman (2003), *Psychological Methods*, 5, 425-433.

Dragalin, V., Fedorov, V., Jones, B., and Rockhold, F. (2001), "Estimation of the combined response to treatment in multicenter trials." *Journal of Biopharmaceutical Statistics*, 11, 75—295.

Fedorov, V. and Jones, B (2005), "The design of multicentre trials," *Statistical Methods in Medical Research*, 14, 205—248.

Hansen, MH, Madow, W.G., and Tepping, BJ (1983), "An evaluation of model-dependent and probability-sampling inferences in sample surveys," *Journal of the American Statistical Association*, **78**, 776-793.

Feaster, D. J., Miklulich-Gilbertson, S., and Brincks, A. M. (2011), "Modeling site effects in the design and analysis of multisite trials," *American Journal of Drug and Alcohol Abuse*, 37, 383-391.

James, D. M. (2010) "Better: Dr. Deming consults on quality for Sir William Osler," Deming Lecture at the 2010 Joint Statistical Meetings.

Judkins, D. R. (2009). "The Hype and Futility of Fidelity Measurement," Presentation at the Annual Meeting of the American Evaluation Association in Orlando.

Kempthorne, O. and Doerfler, T. E. (1959), "The behavior of some significance tests under randomization, *Biometrika*, 56, 231-248.

Kish, L. and Frankel, M. (1974), "Inference for complex samples," *Journal of the Royal Statistical Society B*, 36, 1-37.

Lingsma, H. F., Roozenbeek, B., Perel, P., Roberts, I., Maas, A. I. R., and Steyerberg, E. W. (2011), "Between-centre difference and treatment effects in randomized controlled trials: A case study in traumatic brain injury," *Trials*, 12, 201.

Martindale, C. (1978), "The therapist-as-fixed-effect fallacy in psychotherapy research, *Journal of Consulting & Clinical Psychology*, 46, 1526-1530.

Michael, R. T. and O'Muircheartaigh, C. A, (2008), "Design priorities and disciplinary perspectives: The case of the US National Children's Study, *Journal of the Royal Statistical Society A*, 171, 465-480.

Puma, M., Bell, S., Cook, R., and Heid, C. (2010), *Head Start Impact Study Final Report,* Washington, DC: Administration for Children and Families.

Schochet, P. Z. (2009), "Statistical power for random assignment evaluations of education programs," *Journal of Education and Behavioral Statistics*, 33, 62-87.

Senn, S. J. (2007), *Statistical Issues in Drug Development*, 2nd ed., Wiley: Hoboken.

Walwyn, R. and Roberts, C. (2010), "Therapist variation within randomised trials of psychotherapy: Implications for precision, internal and external validity," *Statistical Methods in Medical Research*, 19, 291-315.