Empirical Education

RESEARCH REPORT

Matched-Pairs Designs and Standard Errors of Impact Estimates:

Lessons From 10 Experiments

Andrew P. Jaciw
Xin Wei
Boya Ma
Denis Newman
Empirical Education Inc.

March 2008

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

The research was funded by a grant (#R305E040031) to Empirical Education Inc. from the U.S. Department of Education. The purpose of this grant is to improve our ability to conduct small scale experiments to assist local decision-makers. The U.S. Department of Education is not responsible for the content of this report.

This report was presented as a paper discussion at the Annual American Education Research Association conference in March 2008 (Division D-Measurement and Research Methodology, Section 2: Quantitative Methods and Statistical Theory).

About Empirical Education Inc.

Empirical Education Inc. was founded to help K–12 school districts, publishers, and the educational R&D community assess new or proposed instructional programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

© 2008 by Empirical Education Inc. All rights reserved.

Abstract

This study investigated how effective a matched-pairs strategy is at reducing the standard errors of estimates of experimental impacts and whether it continues to be effective after modeling pretest and fixed effects for upper-level units when the number of randomized clusters is small. We found that we benefit from a matched-pairs design even if we have a small number of clusters when the intraclass correlation (ICC) is large. For some of the cases, modeling pairs has great value even after controlling for pretest and fixed effects for schools. It decreases the standard error of the treatment effect estimate, leads to a smaller ICC, and, in some cases, improves model fit. Employing a matched-pairs design and modeling pairs lead to a useful gain in precision.

Introduction

The randomized control trial (RCT) is an important method for evaluating the effectiveness of school interventions. By randomizing subjects to treatment or control, we are able to obtain unbiased impact estimates. In school settings it is usually impractical to randomize lower-level units, such as students, to conditions. Instead, we use cluster randomized trials (CRTs), in which upper-level units, such as teachers, are randomly assigned to treatment or control. A CRT is an RCT and has the same benefits; however, because the unit of randomization is the upper-level unit, the effective sample size is the number of upper-level units. This has implications for statistical power and experimental design. If we are randomizing teachers but looking at outcomes for students, then the number of teachers becomes a strong determinant of the standard error of the impact estimate. Because locally-conducted experiments sometimes either don't have or cannot afford large numbers of teachers, strategies such as randomizing units within matched pairs become critical. This work, part of a larger research program on the use of experiments to inform local school district decisions, examines the effectiveness of this strategy for use with small-scale trials. The larger research program, funded by grant #R305E040031 from the U.S. Department of Education, focused on methods for obtaining as precise estimates of program effectiveness as possible in situations where the researcher has to work with available samples. The primary goal was to answer questions about local program effectiveness as identified by districts, as opposed to recruiting large samples across districts that yield less relevant information for any one site. The problem investigated in this paper originated from this more general effort.

The potential effectiveness of using a matched-pairs design has implications beyond the first concern to increase precision. In the experiments we consider in this study, teachers were actively involved in selecting the criteria by which they were paired up. They often made holistic and heuristic judgments about what factors affect performance and the order in which things matter. It is hard to say what the gains in precision are from these somewhat idiosyncratic ways of forming pairs; in particular, it would be hard to mathematically formulate or simulate this process because it is so non-uniform. (Normally, discussions of gains from blocking assume pairs are formed by taking adjacent units along some dimension that is assumed to be predictive of the outcome. In the experiments considered here, it is hard to determine what this composite dimension would be.) Nonetheless, we would like to know the net effect on precision of the pairing strategies used here.

Teacher participation in establishing pairs has the benefit of involving teachers in the experimental process. The hands-on approach is both educational and may promote buy-in. However, it may also be unnecessary, burdensome, and costly, especially if it involves teachers in a process that fails to produce expected gains in precision. This work examines whether pairing increases precision in situations where experiments are small, where pairing is performed using multiple criteria in idiosyncratic but outwardly valid ways, and where teachers are directly involved in deciding what criteria to use in pairing up.

Standard Errors of Impact Estimate from CRTs

We start by considering the standard error for the impact estimate in a two-level experiment with randomization at level two. The formula motivates how we decide whether matching helps. The formula for the standard error of the estimate of the treatment effect (SE) from a CRT is a generalization of the formula for a one-level randomized trial, which allows for the fact that the

variance in the outcome and sample size exist at more than one level. Several authors give the expression for the standard error in two- and three-level experiments (Bloom, 2005; Bloom, Bos & Lee, 1999; Moerbeek, van Breukelen & Berger, 2000; Raudenbush, 1997.)

In a two-level design, the SE for the impact depends on both the variance in outcomes among upper-level units (e.g., teachers) and the variance among lower-level units (e.g., students) within upper-level units. The SE is usually highly dependent on the former of these. Mathematically, for a balanced design, the standard error for the treatment impact is proportional to the following quantity:

$$\sqrt{\frac{\tau^2}{J} + \frac{\sigma^2}{nJ}}$$

- J is the total number of upper-level units
- n is the number of lower-level units per upper-level unit
- Tau-squared is the between-teacher variance in average outcomes (the variance component for teachers)
- Sigma-squared is the variance within teachers (the variance component for students)

We note that the ratio of tau-squared to the total variance is called the intraclass correlation (ICC) (Hedges and Hedberg, 2006). The standard error increases as the ICC increases. This quantity is important to the current work because one way to assess whether matching works is to observe how the ratio of tau-squared to either the sum of tau-squared plus sigma-squared or the total variance (i.e., the ICC) changes as we introduce new effects in our models.

Approaches to Reducing the SE of Impact Estimates from CRTs

Several approaches are typically used to lower the standard error. The general goal is to shrink the variance components by estimating additional effects that account for the variability. For these methods to be successful, they have to more than offset the increase in the standard error that results from degrees of freedom being lost in order to estimate the additional effects (Raudenbush, Martinez & Spybrook, 2007). This tradeoff is especially problematic if there are relatively few randomization units, because there are few degrees of freedom to spare, as is the case in the experiments that we consider in this work.

Next we review three commonly used strategies for lowering the standard error.

Modeling Covariates

The inclusion of covariates, especially a pretest, is a commonly used way to increase precision. With two-level designs, the individual student pretest scores or the teacher average of student scores can be used (Raudenbush, 1997; Bloom, Bos and Lee, 1999). Power calculations in two-level designs often assume that the latter of these is being used (Raudenbush, Spybrook, Liu, & Congdon, 2006). Bloom, Hayes and Black (2005) empirically show the effectiveness of modeling the pretest at reducing the variance in the outcome.

Modeling Fixed Effects

A second way to stratify the analysis is to block on upper-level units such as schools or districts (Schochet, 2005). This essentially removes the between-block variance from the uncertainty contributing to the SE of the impact estimate. For instance, if we model schools, then the variance component for teachers will reflect variation among teachers within schools rather than within and between schools.

Modeling Matched-Pairs

This is a form of blocking. In a matched-pairs design, units of randomization are paired based on one or more criteria and randomization is performed within pairs (Bloom, 2005). The goal is to pair similar units together because it is only the variation between units within pairs (and that is not due

to treatment) that contributes to the SE for the impact estimate; the between-pair variance is eliminated from the equation. This form of blocking can be especially costly because almost half the degrees of freedom that could be used to estimate the impact are used to estimate pair effects.

Bloom (2005) and Raudenbush, Martinez, and Spybrook (2007) show that a matched-pairs strategy can increase power in a CRT if the pairs of clusters are well matched. Also, matching will not be effective if the ICC is very small because there is little variation between groups to be removed by matching. Raudenbush, Martinez, and Spybrook (2007) set a minimum ICC of .05 for matching to be effective. An additional criterion for determining whether matching will be effective involves looking at how similar the members of pairs are on specific outcomes (Donner, Taljaard, & Klar, 2007; Raudenbush, Martinez, & Spybrook, 2007).

Naive pair correlations and adjusted pair correlations are two statistics that are commonly used to check whether matching will be effective. The naive pair correlation is the correlation between group means within pairs. The adjusted pair correlation is the fraction of variance in "true" group means that is explained by matching (Raudenbush, Martinez, & Spybrook, 2007).

In the current work, we will consider how well the third of these strategies, pairing, works both in combination with and separately from the first two strategies.

An Approach to Measuring the Effectiveness of Matching

We considered several ways of assessing whether pairing worked. One approach is to determine whether modeling pairs results in an improvement in model fit. A drawback of this approach is that modeling pairs may absorb much of the variance between units at the level of randomization (i.e., it may shift the variance to a level above the level of randomization). Although according to the formula given above, this will reduce the standard error, the reduction may not be accompanied by an improvement in fit. Also, improvement in fit may result from level-one variance (e.g., due to student variation within teachers) going down, but a reduction in this variance component may not affect the standard error much (i.e., in the formula for the standard error, this variance component is divided by the total sample size and so may not figure into a large reduction in the SE).

Alternatively, we can consider whether pairing decreases the *p* value from a statistically non-significant to a significant level. A limitation of this approach is that if the effect is very small, pairing may be effective at reducing the teacher-level variance, and therefore increasing precision, but this might not drop the *p* value enough.

To assess the effectiveness of pairing, we use two measures. The first measure is the change in ICC, or the change in the ratio of the between-teacher variance to the total variance. (We anticipate that the teacher-level variance will drop as we model combinations of (a) pretest, (b) pairs, and (c) fixed effects for upper-level units such as schools.)

A couple of drawbacks to considering the ICC as a measure of whether pairing works are that (1) the SE is a standard deviation whereas the ICC is a ratio of variances so that changes in the ICC may not convey the degree of change in the SE; and (2) the ICC does not figure in the effect of losing degrees of freedom on the SE. For these reasons we also consider a second measure: the ratio of the SE after adjustment (i.e., after the application of each strategy for increasing precision) to the SE before adjustment, as a way to assess the success of matching.

The purpose of this paper is to test how effective a matched-pairs strategy is for reducing the standard error of the estimate of treatment impact. We will use the results of 10 CRTs to answer this question. Importantly, the CRTs were relatively small – the median number of pairs is 10.5. The tradeoff between degrees of freedom lost and variance accounted for should figure in importantly here. The 10 datasets give us the opportunity to empirically test the effectiveness of matching. We will answer this question per experiment as well as overall. If matching is effective, a secondary purpose is to determine whether it continues to be effective after the two other strategies for increasing precision noted above – modeling pretest and modeling fixed effects for schools – are used. If matching does not help after the pretest and fixed effects for upper-level units are modeled, then there is no point in using the strategy.

Method

Data

Over the last two years we carried out a number of group randomized trials, including the 10 analyzed here, to test the effectiveness of various educational interventions. Among these experiments, we have tested the effectiveness of graphing calculator technology, science curricula, and math interventions. A matched-pairs design was used in each of the experiments. Various matching criteria were used to establish the pairs. A summary of the criteria, as well as the number of pairs in each experiment, is given in Table 1.

Table 1. Matching Strategy for 12 Projects

Project Code	Pairing by	No. of Pairs	Pairing Criteria in Order of Application		
			First	Second	Third
1	Teacher	11	Grade-level	Bilingual or immersion	Approach to classroom teaching
2	Teacher	10	Grade-level	Student group	School
3	Class within teacher	14	Teacher	Class size	- Behavior issues - Enthusiasm - Stronger academics - Time of day for the class period
4	Class within teacher	9	Teacher	Class size or achievement level	
5	Teacher	5	Grade-level	Teaching experience	
6	Teacher	11	School	Grade level	
7	Teacher	8	School	Grade level	
8	Teacher	9	School	Grade level	Teaching experience
9	Teacher	15	School	Grade level	Teacher preparation/ training in math
10	Teacher	32	School/ grade-level[a]	Grade level/ teaching experience	Teaching experience

[a] This represents a composite of 4 experiments on the same intervention.

Analysis 1

The first analysis gives the descriptive statistics used as criteria for deciding if matching may be effective (as specified in Raudenbush, Martinez, & Spybrook, 2007). Results are provided for each of the 10 experiments.

First, we examined the cluster level ICC of each experiment to judge whether it is worthwhile to adopt a matched-pairs design. Matched-pairs designs were considered inappropriate for experiments with ICCs less than .05 (Raudenbush, Martinez, & Spybrook, 2007). Second, we calculated the naive pair correlation and adjusted pair correlation for each experiment. We report the naive pair correlation

because it indicates whether the pairs are well matched. If the correlation is less than .40 or ICCs are less than .05, we coded the experiment as using an ineffective matched-pairs design.

Analysis 2

For each of the 10 experiments, we examined the extent to which modeling pairs reduces the ICC. We also examined changes in the standard error for the impact estimate as a proportion of the standard error before pairing.

As noted above, we were also interested in whether modeling matched pairs reduces the ICC and the SE of the estimated treatment effect even after fixed effects for upper-level units (i.e., schools) and the pretest covariate are modeled. To answer this question we tested an expanded set of 12 models at each site. The specifications of the models are given in Table 2.

Table 2: Effect Estimates at Each Level for Each of the 12 Models

Model	Pairs	Upper-level unit (e.g., school)	Pretest covariate (student level)	Pretest covariate teacher-level (or class-level)
1				
2	X			
3		X		
4	X	X		
5				X
6	X			X
7		X		X
8	X	X		X
9			X	
10	X		X	
11		X	X	
12	X	X	X	

Notes:

Student-level and teacher-level intercepts were modeled as random in each model. Pair-level intercepts were also modeled as random. Upper-level intercepts were modeled as fixed.

We can estimate the variance component for teachers and matched-pairs because we assume a constant treatment effect across pairs (Klar & Donner, 1997).

In some cases SAS yielded an estimate of zero for the variance component, with no *p* value given – SAS imposes this constraint when the variance component is extremely close to zero.

The results for the 12 models were averaged across sites. For the ICC's we calculated the mean and median values. We computed the mean of the proportion change in the SE, though one could also weight the proportions by the precisions of the standard errors.

All analyses were conducted using SAS proc MIXED (Singer, 1998). Following Singer and Willett (2003), we used full maximum likelihood estimation so that we could use the deviance test to compare models.

Results

Analysis 1

Table 3 summarizes the statistics from each site that can be used to predict whether matching will be effective (the criteria were described earlier).

Table 3: Descriptive Statistics for Matching Effectiveness

Project Code	No. of pairs	ICCpre	ICCpost	Naive ρpair on pretest	Naive ρpair on posttest	Matching
1	11	0.37	0.35	0.38	0.82	Effective
2	10	0.29	0.27	0.62	0.61	Effective
3	14	0.04	0.10	0.31	0.25	Ineffective
4	9	0.26	0.37	0.95	0.65	Effective
5	5	0.53	0.48	0.93	0.95	Effective
6	11	0.07	0.07	0.38	0.38	Ineffective
7	8	0.34	0.27	0.96	0.95	Effective
8	9	0.06	0.12	0.44	0.17	Ineffective
9	15	0.47	0.50	0.96	0.91	Effective
10	32	0.61	0.48	0.92	0.95	Effective

Analysis 2

For each of the 10 experiments, we ran the 12 models described in the previous section. For each experiment, we established a profile showing changes in estimates of the variance components across the models. We show examples of the profiles for three experiments in the Appendix with Figures A1, A2, and A3. In these profiles we also tested for changes in fit of the models to the data; however, we stress that, in the event that a significant change in fit is observed, this does not necessarily mean that matching has been effective – improvement in fit can result from a drop in either the student-level variance or the teacher-level variance. With the former of these, the SE might not change much, with the latter, it does. The estimates of variance components also reveal, for example, whether there is any teacher-level variance remaining to be reduced through matching after modeling pretest and fixed effects.

The main results of Analysis 2 are displayed in two sets of figures. The first set, Figures 1, 2, and 3, display the average SE for the treatment impact as a proportion of the SE obtained without adjustment for pairing, blocking, or pretest. Figure 1 includes all 10 experiments. Figure 2 shows the same information but excluding the three experiments where matching was determined to be ineffective in Analysis 1 (i.e., using previously described criteria). Similarly, Figure 3 shows the same information but includes only those experiments where teachers did not use school membership as a pairing criterion. (For these experiments we would expect modeling fixed effects for schools to make a difference over and above pairing, for the other five schools the benefits of pairing and modeling fixed effects for schools are confounded because school membership is used as a matching criterion.)

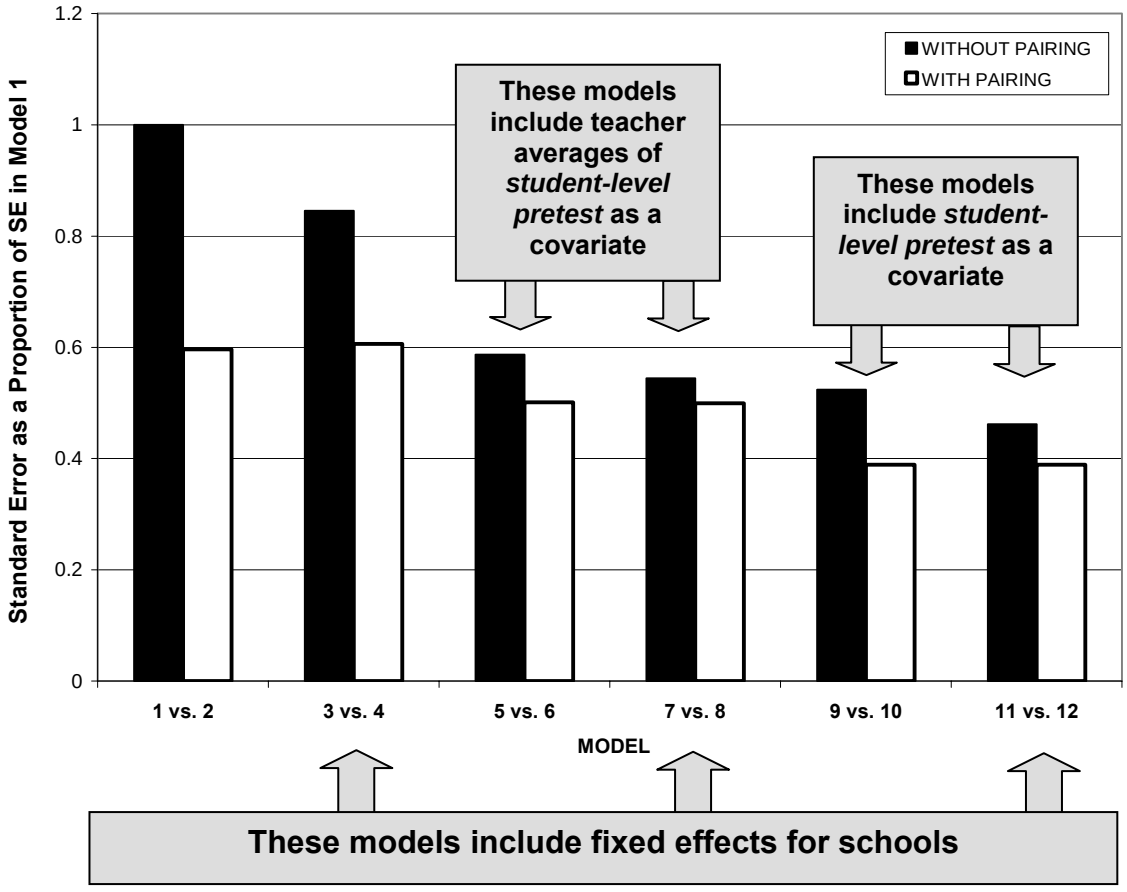


Figure 1. The Effect of Pairing on Proportion Reductions in Standard Error of the Estimated Treatment Effect (Averaged over 10 Group Randomized Trials)

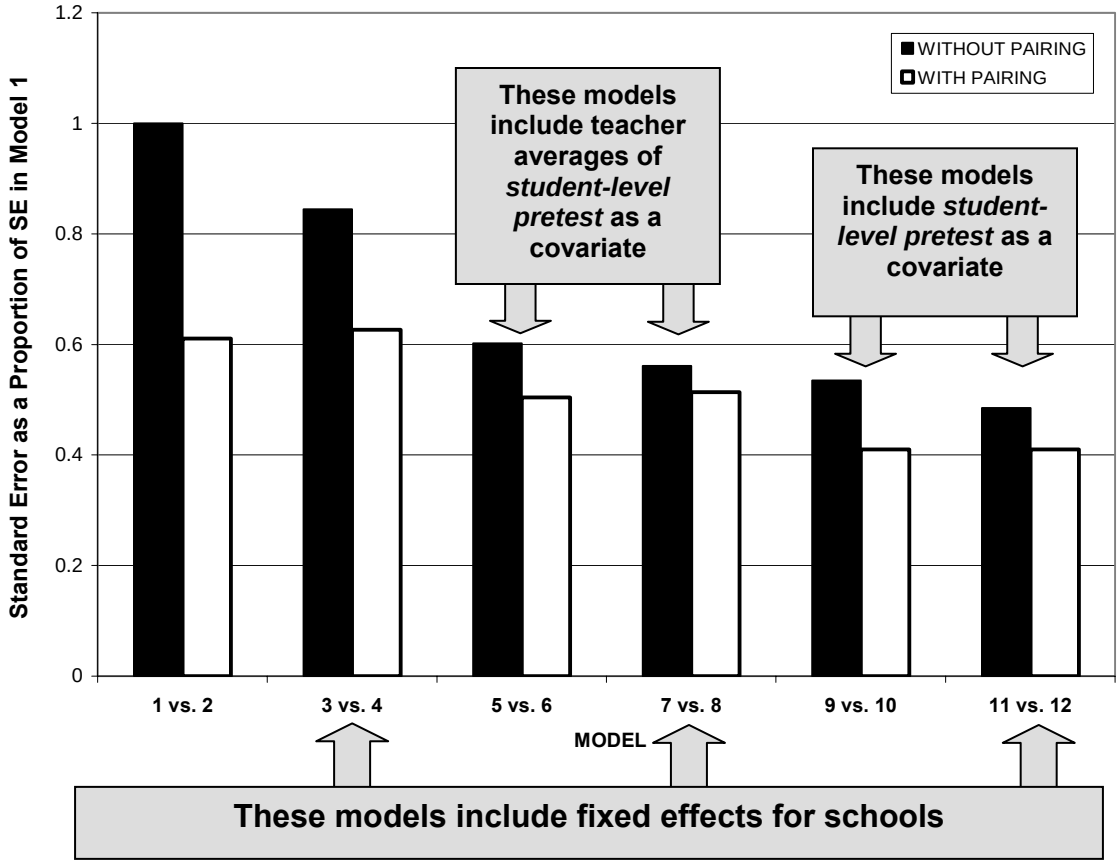


Figure 2. The Effect of Pairing on Proportion Reductions in Standard Error of the Estimated Treatment Effect (Averaged over 7 Group Randomized Trials)

We see that blocking on pairs reduces the standard error when we average across studies. It is an effective strategy even after modeling the pretest and/or the fixed effects for upper-level units. This means that modeling fixed effects for schools and/or using pretest reduce the SE only by so much, and there is an added benefit to pairing. We observe this for outcomes displayed in Figures 1, 2, and 3. (That is, limiting the analysis to experiments for which matching is expected to be effective, or to experiments where matching is not determined using school memberships as a criterion, does not seem to influence changes in precision as measured by the proportion change in the SE.)

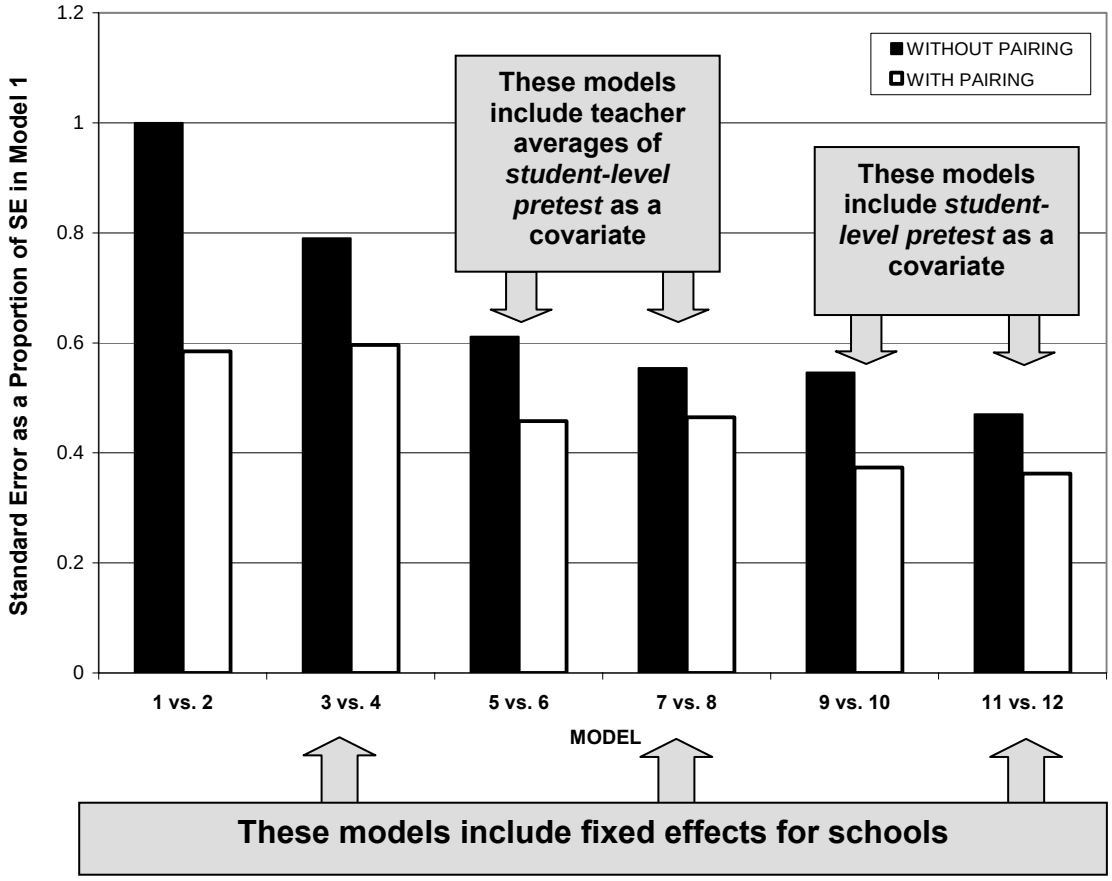
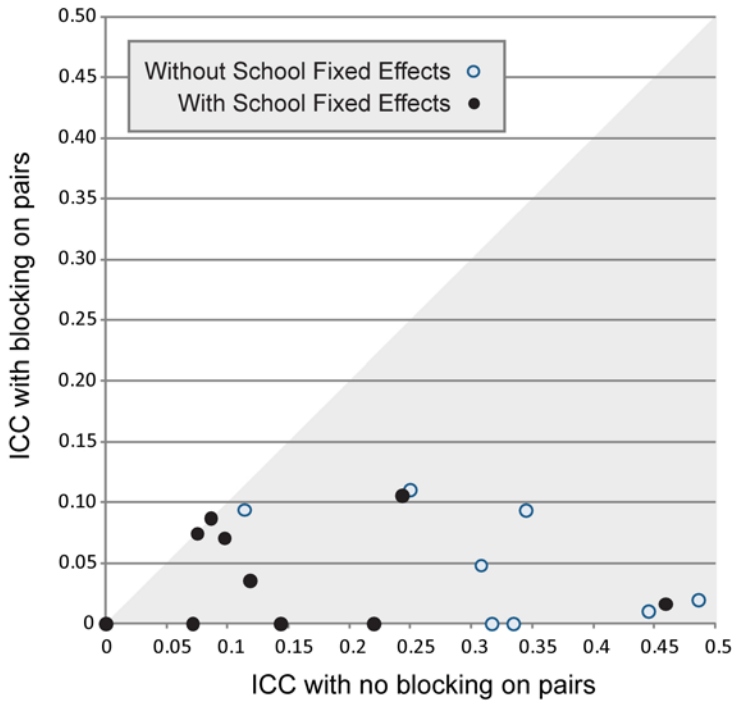


Figure 3. The Effect of Pairing on Proportion Reductions in Standard Error of the Estimated Treatment Effect (Averaged over 5 Group Randomized Trials)

To confirm the results shown in Figures 1, 2, and 3, we also computed the ICC and conditional ICCs for the 10 experiments for each of the 12 models. Figures 4 through 7 display ICC estimates in graphical form (Figures 5 and 6 show the same scatter except Figure 6 includes several arrows that trace changes of results for specific experiments.) Each point represents one experiment: the x-value is the ICC before pairing; the y-value is the ICC after pairing. In Figure 4, the points lie below the y=x line, which means the ICCs dropped as a result of pairing. In this figure we also see that modeling fixed effects resulted in the points shifting to the left, which means that this strategy also reduced the ICCs. Importantly, in spite of this leftward shift, the points remained below the y=x line, indicating the benefit of pairing over and above the modeling of fixed effects. (This study is not concerned with the variations in the leftward shift (i.e., variations in the benefit of modeling fixed effects) we therefore don't link the sets of two dots that represent individual studies. Instead we look at the average degree of leftward shift.)



Note. In some cases points coincide, therefore fewer than ten white and black points are visible. Most of the overlap occurs at the origin. For example, in one study, the ICC was zero for all models.

Figure 4. ICC's for 10 Experiments With and Without Blocking Unadjusted for Pretest

Figures 5 through 7 show the results of similar analyses except that in Figures 5 and 6, average pretest is modeled at the randomization level, whereas in Figure 7 pretest is modeled at the student level. With the effect of pretest figured in, the results are less dramatic than they were in Figure 4. Modeling pretest has the effect of dropping most of the ICCs to .10 or less. Also, the benefit of modeling fixed effects for schools is less conclusive – the leftward shift of points is less obvious. However, especially in Figure 7, it seems that the points are clustered below the line y=x suggesting that there continues to be a benefit due to pairing even after modeling pretest and / or fixed effects for schools. The effect is to further reduce what are already small values of ICCs. We note again that changes in SE relate more closely to the square root of the variance component in the numerator of the ICC so the changes depicted here to some extent underestimate the added value of pairing. (In some cases, the ICCs take on a zero value which is a constraint imposed by SAS when the teacher-level variance components are extremely close to zero.)

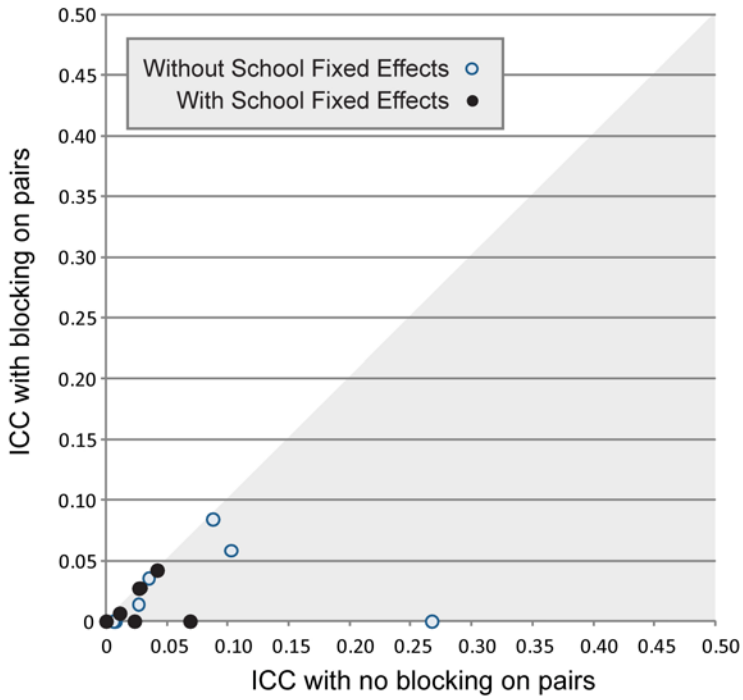


Figure 5. ICC's for 10 Experiments with and without Blocking Controlling for Pretest at the School Level

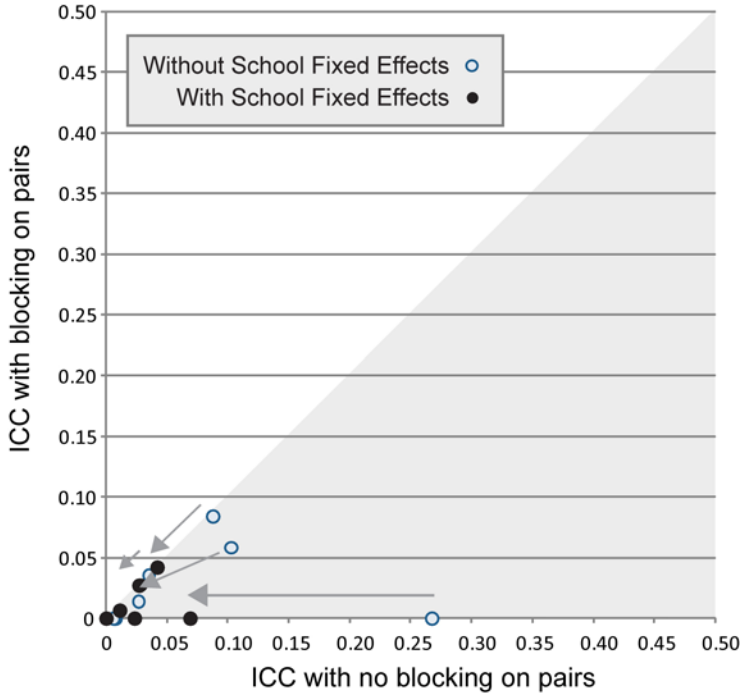


Figure 6. ICC's for 10 Experiments with and without Blocking Controlling for Pretest at the School Level

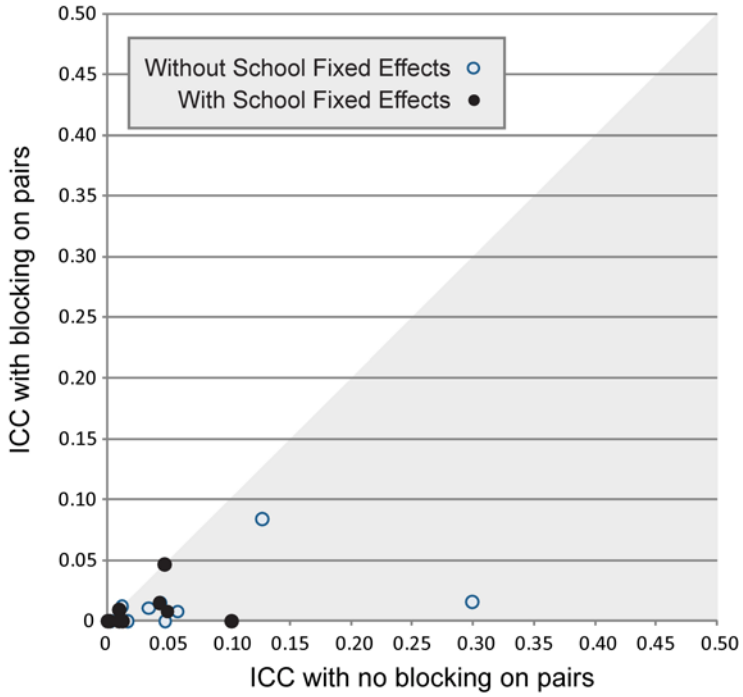


Figure 7. ICC's for 10 Experiments with and without Blocking Controlling for Pretest at the Student Level

We also summarize the ICC results in Table 4, by displaying the mean and median ICC values (across studies) for each of the 12 analyses. We see that pairing reduces the ICC with or without the other strategies; however the benefits are smaller once pretest is modeled because the pretest reduces the ICC by a large amount leaving little room for additional reductions in variability[1].

[1] The studies here reflect various interventions and outcome measures. We debated whether to display summary statistics for our analyses given the relatively small number of experiments and the range of materials that they cover (e.g., it is not obvious that average ICC's from reading interventions are the same as those for math or that the success of the strategies described here would be the same for both subjects.) We decided to include them with the caveat that generalization from this study should be done carefully since the benefits described here may hold to different degrees depending on what the intervention is, and the averages shown may not apply to subdomains of results with bigger samples.

Table 4: Conditional and Unconditional ICC Estimates

Model	Pairs	Upper-level unit (e.g., school)	Pretest covariate (student level)	Pretest covariate teacher-level (or class-level)	Mean ICC (N=10 studies)	Median ICC (N=10 studies)	Minimum	Maximum
1					0.27	0.31	0.00	0.49
2	X				0.04	0.03	0.00	0.11
3		X			0.15	0.11	0.00	0.46
4	X	X			0.04	0.03	0.00	0.11
5				X	0.06	0.03	0.00	0.27
6	X			X	0.02	0.00	0.00	0.08
7		X		X	0.02	0.02	0.00	0.07
8	X	X		X	0.01	0.00	0.00	0.04
9			X		0.06	0.04	0.00	0.30
10	X		X		0.01	0.01	0.00	0.08
11		X	X		0.03	0.01	0.00	0.10
12	X	X	X		0.01	0.00	0.00	0.05

Conclusion

The ICC at the cluster level determines whether or not pairing can work. With a high ICC (which was the case in most experiments) we benefit from a matched-pairs design even if we have a small number of clusters. For some of the cases we considered, modeling pairs has value even after controlling for pretest and/or modeling fixed effects at the school level. It decreases the standard error of the treatment effect estimate, leads to a smaller ICC, and, in some cases, improves goodness-of-fit. The benefits of matching after modeling pretest are small because the variance that is left over, to be potentially further reduced through pairing, is already small. However, it does not hurt to model pairs – the gains outweigh the losses from using up degrees of freedom to model pair effects. For small experiments, any gain in precision is welcome. We conclude that teacher-recommended criteria can be an effective basis for matching, and recommend further use of this practice.

Educational Importance

Matching is an effective strategy for increasing the precision of the impact estimate if the ICC is large. This is true even when there are relatively few pairs. This can make the difference in deciding whether there is adequate power to proceed with an experiment. This applies especially to small-scale experiments designed to answer locally-relevant questions (e.g., within a single district).

With this study we confirm empirically what Raudenbush, Martinez, & Spybrook (2007) demonstrated using a simulation study and thereby inform discussion about the value of CRTs in education.

We also find, in some cases, that matching continues to work even after modeling pretest and fixed effects for upper-level units. This means that efforts to match are not wasted – they translate into a useful gain in precision. This, in turn, leads to more conclusive findings about the effectiveness of educational interventions. Our finding supports involving teachers in the matching activity, which, in addition to increasing precision, benefits teacher knowledge and potentially increases teacher buy-in.

It also supports the intuition that teachers are effective at judging and prioritizing the kinds of factors that affect student performance.

References

Bloom, H. S., (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed), *Learning More from Social Experiments*. New York, NY: Sage.

Bloom, H. S., Bos, J. M., & Lee, S., (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*, 445-469.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R., (November, 2005). Using Covariates to Improve Precision: Empirical Guidance for Studies That Randomize Schools to Measure the Impacts of Educational Interventions. MDRC Working Papers on Research Methodology.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T., (2004). HLM: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International, Inc.

Donner, A. Taljaard, M. & Neil Klar, N. (2007). The merits of breaking the matches: A cautionary tale. *Statistics in Medicine, 26,* 2036–2051.

Hedges, L. V., & Hedberg, E. C., (2006). *Intraclass correlation values for planning group randomized trials in education (WP-06-12).* Evanston, IL: Institute for Public Research Northwestern University.

Klar, N., Donner, A. (1997). The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medici*ne, 16, 1753–1764.

Moerbeek, M., vanBreukelen, G. J. P., & Berger, M. P. F., (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25,* 271-284.

Raudenbush, S. W., (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Raudenbush, S. W., & Bryk, A. S., (2002). *Hierarchical Linear Models (2nd ed).* Thousand oaks, CA: Sage.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the 'Optimal Design' software.* Through version 1.56.

Raudenbush, S. W., Martinez, A. & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis, 29(1),* 5–29.

SAS Institute (2006). *SAS/STAT Software: Changes and Enhancements through Release 9.1.* Cary, NC.

Schochet, P. Z., (2005). *Statistical power for random assignment evaluations of educational programs.* Princeton, NJ: Mathematica Policy Research Inc.

Singer, J. D., (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323-355.

Singer, J. D., & Willett, J. B., (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Appendix

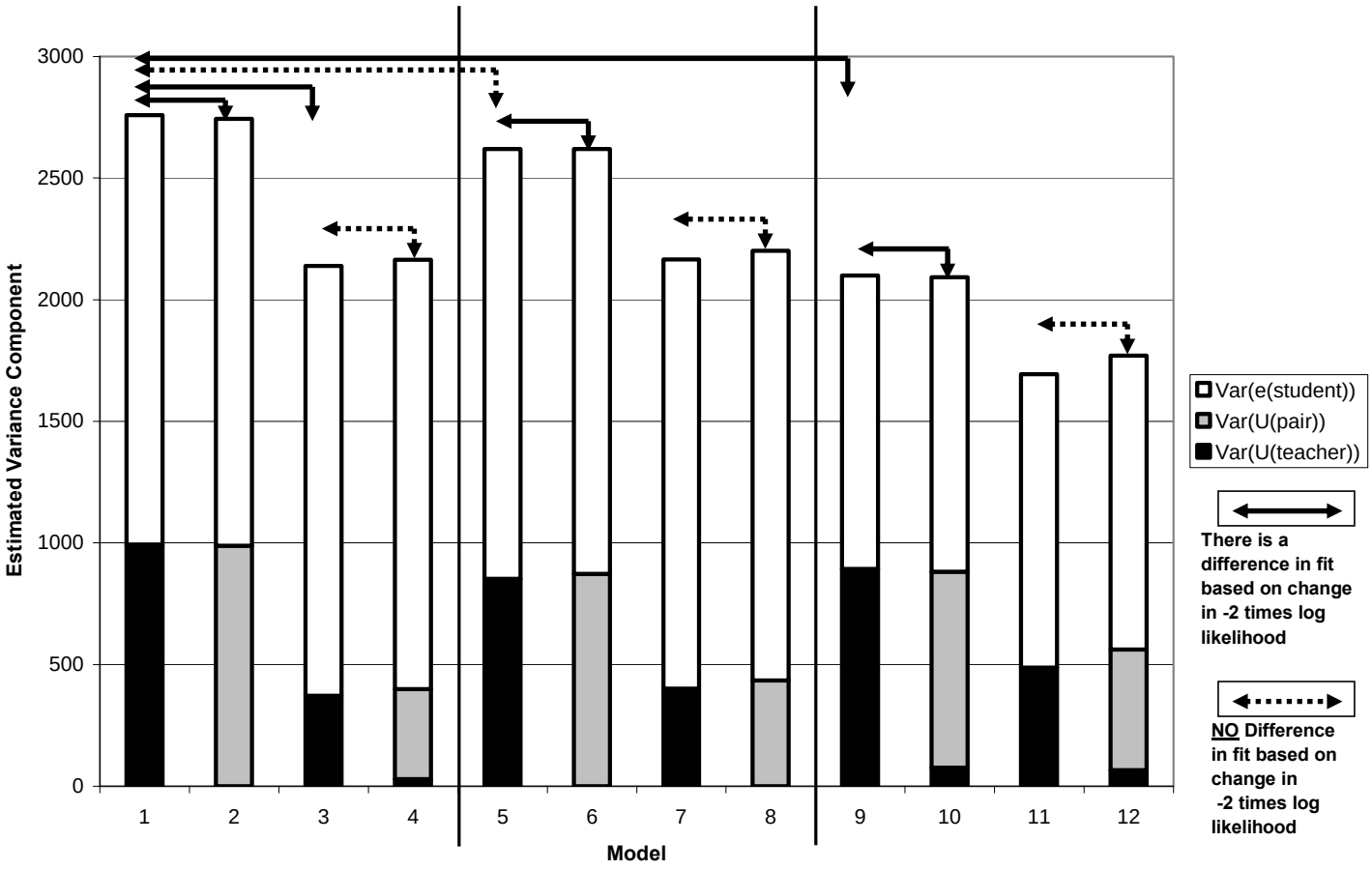


Figure A1. Breakdown of Variance Components (Location 1)

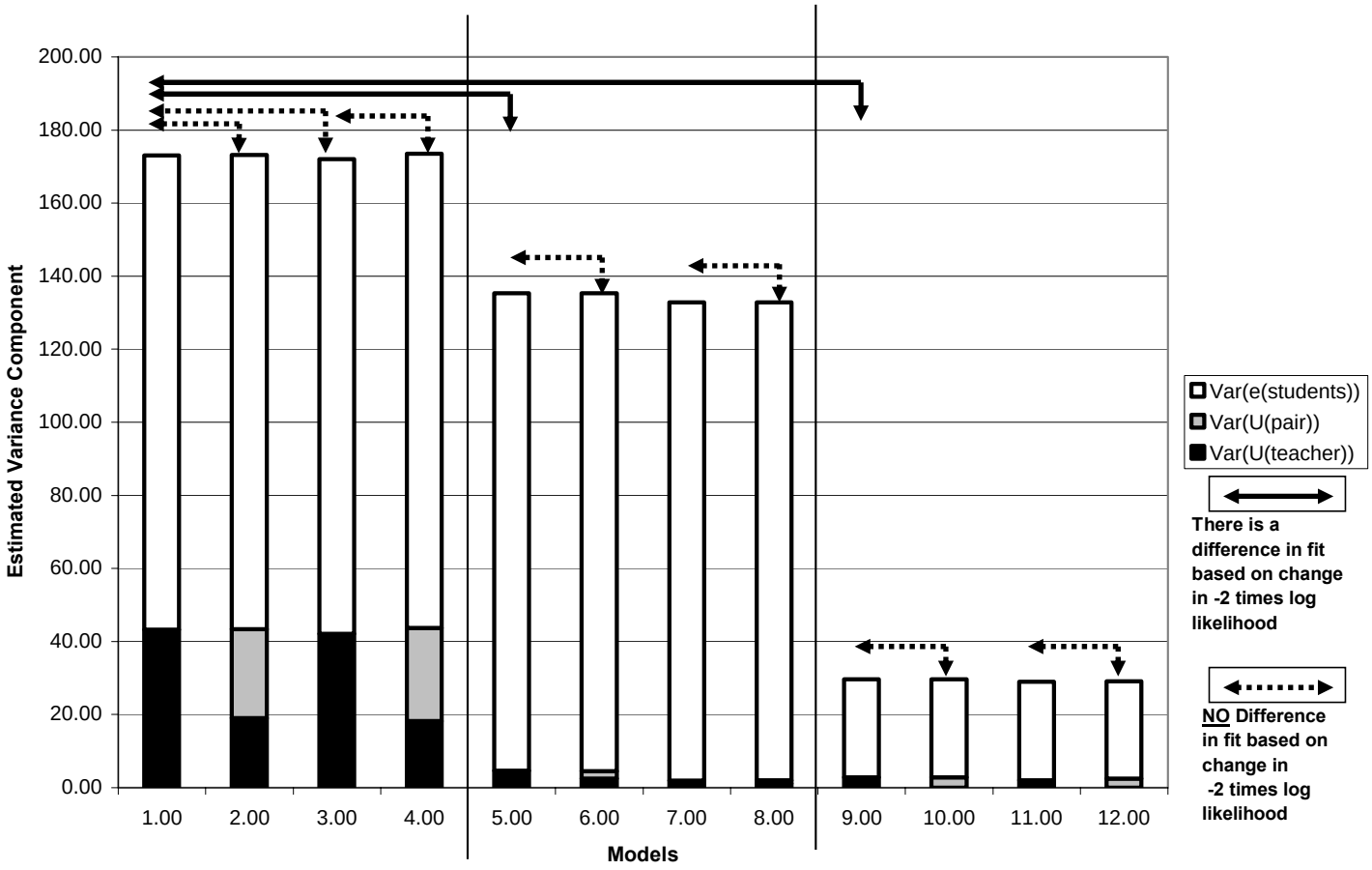


Figure A2. Breakdown of Variance Components (Location 2)

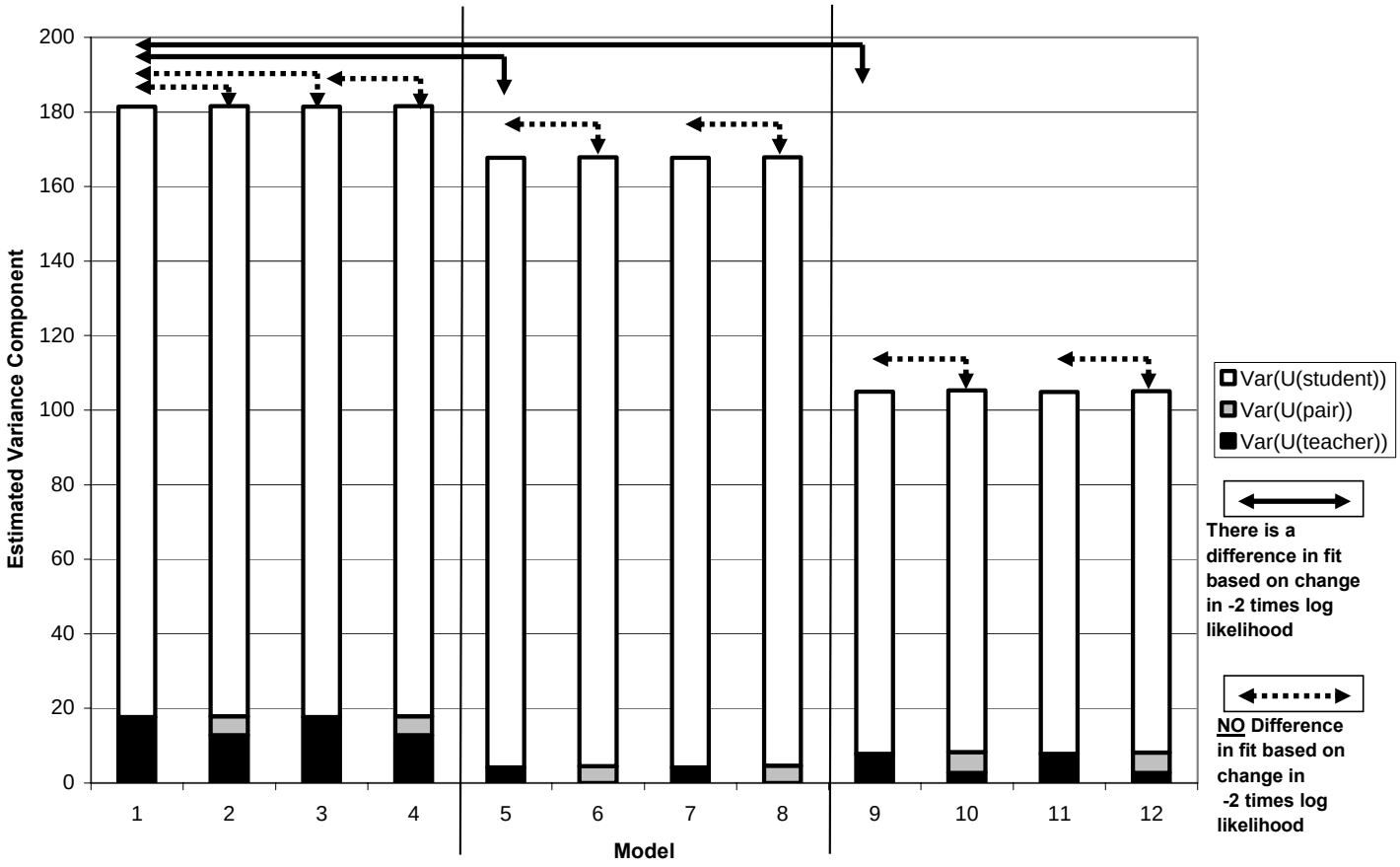


Figure A3. Breakdown of Variance Components (Location 3)