Robust Covariate Control in Cluster-Randomized Trials

Jiaquan Fan and David Judkins
Westat, 1650 Research Blvd., Rockville, MD 20850

1. Introduction and Background

In cluster-randomized trials, it is not necessary to control on either group- or subject-level covariates for valid inference, but controlling on such covariates can increase statistical power (Donner and Klar, 2000). The price for achieving such power is that one must use more sophisticated software that usually entails making parametric assumptions. We had a specific trial from which we believed the data structure would violate a number of common parametric assumptions but where we felt that subject-level controls offered substantial power improvement. We therefore undertook a simulation study to study the robustness of some standard options for covariate control. We also developed and tested a new semi-parametric method. There does not appear to be much relevant prior research in this area. One exception is Cheong, Fotui and Raudenbush (2001). They have simulated a different range of population structures and analysis procedures. Also, Jenkins, et al, (2006) undertook a simulation study of other forms of robustness in hierarchical linear modeling at about the same time.

The application of interest was a randomized experiment with alternate preschool instructional paradigms, loosely referred to as curricula. There were four alternate curricula and one control curriculum. All five arms were assigned to a recruited sample of 120 Even Start schools. The schools were deeply stratified into 24 blocks, each containing five schools. Within a block, the five schools were then randomly assigned to the five arms. The curricula involved instructional materials, instructional strategies, teacher training, teacher observation, and teacher consultation. Within the schools, parents of age-eligible children were recruited into the study. Measurements were conducted in the spring of 2004, prior to the introduction of the new curricula, and repeated at one-year intervals in 2005 and 2006. Measurements involved formal assessments of pre-literacy, social competency (teacher observation), parent interviews, and video-taping and behavior-coding of staged parent-child interactive reading and toy-play sessions to gauge parenting skills. There was considerable turnover in the student-body each year, but there is some overlap of sample across years, and of course, there is considerable organizational and staffing stability. So one set of important covariates involved school-level past performance and child-teacher ratios. Another important set of covariates involved parent socio-economic status, native language, and child demographics (age, race, sex, and disability status). Native language, in particular, has a huge effect on English pre-literacy.

For analysis, we wanted something more powerful than either the very robust Fisher's exact test or the slightly less robust standard two-way ANOVA. However, we wanted something that was robust to unequal student sample sizes per school, school-level nonresponse, deep stratification, heteroscedasticity, non-Gaussian errors and interactions. We therefore developed superpopulations that had the features of interest, generated samples from them, and tested several alternative analysis procedures on them, using type I error rates and statistical power as evaluation criteria.

In section 2, we discuss the superpopulations that we simulated. In section 3, we provide more detail on the analysis methods studied. In section 4, we present results. In section 5, we give some concluding thoughts and ideas for further research.

2. Simulated Superpopulations

Given the application, we built a series of superpopulations with an increasing number of violations of standard models. All shared a common form of having two child-level covariates, one school-level covariate, a random effect at the school level, and student level random error. The project-level covariate was built with a structure similar to the outcome of interest because the way it will be generated in the application is to take the average of students at the school the prior year. All of the superpopulations share a common model structure:

$$y_{ijk} = \mu + \beta_i + \alpha_i + X_{ijk}\theta + Z_{ij}\gamma + u_{ij} + e_{ijk} ,$$

$$Z_{ij} = \beta_i + u_{ij} + \bar{X}_{ij.}\theta + V_{ij}$$

where:

The indices stands respectively for block (i), treatment (j), child (k);
y_{ijk} is the outcome variable;
μ is the overall mean;
β_i is the (fixed) block effect;
α_i is the treatment effect;
X_{ijk} is a vector of two child level covariates (X1=FamilyIncome, X2=MothersEducation);
Z_{ij} is the baseline school-level average of the outcome variable measured on a different set of students prior to the intervention;
u_{ij} is the school level-random effect;
e_{ijk} is a child level random error;
$\bar{X}_{ij.}$ is a vector of school-level averages of child level covariates;
V_{ij} is a normally distributed random error term reflecting the error caused by basing the project-level fixed covariate on a small sample from the prior year rather than a long-run average;
u_{ij}, e_{ijk}, and V_{ij} are mutually independent.

Because the theory is better developed for balanced designs, we introduced imbalance both at the school and the child level. Note that standard multi-level software assumes that all the random errors are normal and homoscedastic. So we developed superpopulations that violated those assumptions. Finally, we allowed interactions. We simulated a series of superpopulations that violated various combinations of these standard assumptions to various degrees while generally keeping the violations within the range that we think might reasonably occur in our application.

Seven different superpopulations with no treatment effect ($\alpha_i = 0$) were generated to test robustness of type 1 error rates. Superpopulation 1 satisfies most of the standard assumptions. The numbering of superpopulations 2 through 7 generally reflects increasing severe violations of standard assumptions:

Superpopulation 1: There are 24 blocks with five schools per block and each school contains exactly 12 children. There is no school-level nonresponse and the school- and child-level random errors are normally distributed. Residual variances are constant with var(u_{ijk}) =12.81 and var(e_{ijk}) =55.26. The block effect is very large with $\beta_i = 2i$.

Superpopulation 2: Same as superpopulation 1 except that the number of children per school is allowed to vary. The number of children per school follows a Poisson distribution with mean 12.

Superpopulation 3: Same as Superpopulation 2 except that there are two schools missing at random (for a total of 118 schools). The missing schools are from different blocks.

Superpopulation 4: Same as Superpopulation 3 except that the school- and child-level random errors have different variances in different blocks:

Block 1 – 6 has u_{ij} and e_{ijk} with variances 3 and 56,
Block 7 – 12 has u_{ij} and e_{ijk} with variances 6 and 42,
Block 13 – 18 has u_{ij} and e_{ijk} with variances 9 and 28,
Block 19 – 24 has u_{ij} and e_{ijk} with variances 12 and 14.

Superpopulation 5: Same as Superpopulation 3 except that the school- and child-level random errors have different variances in different treatment groups:

Treatment 1 has u_{ij} and e_{ijk} with variances 3 and 70,
Treatment 2 has u_{ij} and e_{ijk} with variances 6 and 56,
Treatment 3 has u_{ij} and e_{ijk} with variances 9 and 42,
Treatment 4 has u_{ij} and e_{ijk} with variances 12 and 28.
Control has u_{ij} and e_{ijk} with variances 15 and 14.

Superpopulation 6: Same as Superpopulation 3 except that school- and child-level random errors have Gamma distributions. u_{ij} has shape parameter α =2 and scale parameter β=0.395. For e_{ijk}, α =3 and β=0.233. Note that in this population, the school-level errors are more seriously non-normal than the student-level errors. Both skew and kurtosis are stronger for the school-level errors.

Superpopulation 7: Same as Superpopulation 4 except that there are treatment group effects for individual blocks but no effect on average. That is, within each single block there are significant differences between the treatment groups, but when schools are aggregated to the treatment level, these differences average out.

Another three superpopulations with treatment effect were generated to compare type II error rates. For each of these superpopulations, all four experimental arms are assumed to be equally effective with the magnitude of 1.5. This number was picked to give power in the range of 0.5 to 0.6, a range where we thought we might see the largest differences in power among the techniques.

Superpopulation 8: Model is the same as Superpopulations 4 except that treatment effect is added.

Superpopulation 9: Same as Superpopulations 5 except that treatment effect is added.

Superpopulation 10: Same as Superpopulation 6 except that treatment effect is added.

The components of variance in the model for the superpopulations are shown in Table 1. Naturally, there positive variance between treatment arms only for superpopulations 8, 9 and 10. All other variance components are constant across superpopulations. Also note that the between-block variance is very large. This was done with the aim of making it large enough to matter.

Table 1. Components of variances

Component	Magnitude
Between block (fixed)	192
Between arm (fixed)	0 or 0.36
Child-level covariates (fixed)	3.4
School-level covariates (fixed)	18
School-level random effect (random)	13
Child level error (random)	55

3. Analysis Methods

The analysis methods we studied included HLM (Raudenbush, et al, 2004), SAS PROC MIXED (SAS Institute Inc., 2006), SUDAAN (Research Triangle Institute, 2001), and a new variant of Koch's nonparametric ANCOVA that we called semi-parametric ANOVA. The applications of HLM and MIXED were straightforward and are thus only briefly described. We give more detail on SUDAAN and the semi-parametric ANOVA.

For PROC MIXED, we used school as subject, block as fixed effects and the restricted maximum likelihood option. An example of the code used is shown below:

```
proc mixed data=population method=REML;
class block schoolid treatment;
model y=treatment block FamilyIncome
MothersEducation Z/ solution ;
random int/ type = un subject = schoolid; run;
```

For HLM, we used a 2 level HLM model with student as the first level and school as the second level. Indicator variables for four of the five treatments and 23 of the 24 blocks were entered as fixed effects in addition to FamilyIncome, MothersEducation, and Z. The estimation method was also restricted maximum likelihood. More details on the HLM code are given in the appendix.

For SUDAAN, the knowledge that SUDAAN gives badly biased variance estimates for domain means where there is only one variance unit selected per domain led us to experiment with several alternative SUDAAN setup options. The first option was to treat the sample as a simple random sample of schools but to include dummy variables for the blocks as fixed effects in a linear model. The second option was to treat the sample as a simple random sample of blocks, where each block contains five schools. The third option was to treat the sample as a stratified simple random sample of schools. This was, of course, the correct design, but given that SUDAAN generally requires two variance units from the same stratum and analysis domain in order to give sensible variance estimates for the analysis domain, we thought that the first two options might show some advantages. Also obviously, dummy effects for blocks were not required for the second and third options since blocks were specified on the NEST statements. Examples for the SUDAAN programs are shown below:

SUDAAN Option 1. Simple Random Sample of Schools with Block as Fixed Effect:

```
proc regress data=population1 design=wr
R=exchange ;
weight _ONE_;
nest _ONE_      schoolid ;
subgroup treatment block;
levels 5  24;
model y=treatment block FamilyIncome
MothersEducation Z; run;
```

SUDAAN Option 2. Simple random sample of blocks:

```
proc regress data= population1 design=wr
R=exchange ;
weight _ONE_;
```

```
        nest _ONE_        block ;
        subgroup treatment;
        levels 5;
        model    y= treatment FamilyIncome
        MothersEducation Z; run;
```

SUDAAN Option 3. Stratified simple random sample of schools:

```
        proc regress data=population1 design=wr
        R=exchange ;
        weight _ONE_;
        nest block          schoolid ;
        subgroup treatment ;
        levels 5;
        model y= treatment FamilyIncome
        MothersEducation Z; run;
```

The new semi-parametric ANOVA was inspired by Rosenbaum (2002) but has much in common with a line of papers mostly by Gary Koch and coauthors (Koch, et al, 1982 and 1998; Stokes, Davis, and Koch, 2000; Lavange, Durham, and Koch, 2005) that was launched by Quade (1967).

The procedure involves three steps. In the first, a parametric model for the outcome is developed. (Linear in our case, but other parametric models could be used.) It is important that the putative causal agent (in this case treatment arm) be omitted from the model. Theoretically, it seemed to us that we should include block in this model, but we seemed to get worse results that way and so left block out of the model. Instead, other school-level and subject-level covariates were allowed into the model. In the second step, the residuals from the parametric model are averaged to the school level. In the third step, a randomization based permutation test is run on the average school-level residuals. Theoretically, we should have used a stratified permutation test, but we used a simple unstratified permutation test. Using WILCOXON option in PROC NPAR1WAY, we did Wilcoxon rank-sum test for constrasts and Kruskal-Wallis test for the overall F-test. Example of program is given below:

```
proc glm data=population1;
model y= FamilyIncome MothersEducation  Z;
output out=R residual=residual; run;

proc summary data= R (keep=schoolid
    treatment residual );
var residual ;
output   out=Raver mean=mresid;
class schoolid      treatment; run;

data Raver; set Raver;
if treatment=1 then contrast11=1;
else if treatment=2 then contrast11=2;
else contrast11=.;
if treatment in (1,2) then contrast22=1;
else if treatment in (3,4) then contrast22=2;
else contrast22=.;
if treatment in (1,2) then contrast21=1;
else if treatment in (3) then contrast21=2;
else contrast21=.;
if treatment in (1,2,3,4) then contrast41=1;
else if treatment in (5) then contrast41=2;
else contrast41=.; run;

proc npar1way data=Raver (where=(_type_=3))
wilcoxon; class treatment; var mresid;

proc npar1way data=Raver (where=(_type_=3))
wilcoxon; class contrast11; var mresid;

proc npar1way data=Raver (where=(_type_=3))
wilcoxon; class contrast21; var mresid;

proc npar1way data=Raver (where=(_type_=3))
wilcoxon; class contrast22; var mresid;

proc npar1way data=Raver (where=(_type_=3))
wilcoxon; class contrast41; var mresid; run;
```

In each analysis method, five tests were conducted. One is the overall F-test for any differences among the five arms. The other four tests are contrasts between treatment groups based on research questions in our application:

Treatment 1 vs. control
Treatment 1 plus 2 vs. control
Treatment 1 plus 2 vs. Treatment 3 plus 4
Sum of treatment 1, 2, 3, 4 vs. control

In the simulation, 1000 populations were generated in each superpopulations and the six analysis methods were used to do the five tests above. For type-1 error simulation, the null hypothesis rejection rate is at nominal size of 0.05. The percentage of the significant tests among the 1000 tests is treated as the simulated type-1 error. For power analysis, the percentage of the significant tests is treated as the simulated power of the test.

4. Simulation Results

The simulation results are shown in Figures 1 through 4. In each of these, the horizontal axis reflects the various superpopulations. Type 1 error rates for the overall treatment effect is shown in Figure 1 (for the seven populations with no treatment effect). There is a separate curve for each of the six analysis methods. The horizontal dotted line stands for the significance threshold for a test of 1000 independent p-values to be significantly larger than 0.05. Figure 2 shows the type-1 error simulation results for contrasts. Since the four contrasts do not differ much, the four tests were pooled together to be represented by a single line for each analysis method. Power for detecting overall treatment effect is shown in Figure 3 for the three populations with treatment effects. Similarly, power levels for the contrasts are shown in Figure 4, which shows only the average of the four contracts for each method.

Note that the results for SAS PROC MIXED and HLM are nearly identical, so results for them will be referred to as simple MIXED/HLM results. Only the semi-parametric procedure preserved type I error rates for every superpopulation (i.e., had a rejection rate under the null hypothesis of no treatment effect within sampling error of the nominal rate). However, MIXED/HLM was surprisingly robust. It only clearly exceeded the nominal type I error rate for one superpopulation. The SUDAAN variants all performed poorly for type I error rates on the overall F-test. The second option performed well for contrasts, but this was not good enough to redeem it in our eyes. The third option (which is the standard SUDAAN setup) did not perform well even for contrasts.

Some would argue that because only the semi-parametric procedure consistently preserved type I error rates, it is the only principled choice, and that any consideration of power levels is irrelevant. However, the single MIXED/HLM violation seemed mild enough to allow us to consider statistical power. With regard to power, MIXED/HLM is a clear winner over the semi-parametric procedure.

Based on these calculations and the fact that semi-parametric procedure is new while the others are well-known, we decided to use either MIXED or HLM. The final decision between them came down to convenience. Although HLM is very difficult to fully integrate into the batch-driven analytic processes of an SAS shop, it does have some nice graphics and so we decided to use it.

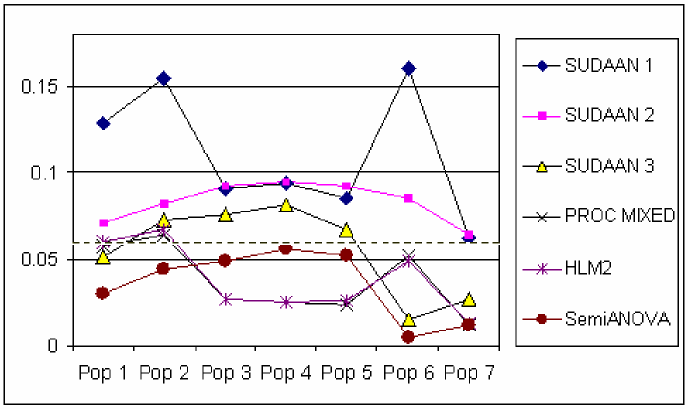

Figure 1. Type I error simulation for the test for overall treatment effect

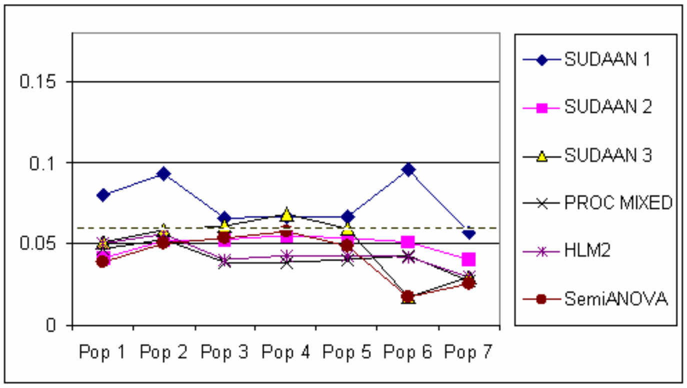

Figure 2. Type I error simulation: Test for contrasts

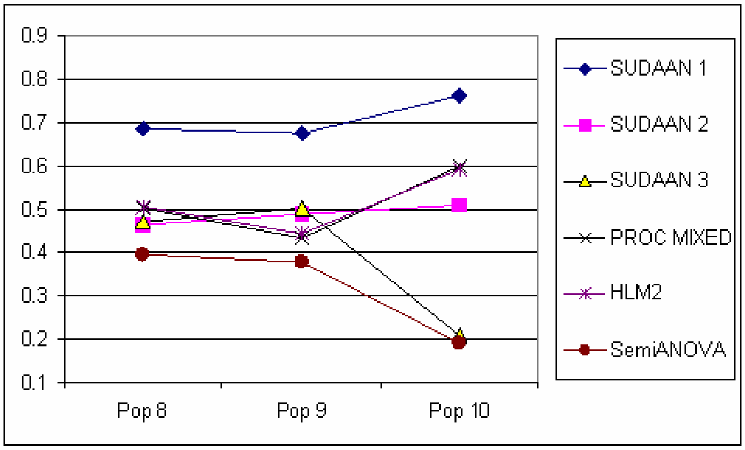

Figure 3. Power simulation: Test for overall treatment effect

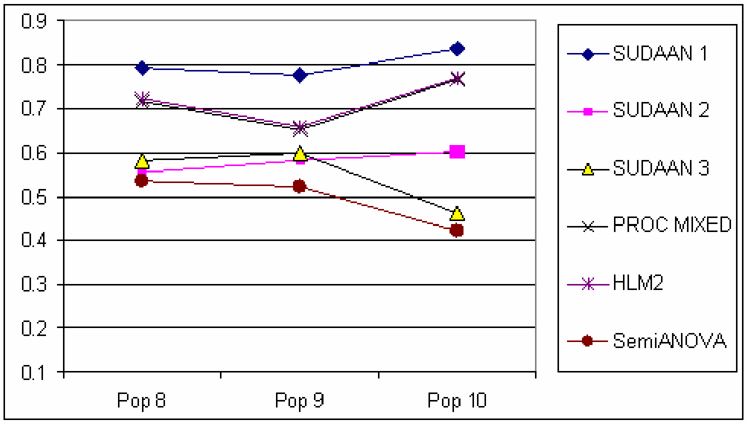

Figure 4. Power simulation: Test for contrast

5. Conclusions and Further Study

We fully expected one of the SUDAAN options to be more robust than MIXED/HLM. The fact that this expectation was not born out is perhaps due to the superpopulation structure. The between-block component is extremely large. It could be that with a more reasonable value for this component, SUDAAN would perform better.

We also did not expect the semi-parametric procedure to surrender as much power as it did. This may also be caused by the very large between-block variance component since we did not find a way to reflect block in the semi-parametric procedure. We tried using block indicators in the parametric model and met with poor results. After the meetings, a reviewer suggested that if we had used a stratified randomization test on the school-averaged residuals, we would have preserved more of the power. That would be an interesting possibility to explore in further research.

References

Cheong, Y.F., Fotui, R.P., and Raudenbush, S.W. (2001). Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of Education and Behavioral Statistics*, 26, 411-429.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Quade, D. (1967). Rank Analysis of Covariance. *Journal of the American Statistical Association*, 62, 1187-1200.

Jenkins, F., Lee, H., Cheah, B. and Leytush, O. (2006). Robustness of Hierarchical Linear Modeling of Complex Survey Data When Higher Levels of Aggregation are Left out of the Model. *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

Koch, G.G., Amara, I.A., Davis, G.W., and Gillings, D.B. (1982). A Review of Some Statistical Methods for Covariance Analysis of Categorical Data, *Biometrics,* 38, 563-595.

Koch, G.G., Tangen, C.M., Jung, J-W, and Amara, I.A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from andomized clinical trials and non-parametric strategies for addressing them, *Statistics in Medicine*, 17, 1863-92.

LaVange, L.M., Durham, T.A. and Koch, G.G. (2005). Randomization-based nonparametrick methods for the analysis of multicentre trials. *Statistical Methods in Medical Research*, 14, 281-301.

Stokes, M.E., Davis, C.E., and Koch, G.G. (2000), *Categorical Data Analysis Using the SAS System*, 2nd edition. Cary, NC: SAS Institute Inc.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.K. and Congdon, R.T., Jr. (2004). *HLM6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Research Triangle Institute (2001). *SUDAAN User's Manual, Release 8.0.* Research Triangle Park, NC: Research Triangle Institute.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies, *Statistical Science*, 286-303.

SAS Institute Inc. 2006. *SAS OnlineDoc® 9.1.3.* Cary, NC: SAS Institute, Inc.

Appendix. HLM Program Code

```
#WHLM CMD FILE FOR population1
nonlin:n
numit:100,y
stopval:0.0000010000
level1:Y=INTRCPT1+INCOME+MGRADE+BASE
LINE+TREAT1+TREAT2+TREAT3+TREAT4+BL
KID1+BLKID2+BLKID3+BLKID4+BLKID5+BLK
ID6+BLKID7+BLKID8+BLKID9+BLKID10+BLKI
D11+BLKID12+BLKID13+BLKID14+BLKID15+B
LKID16+BLKID17+BLKID18+BLKID19+BLKID2
0+BLKID21+BLKID22+BLKID23+ +RANDOM
level2:INTRCPT1=INTRCPT2+random/
level2:INCOME=INTRCPT2/
level2:MGRADE=INTRCPT2/
level2:BASELINE=INTRCPT2/
level2:TREAT1=INTRCPT2/
level2:TREAT2=INTRCPT2/
level2:TREAT3=INTRCPT2/
level2:TREAT4=INTRCPT2/
level2:BLKID2=INTRCPT2/
level2:BLKID3=INTRCPT2/
level2:BLKID4=INTRCPT2/
level2:BLKID5=INTRCPT2/
level2:BLKID6=INTRCPT2/
level2:BLKID7=INTRCPT2/
level2:BLKID8=INTRCPT2/
level2:BLKID9=INTRCPT2/
level2:BLKID10=INTRCPT2/
level2:BLKID11=INTRCPT2/
level2:BLKID12=INTRCPT2/
level2:BLKID13=INTRCPT2/
level2:BLKID14=INTRCPT2/
level2:BLKID15=INTRCPT2/
level2:BLKID16=INTRCPT2/
level2:BLKID17=INTRCPT2/
level2:BLKID18=INTRCPT2/
level2:BLKID19=INTRCPT2/
level2:BLKID20=INTRCPT2/
level2:BLKID21=INTRCPT2/
level2:BLKID22=INTRCPT2/
level2:BLKID23=INTRCPT2/
level2:BLKID24=INTRCPT2/
fixtau:3
lev1ols:10
accel:5
level1weight:none
level2weight:none
varianceknown:none
hypoth:n
resfil1:n
resfil2:n
homvar:n
constrain:N
heterol1var:n
graphgammas:grapheq.geq
lvr:n
title:no title
output:population1.t
```