CALDER

NATIONAL CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

Urban Institute

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

New Estimates of Design Parameters for Clustered Randomization Studies

Findings from North Carolina and Florida

ZEYU XU AND AUSTIN NICHOLS

WORKING PAPER 43 • MAY 2010

New Estimates of Design Parameters for Clustered Randomization Studies

Findings from North Carolina and Florida

Zeyu Xu
ZXu@urban.org
CALDER, The Urban Institute

Austin Nichols
CALDER, The Urban Institute

Contents

Acknowledgements	ii
Abstract	iii
Introduction	1
Research Design	2
Findings from North Carolina	6
Findings from Florida	17
Compare with Earlier Findings BRB 2007	26
Summary of Findings and Conclusion	32
References	34
Tables	35

Acknowledgements

The authors gratefully acknowledge helpful comments from Howard S. Bloom. This research was supported by a grant from the W. T. Grant Foundation as well as by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) supported through Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. The authors are also grateful to the North Carolina Education Research Data Center, housed at Duke University, and the Florida Department of Education for providing access to the data for this project.

CALDER, The Urban Institute
2100 M Street N.W., Washington, D.C. 20037
202-261-5739 • www.caldercenter.org

New Estimates of Design Parameters for Clustered Randomization Studies: Findings from North Carolina and Florida
Zeyu Xu and Austin Nichols
CALDER Working Paper No. 43
May 2010

Abstract

The gold standard in making causal inference on program effects is a randomized trial. Most randomization designs in education randomize classrooms or schools rather than individual students. Such "clustered randomization" designs have one principal drawback: They tend to have limited statistical power or precision. This study aims to provide empirical information needed to design adequately powered studies that randomize schools using data from Florida and North Carolina. The authors assess how different covariates contribute to improving the statistical power of a randomization design and examine differences between math and reading tests; differences between test types (curriculum-referenced tests versus norm-referenced tests); and differences between elementary school and secondary school, to see if the test subject, test type, or grade level makes a large difference in the crucial design parameters. Finally they assess bias in 2-level models that ignore the clustering of students in classrooms.

Introduction

The gold standard in making causal inference on program effects is a randomized trial. Depending on the nature of an intervention (whether it is a school-level program or a classroom-level program) and practical feasibility, most randomization designs in the education field randomize classrooms or schools rather than individual students. Such “clustered randomization” designs have one principal drawback: They tend to have limited statistical power or precision. The implication of this drawback is the need to randomize many schools or classrooms.

It is a widespread perception that the standard error of the treatment contrast will typically depend more heavily on the number of clusters than on the number of participants per cluster. The precision can be substantially improved by using pretreatment covariates. How effective this approach is in improving the precision of estimates in a clustered randomization study is an empirical question. Answers to this question will guide how a clustered randomized trial should be designed; particularly the number of schools and classrooms that should be randomized given various sets of available covariates. Research on this topic has been limited to a small number of school districts, most of which are large urban districts.

Additionally, most existing information is based on two-level data which only account for the clustering of students within schools and ignore the clustering of students at the classroom level (for example, Bloom et al., 2007; Hedges and Hedberg, 2007). Because students in the same classroom are taught by the same teacher and share their educational experiences on a daily basis, and because of non-random teacher-student sorting, there could be significant clustering at the classroom level. Ignoring classroom-level variance components could lead to inaccurate estimates of the standard error of the treatment effect; more importantly, estimates of classroom-level variance components are required for designing studies that randomize at the classroom level.

The goal of this study is to provide empirical information needed to design adequately powered studies that randomize schools. Of particular importance are estimates of three-level variance components for measures of student achievement that account for the clustering of students within classrooms within schools. There are a number research questions addressed using these estimates. First, we assess how different covariates contribute to improving statistical power of a randomization design. We also examine differences between math and reading tests, differences between test types (curriculum-referenced tests versus norm-referenced tests), and differences between elementary school and secondary school, to see if the test subject, test type, or grade level makes a large difference in the crucial design parameters. Finally we assess bias in 2-level models that ignore the clustering of students within classrooms.

Research Design

Statistical power is an important part to any evaluation design. It demonstrates how well a study will be able to distinguish real impacts of a program or intervention from differences by chance. Statistical power is defined as the probability of detecting an effect when the effect is real. Everything else held constant, the more schools we randomize, the more powerful a study will be in detecting an effect of a given size. In practice, the significance level, power, and the expected effect size of a program or intervention are predetermined. Under these conditions, researchers designing a study will vary the sample sizes and calculate their corresponding detectable effect sizes to check when they will fall below the expected effect size. Specifically, the minimum detectable effect size (MDES) is calculated as follows:

$$MDES = Factor(\alpha, \beta, df) * \sqrt{Var(impact)} / \sigma \qquad \text{(Schochet, 2008)}$$

where *Factor* is a constant based on the *t*-distribution, and it is a function of the significance level (α), statistical power (β), and the number of degrees of freedom (*df*). Bloom (1995) shows that when the number of degrees of freedom exceeds about 20, the value of the *Factor* is about 2.8 for a two-tailed test and 2.5 for a one-tailed test, given a significance level of 0.05 and statistical power of 0.80. σ is the standard deviation of the outcome measure and is used to standardize the minimum detectable effect into effect size units. *Var(impact)* is the key parameter that needs to be empirically estimated. In a setting where students are nested within classrooms within schools, this parameter is composed of within-classroom variance, between-classroom variance, and between-school variance.

Our study uses longitudinal administrative data from North Carolina and Florida to estimate the MDES for student achievement. At the center of our analysis is the estimation of 3-level models with student, classroom, and school at each level for each school district in these two states over multiple years. We focus on how the inclusion of various covariates in the model can improve the precision of student achievement estimates in studies that randomize schools. Previous research shows that pretest scores (either individual test scores or aggregate scores at the school level) and student demographic characteristics are some of the most powerful covariates that can significantly improve estimation precision. In addition, we also explore how estimated MDES varies by subject, test type (for Florida only), and school level.

As a result, in order to provide practical guidance for future clustered randomization design, we estimate design parameters with a variety of model specifications and sample restrictions. First, we estimate the following 3-level model:

$$y_{ics} = \alpha + \beta_0 T_s + x\beta_2 + e_s + u_{cs} + \varepsilon_{ics}$$

where y_{ijk} denotes test scores for student *i* in classroom *c* at school *s*. *T*=1 if school *s* is in the intervention group and 0 if in the control group. e_s, u_{cs} and ϵ_{ics} denote school level, classroom level and individual level random errors respectively. Vector *x* represents a number of covariate sets that may be available to researchers to improve the precision of studies using randomized data. The following model specifications are estimated in this analysis:

Model 0: unconditional model with no covariates

Model 1: *x* includes student level test scores lagged 1 year

Model 2: *x* includes student level test scores lagged 2 years

Model 3: *x* includes student level test scores from two previous time periods

Model 4: *x* includes school average test scores lagged 1 year

Model 5: *x* includes school average test scores lagged 2 years

Model 6: *x* includes school average test scores from two previous time periods

Model 7: *x* includes student demographic characteristics

Model 8: *x* includes student demographic characteristics and student level test scores lagged 1 year

Model 9: *x* includes student demographic characteristics and school average test scores lagged 1 year

In North Carolina, the outcome variable includes test scores on the following subjects: 5^{th}-grade End-of-Grade (EOG) math, 5^{th}-grade EOG reading, and secondary school End-of-Course (EOC) tests including algebra II, biology, chemistry, and geometry. The study period for the two elementary school subjects are from school years 1999-2000 to 2005-06, and the study period for high school subjects are from 2000-01 to 2005-06. As students were tested repeatedly in math and reading each year in elementary school, test scores on the same subject from previous years are used as covariates. By comparison, tests on secondary school subjects were not repeated. Each student took an EOC test in a particular subject only once (unless he failed the first time and had to retake the test). As a result, based on the typical course/test sequence in North Carolina, we take EOC algebra I test scores and 8th-grade EOG math scores as student pretest scores in our secondary school analysis (model 1 and model 2). For model 4 through 5, however, we always use the school average scores on the *same* test subject from earlier years as the covariate at both school levels. All test scores are normalized by subject and year.

In Florida, we have student test scores in math and reading from 2002-03 through 2005-06 for grades 3 through 11. During these years, Florida students typically took two types of tests: the high-stakes Florida Comprehensive Assessment Test (FCAT), which consists of criterion-referenced tests (CRT) that measure student progress toward meeting the Sunshine State Standards (SSS) benchmarks, and norm-referenced tests (NRT), which are based on the Stanford 9 test series.[1] As a result, Florida offers us a unique opportunity to investigate how test types (CRT versus NRT) may affect clustered randomization designs. For this purpose, we construct our samples that contain students who took both the CRT and NRT math and reading tests.

In addition, both types of tests should have been taken in the same schools. Our final samples retain 97 percent of all FCAT test-takers in both math and reading at the elementary level (5th grade). At the secondary level (10th grade), 87.6 percent of students who took the FCAT math test also took the NRT math test, and 85.7 percent of students who took the FCAT reading test also took the NRT reading test. Florida data allow us to construct two 5th-grade cohorts (the 2004-05 cohort and the 2005-06 cohort) for which two years of lagged test scores are available. At the secondary level, we construct two 10th-grade cohorts for which their 8th and 9th grade test scores in the same subjects can be used as baseline performance in math and reading.

[1] Stanford 10 test series was used since March 2005 for the NRT. Florida NRT testing ends in school year 2007-08.

The next step is to calculate intraclass correlations ρ_s and ρ_c (the proportion of total variance that is *between schools* and *between classrooms* respectively) using variance components estimated from the unconditional model (Model 0), and to calculate the percentage of variance reduced at the school, classroom, and student level (R^2_s, R^2_c, and R^2_i respectively) by the inclusion of covariates (by comparing variance components in the unconditional model and those in models with covariates).

Assume we wish to use a two-tailed test where a normal approximation is justified, usually true with more than 20 degrees of freedom. With a desired statistical power of 0.80 and a significance level of 0.05, MDES can then be approximated by the estimator *M* given by:

$$M = 2.8\sqrt{\frac{\rho_s(1-R_s^2)}{p(1-p)J} + \frac{\rho_c(1-R_c^2)}{p(1-p)JK} + \frac{(1-\rho_s-\rho_c)(1-R_i^2)}{p(1-p)JKN}}$$

where *p* is the proportion of schools randomized to treatment and *J, K, N* represent the number of schools, the average number of classrooms in school, and the average class size respectively. Assuming half of the randomized schools are in the treatment group and the other half in the control group, *M* is calculated for various combinations of sample sizes at the various levels, so we construct points on a surface *M(J,K,N)*.

Findings from North Carolina

The overall finding is that past individual level test scores outperform alternative controls at reducing variance and improving precision of estimates, but that multiple years of school averages may perform adequately when individual level scores are unavailable. Following the suggestion in Bloom et al. (2007, henceforth referred to as BRB 2007), an effect size of 0.20 is considered an effect of substantive importance in education research, and estimates of *M* below this target level are highlighted in tables and discussion below.

Elementary school subjects

For the purpose of comparing the effectiveness of various covariates in increasing the precision of clustered randomization designs, results are presented for districts where all model specifications outlined above have converged. Over the study period of seven years, 143 and 73 district-years have achieved convergence for 5^{th}-grade math and reading tests respectively. On average, there are about 20 schools per district, three classrooms per school and 17 individual students per classroom (Table NC-1).

Intraclass correlations

On average, about 11 percent of the total variation in math test scores is between schools and seven percent of the total variation is between classrooms. Reading scores vary slightly less than math scores both between schools (10 percent) and between classrooms (six percent) (Table NC-2). Everything else equal, higher intraclass correlations result in larger *M*, as demonstrated in BRB 2007. Therefore, without any baseline period covariates, we expect *M* to be slightly larger with math scores than with reading scores as the outcome measure. In other words, in order to detect the same effect size with the same statistical power, more schools may need to be randomized in a clustered randomization design that examines student math performance than in a design that examines student reading performance.

Explanatory power of covariates

The bottom panel of table NC-2 presents R^2_i measures (at different levels *i*), i.e. the proportion of variance "explained" (proportional reduction in variance) at the school, classroom, and individual level due to including various covariates. Individual-level test scores lagged one year have the strongest predictive power (where the outcome variable is current year test scores) for both math and reading.

They explain 68 percent of the within-classroom variance of 5th grade math scores and 58 percent of the within-classroom variance of 5th grade reading scores. By comparison, student characteristics (including race, gender, free/reduced-price lunch status, age and grade repetition) explain about 14 to 15 percent of the within-classroom variance for both math and reading.

But probably more important than R^2 at the individual level is R^2 at the clustered level. As an illustration, BRB 2007 demonstrates that even though increasing the individual-level R^2 from zero percent (using no covariates) to 80 percent only reduces *M* by 0.01 to 0.03 standard deviations, increasing the school-level R^2 from zero to 80 percent cuts *M* by roughly half. Those authors reckon that "this improvement in precision is equivalent to that which would be produced by a fourfold increase in the number of schools randomized" (p.36).

Table NC-2 shows that individual test scores lagged one year explain a large proportion of the classroom and school level variance. For 5th grade math, 53 percent of the between-classroom variance is explained by individual test scores from the previous year. In addition, individual test scores also explain 65 percent of the between-school variance. Lagged individual test scores have even stronger predictive power for 5th grade reading: they explain 65 percent of the classroom level variance and 80 percent of the school level variance. All these indicate that individual student test score lagged one year may be the most effective single covariate that will significantly improve the precision of clustered randomization designs, thus dramatically reducing the number of schools that need to be randomized.

Individual student test scores lagged two years have smaller R^2 at all levels than scores lagged one year, indicating decreasing influence of earlier test scores on students' current performance. Compared with using one year of previous test scores, using multiple years of earlier test scores further increases the R^2 at all levels, but only marginally. Similarly, controlling for student characteristics in addition to one year of earlier student test scores does not further increase the R-squared by much.

Individual-level test scores from earlier years, however, may be more difficult and costly to obtain. School level aggregate scores are more readily available, in many cases as publically accessible information downloadable from the Internet. As a result, we investigate whether school aggregate scores of the same grade from previous years are as strong as individual lagged scores in explaining cluster-level variance, the key factor in reducing *M*.

Table NC-2 shows that, with the EOG tests in North Carolina, school aggregate scores from the previous year explain less math score variance at the school level by 10 percentage points and less reading score variance at the school level by 17 percentage points than individual student scores lagged one year could. However, using school averages from two prior years reduces the school-level variance of math test scores by a comparable amount (66 percent) to using one year of lagged individual student test scores, but it is still less effective for reading performance (64 percent). Alternatively, supplementing one year of school lagged average scores with student characteristics reduces the variance at the school level by 67 percent for math performance and 69 percent for reading performance.

Estimates of MDES

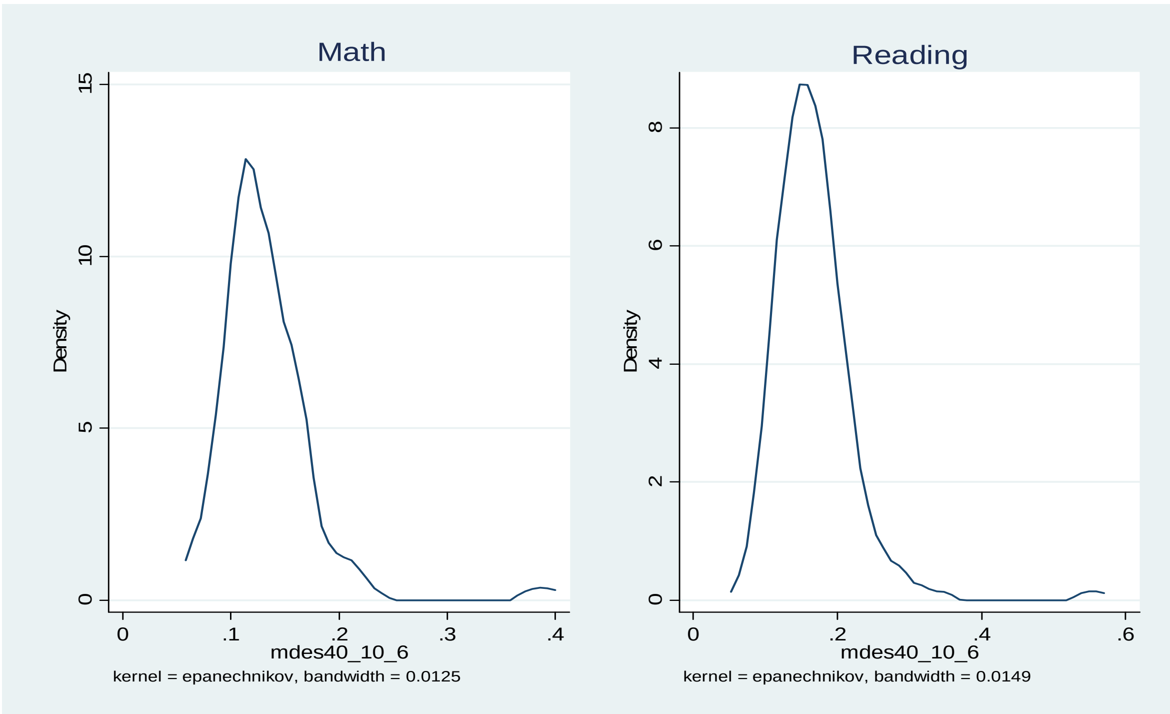
Based on the parameters estimated above, we calculate *M* assuming 20, 40, or 60 schools randomized, with 60 students within each school. These students could be distributed into three, five or 10 classrooms. We further assume that half of the schools are assigned to the treatment group and the other half assigned to the control group. The results are summarized in tables NC-3 and NC-4. *M* values below the target of 0.20 standard deviations (that are detectable with statistical power of 0.80 at a significance level of 0.05 in two-tailed tests) are shown in bold. The first pattern we notice is that the more widely distributed the students to classrooms, the lower is *M*, the estimate of MDES.

Without using any covariates, the target detectable effect size of 0.20 cannot be achieved even with 60 schools randomized. Controlling for individual student test scores lagged one year reduces *M* by about half. For 5th grade math, randomizing 40 schools, regardless the number of classrooms per school produces an *M* between 0.17 and 0.18 on average while controlling for individual lagged scores. Adding another year of individual student lagged score or adding student characteristics further reduces *M*, but only marginally. By comparison, similar *M* can be achieved with 60 schools randomized when school aggregate lagged scores are used in combination with either another year of lagged school aggregate scores or student characteristics.

Because of lower intraclass correlations and higher R^2 at the cluster level, *M* for the 5th grade reading is lower than that for math. With one year of individual lagged reading scores as the covariate, effect sizes of 0.20 or lower can be detected with statistical power of 0.80 at the 0.05 significance level when 20 schools are randomized. By comparison, 40 schools are needed to achieve similar levels of *M* when school aggregate scores lagged one year are used in combination with student characteristics or with an additional year of lagged aggregate scores. When only one year of school aggregate scores are available, in general 60 schools are needed to achieve an estimated minimum detectable effect of 0.20 standard deviations or lower.

Even though the average *M* provides valuable information for future clustered randomization designs, it should be noted that there is variation from district to district. Table NC-5 presents an example of how much variation there is between district *M'*s. The table presents the 5th and the 95th percentile *M* while assuming 40 schools are randomized, each with 10 classrooms and an average class size of six. The distribution of *M*, when one year of lagged individual scores are used as the covariate, is shown in figure 1. It shows that the distribution of M is skewed, with the vast majority of districts attaining minimum detectable effects lower than 0.20 and a few with *M*'s that are much higher than 0.20.

Figure 1. Distributions of *M*, estimated with one year of lagged individual scores as the covariate (assuming 40 schools, 10 classrooms and six students): 5th grade math and reading in North Carolina



Secondary school subjects

At the secondary level, we examine four End-of-Course subjects in North Carolina: algebra II, biology, chemistry and geometry. Since these subjects are not tested repeatedly, earlier test scores on the same subjects are not available. As a result, we choose test results from two most closely related subjects: algebra I and the 8th grade End-of-Grade (EOG) math as approximate measures of students' preparedness for these EOC test subjects in baseline years.

During our study period from 2000-01 through 2005-06, 37 district-years have achieved convergence on all model specifications for algebra II, and 32, 19 and 29 district-years have achieved convergence on all model specifications for biology, chemistry and geometry respectively (Table NC-6).

On average, there are about eight to 11 schools per district for all these subjects, with an average of three to four classrooms per school. The average class sizes are about 18 to 19 students. Compared with elementary school analysis, secondary school districts have fewer schools.

Intraclass correlations

The proportion of total variance that is between schools or between classrooms is much higher for secondary school test outcomes than those at the elementary level, probably indicating more school-student sorting and teacher-student sorting at the secondary level. About 23 percent of the total variance in algebra II test scores is between schools (Table NC-7). The intraclass correlations at the school level for biology, chemistry and geometry are 20 percent, 21 percent and 28 percent respectively. The intraclass correlations at the classroom level are even higher than those at the school level for all four subjects except for geometry.

Overall, 22 percent or more of the total variance in secondary school test outcomes is between classrooms, substantially higher than the six to seven percent for 5th grade math and reading outcomes. Such high intraclass correlations for secondary school test results indicate that in order to detect an effect of 0.20 under the same power and significance level requirements, clustered randomization studies that examine secondary school math and science performance may require a larger number of schools to be randomized than studies on elementary school student performance.

Explanatory power of covariates

Not surprisingly, using earlier test scores on a different although related subject leads to lower R^2. Nevertheless, in the absence of repeated measures on the same secondary school subjects, earlier test scores on a related subject may still be used to significantly improve the precision of a clustered randomization design. Using individual student algebra I test scores explains 60 percent of the between-school variance of algebra II and biology test scores, 52 percent of the between-school variance of chemistry test scores, and 76 percent of the between-school variance of geometry test scores.

Similarly, individual student math scores at the end of 8th grade explain 47 percent, 68 percent, 63 percent and 75 percent of the between-school variance of algebra II, biology, chemistry and geometry test scores respectively. The higher R^2 achieved when using 8th grade math scores as the covariate for biology and chemistry implies that 8th grade math performance may be the better predictor of between-school variations in these two subjects than algebra I is. Combining individual algebra I scores with 8th grade math scores further increases the R^2 at the school level, substantially for biology, chemistry and geometry and slightly for algebra II. Similarly, adding student characteristics to individual algebra I scores also increases the R^2 at the school level.

Both individual algebra I scores and 8th grade math scores explain a significant amount of between-classroom variance as well. Interestingly, 8th grade math scores have higher R^2 than algebra I scores for all subjects (64-79 percent versus 52-70 percent). One possible explanation, though not explored here, is that the assignment of students to classrooms within a school may depend more directly on students' 8th grade math performance than on their performance on the algebra I test, which students can take anytime between the 6th and the 9th grade.

An alternative to using individual student test scores on a related but different subject as the covariate is using the school average test scores on the *same* secondary school subject from earlier years. Table NC-7 shows that these aggregate measures are not as effective in reducing between-school variances as individual student scores, with the exception of chemistry. School average scores lagged one year explain 47 percent of the total variance for algebra II, 45 percent for biology, 74 percent for chemistry, and 59 percent for geometry. School averages lagged two years explain even less. However, adding student characteristics to one year of lagged aggregate scores substantially raises the R^2, to 61 percent, 62 percent, 76 percent and 68 percent for algebra II, biology, chemistry and geometry respectively.

Estimates of MDES

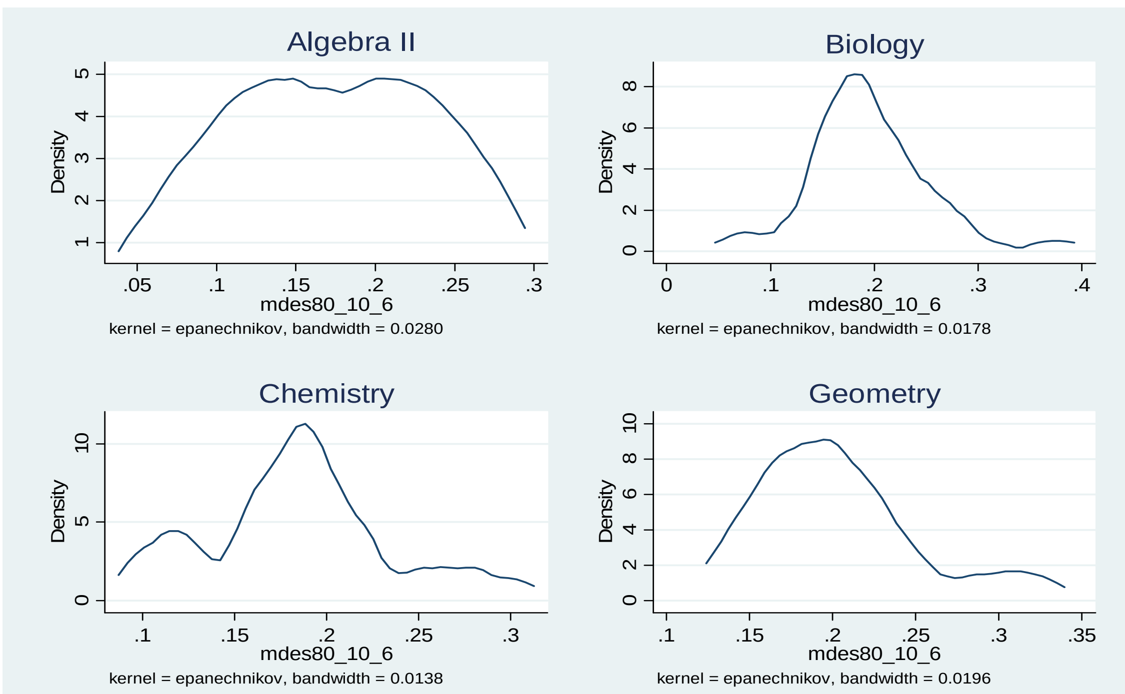
With these estimated parameters, tables NC-8 through NC-11 present the estimated MDES, or *M*, when assuming 40, 60, or 80 schools to be randomized. In all scenarios half of the schools are assigned to the treatment, and each school is assumed to have 60 students who are assigned to three, five or 10 classrooms.

As predicted by the high intraclass correlations, *M* for secondary subjects is 0.30 or higher when no covariates are used and 80 schools are randomized. When covariates are used, the number of schools that need to be randomized in order to detect an effect of 0.20 varies by test subjects. For a clustered randomization study that investigates student performance on algebra II, 80 schools are needed when the study either controls for both individual algebra I scores and 8^{th} grade math scores, or controls for both individual algebra I scores and student characteristics. With the same set of covariates, 60 schools are needed if the subject of study is student performance on biology.

For studies on student chemistry performance, the target MDES of 0.20 can be reached by randomizing 80 schools when controlling for either individual 8^{th} grade math scores or both individual algebra I and 8^{th} grade math scores. As for studies focusing on student geometry performance, if both individual algebra I and 8^{th} grade math scores are available, randomizing 40 schools is sufficient to reduce *M* to 0.20 or lower. When only 8^{th} grade math scores are available, 60 schools are needed; and when only algebra I scores are available, 80 schools are needed to detect an effect of 0.20 or lower.

In general, between 60 and 80 schools are needed for studies on student performance on secondary school test subjects when students' algebra I and 8^{th} grade math scores are used as covariates. By comparison, more than 80 schools need to be randomized when school average scores on the same test subjects from earlier years are used as covariates. Finally, table NC-12 shows an example of how much variation there is in *M* between districts. It seems that, for geometry, 90 percent of the districts are able to achieve an *M* of 0.20 or lower when 80 schools are randomized when individual Algebra I and 8th grade math scores are used as covariates. For all other subjects, *M* is lower than 0.24 to 0.28 in 90 percent of the districts under the same scenario. The distribution of the *M,* when individual algebra I scores are used as the covariate, is shown in figure 2:

Figure 2. Distributions of *M*, estimated with individual algebra I scores as the covariate (assuming 80 schools, 10 classrooms and six students): by secondary school test subject in North Carolina:



Summary of North Carolina findings

Using state standardized test results from North Carolina, we find that

1. The intraclass correlations at the secondary school level are higher than those at the elementary school level, probably a result of more teacher-student and school-student sorting at the secondary level. The intraclass correlation at the classroom level is particularly high at the secondary level.
2. The correlation between available measures of student skills in the baseline years and student performance in the current period is weaker at the secondary school level than it is at the elementary school level, likely due to the lack of repeated measures of subject-specific student performance at the secondary level.
3. Because of 1 and 2, studies that focus on student performance at the secondary school level will need to randomize more schools than studies on elementary school student performance in order to detect the same effect size with the same statistical power. Indeed, findings from North Carolina suggest that in order to detect a program effect of 0.20 standard deviations or lower, with statistical power of 0.80 at a significance level of 0.05 in two-tailed tests, an elementary school study will require randomizing about 20 to 40 schools whereas a secondary school study will require 60 to 80 schools. These estimates are based on models that control for individual pretest scores.
4. At the elementary school level, there is less variation between schools in reading performance than in math performance. As a result, fewer schools need to be randomized in a study that focuses on student reading performance (20 schools when controlling for individual pretest reading scores, as compared with 40 schools for math).
5. Finally, we find that lagged school aggregate scores do not improve the precision of a randomization study as much as lagged individual scores can. However, there is a tradeoff between the cost of randomizing more schools and the cost of obtaining individual-level test data from earlier years. In cases where the cost or time of obtaining individual-level data becomes prohibitive, or in cases

where individual-level data are not available for baseline years, the use of school aggregate scores as covariates provides a valuable alternative that can still substantially reduce the number of schools to be randomized.

Findings from Florida

Data from Florida give us an opportunity to check whether findings from North Carolina will hold in a different education context. Additionally, Florida data also provide us with a unique opportunity to examine whether the type of tests (curriculum-referenced test versus norm-referenced test) used to measure student performance should be considered as a factor in designing clustered randomization studies.

Elementary school subjects

At the elementary level, we focus on 5th grade math and reading performance. The study period is shorter in Florida than that in North Carolina, consisting of two school years from 2004-05 to 2005-06. Again, our report includes only those district-years for which all model specifications have converged. This results in 37 district-years for FCAT math, 34 for NRT math, 22 for FCAT reading, and 26 for NRT reading (Table FL-1). On average there are between 35 and 47 schools per district, each school with four to five classrooms and a class size of 18 students. Compared to districts in North Carolina, Florida school districts are on average larger.

Intraclass correlations

About 10 to 11 percent of the total variance in math test scores is between schools and another six to seven percent is between classrooms (Table FL-2). This is comparable to the intraclass correlations found in North Carolina. Also similarly in both states, the intraclass correlations of reading test scores appear to be lower than those of math scores. In Florida, about eight percent of the total variance in reading scores is between schools and four to five percent is between classrooms.

Explanatory power of covariates

The bottom panel of table FL-2 reports the R^2 at the school, classroom and student level when various covariates are used. It appears that the correlations of lagged test scores and current scores are comparable to those found in North Carolina. Individual student test scores lagged one year explain 70 percent of the school-level variance in math outcomes and 80 to 82 percent of the school-level variance in reading outcomes. Individual student scores lagged one year explain an additional 55 to 59 percent of the between-classroom variance in math scores and 62 to 79 percent of the between-classroom variance in reading scores. Further lagged individual scores are less correlated with current test performance, and using two years of prior test performance data only marginally improves the R^2 at all levels over just using one year of lagged scores.

Student characteristics can also be used to reduce the variances at all levels. However, they are less effective than student lagged test scores. For math outcomes, student characteristics explain about 42 percent of the school-level variance and 11 to 15 percent of the classroom-level variance. For reading outcomes, they explain 55 to 60 percent of the school-level variance and 22 to 29 percent of the classroom-level variance. The R^2 achieved when using both student characteristics and individual lagged test scores are not significantly higher than the R^2 achieved when individual scores are used alone as the covariate.

Similar to findings from North Carolina, lagged school aggregate scores are less effective than individual lagged scores in reducing the variance at the school level. School aggregate scores lagged one year explain about 63 to 68 percent of the school-level variance in math and reading. Adding school aggregate scores lagged two years explains three to four percentage points more school-level variance for math and about nine percentage points more school-level variance for reading.

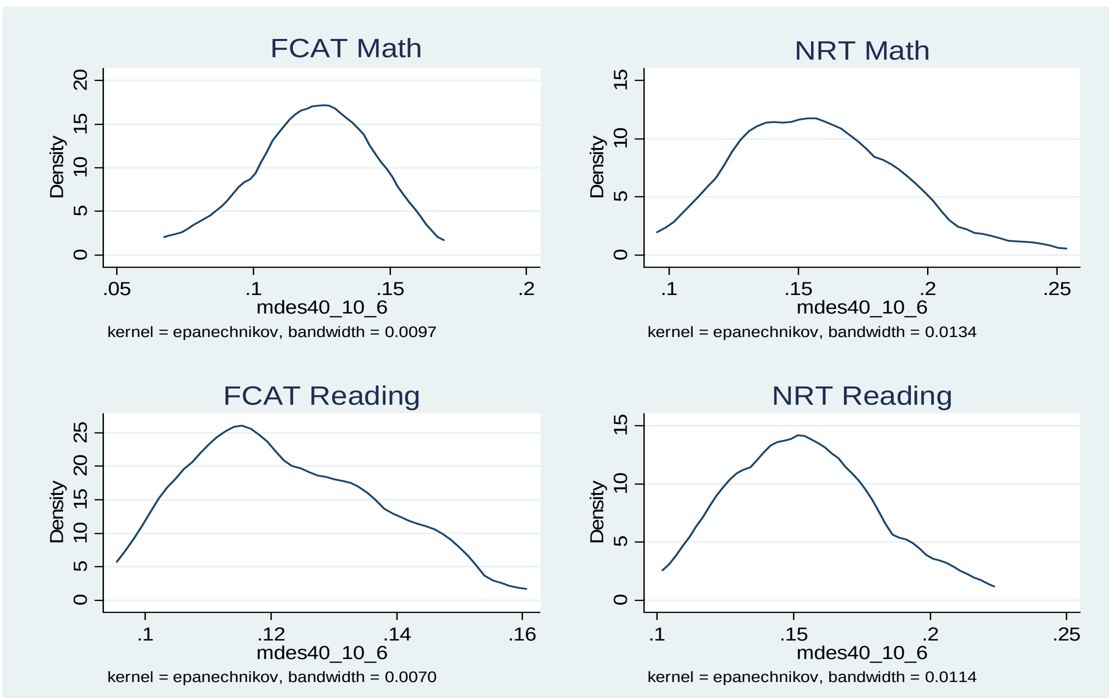
Estimates of MDES

Based on these estimated parameters, tables FL-3 to FL-6 present the estimated MDES that can be achieved when 20, 40, or 60 schools are randomized, with each school having 60 students who are distributed into three, five or 10 classrooms. The use of covariates, particularly the use of individual lagged test scores, substantially reduces *M* as compared with when no covariates are used. When 5th grade math performance is the outcome variable, an effect of 0.20 or lower can be detected with statistical power of 0.80 and a significance level of 0.05 when 40 schools are randomized while controlling for individual test scores lagged one year. Adding another year of individual lagged scores or student characteristics further reduces *M* to 0.16 or lower with 40 schools randomized. It is also interesting to note that, when there are five or more classrooms in each school, the target MDES can be reached with only 40 schools randomized when lagged school aggregate scores are used as covariates in combination with student characteristics.

Fewer schools have to be randomized to produce an *M* of 0.20 or lower when 5th grade reading performance is the outcome under study. Because of the higher R-squared attained using various covariates in explaining variances in reading outcomes than in explaining variances in math outcomes, the use of covariates reduces *M* that could have been achieved when no covariates are used more dramatically, by more than half. As a result, when we control for individual lagged test scores, only 20 schools have to be randomized to reach an *M* 0.18 or lower. And when we control for lagged school aggregate scores, 40 schools are needed to reach an *M* 0.19 or lower. Additionally, 40 schools would also be sufficient to detect an effect size 0.20 or lower when we control for student characteristics.

Table FL-7 presents one way to examine the variation of *M* across districts and years. The table assumes that 40 schools are randomized. It shows that, for example, when two years of individual lagged test scores are used, in 90 percent of the time *M* for math is between 0.11 and 0.20. As another example, when two years of lagged school aggregate scores are used, in 90 percent of the time *M* for reading is between 0.13 and 0.20. The distribution of *M*, estimated with one year of lagged individual test scores as the covariates, is shown in figure 3:

Figure 3. Distributions of *M*, estimated with one year of lagged individual scores as the covariate (assuming 40 schools, 10 classrooms and six students): by elementary school subject in Florida



Comparing test types

Finally, as pointed out above, one advantage of examining the Florida data is the opportunity to compare whether test type should be factored into clustered randomization designs. In order to make this comparison, we need to restrict our results to those district-years in which all models have converged for both the FCAT and NRT tests. We are able to identify 29 district-years in which all models have converged on the FCAT and the NRT math tests, and 17 district-years in which all models have converged on the FCAT and the NRT reading tests (Table FL-8).

Table FL-9 shows that *M* for NRT math is almost the same as that for FCAT math. The effectiveness of various covariates in reducing *M* is also similar between the two test types, with *M* being slightly lower for NRT math than for FCAT math. For both test types, 40 schools have to be randomized to achieve an *M* of 0.20 or lower when lagged individual test scores are used as covariates, or when lagged school aggregate scores and student characteristics are used as covariates. The same pattern is found with the FCAT-NRT reading comparisons (Table FL-10).

Secondary school subjects

Different from North Carolina, Florida has the End-of-Grade tests in math and reading from the third through the 11th grade. In our secondary school analysis, we use the 10th grade math and reading scores as the outcome measures and earlier test scores on the same subjects are used as lagged scores. Because of this, we expect higher correlations between these baseline year performance measures and current student performance than those found in the secondary school analyses for North Carolina.

Another important difference between the secondary school analyses for North Carolina and Florida is that Florida districts are larger. Across all test subjects, there are about 20 schools per districts, and 28 math classrooms and slightly fewer than 20 reading classrooms per school (Table FL-11). The class sizes are, however, smaller than those found in North Carolina, at about eight to 10 students in each classroom.

Intraclass correlations

Similar to North Carolina, in Florida there is substantially higher variation between schools and between classrooms at the secondary level than at the elementary level. For FCAT math and reading, about 20 percent of the total variance in student scores is between schools. For NRT math and reading, about 12 percent is between schools. It should be noted that the differences in the school-level intraclass correlations between the FCAT and the NRT tests may not be over-interpreted, as they are summarized based on different sets of districts. Direct comparisons between test types are presented in tables FL-18 to FL-20 and will be discussed later.

What probably is more interesting here is the magnitude of classroom-level intraclass correlations. Although it is also the case in North Carolina that there is more variance between classrooms than the variance between schools, in Florida the between-classroom variances are 1.5 to three times the size of between-school variance. The large size of the classroom-level intraclass correlation relative to the school-level intraclass correlation may have important implications on study designs that ignore the clustering at the classroom level. This is discussed in more detail later in this paper.

Explanatory power of covariates

As expected, the R^2 achieved when using previous test scores on the same subjects as the covariate are much higher than those attained in the secondary school analyses for North Carolina. Individual test scores lagged one year explain about 80 percent of school-level variance in all subjects, and adding another year of individual lagged scores increases the school-level R^2 further to about 85 percent. Individual lagged test scores also explain 84 to 90 percent of the classroom-level variance, and two years of individual scores explain 92 to 96 percent of the classroom-level variance, higher than any R^2 we have seen so far. This indicates that individual lagged scores may substantially improve the precision of clustered randomization studies on secondary school student performance in Florida.

The use of lagged school aggregate scores is also promising. The R^2 associated with using lagged school aggregate scores are just slightly lower than the R^2 achieved with using lagged individual scores.

The use of two years of lagged school aggregate scores is as effective as the use of one year of lagged individual student scores in reducing the school-level variance.

Estimates of MDES

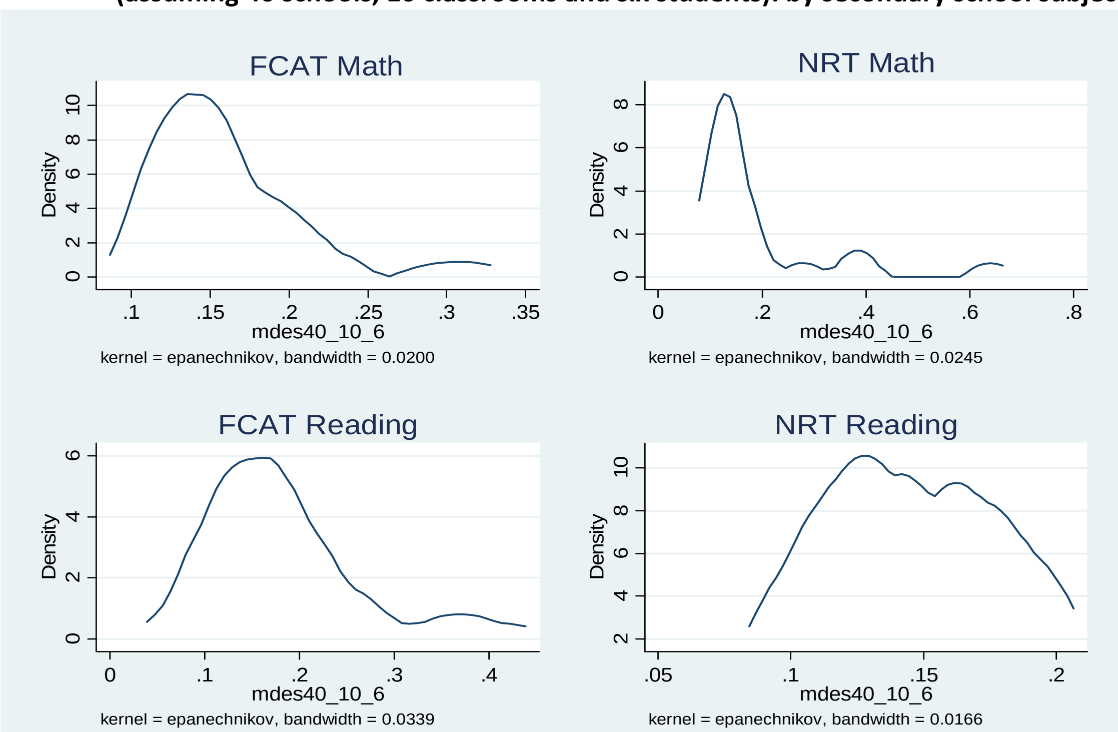
However, because school aggregate scores do not help reducing the classroom-level variance, which is very large among Florida's secondary school districts, high R^2 at the school level alone does not necessarily translate into low *M*. Tables FL-13 through FL-16 summarize the estimated MDES. Without any covariates, *M* for both math and reading is about 0.30 or higher even when 60 schools are randomized. This is hardly surprising given the large intraclass correlations in these districts. It is also not surprising that, given the high R^2 at both the school and the classroom-levels, the use of individual lagged test scores reduces *M* by more than half.

When individual student scores lagged one year are used as the covariate, in general 40 schools have to be randomized in order to detect an effect of 0.20 or lower. When two years of individual lagged scores are available, 20 schools would be sufficient to achieve *M* between 0.17 and 0.21 if the outcome under study is either student performance on NRT math or NRT reading, and between 0.21 and 0.24 if the outcome under study is either FCAT math or FCAT reading. By comparison, when lagged school aggregate scores are used as covariates, 60 schools are needed to reduce *M* to 0.20 or lower and this could be achieved only when there are ten or more classrooms in each school.

Table FL-17 demonstrates how *M* varies across districts. With the exception of NRT math, for which in 90 percent of the time *M* is between 0.10 and 0.19 when 40 schools are randomized and individual test scores lagged one year are used as covariates, there is large variation in *M* when lagged individual scores are used as covariates. For both FCAT math and FCAT reading, for example, *M* with one year of lagged individual scores ranges from 0.10 to 0.40. This range is larger than the range of *M* when lagged school aggregate scores are used as the covariates. This is different from findings from the North

Carolina elementary and secondary school analyses and from the Florida elementary school analyses, where the ranges of *M* are more or less comparable no matter whether individual scores or school averages are used as covariates. Figure 4 shows the distribution of the *M* when individual test scores lagged one year are used as the covariates with 40 schools randomized and 10 classrooms within each school:

Figure 4. Distributions of *M*, estimated with one year of lagged individual scores as the covariate (assuming 40 schools, 10 classrooms and six students): by secondary school subject in Florida



Comparing test types

Finally, there appears to be some differences in the estimated design parameters and *M* between the test types under investigation. In order to make comparisons between test types, we restrict our district-years to those for which all models have converged for both the FCAT and the NRT tests. Our comparison groups include nine district-years for math tests and 14 for reading tests (Table FL-18).

Comparisons show that, even though the classroom-level intraclass correlations are similar for both test types, the intraclass correlations at the school level are much lower for NRT tests than for FCAT tests (12 percent for NRT math versus 23 percent for FCAT math, and 15 percent for NRT reading versus 24 percent for FCAT reading).

Such differences are not seen when we compare the two types of test at the elementary school level. There could be a number of explanations for these differences. One possibility is that the nature of the 10th grade tests is very different between the FCAT and the NRT. Another possibility is that, under the school accountability pressure, schools, teachers and students treat the FCAT tests, the high-stakes tests in Florida, differently from the NRT tests that have low stakes. Unfortunately, a thorough investigation of these possible explanations is beyond the scope of the current report.

The comparisons of *M* between the FCAT and the NRT tests at the secondary level follow the same patterns as those found at the elementary level. In general, when the outcome of interest is test scores on the NRT tests, *M* is lower than when the outcome of interest is student performance on FCAT tests. For example, for 10th grade math performance, the use of two years of lagged individual test scores reduces *M* for FCAT math to 0.14 to 0.16 standard deviations when 40 schools are randomized. Under the same condition, *M* for NRT math ranges between 0.11 and 0.14. The comparisons between the two types of reading tests reveal similar patterns.

Summary of Florida findings

The findings from Florida are generally consistent with those from North Carolina, although there are a few new patterns:

1. At the elementary school level, in order to detect an effect of 0.20 or lower, about 40 schools are needed for math and 20 for reading when individual pretest scores are controlled for. Slightly different from findings from North Carolina, however, lagged school aggregate scores appear to be stronger predictors of current student performance in Florida. When school aggregate scores are used in combination with student characteristics, 40 schools have to be randomized in order to reach the target MDES of 0.20.
2. At the secondary school level, the intraclass correlation at the classroom level is higher than it is in North Carolina secondary schools. However, because Florida has repeated measures in math and reading performance at the secondary level, those pretest scores prove to be highly correlated with current student performance. As a result, when lagged individual test scores are controlled for, 40 schools need to be randomized, as opposed to 60-80 in North Carolina.
3. Finally, our findings show that test types have small impact on a clustered randomization design. *M* for NRT tests is slightly lower than *M* for the FCAT tests. Therefore, when the outcome measure is student scores on NRT tests, slightly fewer schools have to be randomized compared to an outcome measure based on CRT test scores.

Compare with Earlier Findings in BRB 2007

The recommended number of schools that need to be randomized in order to detect an effect size of 0.20 standard deviations with a significance level of 0.05 and power of 0.80 differs somewhat from the empirical findings from BRB 2007. For example, BRB 2007 finds that, using lagged test scores (either individual scores or school averages) as covariates, 60 schools have to be randomized to produce an

MDES of about 0.20 at the elementary school level. To produce an MDES of similar sizes at the secondary level, only 20 schools are needed while using the same covariates. Our findings suggest fewer (20-40) schools required at the elementary level and more (40-80) schools at the secondary level.[2] Another notable difference between our findings and those in BRB 2007 is the effectiveness of using lagged school aggregate scores as the covariates in reducing *M*. BRB 2007 suggests that in most cases the use of aggregate scores and the use of individual student scores are equally effective in improving the precision of clustered randomization designs. By contrast, our results have consistently shown that individual student test scores from earlier years are more effective than lagged school aggregate scores in reducing *M*.

Differences between 2-level and 3-level models

These differences may be explained by a number of factors. The most obvious possible explanation is that in this report we explicitly consider the clustering of students within classrooms. In order to determine whether controlling for classroom-level clustering has led to the differences, we estimate 2-level models using the same samples, ignoring the classroom level, and compare the results with the results based on 3-level models. To make such comparisons we include district-years that have achieved convergence on all model specifications with both 2- and 3-level models. The comparisons are summarized in tables NC-3a and NC-4a for North Carolina elementary school tests, tables NC-8a to NC-11a for North Carolina secondary school tests, tables FL-3a to FL-6a for Florida elementary school tests, and tables FL-13a to FL-16a for Florida secondary school tests. All comparisons are based on *M* when assuming 40 schools are randomized, with the exception of comparisons for North Carolina secondary school tests where *M* is calculated assuming 80 schools are randomized.

[2] We assume 60 students per school in both our elementary and secondary level analyses. By comparison, BRB 2007 assumes 60 students per school at the elementary level, and 250 students per school at the secondary level. To explore whether different assumptions about school sizes at the secondary level could have explained some of the differences found in our secondary school results, we recalculated *M* with 250 students per school. This change only led to slight decreases in *M.*

At the elementary level, between nine to 14 percent of the total variance is between schools when the model only considers the school and student level. This is in line with the range of intraclass correlation at the school level found in the literature (Bloom et al., 1999; Hedges and Hedberg, 2007; Schochet, 2008). At the secondary level, estimates from 2-level models show that the between-school variations are much larger, especially when state standardized tests (curriculum-referenced tests) are used. The intraclass correlation at the school level ranges from .27 to .39. This finding is similar to what we have seen when 3-level models are estimated, and it could be the result of stronger student-school sorting at the secondary level.

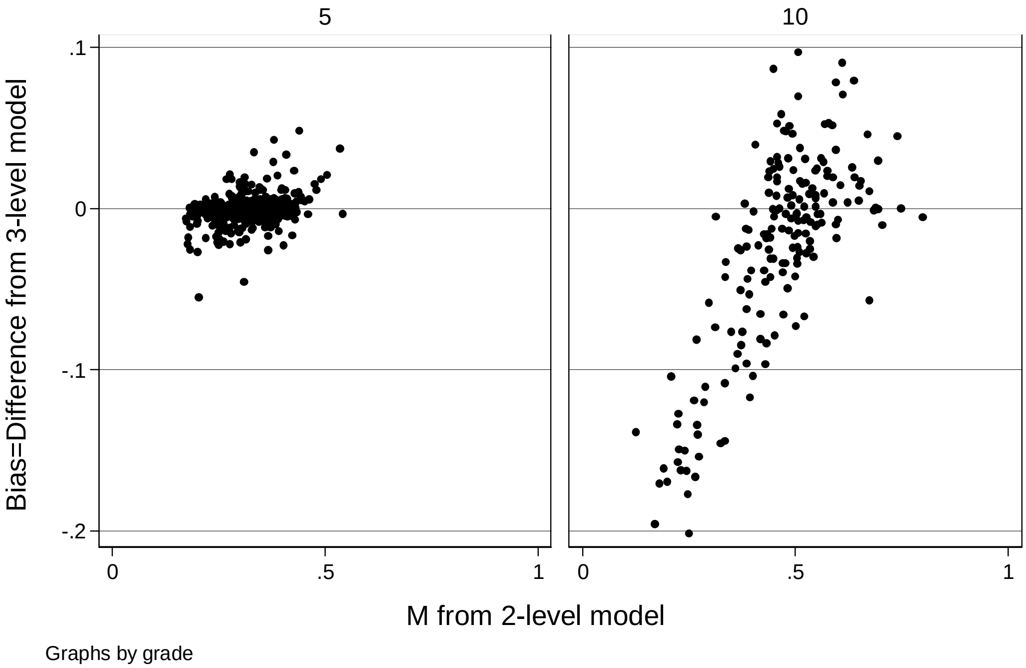
In all comparisons except for those for Florida secondary school analyses, *M* estimated under the 2-level model are generally comparable to those estimated under the 3-level model. In cases where students are distributed into three classrooms in each school, *M* based on 2-level models tend to be slightly lower than *M* based on 3-level models. This indicates that ignoring the student clustering at the classroom level is unlikely to lead to a significantly underpowered clustered randomization design. The comparisons of 2- and 3-level models at the secondary level in Florida, by contrast, tell a different story. Here, ignoring classroom-level clustering will lead to serious under-estimation of *M*. For example, assuming three classrooms per school and using individual test scores lagged one year as the covariate, *M* calculated based on estimates from 3-level models range between 0.18 and 0.22 standard deviations; the corresponding *M* based on 2-level model parameters are between 0.12 and 0.16. These differences imply that ignoring the classroom-level clustering may misguide a clustered randomization design, leading to insufficient number of schools included in the experiment and resulting in an underpowered study.

It is not immediately clear whether the bias in *M* resulting from ignoring the classroom-level clustering should be positive or negative. On the one hand, some of the between-student variance could be explained by differences between classrooms. Therefore ignoring the variance at the classroom level will lead to downwardly biased *M*. On the other hand, differences between schools could be partially explained by the differences in the classrooms that each school has. Therefore the school-level intraclass correlation estimated by a 3-level model will be lower than that estimated by a 2-level model, thus leading to upward biases in *M* calculated based on 2-level models. The direction and magnitude of bias depend on how strongly students are clustered within classrooms.

For example, for 10th grade FCAT math performance, about 72 percent of the total variance is between students (Table FL-13a) when the classroom level is ignored. Part of this between-individual variance can be explained by between-classroom variance. As a result, when the classroom level clustering is considered, the percent of variance that is between students is reduced to 35 percent. In other words, about 51 percent of the between-individual variance is explained by between-classroom variations (1-.35/.72). Similarly, for the other three 10-grade tests examined in Florida, a large percentage of the between-individual variance is explained by between-classroom variations (between 30 and 48 percent). By contrast, in all other tests examined in this report, classroom-level variance never explains more than 30 percent of the individual-level variance estimated from 2-level models. In fact, excluding cases where End-of-Course test scores are used as the outcome, classroom-level variance never explains more than five percent of the individual-level variance.

In terms of its distribution, bias in *M* from 2-level model is negatively skewed and so the chances of having serious bias are small. Probability of a really poor estimate (severe negative bias in M) is much higher if *M* is low in a 2-level model, and much higher in secondary school (Figure 5).

Figure 5. Scatter plot of bias in *M* and *M* from 2-level models, by grade



Differences in explanatory power of covariates

The biases in *M* created by ignoring the student clustering at the classroom level do not explain all the differences between the findings of this report and those from BRB 2007. The substantially downwardly biased 2-level *M* in the Florida secondary school analyses may explain the smaller number of schools suggested in BRB 2007, but the biases do not explain why BRB 2007 suggests more schools that need to be randomized at the elementary level. This can be explained by two factors. First, the intraclass correlations at the cluster level in our study are on average lower than those found in BRB 2007. Tables NC-3a, 4a and FL-3a through 6a show that, when a 2-level model is estimated, the school-level intraclass correlations range between 0.09 and 0.14. The corresponding intraclass correlations in the five districts investigated in BRB 2007 range between 0.15 and 0.22.

Second, when covariates are used, it appears that state test scores in our study are more closely correlated across years. In the current study, individual student test scores from the previous year explain from 65 to 82 percent of the between-school variation in current test scores. They explain an additional 52 to 79 percent of the between-classroom variation and 50 to 68 percent of the within-classroom variation. By comparison, individual test scores from the previous year only explain 30 to 73 percent of the between-school variation and 22 to 52 percent of the within-school variation in BRB 2007. Because of these differences, districts in our study require fewer schools to be randomized when no covariates are used, and baseline test scores help further reduce the required number of schools by more than they do in the BRB 2007 districts.

The differences in the relationship between lagged test scores and current test scores also help explain another discrepancy between our study and BRB 2007: the effectiveness of using lagged school aggregate scores in reducing *M*. As pointed out in BRB 2007 and Bloom (2005), increasing the number of individuals per cluster often has little effect on precision. However, Raudenbush (1997) and Bloom, et al. (2007) also show that the relative importance of these two sample sizes will depend on whether covariates are used. If an individual-level covariate can reduce the cluster-level variance more than a cluster-level covariate, as is the case in the current study, using individual-level covariate should be more effective in reducing *M* for any given sample sizes.

In our report, individual-level test score lagged one year reduces the school-level variance more than school average scores from the previous year: 65 percent versus 59 percent for 5th grade math and 80 percent versus 62 percent for 5th grade reading in North Carolina, and 70 percent versus 63 percent for 5th grade FCAT math and 80 percent versus 68 percent for 5th grade FCAT reading in Florida. By contrast, in all five districts investigated in BRB 2007, individual student's lagged test scores always explain less school-level variance than lagged school aggregate scores (on average, 56 percent versus 62 percent). As a result, it is not surprising that BRB 2007 concludes that school aggregate scores from previous years are as effective as individual level lagged scores in reducing *M*, whereas we find that individual level lagged scores are always more effective than school aggregate scores.

Summary of Findings and Conclusion

The results indicate that, even with two years of individual student pretest scores to condition on in a school-randomized design, about 60 to 80 schools are required to achieve a minimum detectable effect of 0.20 with 0.80 power in secondary schools in North Carolina, while as few as 40 schools could suffice in Florida. The differences in the required number of schools arise in large part because of the lack of repeated measures of student performance in subject-specific tests in North Carolina secondary schools.

At the elementary school level, about 20 schools are required for studies on student reading performance and 40 schools for studies on student math performance in both states while controlling for individual student pretest scores. Without individual student pretest scores, the required sample sizes are larger, typically requiring 20 more schools to be randomized when lagged school average scores are controlled for. School average scores do capture a substantial fraction of pretreatment variability, since student scores vary across schools in addition to varying within school. In practice, there is a tradeoff between randomizing more schools when only school average scores from baseline years are available, and the higher cost and challenge in obtaining individual level pretest scores.

The minimum detectable effect size estimates we present for designs randomizing across 40 or 80 schools exhibit considerable positive skewness, which suggests that in the majority of settings the minimum detectable effect will be substantially below the mean, but in a few cases, the minimum detectable effect will be surprisingly high. Presumably, heterogeneity across districts that we do not model explicitly explains some of this "right tail" in estimated minimum detectable effect. However, it is important to bear in mind that a small group of districts may exhibit surprisingly large minimum detectable effects, perhaps due to increased clustering at the school level, so a post hoc power analysis may be in order using pretest scores where possible to rule out the most extreme right tail outcomes.

Taking into consideration the clustering of students within classrooms is important in some samples and not others, which may reflect heterogeneity in the degree to which students are "tracked" by ability level, since tracking would greatly increase the within-classroom clustering of student scores. This is an important factor to consider when designing randomization studies especially at the secondary school level, where student-teacher sorting is more frequent than it is at the elementary school level.

Finally, we find the type of test has some implications for designing a clustered randomization study. Specifically, a student performance measure based on norm-referenced assessments may require fewer schools to be randomized as compared with an outcome measure that is based on criteria-referenced assessments. However, the differences are small. It should be noted that the differences between the NRT and the CRT may arise not only from differences in test properties but also from the fact that they are meant for different policy use in Florida. Had the NRT test been used to make high-stakes decisions instead of the CRT test, our finding might have been different and therefore it should not be immediately generalized to other contexts.

References

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5): 547–56.

Bloom, Howard S. 2005. "Randomizing Groups to Evaluate Place-based Programs." In *Learning More from Social Experiments: Evolving Analytic Approaches* (115–72), edited by Howard S. Bloom. New York: Russell Sage.

Bloom, Howard S., Johannes M. Bos, and Suk-Won Lee. 1999. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs." *Evaluation Review* 23(4): 445–69.

Bloom, Howard S., Lashawn Richburg-Hayes, and Alison R. Black. 2007. "Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis* 29(1): 30–59.

Hedges, Larry V., and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29(1): 60–87.

Raudenbush, S. W. 1997. "Statistical Analysis and Optimal Design for Group Randomized Trials." *Psychological Methods* 2(2): 173–85.

Schochet, Peter Z. 2008. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics* 33(1): 62–87.

Schochet, Peter Z. 2009. "Statistical Power for Regression Discontinuity Designs in Education Evaluations." *Journal of Educational and Behavioral Statistics* 34(2): 238–66.

Tables

Table NC-1. Number of districts, average number of schools per district, average number of classrooms per school, and average number of students per classroom: by test subject, North Carolina 5th grade

	MATH	READING
No. of district-year	143	73
Avg # of schools per district	20	22
Avg # of classes per school	3	3
Avg # of students per class	17	17

Note: Numbers are based on districts for which all models have converged.

Table NC-2. Intraclass correlations and R-squared at various levels, by test subject and model, North Carolina 5th grade

	MATH	READING
Intraclass correlation (no covariates)		
School level	0.114	0.098
Classroom level	0.070	0.057
R-squared		
School level		
ylag1	0.647	0.797
ylag2	0.588	0.695
ylag12	0.686	0.825
sylag1	0.587	0.623
sylag2	0.535	0.560
sylag12	0.661	0.636
x	0.446	0.540
xylag1	0.652	0.827
xsylag1	0.665	0.689
Classroom level		
ylag1	0.525	0.649
ylag2	0.464	0.587
ylag12	0.529	0.655
sylag1	0.008	-0.030
sylag2	-0.006	-0.001
sylag12	0.004	-0.012
x	0.197	0.317
xylag1	0.528	0.662
xsylag1	0.204	0.294
Student level		
ylag1	0.680	0.580
ylag2	0.608	0.508
ylag12	0.725	0.629
sylag1	0.000	0.000
sylag2	0.000	0.000
sylag12	0.000	0.000
x	0.146	0.140
xylag1	0.687	0.591
xsylag1	0.146	0.140

Note: Numbers are based on districts for which all models have converged.
Model specifications:
ylag1: individual student scores lagged 1 year
ylag2: individual student scores lagged 2 years
ylag12: individual student scores lagged 1 and 2 years
sylag1: Mean school scores for the same grade lagged 1 year
sylag2: Mean school scores for the same grade lagged 2 years
sylag12: Mean school scores for the same grade lagged 1 and 2 years
x: race, gender, FRPL, LEP, age and grade repetition
xylag1: all demographic variables and individual student scores lagged 1 year
xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-3. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina 5th-grade End-of-Grade (EOG) math, school years 1999-2000 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.48	0.46	0.45	0.34	0.33	0.32	0.28	0.27	0.26
ylag1	0.25	0.24	0.24	**0.18**	**0.17**	**0.17**	**0.15**	**0.14**	**0.14**
ylag2	0.29	0.28	0.27	**0.20**	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**
ylag12	0.24	0.23	0.22	**0.17**	**0.16**	**0.16**	**0.14**	**0.13**	**0.13**
sylag1	0.34	0.32	0.30	0.24	0.22	0.21	**0.19**	**0.18**	**0.17**
sylag2	0.35	0.33	0.31	0.25	0.23	0.22	**0.20**	**0.19**	**0.18**
sylag12	0.32	0.30	0.28	0.23	0.21	**0.20**	**0.19**	**0.17**	**0.16**
x	0.37	0.35	0.34	0.26	0.25	0.24	0.21	**0.20**	**0.19**
xylag1	0.25	0.24	0.23	**0.18**	**0.17**	**0.16**	**0.14**	**0.14**	**0.13**
xsylag1	0.30	0.28	0.27	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP, age and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-4. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina 5th-grade End-of-Grade (EOG) reading, school years 1999-2000 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.44	0.43	0.42	0.31	0.30	0.30	0.26	0.25	0.24
ylag1	**0.20**	**0.19**	**0.19**	**0.14**	**0.14**	**0.13**	**0.12**	**0.11**	**0.11**
ylag2	0.23	0.22	0.22	**0.16**	**0.16**	**0.15**	**0.13**	**0.13**	**0.12**
ylag12	**0.18**	**0.18**	**0.17**	**0.13**	**0.12**	**0.12**	**0.11**	**0.10**	**0.10**
sylag1	0.31	0.29	0.28	0.22	0.21	**0.20**	**0.18**	**0.17**	**0.16**
sylag2	0.32	0.30	0.29	0.23	0.21	**0.20**	**0.19**	**0.17**	**0.17**
sylag12	0.30	0.28	0.26	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**
x	0.31	0.30	0.28	0.22	0.21	**0.20**	**0.18**	**0.17**	**0.16**
xylag1	**0.18**	**0.18**	**0.17**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**	**0.10**
xsylag1	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.15**	**0.15**	**0.14**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP, age and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-5. Variation of minimum detectable effect size (MDES), by 5th grade test subject (J=40, K=10, N=6): North Carolina

	MATH			READING		
Model	p5	p95	Range	p5	p95	Range
unconditional	0.21	0.44	0.23	0.18	0.41	0.23
ylag1	0.10	0.25	0.15	0.08	0.20	0.12
ylag2	0.11	0.29	0.18	0.09	0.22	0.13
ylag12	0.09	0.24	0.15	0.07	0.18	0.11
sylag1	0.13	0.34	0.21	0.13	0.27	0.14
sylag2	0.15	0.32	0.17	0.13	0.30	0.17
sylag12	0.13	0.30	0.16	0.13	0.29	0.16
x	0.16	0.33	0.17	0.14	0.26	0.12
xylag1	0.10	0.25	0.15	0.08	0.18	0.10
xsylag1	0.12	0.28	0.16	0.11	0.22	0.11

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP, age and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-3a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: North Carolina 5th-grade End-of-Grade (EOG) math, school years 1999-2000 through 2005-06

Intra-class correlation and model	3-level model *J=40, K=3, N=20*	3-level model *J=40, K=5, N=12*	3-level model *J=40, K=6, N=10*	2-level model *J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.115		0.139
Classroom level		0.069		--
MDES by model specification				
unconditional	0.34	0.33	0.32	0.34
ylag1	**0.18**	**0.17**	**0.17**	**0.18**
ylag2	**0.20**	**0.20**	**0.19**	0.21
ylag12	**0.17**	**0.16**	**0.16**	**0.17**
sylag1	0.24	0.22	0.21	0.24
sylag2	0.25	0.23	0.22	0.25
sylag12	0.23	0.21	**0.20**	0.23
x	0.26	0.25	0.24	0.26
xylag1	**0.18**	**0.17**	**0.16**	**0.18**
xsylag1	0.22	**0.20**	**0.19**	0.22

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 139 district-years for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP, age and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-4a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized:: Florida 5th-grade End-of-Grade (EOG) reading, school years 1999-2000 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.098		0.116
Classroom level		0.051		--
MDES by model specification				
unconditional	0.31	0.30	0.30	0.31
ylag1	**0.14**	**0.14**	**0.13**	**0.14**
ylag2	**0.16**	**0.16**	**0.15**	**0.16**
ylag12	**0.13**	**0.13**	**0.12**	**0.13**
sylag1	0.22	0.21	**0.20**	0.22
sylag2	0.22	0.21	**0.20**	0.23
sylag12	0.21	**0.20**	**0.19**	0.21
x	0.22	0.21	**0.20**	0.22
xylag1	**0.19**	**0.18**	**0.17**	**0.19**
xsylag1	**0.19**	**0.18**	**0.17**	**0.19**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 69 district-years for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP, age and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table NC-6. Number of districts, average number of schools per district, average number of classrooms per school, and average number of students per classroom: by End-of-Course (EOC) test subject, North Carolina

	ALGEBRA II	BIOLOGY	CHEMISTRY	GEOMETRY
No. of district-year	37	32	19	29
Avg # of schools per district	8	8	10	11
Avg # of classes per school	4	3	3	3
Avg # of students per class	19	18	18	19

Note: Numbers are based on districts for which all models have converged

Table NC-7. Intraclass correlations and R-squared at various levels, by End-of-Course (EOC) test subject and model, North Carolina

	ALGEBRA II	BIOLOGY	CHEMISTRY	GEOMETRY
Intraclass correlation (no covariates)				
School level	0.231	0.196	0.208	0.284
Classroom level	0.299	0.218	0.265	0.237
R-squared				
School level				
ylag1	0.595	0.597	0.519	0.756
ylag2	0.472	0.683	0.628	0.745
ylag12	0.628	0.730	0.621	0.840
sylag1	0.472	0.447	0.737	0.588
sylag2	0.345	0.479	0.542	0.484
sylag12	0.477	0.565	0.617	0.557
x	0.302	0.467	0.345	0.434
xylag1	0.638	0.691	0.587	0.798
xsylag1	0.609	0.620	0.762	0.682
Classroom level				
ylag1	0.655	0.690	0.524	0.704
ylag2	0.686	0.787	0.638	0.777
ylag12	0.771	0.814	0.668	0.820
sylag1	0.004	0.007	-0.029	0.028
sylag2	0.005	0.024	-0.005	0.026
sylag12	0.004	0.032	-0.002	0.038
x	0.417	0.322	0.235	0.436
xylag1	0.741	0.740	0.566	0.773
xsylag1	0.413	0.321	0.214	0.442
Student level				
ylag1	0.307	0.286	0.251	0.386
ylag2	0.292	0.350	0.298	0.454
ylag12	0.372	0.388	0.336	0.512
sylag1	0.000	0.000	0.000	0.000
sylag2	0.000	0.000	0.000	0.000
sylag12	0.000	0.000	0.000	0.000
x	0.091	0.098	0.066	0.143
xylag1	0.331	0.320	0.269	0.426
xsylag1	0.092	0.098	0.066	0.143

Note: Numbers are based on districts for which all models have converged
Model specifications:
ylag1: individual student Algebra 1 scores
ylag2: individual student 8th-grade math scores
ylag12: individual student Algebra 1 and 8th-grade math scores
sylag1: Mean school scores on the same outcome subject lagged 1 year
sylag2: Mean school scores on the same outcome subject lagged 2 years
sylag12: Mean school scores on the same outcome subject lagged 1 and 2 years
x: race, gender, LEP, grade repetition and age
xylag1: all demographic variables and individual student Algegra 1 scores
xsylag1: all demographic variables and mean school scores on the same outcome subject lagged 1year

Table NC-8. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina End-of-Course (EOC) algebra II, school years 2000-01 through 2005-06

	J=40			*J=60*			*J=80*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.51	0.47	0.45	0.41	0.39	0.36	0.36	0.34	0.32
ylag1	0.31	0.29	0.28	0.25	0.24	0.23	0.22	0.21	**0.19**
ylag2	0.33	0.31	0.30	0.27	0.25	0.24	0.23	0.22	0.21
ylag12	0.28	0.26	0.25	0.23	0.21	**0.20**	**0.20**	**0.19**	**0.18**
sylag1	0.40	0.36	0.32	0.33	0.29	0.26	0.29	0.25	0.23
sylag2	0.44	0.40	0.36	0.36	0.32	0.29	0.31	0.28	0.25
sylag12	0.40	0.36	0.32	0.33	0.29	0.26	0.28	0.25	0.23
x	0.41	0.39	0.37	0.34	0.32	0.30	0.29	0.27	0.26
xylag1	0.28	0.27	0.26	0.23	0.22	0.21	**0.20**	**0.19**	**0.18**
xsylag1	0.33	0.30	0.28	0.27	0.25	0.23	0.24	0.21	**0.20**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student Algebra 1 scores

ylag2: individual student 8th-grade math scores

ylag12: individual student Algebra 1 and 8th-grade math scores

sylag1: Mean school scores on Algebra 2 lagged 1 year

sylag2: Mean school scores on Algebra 2 lagged 2 years

sylag12: Mean school scores on Algebra 2 lagged 1 and 2 years

x: race, gender, LEP, grade repetition and age

xylag1: all demographic variables and individual student Algegra 1 scores

xsylag1: all demographic variables and mean school scores on Algebra 2 lagged 1year

Table NC-9. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina End-of-Course (EOC) biology, school years 2000-01 through 2005-06

	J=40			*J=60*			*J=80*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.46	0.44	0.42	0.38	0.36	0.34	0.33	0.31	0.30
ylag1	0.29	0.27	0.26	0.23	0.22	0.21	**0.20**	**0.19**	**0.18**
ylag2	0.25	0.24	0.23	0.21	**0.20**	**0.19**	**0.18**	**0.17**	**0.17**
ylag12	0.23	0.23	0.22	**0.19**	**0.18**	**0.18**	**0.17**	**0.16**	**0.15**
sylag1	0.35	0.32	0.29	0.29	0.26	0.23	0.25	0.22	**0.20**
sylag2	0.36	0.33	0.30	0.29	0.27	0.24	0.26	0.23	0.21
sylag12	0.34	0.30	0.27	0.28	0.25	0.22	0.24	0.22	**0.19**
x	0.35	0.33	0.31	0.29	0.27	0.25	0.25	0.23	0.22
xylag1	0.25	0.24	0.23	**0.20**	**0.19**	**0.19**	**0.18**	**0.17**	**0.16**
xsylag1	0.30	0.27	0.25	0.24	0.22	**0.20**	0.21	**0.19**	**0.18**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student Algebra 1 scores

ylag2: individual student 8th-grade math scores

ylag12: individual student Algebra 1 and 8th-grade math scores

sylag1: Mean school scores on Biology lagged 1 year

sylag2: Mean school scores on Biology lagged 2 years

sylag12: Mean school scores on Biology lagged 1 and 2 years

x: race, gender, LEP, grade repetition and age

xylag1: all demographic variables and individual student Algegra 1 scores

xsylag1: all demographic variables and mean school scores on Biology lagged 1year

Table NC-10. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina End-of-Course (EOC) chemistry, school years 2000-01 through 2005-06

	J=40			*J=60*			*J=80*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.48	0.45	0.43	0.39	0.37	0.35	0.34	0.32	0.30
ylag1	0.32	0.30	0.29	0.26	0.25	0.23	0.23	0.21	**0.20**
ylag2	0.29	0.27	0.25	0.23	0.22	0.21	**0.20**	**0.19**	**0.18**
ylag12	0.27	0.26	0.24	0.22	0.21	**0.20**	**0.19**	**0.18**	**0.17**
sylag1	0.35	0.30	0.26	0.29	0.25	0.22	0.25	0.22	**0.19**
sylag2	0.37	0.33	0.29	0.30	0.27	0.24	0.26	0.23	0.21
sylag12	0.37	0.33	0.29	0.30	0.27	0.24	0.26	0.23	0.21
x	0.41	0.38	0.35	0.33	0.31	0.29	0.29	0.27	0.25
xylag1	0.30	0.28	0.27	0.25	0.23	0.22	0.21	**0.20**	**0.19**
xsylag1	0.32	0.28	0.24	0.26	0.23	**0.20**	0.22	**0.20**	**0.17**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student Algebra 1 scores

ylag2: individual student 8th-grade math scores

ylag12: individual student Algebra 1 and 8th-grade math scores

sylag1: Mean school scores on Chemistry lagged 1 year

sylag2: Mean school scores on Chemistry lagged 2 years

sylag12: Mean school scores on Chemistry lagged 1 and 2 years

x: race, gender, LEP, grade repetition and age

xylag1: all demographic variables and individual student Algegra 1 scores

xsylag1: all demographic variables and mean school scores on Chemistry lagged 1year

Table NC-11. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: North Carolina End-of-Course (EOC) geometry, school years 2000-01 through 2005-06

	J=40			*J=60*			*J=80*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.53	0.51	0.48	0.43	0.41	0.39	0.38	0.36	0.34
ylag1	0.27	0.26	0.24	0.22	0.21	**0.20**	**0.19**	**0.18**	**0.17**
ylag2	0.23	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.17**	**0.16**	**0.15**
ylag12	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**	**0.15**	**0.14**	**0.13**	**0.13**
sylag1	0.37	0.33	0.30	0.30	0.27	0.25	0.26	0.24	0.21
sylag2	0.40	0.37	0.34	0.33	0.30	0.27	0.28	0.26	0.24
sylag12	0.36	0.33	0.29	0.30	0.27	0.24	0.26	0.23	0.21
x	0.38	0.36	0.35	0.31	0.30	0.28	0.27	0.26	0.24
xylag1	0.24	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.17**	**0.16**	**0.15**
xsylag1	0.30	0.27	0.25	0.24	0.22	**0.20**	0.21	**0.19**	**0.18**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student Algebra 1 scores

ylag2: individual student 8th-grade math scores

ylag12: individual student Algebra 1 and 8th-grade math scores

sylag1: Mean school scores on Geometry lagged 1 year

sylag2: Mean school scores on Geometry lagged 2 years

sylag12: Mean school scores on Geometry lagged 1 and 2 years

x: race, gender, LEP, grade repetition and age

xylag1: all demographic variables and individual student Algegra 1 scores

xsylag1: all demographic variables and mean school scores on Geometry lagged 1year

Table NC-12. Variation of minimum detectable effect size (MDES), by End-of-Course (EOC) test subject (J=80, K=10, N=6): North Carolina

	ALGEBRA II			BIOLOGY			CHEMISTRY			GEOMETRY		
Model	p5	p95	Range	p5	p95	Range	p5	p95	Range	p5	p95	Range
unconditional	0.22	0.47	0.25	0.23	0.35	0.12	0.19	0.42	0.23	0.20	0.47	0.27
ylag1	0.09	0.30	0.21	0.11	0.29	0.18	0.14	0.32	0.18	0.08	0.26	0.18
ylag2	0.13	0.34	0.21	0.09	0.25	0.17	0.10	0.25	0.15	0.07	0.24	0.18
ylag12	0.09	0.28	0.19	0.09	0.24	0.15	0.11	0.25	0.14	0.05	0.19	0.14
sylag1	0.11	0.38	0.26	0.10	0.33	0.24	0.11	0.32	0.21	0.11	0.44	0.33
sylag2	0.12	0.49	0.37	0.10	0.32	0.22	0.11	0.29	0.18	0.12	0.42	0.30
sylag12	0.11	0.40	0.29	0.11	0.31	0.20	0.12	0.32	0.19	0.10	0.47	0.37
x	0.14	0.43	0.29	0.14	0.33	0.19	0.11	0.38	0.27	0.08	0.36	0.28
xylag1	0.08	0.31	0.22	0.09	0.27	0.18	0.14	0.32	0.19	0.06	0.24	0.18
xsylag1	0.10	0.35	0.25	0.09	0.28	0.18	0.11	0.35	0.24	0.08	0.44	0.36

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student Algebra 1 scores

ylag2: individual student 8th-grade math scores

ylag12: individual student Algebra 1 and 8th-grade math scores

sylag1: Mean school scores on the same outcome subject lagged 1 year

sylag2: Mean school scores on the same outcome subject lagged 2 years

sylag12: Mean school scores on the same outcome subject lagged 1 and 2 years

x: race, gender, LEP, grade repetition and age

xylag1: all demographic variables and individual student Algegra 1 scores

xsylag1: all demographic variables and mean school scores on the same outcome subject lagged 1year

Table NC-8a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 80 schools randomized: North Carolina End-of-Course (EOC) algebra II, school years 2000-01 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=80, K=3, N=20*	*J=80, K=5, N=12*	*J=80, K=6, N=10*	*J=80, N=60*
Intra-class correlation (no covariates)				
School level		0.228		0.334
Classroom level		0.309		--
MDES by model specification				
unconditional	0.36	0.34	0.32	0.36
ylag1	0.22	0.21	**0.20**	**0.20**
ylag2	0.23	0.22	0.21	0.21
ylag12	**0.20**	**0.19**	**0.18**	**0.18**
sylag1	0.30	0.27	0.24	0.30
sylag2	0.31	0.28	0.26	0.31
sylag12	0.30	0.27	0.24	0.30
x	0.30	0.28	0.27	0.26
xylag1	**0.20**	**0.19**	**0.18**	**0.18**
xsylag1	0.24	0.22	**0.20**	0.23

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 31 district-years for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student Algebra 1 scores
ylag2: individual student 8th-grade math scores
ylag12: individual student Algebra 1 and 8th-grade math scores
sylag1: Mean school scores on Algebra 2 lagged 1 year
sylag2: Mean school scores on Algebra 2 lagged 2 years
sylag12: Mean school scores on Algebra 2 lagged 1 and 2 years
x: race, gender, LEP, grade repetition and age
xylag1: all demographic variables and individual student Algegra 1 scores
xsylag1: all demographic variables and mean school scores on Algebra 2 lagged 1year

Table NC-9a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 80 schools randomized: North Carolina End-of-Course (EOC) biology, school years 2000-01 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=80, K=3, N=20*	*J=80, K=5, N=12*	*J=80, K=6, N=10*	*J=80, N=60*
Intra-class correlation (no covariates)				
School level		0.191		0.273
Classroom level		0.216		--
MDES by model specification				
unconditional	0.33	0.31	0.29	0.33
ylag1	**0.20**	**0.19**	**0.18**	**0.19**
ylag2	**0.17**	**0.17**	**0.16**	**0.16**
ylag12	**0.16**	**0.16**	**0.15**	**0.16**
sylag1	0.25	0.22	**0.20**	0.25
sylag2	0.26	0.24	0.22	0.26
sylag12	0.24	0.22	**0.20**	0.24
x	0.25	0.23	0.22	0.23
xylag1	**0.18**	**0.17**	**0.16**	**0.16**
xsylag1	0.21	**0.19**	**0.17**	**0.19**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 29 district-years for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student Algebra 1 scores
ylag2: individual student 8th-grade math scores
ylag12: individual student Algebra 1 and 8th-grade math scores
sylag1: Mean school scores on Biology lagged 1 year
sylag2: Mean school scores on Biology lagged 2 years
sylag12: Mean school scores on Biology lagged 1 and 2 years
x: race, gender, LEP, grade repetition and age
xylag1: all demographic variables and individual student Algegra 1 scores
xsylag1: all demographic variables and mean school scores on Biology lagged 1year

Table NC-10a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 80 schools randomized: North Carolina End-of-Course (EOC) chemistry, school years 2000-01 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=80, K=3, N=20*	*J=80, K=5, N=12*	*J=80, K=6, N=10*	*J=80, N=60*
Intra-class correlation (no covariates)				
School level		0.208		0.291
Classroom level		0.265		--
MDES by model specification				
unconditional	0.34	0.32	0.30	0.34
ylag1	0.23	0.21	**0.20**	0.22
ylag2	**0.20**	**0.19**	**0.18**	**0.19**
ylag12	**0.19**	**0.18**	**0.17**	**0.18**
sylag1	0.25	0.22	**0.19**	0.27
sylag2	0.26	0.23	0.21	0.26
sylag12	0.26	0.23	0.21	0.27
x	0.29	0.27	0.25	0.27
xylag1	0.21	**0.20**	**0.19**	**0.20**
xsylag1	0.22	**0.20**	**0.17**	0.23

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 19 district-years for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student Algebra 1 scores
ylag2: individual student 8th-grade math scores
ylag12: individual student Algebra 1 and 8th-grade math scores
sylag1: Mean school scores on Chemistry lagged 1 year
sylag2: Mean school scores on Chemistry lagged 2 years
sylag12: Mean school scores on Chemistry lagged 1 and 2 years
x: race, gender, LEP, grade repetition and age
xylag1: all demographic variables and individual student Algegra 1 scores
xsylag1: all demographic variables and mean school scores on Chemistry lagged 1year

Table NC-11a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 80 schools randomized: North Carolina End-of-Course (EOC) geometry, school years 2000-01 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=80, K=3, N=20*	*J=80, K=5, N=12*	*J=80, K=6, N=10*	*J=80, N=60*
Intra-class correlation (no covariates)				
School level		0.283		0.387
Classroom level		0.247		--
MDES by model specification				
unconditional	0.38	0.36	0.34	0.39
ylag1	**0.20**	**0.19**	**0.18**	**0.18**
ylag2	**0.17**	**0.16**	**0.16**	**0.16**
ylag12	**0.14**	**0.14**	**0.13**	**0.13**
sylag1	0.26	0.24	0.21	0.26
sylag2	0.28	0.26	0.24	0.28
sylag12	0.26	0.24	0.21	0.25
x	0.28	0.27	0.25	0.25
xylag1	**0.17**	**0.16**	**0.16**	**0.16**
xsylag1	0.22	**0.20**	**0.18**	**0.20**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 27 district-years for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student Algebra 1 scores
ylag2: individual student 8th-grade math scores
ylag12: individual student Algebra 1 and 8th-grade math scores
sylag1: Mean school scores on Geometry lagged 1 year
sylag2: Mean school scores on Geometry lagged 2 years
sylag12: Mean school scores on Geometry lagged 1 and 2 years
x: race, gender, LEP, grade repetition and age
xylag1: all demographic variables and individual student Algegra 1 scores
xsylag1: all demographic variables and mean school scores on Geometry lagged 1year

Table FL-1. Number of districts, average number of schools per district, average number of classrooms per school, and average number of students per classroom: by test subject, Florida 5th grade

	FCAT-MATH	NRT-MATH	FCAT-READING	NRT-READING
No. of district-year	37	34	22	26
Avg # of schools per district	39	35	47	39
Avg # of classes per school	4	4	5	4
Avg # of students per class	18	18	18	18

Note: Numbers are based on districts for which all models have converged.

Table FL-2. Intraclass correlations and R-squared at various levels, by test subject and model, Florida 5th grade

	FCAT-MATH	NRT-MATH	FCAT-READING	NRT-READING
Intraclass correlation (no covariates)				
School level	0.111	0.099	0.078	0.085
Classroom level	0.066	0.064	0.053	0.040
R-squared				
School level				
ylag1	0.697	0.705	0.797	0.818
ylag2	0.597	0.582	0.792	0.792
ylag12	0.722	0.714	0.866	0.867
sylag1	0.629	0.667	0.684	0.682
sylag2	0.587	0.670	0.721	0.683
sylag12	0.657	0.704	0.776	0.767
x	0.428	0.423	0.604	0.551
xylag1	0.712	0.724	0.838	0.855
xsylag1	0.638	0.674	0.777	0.761
Classroom level				
ylag1	0.592	0.550	0.791	0.616
ylag2	0.539	0.508	0.748	0.666
ylag12	0.606	0.561	0.811	0.692
sylag1	0.013	0.011	0.028	0.003
sylag2	0.001	-0.006	0.017	0.003
sylag12	0.007	0.000	0.027	0.005
x	0.145	0.109	0.222	0.286
xylag1	0.599	0.550	0.792	0.663
xsylag1	0.151	0.109	0.222	0.273
Student level				
ylag1	0.619	0.531	0.529	0.503
ylag2	0.547	0.479	0.450	0.476
ylag12	0.674	0.605	0.578	0.573
sylag1	0.000	0.000	0.000	0.000
sylag2	0.000	0.000	0.000	0.000
sylag12	0.000	0.000	0.000	0.000
x	0.089	0.070	0.077	0.084
xylag1	0.629	0.539	0.539	0.512
xsylag1	0.089	0.070	0.077	0.084

Note: Numbers are based on districts for which all models have converged.
Model specifications:
ylag1: individual student scores lagged 1 year
ylag2: individual student scores lagged 2 years
ylag12: individual student scores lagged 1 and 2 years
sylag1: Mean school scores for the same grade lagged 1 year
sylag2: Mean school scores for the same grade lagged 2 years
sylag12: Mean school scores for the same grade lagged 1 and 2 years
x: race, gender, FRPL, LEP and grade repetition
xylag1: all demographic variables and individual student scores lagged 1 year
xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-3. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 5th-grade FCAT math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.47	0.46	0.44	0.33	0.32	0.31	0.27	0.26	0.26
ylag1	0.24	0.23	0.22	**0.17**	**0.16**	**0.16**	**0.14**	**0.13**	**0.13**
ylag2	0.28	0.27	0.26	**0.20**	**0.19**	**0.19**	**0.16**	**0.16**	**0.15**
ylag12	0.23	0.22	0.21	**0.16**	**0.16**	**0.15**	**0.13**	**0.13**	**0.12**
sylag1	0.32	0.30	0.28	0.23	0.21	**0.20**	**0.19**	**0.17**	**0.16**
sylag2	0.33	0.31	0.29	0.23	0.22	0.21	**0.19**	**0.18**	**0.17**
sylag12	0.31	0.29	0.27	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.16**
x	0.36	0.35	0.33	0.26	0.24	0.24	0.21	**0.20**	**0.19**
xylag1	0.23	0.22	0.21	**0.16**	**0.16**	**0.15**	**0.13**	**0.13**	**0.12**
xsylag1	0.30	0.28	0.26	0.21	**0.20**	**0.18**	**0.17**	**0.16**	**0.15**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-4. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 5th-grade NRT math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.45	0.44	0.42	0.32	0.31	0.30	0.26	0.25	0.24
ylag1	0.24	0.23	0.22	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.13**
ylag2	0.28	0.27	0.26	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**	**0.15**
ylag12	0.23	0.22	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**	**0.12**
sylag1	0.31	0.29	0.27	0.22	**0.20**	**0.19**	**0.18**	**0.17**	**0.16**
sylag2	0.31	0.29	0.27	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.16**
sylag12	0.30	0.28	0.26	0.21	**0.20**	**0.18**	**0.17**	**0.16**	**0.15**
x	0.35	0.34	0.32	0.25	0.24	0.23	**0.20**	**0.19**	**0.19**
xylag1	0.23	0.22	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**	**0.12**
xsylag1	0.29	0.27	0.25	0.21	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-5. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 5th-grade FCAT reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.41	0.39	0.38	0.29	0.28	0.27	0.24	0.23	0.22
ylag1	**0.18**	**0.18**	**0.17**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**	**0.10**
ylag2	0.21	**0.20**	**0.19**	**0.15**	**0.14**	**0.14**	**0.12**	**0.12**	**0.11**
ylag12	**0.17**	**0.16**	**0.16**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**	**0.09**
sylag1	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
sylag2	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
sylag12	0.26	0.24	0.22	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
x	0.29	0.27	0.26	**0.20**	**0.19**	**0.19**	**0.17**	**0.16**	**0.15**
xylag1	**0.17**	**0.16**	**0.16**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**	**0.09**
xsylag1	0.24	0.22	0.21	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-6. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 5th-grade NRT reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.41	0.40	0.39	0.29	0.28	0.28	0.24	0.23	0.23
ylag1	**0.18**	**0.18**	**0.17**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**	**0.10**
ylag2	**0.20**	**0.19**	**0.19**	**0.14**	**0.13**	**0.13**	**0.11**	**0.11**	**0.11**
ylag12	**0.16**	**0.15**	**0.15**	**0.11**	**0.11**	**0.11**	**0.09**	**0.09**	**0.09**
sylag1	0.26	0.25	0.23	**0.19**	**0.17**	**0.17**	**0.15**	**0.14**	**0.14**
sylag2	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
sylag12	0.25	0.23	0.22	**0.18**	**0.16**	**0.15**	**0.14**	**0.13**	**0.13**
x	0.29	0.28	0.27	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.16**
xylag1	**0.17**	**0.16**	**0.16**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**	**0.09**
xsylag1	0.24	0.23	0.21	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-7. Variation of minimum detectable effect size (MDES), by 5th grade test subject (J=40, K=10, N=6): Florida

	5th grade math						5th grade reading					
	FCAT			NRT			FCAT			NRT		
Model	p5	p95	Range	p5	p95	Range	p5	p95	Range	p5	p95	Range
unconditional	0.17	0.40	0.23	0.17	0.40	0.23	0.18	0.34	0.16	0.18	0.35	0.18
ylag1	0.11	0.22	0.11	0.11	0.21	0.09	0.11	0.14	0.04	0.08	0.15	0.08
ylag2	0.13	0.27	0.14	0.14	0.25	0.11	0.10	0.17	0.07	0.09	0.16	0.07
ylag12	0.11	0.20	0.10	0.11	0.19	0.08	0.09	0.13	0.04	0.08	0.14	0.06
sylag1	0.14	0.30	0.16	0.16	0.27	0.11	0.14	0.21	0.07	0.13	0.25	0.12
sylag2	0.14	0.30	0.15	0.14	0.30	0.16	0.13	0.21	0.08	0.14	0.24	0.10
sylag12	0.14	0.35	0.20	0.14	0.32	0.18	0.13	0.21	0.08	0.13	0.20	0.07
x	0.17	0.32	0.15	0.17	0.31	0.14	0.13	0.22	0.09	0.14	0.24	0.10
xylag1	0.10	0.22	0.12	0.11	0.20	0.09	0.09	0.14	0.05	0.09	0.13	0.04
xsylag1	0.12	0.29	0.17	0.14	0.25	0.11	0.12	0.19	0.07	0.12	0.20	0.07

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-3a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 5th-grade FCAT math, school years 2004-05 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.111		0.129
Classroom level		0.066		--
MDES by model specification				
unconditional	0.33	0.32	0.31	0.33
ylag1	**0.17**	**0.16**	**0.16**	**0.17**
ylag2	**0.20**	**0.19**	**0.19**	**0.19**
ylag12	**0.16**	**0.16**	**0.15**	**0.16**
sylag1	0.23	0.21	**0.20**	0.22
sylag2	0.23	0.22	0.21	0.23
sylag12	0.22	0.21	**0.19**	0.21
x	0.26	0.24	0.24	0.25
xylag1	**0.16**	**0.16**	**0.15**	**0.16**
xsylag1	0.21	**0.20**	**0.18**	**0.20**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 37 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-4a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 5th-grade NRT math, school years 2004-05 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.099		0.117
Classroom level		0.064		--
MDES by model specification				
unconditional	0.32	0.31	0.30	0.31
ylag1	**0.17**	**0.16**	**0.15**	**0.16**
ylag2	**0.20**	**0.19**	**0.18**	**0.19**
ylag12	**0.16**	**0.15**	**0.15**	**0.16**
sylag1	0.22	**0.20**	**0.19**	0.21
sylag2	0.22	0.21	**0.19**	0.21
sylag12	0.21	**0.20**	**0.18**	**0.20**
x	0.25	0.24	0.23	0.24
xylag1	**0.16**	**0.15**	**0.15**	**0.16**
xsylag1	0.21	**0.19**	**0.18**	**0.20**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 34 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-5a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized:: Florida 5th-grade FCAT reading, school years 2004-05 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.078		0.091
Classroom level		0.053		--
MDES by model specification				
unconditional	0.29	0.28	0.27	0.28
ylag1	**0.13**	**0.13**	**0.12**	**0.13**
ylag2	**0.15**	**0.14**	**0.14**	**0.14**
ylag12	**0.12**	**0.11**	**0.11**	**0.11**
sylag1	**0.19**	**0.18**	**0.17**	**0.18**
sylag2	**0.19**	**0.18**	**0.17**	**0.18**
sylag12	**0.18**	**0.17**	**0.16**	**0.17**
x	**0.20**	**0.19**	**0.19**	**0.20**
xylag1	**0.17**	**0.16**	**0.15**	**0.16**
xsylag1	**0.17**	**0.16**	**0.15**	**0.16**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 22 districts for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student scores lagged 1 year
ylag2: individual student scores lagged 2 years
ylag12: individual student scores lagged 1 and 2 years
sylag1: Mean school scores for the same grade lagged 1 year
sylag2: Mean school scores for the same grade lagged 2 years
sylag12: Mean school scores for the same grade lagged 1 and 2 years
x: race, gender, FRPL, LEP and grade repetition
xylag1: all demographic variables and individual student scores lagged 1 year
xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-6a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 5th-grade NRT reading, school years 2004-05 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.085		0.095
Classroom level		0.040		--
MDES by model specification				
unconditional	0.29	0.28	0.28	0.29
ylag1	**0.13**	**0.13**	**0.12**	**0.13**
ylag2	**0.14**	**0.13**	**0.13**	**0.14**
ylag12	**0.11**	**0.11**	**0.11**	**0.11**
sylag1	**0.19**	**0.17**	**0.17**	**0.18**
sylag2	**0.19**	**0.18**	**0.17**	**0.18**
sylag12	**0.18**	**0.16**	**0.15**	**0.17**
x	0.21	**0.20**	**0.19**	**0.20**
xylag1	**0.12**	**0.11**	**0.11**	**0.12**
xsylag1	**0.17**	**0.16**	**0.15**	**0.16**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom
Estimates are based on 26 districts for which both the 3-level and the 2-level models have converged for all model specifications.
Model specifications:
ylag1: individual student scores lagged 1 year
ylag2: individual student scores lagged 2 years
ylag12: individual student scores lagged 1 and 2 years
sylag1: Mean school scores for the same grade lagged 1 year
sylag2: Mean school scores for the same grade lagged 2 years
sylag12: Mean school scores for the same grade lagged 1 and 2 years
x: race, gender, FRPL, LEP and grade repetition
xylag1: all demographic variables and individual student scores lagged 1 year
xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-8. Compare test types--Number of districts, average number of schools per district, average number of classrooms per school, average number of students per classroom, and intraclass correlations, by test subject and type: Florida 5th grade

	5th grade math		5th grade reading	
	FCAT	NRT	FCAT	NRT
No. of district-year	29		17	
Avg # of schools per district	40		54	
Avg # of classes per school	4		4	
Avg # of students per class	17		18	
Intraclass correlation (no covariates)				
School level	0.116	0.104	0.084	0.089
Classroom level	0.067	0.061	0.047	0.044

Note: Numbers are based on districts for which all models have converged for both FCAT and NRT test types.

Table FL-9. Compare test types--Average minimum detectable effect size (MDES), by test type, number of schools, classrooms, and students: Florida 5th-grade math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
FCAT									
unconditional	0.48	0.47	0.45	0.34	0.33	0.32	0.28	0.27	0.26
ylag1	0.25	0.24	0.23	**0.18**	**0.17**	**0.16**	**0.14**	**0.14**	**0.13**
ylag2	0.28	0.27	0.26	**0.20**	**0.19**	**0.19**	**0.16**	**0.16**	**0.15**
ylag12	0.23	0.22	0.21	**0.17**	**0.16**	**0.15**	**0.13**	**0.13**	**0.12**
sylag1	0.33	0.31	0.29	0.23	0.22	**0.20**	**0.19**	**0.18**	**0.17**
sylag2	0.33	0.31	0.29	0.24	0.22	0.21	**0.19**	**0.18**	**0.17**
sylag12	0.32	0.30	0.28	0.22	0.21	**0.20**	**0.18**	**0.17**	**0.16**
x	0.36	0.34	0.33	0.25	0.24	0.23	0.21	**0.20**	**0.19**
xylag1	0.24	0.23	0.22	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.13**
xsylag1	0.30	0.28	0.26	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**
NRT									
unconditional	0.46	0.45	0.43	0.33	0.31	0.31	0.27	0.26	0.25
ylag1	0.24	0.23	0.22	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.13**
ylag2	0.28	0.27	0.26	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**	**0.15**
ylag12	0.22	0.21	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**	**0.12**
sylag1	0.31	0.29	0.27	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.16**
sylag2	0.31	0.29	0.28	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.16**
sylag12	0.30	0.28	0.26	0.22	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**
x	0.35	0.33	0.32	0.25	0.24	0.23	**0.20**	**0.19**	**0.19**
xylag1	0.23	0.22	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**	**0.12**
xsylag1	0.29	0.27	0.25	**0.20**	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Numbers are based on districts for which all models have converged for both FCAT and NRT test types.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-10. Compare test types--Average minimum detectable effect size (MDES), by test type, number of schools, classrooms, and students: Florida 5th-grade reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
FCAT									
unconditional	0.42	0.41	0.40	0.30	0.29	0.28	0.24	0.23	0.23
ylag1	**0.19**	**0.18**	**0.18**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**	**0.10**
ylag2	0.21	**0.20**	**0.20**	**0.15**	**0.14**	**0.14**	**0.12**	**0.12**	**0.11**
ylag12	**0.17**	**0.16**	**0.16**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**	**0.09**
sylag1	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
sylag2	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
sylag12	0.26	0.24	0.23	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
x	0.29	0.28	0.27	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.16**
xylag1	**0.17**	**0.16**	**0.16**	**0.12**	**0.12**	**0.11**	**0.10**	**0.09**	**0.09**
xsylag1	0.24	0.23	0.21	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**
NRT									
unconditional	0.42	0.41	0.40	0.30	0.29	0.29	0.24	0.24	0.23
ylag1	**0.19**	**0.18**	**0.17**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**	**0.10**
ylag2	**0.20**	**0.19**	**0.19**	**0.14**	**0.14**	**0.13**	**0.12**	**0.11**	**0.11**
ylag12	**0.16**	**0.16**	**0.15**	**0.11**	**0.11**	**0.11**	**0.09**	**0.09**	**0.09**
sylag1	0.26	0.24	0.23	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
sylag2	0.26	0.25	0.23	**0.19**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
sylag12	0.25	0.23	0.22	**0.18**	**0.17**	**0.15**	**0.15**	**0.13**	**0.13**
x	0.29	0.28	0.27	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.16**
xylag1	**0.17**	**0.16**	**0.15**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**	**0.09**
xsylag1	0.23	0.22	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.13**	**0.12**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Numbers are based on districts for which all models have converged for both FCAT and NRT test types.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-11. Number of districts, average number of schools per district, average number of classrooms per school, and average number of students per classroom: by test subject, Florida 10th grade

	FCAT-MATH	NRT-MATH	FCAT-READING	NRT-READING
No. of district-year	21	12	19	19
Avg # of schools per district	19	20	22	18
Avg # of classes per school	28	28	19	17
Avg # of students per class	8	8	10	10

Note: Numbers are based on districts for which all models have converged

Table FL-12. Intraclass correlations and R-squared at various levels, by test subject and model, Florida 10th grade

	FCAT-MATH	NRT-MATH	FCAT-READING	NRT-READING
Intraclass correlation (no covariates)				
School level	0.214	0.114	0.203	0.121
Classroom level	0.435	0.446	0.331	0.286
R-squared				
School level				
ylag1	0.788	0.832	0.802	0.790
ylag2	0.765	0.803	0.724	0.775
ylag12	0.843	0.883	0.855	0.873
sylag1	0.722	0.825	0.795	0.748
sylag2	0.667	0.803	0.814	0.673
sylag12	0.810	0.829	0.839	0.802
x	0.193	0.247	0.325	0.318
xylag1	0.796	0.836	0.861	0.823
xsylag1	0.745	0.842	0.846	0.768
Classroom level				
ylag1	0.892	0.853	0.882	0.839
ylag2	0.888	0.808	0.897	0.839
ylag12	0.946	0.911	0.956	0.926
sylag1	0.002	0.011	0.005	0.014
sylag2	-0.004	0.009	0.007	0.010
sylag12	0.003	0.011	0.009	0.007
x	0.051	0.065	0.201	0.212
xylag1	0.888	0.853	0.896	0.863
xsylag1	0.051	0.072	0.206	0.217
Student level				
ylag1	0.373	0.287	0.309	0.268
ylag2	0.347	0.256	0.308	0.223
ylag12	0.444	0.351	0.396	0.324
sylag1	-0.001	0.000	-0.001	-0.001
sylag2	-0.001	0.000	-0.001	-0.001
sylag12	0.000	0.000	-0.001	-0.001
x	0.058	0.034	0.026	0.020
xylag1	0.381	0.293	0.320	0.274
xsylag1	0.058	0.034	0.025	0.020

Note: Numbers are based on districts for which all models have converged

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-13. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 10th-grade FCAT math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.74	0.67	0.61	0.52	0.47	0.43	0.43	0.39	0.35
ylag1	0.31	0.28	0.27	0.22	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**
ylag2	0.31	0.29	0.28	0.22	0.21	**0.19**	**0.18**	**0.17**	**0.16**
ylag12	0.24	0.23	0.22	**0.17**	**0.16**	**0.16**	**0.14**	**0.13**	**0.13**
sylag1	0.53	0.43	0.34	0.37	0.31	0.24	0.30	0.25	**0.20**
sylag2	0.52	0.42	0.33	0.37	0.30	0.23	0.30	0.24	**0.19**
sylag12	0.51	0.42	0.33	0.36	0.30	0.23	0.30	0.24	**0.19**
x	0.69	0.62	0.56	0.49	0.44	0.39	0.40	0.36	0.32
xylag1	0.30	0.28	0.26	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**
xsylag1	0.51	0.42	0.33	0.36	0.30	0.24	0.30	0.24	**0.19**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-14. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 10th-grade NRT math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.65	0.57	0.50	0.46	0.40	0.35	0.37	0.33	0.29
ylag1	0.26	0.23	0.21	**0.18**	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**
ylag2	0.29	0.26	0.23	**0.20**	**0.18**	**0.16**	**0.17**	**0.15**	**0.13**
ylag12	0.21	**0.19**	**0.17**	**0.15**	**0.13**	**0.12**	**0.12**	**0.11**	**0.10**
sylag1	0.51	0.41	0.31	0.36	0.29	0.22	0.29	0.24	**0.18**
sylag2	0.51	0.41	0.32	0.36	0.29	0.22	0.30	0.24	**0.18**
sylag12	0.51	0.41	0.31	0.36	0.29	0.22	0.29	0.24	**0.18**
x	0.60	0.52	0.45	0.42	0.37	0.32	0.35	0.30	0.26
xylag1	0.25	0.22	**0.20**	**0.18**	**0.16**	**0.14**	**0.15**	**0.13**	**0.12**
xsylag1	0.49	0.40	0.30	0.35	0.28	0.21	0.28	0.23	**0.17**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-15. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 10th-grade FCAT reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.70	0.64	0.59	0.49	0.45	0.42	0.40	0.37	0.34
ylag1	0.28	0.27	0.26	**0.20**	**0.19**	**0.18**	**0.16**	**0.16**	**0.15**
ylag2	0.33	0.31	0.30	0.23	0.22	0.21	**0.19**	**0.18**	**0.17**
ylag12	0.22	0.22	0.21	**0.16**	**0.15**	**0.15**	**0.13**	**0.12**	**0.12**
sylag1	0.48	0.40	0.33	0.34	0.28	0.23	0.28	0.23	**0.19**
sylag2	0.47	0.39	0.31	0.33	0.27	0.22	0.27	0.22	**0.18**
sylag12	0.46	0.38	0.30	0.32	0.27	0.21	0.26	0.22	**0.17**
x	0.60	0.55	0.50	0.42	0.39	0.36	0.35	0.32	0.29
xylag1	0.25	0.24	0.22	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
xsylag1	0.43	0.36	0.29	0.30	0.25	0.21	0.25	0.21	**0.17**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-16. Average minimum detectable effect size (MDES), by number of schools, classrooms, and students: Florida 10th-grade NRT reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
unconditional	0.59	0.53	0.48	0.42	0.38	0.34	0.34	0.31	0.28
ylag1	0.26	0.24	0.23	**0.19**	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**
ylag2	0.27	0.25	0.23	**0.19**	**0.18**	**0.16**	**0.15**	**0.14**	**0.13**
ylag12	**0.20**	**0.19**	**0.18**	**0.14**	**0.13**	**0.13**	**0.11**	**0.11**	**0.10**
sylag1	0.44	0.36	0.30	0.31	0.26	0.21	0.25	0.21	**0.17**
sylag2	0.44	0.37	0.30	0.31	0.26	0.21	0.25	0.21	**0.17**
sylag12	0.43	0.35	0.29	0.30	0.25	**0.20**	0.25	**0.20**	**0.16**
x	0.50	0.45	0.40	0.35	0.32	0.29	0.29	0.26	0.23
xylag1	0.24	0.22	0.21	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**
xsylag1	0.40	0.33	0.27	0.28	0.23	**0.19**	0.23	**0.19**	**0.16**

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-13a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 10th-grade FCAT math, school years 2004-05 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.223		0.275
Classroom level		0.424		--
MDES by model specification				
unconditional	0.53	0.48	0.44	0.44
ylag1	0.22	**0.20**	**0.19**	**0.16**
ylag2	0.22	0.21	**0.20**	**0.17**
ylag12	**0.17**	**0.17**	**0.16**	**0.15**
sylag1	0.37	0.30	0.24	**0.19**
sylag2	0.36	0.30	0.23	0.21
sylag12	0.36	0.29	0.23	**0.17**
x	0.49	0.44	0.40	0.37
xylag1	0.22	**0.20**	**0.19**	**0.16**
xsylag1	0.36	0.29	0.23	**0.19**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 20 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-14a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 10th-grade NRT math, school years 2004-05 through 2005-06

	3-level model			2-level model
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.116		0.145
Classroom level		0.440		--
MDES by model specification				
unconditional	0.46	0.40	0.36	0.35
ylag1	**0.18**	**0.16**	**0.15**	**0.12**
ylag2	**0.20**	**0.18**	**0.16**	**0.13**
ylag12	**0.15**	**0.13**	**0.12**	**0.11**
sylag1	0.36	0.29	0.22	**0.15**
sylag2	0.36	0.29	0.22	**0.16**
sylag12	0.36	0.29	0.22	**0.15**
x	0.42	0.37	0.32	0.27
xylag1	**0.18**	**0.16**	**0.14**	**0.12**
xsylag1	0.35	0.28	0.21	**0.16**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 11 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-15a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 10th-grade FCAT reading, school years 2004-05 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.203		0.251
Classroom level		0.331		--
MDES by model specification				
unconditional	0.49	0.45	0.42	0.43
ylag1	**0.20**	**0.19**	**0.18**	**0.16**
ylag2	0.23	0.22	0.21	**0.18**
ylag12	**0.16**	**0.15**	**0.15**	**0.14**
sylag1	0.34	0.28	0.23	**0.19**
sylag2	0.33	0.27	0.22	**0.18**
sylag12	0.32	0.27	0.21	**0.16**
x	0.42	0.39	0.36	0.33
xylag1	**0.18**	**0.17**	**0.16**	**0.14**
xsylag1	0.30	0.25	0.21	**0.17**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 19 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-16a. Compare average minimum detectable effect size (MDES) in 3-level and 2-level models with 40 schools randomized: Florida 10th-grade NRT reading, school years 2004-05 through 2005-06

	3-level model			**2-level model**
Intra-class correlation and model	*J=40, K=3, N=20*	*J=40, K=5, N=12*	*J=40, K=6, N=10*	*J=40, N=60*
Intra-class correlation (no covariates)				
School level		0.125		0.153
Classroom level		0.283		--
MDES by model specification				
unconditional	0.42	0.38	0.35	0.34
ylag1	**0.18**	**0.17**	**0.16**	**0.15**
ylag2	**0.19**	**0.18**	**0.17**	**0.15**
ylag12	**0.14**	**0.13**	**0.13**	**0.12**
sylag1	0.31	0.26	0.21	**0.17**
sylag2	0.31	0.26	0.21	**0.17**
sylag12	0.30	0.25	**0.20**	**0.16**
x	0.36	0.32	0.29	0.27
xylag1	**0.17**	**0.16**	**0.15**	**0.14**
xsylag1	0.28	0.24	**0.19**	**0.16**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Estimates are based on 18 districts for which both the 3-level and the 2-level models have converged for all model specifications.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-17. Variation of minimum detectable effect size (MDES), by 10th grade test subject (J=40, K=10, N=6): Florida

	10th grade math						10th grade reading					
	FCAT			NRT			FCAT			NRT		
Model	p5	p95	Range	p5	p95	Range	p5	p95	Range	p5	p95	Range
unconditional	0.26	0.70	0.44	0.26	0.53	0.27	0.25	0.73	0.48	0.23	0.51	0.28
ylag1	0.10	0.40	0.29	0.10	0.19	0.09	0.07	0.40	0.33	0.11	0.31	0.20
ylag2	0.10	0.45	0.35	0.12	0.23	0.11	0.10	0.58	0.48	0.09	0.27	0.18
ylag12	0.08	0.37	0.29	0.09	0.17	0.08	0.08	0.47	0.38	0.09	0.20	0.11
sylag1	0.17	0.40	0.22	0.19	0.31	0.11	0.16	0.37	0.21	0.15	0.29	0.14
sylag2	0.15	0.34	0.19	0.20	0.30	0.10	0.17	0.29	0.12	0.16	0.28	0.12
sylag12	0.19	0.34	0.16	0.19	0.30	0.11	0.16	0.28	0.12	0.15	0.29	0.14
x	0.25	0.70	0.44	0.24	0.47	0.23	0.21	0.68	0.47	0.19	0.40	0.21
xylag1	0.10	0.40	0.30	0.10	0.19	0.09	0.07	0.36	0.29	0.10	0.25	0.15
xsylag1	0.16	0.32	0.16	0.18	0.25	0.07	0.16	0.32	0.16	0.15	0.31	0.17

Note: Numbers are based on districts for which all models have converged. J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-18. Compare test types--Number of districts, average number of schools per district, average number of classrooms per school, average number of students per classroom, and intraclass correlations, by test subject and type: Florida 10th grade

	10th grade math		10th grade reading	
	FCAT	NRT	FCAT	NRT
No. of district-year	9		14	
Avg # of schools per district	23		21	
Avg # of classes per school	29		19	
Avg # of students per class	7		10	
Intraclass correlation (no covariates)				
School level	0.234	0.118	0.238	0.147
Classroom level	0.424	0.441	0.315	0.276

Note: Numbers are based on districts for which all models have converged for both FCAT and NRT test types.

Table FL-19. Compare test types--Average minimum detectable effect size (MDES), by test type, number of schools, classrooms, and students: Florida 10th-grade math, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Level and model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
FCAT									
unconditional	0.76	0.70	0.65	0.54	0.50	0.46	0.44	0.40	0.37
ylag1	0.29	0.27	0.25	**0.20**	**0.19**	**0.18**	**0.16**	**0.15**	**0.14**
ylag2	0.31	0.29	0.27	0.22	**0.20**	**0.19**	**0.18**	**0.17**	**0.16**
ylag12	0.23	0.21	**0.20**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**	**0.12**
sylag1	0.55	0.46	0.38	0.39	0.33	0.27	0.32	0.27	0.22
sylag2	0.53	0.43	0.35	0.37	0.31	0.25	0.30	0.25	**0.20**
sylag12	0.51	0.41	0.32	0.36	0.29	0.23	0.29	0.24	**0.19**
x	0.69	0.62	0.56	0.49	0.44	0.40	0.40	0.36	0.33
xylag1	0.28	0.26	0.24	**0.20**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
xsylag1	0.52	0.44	0.35	0.37	0.31	0.25	0.30	0.25	**0.20**
NRT									
unconditional	0.65	0.57	0.50	0.46	0.40	0.36	0.37	0.33	0.29
ylag1	0.25	0.22	**0.20**	**0.17**	**0.15**	**0.14**	**0.14**	**0.13**	**0.11**
ylag2	0.27	0.24	0.21	**0.19**	**0.17**	**0.15**	**0.16**	**0.14**	**0.12**
ylag12	**0.20**	**0.18**	**0.16**	**0.14**	**0.12**	**0.11**	**0.11**	**0.10**	**0.09**
sylag1	0.51	0.41	0.31	0.36	0.29	0.22	0.29	0.24	**0.18**
sylag2	0.51	0.42	0.32	0.36	0.29	0.23	0.30	0.24	**0.19**
sylag12	0.51	0.41	0.31	0.36	0.29	0.22	0.29	0.24	**0.18**
x	0.59	0.51	0.44	0.42	0.36	0.31	0.34	0.30	0.26
xylag1	0.24	0.21	**0.19**	**0.17**	**0.15**	**0.13**	**0.14**	**0.12**	**0.11**
xsylag1	0.49	0.39	0.30	0.34	0.28	0.21	0.28	0.23	**0.17**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Numbers are based on districts for which all models have converged for both FCAT and NRT test types.

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

Table FL-20. Compare test types--Average minimum detectable effect size (MDES), by test type, number of schools, classrooms, and students: Florida 10th-grade reading, school years 2004-05 through 2005-06

	J=20			*J=40*			*J=60*		
Level and model	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*	*K=3, N=20*	*K=5, N=12*	*K=10, N=6*
FCAT									
unconditional	0.73	0.68	0.63	0.51	0.48	0.45	0.42	0.39	0.37
ylag1	0.30	0.28	0.27	0.21	**0.20**	**0.19**	**0.17**	**0.16**	**0.16**
ylag2	0.36	0.34	0.33	0.25	0.24	0.24	0.21	**0.20**	**0.19**
ylag12	0.24	0.23	0.23	**0.17**	**0.17**	**0.16**	**0.14**	**0.13**	**0.13**
sylag1	0.48	0.41	0.34	0.34	0.29	0.24	0.28	0.24	**0.20**
sylag2	0.46	0.39	0.32	0.33	0.27	0.23	0.27	0.22	**0.18**
sylag12	0.46	0.38	0.31	0.32	0.27	0.22	0.26	0.22	**0.18**
x	0.63	0.58	0.54	0.45	0.41	0.38	0.36	0.34	0.31
xylag1	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.15**	**0.15**	**0.14**
xsylag1	0.43	0.36	0.30	0.30	0.26	0.22	0.25	0.21	**0.18**
NRT									
unconditional	0.62	0.57	0.52	0.44	0.40	0.37	0.36	0.33	0.30
ylag1	0.27	0.25	0.24	**0.19**	**0.18**	**0.17**	**0.16**	**0.15**	**0.14**
ylag2	0.28	0.26	0.25	**0.20**	**0.19**	**0.17**	**0.16**	**0.15**	**0.14**
ylag12	**0.20**	**0.19**	**0.18**	**0.14**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**
sylag1	0.43	0.36	0.30	0.31	0.26	0.21	0.25	0.21	**0.17**
sylag2	0.43	0.36	0.29	0.30	0.25	0.21	0.25	0.21	**0.17**
sylag12	0.43	0.35	0.29	0.30	0.25	**0.20**	0.25	**0.20**	**0.17**
x	0.52	0.47	0.43	0.37	0.33	0.31	0.30	0.27	0.25
xylag1	0.24	0.23	0.22	**0.17**	**0.16**	**0.15**	**0.14**	**0.13**	**0.12**
xsylag1	0.39	0.33	0.27	0.28	0.23	**0.19**	0.23	**0.19**	**0.16**

Note: J=the number of schools, K=the number of classrooms within each school, and N=the number of students within each classroom

Model specifications:

ylag1: individual student scores lagged 1 year

ylag2: individual student scores lagged 2 years

ylag12: individual student scores lagged 1 and 2 years

sylag1: Mean school scores for the same grade lagged 1 year

sylag2: Mean school scores for the same grade lagged 2 years

sylag12: Mean school scores for the same grade lagged 1 and 2 years

x: race, gender, FRPL, LEP and grade repetition

xylag1: all demographic variables and individual student scores lagged 1 year

xsylag1: all demographic variables and mean school scores lagged 1year

CALDER