NCEE 2011-4033

U. S. DEPARTMENT OF EDUCATION

Variability in Pretest-Posttest Correlation Coefficients by Student Achievement Level

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

Variability in Pretest-Posttest Correlation Coefficients by Student Achievement Level

September 2011

Russell Cole
Joshua Haimson
Irma Perez-Johnson
Mathematica Policy Research

Henry May
University of Pennsylvania
Consortium for Policy Research in Education

Abstract

State assessments are increasingly used as outcome measures for education evaluations. The scaling of state assessments produces variability in measurement error, with the conditional standard error of measurement increasing as average student ability moves toward the tails of the achievement distribution. This report examines the variability in pretest-posttest correlation coefficients of state assessment data for samples of low-performing, average-performing, and proficient students to illustrate how sample characteristics (including the measurement error of observed scores) affect pretest-posttest correlation coefficients. As an application, this report highlights how statistical power can be attenuated when correlation coefficients vary according to sample characteristics. Population achievement data from four states and two large districts in both English/Language Arts and Mathematics for three recent years are examined. The results confirm that pretest-posttest correlation coefficients are smaller for samples of low performers, reducing statistical power for impact studies. We also find substantial variation across state assessments. These findings suggest that it may be useful to assess the pretest-posttest correlation coefficients of state assessments for an intervention's target population during the planning phase of a study.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

NCEE 2011-4033
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES), under Contract ED-04-CO-0112/0006.

Disclaimer

The Institute of Education Sciences at the U.S. Department of Education contracted with Mathematica Policy Research to develop a report documenting the variability in the pretest-posttest correlation coefficients estimated from samples of students who had taken state proficiency assessments and showing how attenuation in pretest-posttest correlation coefficients could affect statistical power in education experiments targeted at low-performing students. The views expressed in this report are those of the authors, and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Rebecca A. Maynard
Commissioner

September 2011

Cole, Russell, Joshua Haimson, Irma Perez-Johnson, and Henry May. “Variability in Pretest-Posttest Correlation Coefficients by Student Achievement Level.” NCEE Reference Report 2011-4033. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2011.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department’s Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

IES contracted with Mathematica Policy Research to develop the discussion of issues presented in this report. Drs. Russell Cole, Joshua Haimson, and Irma Perez-Johnson are employees of Mathematica Policy Research. Dr. Henry May is an employee of the University of Pennsylvania. The authors and other staff of Mathematica and the University of Pennsylvania do not have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) within the Institute of Education Sciences (IES) is responsible for (1) conducting evaluations of federal education programs and other programs of national significance to determine their impacts, particularly on student achievement; (2) encouraging the use of scientifically valid education research and evaluation throughout the United States; (3) providing technical assistance in research and evaluation methods; and (4) supporting the synthesis and wide dissemination of the results of evaluation, research, and products developed.

In line with its mission, NCEE supports the expert appraisal of methodological and related education evaluation issues and publishes the results through two report series: the *NCEE Technical Methods Report* series that offers solutions and/or contributes to the development of specific guidance on state of the art practice in conducting rigorous education research, and the *NCEE Reference Report* series that is designed to advance the practice of rigorous education research by making available to education researchers and users of education research focused resources to facilitate the design of future studies and to help users of completed studies better understand their strengths and limitations.

This *NCEE Reference Report* focuses on the extent to which state tests have increased measurement error for the low-performing students who are commonly the focus of educational interventions, and related considerations for evaluation design (e.g., implications for statistical power to detect program impacts). May et al. (2009) posit that experiments focused on low-performing students may have attenuated study power to detect program impacts when state tests are used, and this paper provides empirical evidence to support this statement. To explore this issue, study authors drew samples of students with different achievement profiles from 6 population datasets and examined the extent to which sample average achievement level attenuates correlation coefficients and, hence, statistical power in pretest-posttest evaluation designs. Study results, which control for the effects of restriction of range in samples, do suggest that pretest-posttest correlation coefficients are reduced for low-performing students, which increases the minimum detectable effect sizes for impact evaluations. The report also provides guidance on how much attenuation to expect in pretest-posttest correlation coefficients due to increased conditional standard errors of measurement in state tests for low-performing students. This information will enable education researchers to increase sample sizes to compensate for the reduced precision due to lower pretest-posttest correlation coefficients.

May, Henry, Irma Perez-Johnson, Joshua Haimson, Samina Sattar, and Phil Gleason. *Using State Tests in Education Experiments: A Discussion of the Issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Acknowledgments

The authors would like to thank the states and districts that allowed their data to be used for this empirical investigation, including the Louisiana Division of Assessments and Accountability. The population data sets used for this study were obtained from three separate, ongoing Mathematica research studies, conducted on behalf of two private organizations and a large school district. Permission from states and sponsors of these research studies was obtained before the use of the population data for this report. To maintain the anonymity of the participating states and districts, we do not reference the studies by name.

The authors also thank Mathematica staff, including Drs. Neil Seftor and Brian Gill, for their careful review of the draft documents. The editing and production staff included Cindy George, Amanda Bernhardt, and Linda Heath. We thank these individuals, and others we may have omitted unintentionally, for their contributions to this report.

The views expressed herein are those of the authors and do not reflect the policies or opinions of the U.S. Department of Education. Any errors or omissions are the responsibility of the authors.

CONTENTS

Chapter **Page**

TABLES

FIGURES

Figure **Page**

I. INTRODUCTION

Over the past decade, there has been increased interest in rigorously evaluating education interventions with randomized controlled trials (RCTs), acknowledged to be the most rigorous method to estimate a program's effects (Boruch 1997; Spybrook and Raudenbush 2009). Researchers commonly follow samples of low-performing students in their RCT evaluations to assess the effectiveness of programs or interventions intended to improve achievement for the neediest student populations (see Chambers et al. 2008; Klein et al. 2008; Vaughn et al. 2009 for recent examples).

Concurrent with the increasing prevalence of RCTs in evaluations of interventions targeting low-performing students has been the increased use of state assessments as outcome measures for such evaluations (May et al. 2009). In their discussion paper describing the use of state assessments for education experiments, May et al. (2009) identify a number of factors that have made state assessments appealing for education research: (1) state assessments in reading and mathematics are administered nearly universally in grades 3 through 8 and in at least one grade in high school; (2) the results from state assessments have significant stakes for students, teachers, and schools, which creates a testing environment in which students will take the test seriously so that the test scores will be an accurate reflection of student achievement; (3) the costs for securing state test data from district and state electronic databases are considerably lower than administering study-specific assessments; and (4) the prevalence of electronic databases creates an opportunity to use linked pretest data for each child to increase statistical power for detecting program impacts.

This report provides insight on how the precision of the impact estimate depends upon properties of the state assessments, in particular, the extent to which earlier test scores (those of the "pretests") are effective predictors of subsequent scores (the "posttests") for the students participating in the study. In particular, this report provides information about the variability in pretest-posttest correlation coefficients for different samples of students and then contextualizes those findings in terms of how those correlation coefficients can affect precision in impact evaluations. The application of using pretest-posttest correlation coefficients to inform power calculations is one possible illustration of how the information from this report can be used, and this focus on power calculations for study designs using state assessments was motivated by the May et al. (2009) discussion paper. In the remainder of this chapter, we provide a brief summary of the framing, methods, and results of this report.

A. REPORT SUMMARY

Assessments that occur prior to random assignment (that is, pretests) are typically the single best covariate for explaining posttest variation and are recommended covariates to improve precision in prospective study designs (Schochet 2008; Bloom et al. 2005). The magnitude of the pretest-posttest correlation coefficient is directly related to the amount of variance in the outcome that is explained by the inclusion of the pretest in a regression analysis, and thus, to the statistical power of the design.

May et al. (2009) point out a number of potential pitfalls that can occur when state assessments are used as outcomes in education experiments. Notably, the authors indicate that many state tests are less precise (have greater measurement error) for students whose test scores are at the low or high tails of the achievement distribution, relative to the center. The increased measurement error for samples of students in the tails of the achievement distribution has the potential to reduce pretest-posttest correlations (Spearman 1904). In addition, when samples of students are drawn from the tails of the distribution, the variation in test scores is reduced relative to the population, which can also attenuate pretest-posttest correlation coefficients. Each of these issues will reduce the amount of variation in the outcome that is explainable by a pretest and will reduce statistical power to detect program impacts in pretest-posttest RCT designs using state assessments as outcomes.

Despite the increasing prevalence of the use of state assessments as pretest and outcome measures for education experiments, evaluation researchers lack descriptions of pretest-posttest correlation coefficients for student samples of various characteristics. To help address this gap, this report examines the variation in pretest-posttest correlation coefficients using population data sets from four states and two large districts. It examines this variation for samples of (1) low-performing students, (2) students whose performance is distributed around their state’s proficiency threshold, and (3) students whose performance is distributed around their state’s population average. In addition, we examine pretest-posttest correlation coefficients for homogeneous and heterogeneous samples to study the role of reduced variation in power calculations. Using achievement test scores from three years of administrative population data sets in each state and district, we examine the variability in correlation coefficients across grades 3 through 8, within and between states, and in both English/Language Arts (ELA) and Mathematics. Although some education experiments randomize intact clusters of students to conditions (e.g., classrooms or schools), this study focuses on implications for experiments that focus on students as the units of assignment.

Our analyses are guided by two broad research questions, which are elaborated upon in Chapter 2***:***

1. *Which characteristics of samples (average achievement level, heterogeneity level, subject, grade, year, and state assessment) are associated with differences in pretest-posttest correlation coefficients?*
2. *How does attenuation in pretest-posttest correlation coefficients affect the statistical power) for experiments focused on low-performing students?*

For our convenience sample of four states and two districts, our analyses confirm that state tests can have attenuated pretest-posttest correlation coefficients for relatively low-performing subgroups of students. On average (across multiple states, years, grade levels, and subjects), the pretest-posttest correlation coefficient for low performers was significantly lower than the correlation coefficients for students who score in the vicinity of the proficiency threshold or students who score near the population average. In addition, for a given mean level of achievement, more homogeneous samples of students had attenuated pretest-posttest correlation coefficients relative to more heterogeneous samples of students. When we compared pretest-posttest correlations across multiple factors, between-state differences explained the greatest proportion of the variance in the pretest-posttest correlation coefficients. Finally, our results for the individual states and districts suggest that pretest-posttest correlation coefficients are relatively consistent across subject areas (ELA and Mathematics), across grade levels, and over time.

From our results, we conclude that statistical power *can* be reduced when state tests are used as outcome measures in evaluations focused on the lowest-performing students. However, such attenuation is sometimes modest and partly offset by other study benefits (such as potential cost savings) from the use of state tests. Importantly, our results may not be generalizable to other states. Hence, we recommend that researchers replicate the type of analysis presented in this report for each state in their study's proposed sample.

The remainder of this report is organized as follows. In Chapter II, we discuss pretest-posttest correlation coefficients and how they relate to power analysis for student-level RCTs in which state assessments are used as outcome and pretest measures. Chapter III describes the data and methods used to estimate variation in the size of the pretest-posttest correlation for various subgroups of students. Chapter IV presents the results of our empirical analyses, which are then extended in examples of prospective power analysis. Chapter V discusses the implications and limitations of this work. An appendix provides summary statistics across all years of available data.

II. FACTORS THAT INFLUENCE CORRELATION COEFFICIENTS IN STATE ASSESSMENTS AND IMPLICATIONS FOR STUDY DESIGN

The magnitude of observed pretest-posttest correlation coefficients depends not only on the choice of the pretest and posttest assessments, but also on the characteristics of the groups of students taking the assessments. Assessments that have a large pretest-posttest correlation in the overall student population may have considerably smaller correlations for subgroups of students, for two key reasons: (1) samples of students drawn from different areas of the achievement distribution will have varying amounts of measurement error in the observed scores and (2) samples will typically have lower variances than the variance of the population.

In this section, we discuss these and other factors that can influence the magnitude of pretest-posttest correlation coefficients in state assessments. After enumerating the key factors, we present an illustration of how the magnitude of a pretest-posttest correlation coefficient can inform study design, by illustrating how including a pretest as a covariate can reduce the Minimum Detectable Effect Size (MDES) in a prospective study. Research questions are presented at the end of the chapter.

A. MEAN ACHIEVEMENT LEVEL OF THE SAMPLE

Assessments of achievement, including state proficiency tests, may be less precise (that is, have greater measurement error) for samples of students selected at the tails of the distribution relative to samples of students drawn from the center of the distribution, which has the potential to attenuate pretest-posttest correlation coefficients. With respect to state assessments in particular, this may reflect the fact that a growing number of states are using item response theory (IRT) methods to assign scale scores to individual students based on their responses to test items (Tong and Kolen 2010). The scaling of these assessments produces heterogeneous conditional standard errors of measurement (CSEM), with greater CSEMs located in the tails of the achievement distribution (Peterson et al. 1989; du Toit 2003; Lee et al. 2000; Raju 2007; Hambleton et al. 1991; May et al. 2009).

The *Standards for Educational and Psychological Testing* of the American Educational Research Association (AERA) acknowledge this heterogeneity of CSEMs—in standard 2.2—and require that states report it: “The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation” (AERA 2002, p. 31). As their name suggests, the principal purpose of state proficiency exams is to define proficiency; this is therefore the score at which CSEMs are generally reported. The technical manuals that we were able to review for this report all confirmed that CSEMs were lower in the center of the distribution (near the relevant proficiency threshold) than at the tails of the distribution.[1]

Quantifying the precision of an assessment (or lack thereof) in a sample requires knowledge of the amount of measurement error in that sample. It is possible to characterize the measurement error of a group or sample by taking into account the variability of the individual measurement errors of the sample that comprises it. Each estimated scale score for an assessment $\left(\hat{\theta}\right)$ is associated with a corresponding conditional standard error $\left(SE\left[\hat{\theta}\right]\right)$ that varies depending on

[1] We were able to examine technical manuals for four of the six state proficiency assessments examined in our analysis. To obtain these manuals, we visited state department websites and downloaded available technical manuals for the years in question. For two states, these technical manuals were not available on the state department websites. We do not cite the technical manuals reviewed in this paper in order to maintain the anonymity of the states and districts that provided data for this project.

the value of $\hat{\theta}$. Within a sample, the individual-level errors in measurement (examinee-level *SEM*$_i$) can be operationalized as the reported conditional standard error for each individual *i's* scale score:

(1) $$SEM_i = SE\left(\hat{\theta}_i\right)$$

The group- or sample-level measurement error variance (*SEM*^{2}*)* can then be considered as the average or expectation of the examinee-level *SEM*$_i^2$ (Lord and Novick 1968):

(2) $$SEM^2 = E\left(SEM_i^{\,2}\right)$$

The relationship between measurement error on an outcome in a sample and power for education experiments can be understood by first partitioning the variance in the outcome. The total observed variance of an outcome in a sample $\left(s_Y^2\right)$ can be partitioned into true, or explainable variance, $\left(s_T^2\right)$and error, or unexplainable variance $\left(s_e^2\right)$:

(3) $$s_Y^2 = s_T^2 + s_e^2$$

Given a group with observed outcome variance $\left(s_Y^2\right)$ and an estimate of the group error variance $\left(SEM^2 = s_e^2\right)$, then by simple subtraction, we can estimate an upper boundary on the amount of variance in the outcome that can be explained $\left(s_T^2\right)$. The proportion $\left(s_T^2 / s_Y^2\right)$indicates the maximum amount of variance in the outcome that can be explained—that is, the proportion of variance in the outcome that is non-error.

To elaborate further, consider two groups of students with equal observed variance on an outcome measure; however, each group was sampled from two different areas of the population achievement distribution on a state proficiency assessment. In order to isolate the contribution of measurement error in different samples to pretest-posttest correlation coefficients, it is necessary to hold constant the variance of these samples, as sample variance is also related to correlation coefficients (see the section below titled “Restriction in Range”). Consider group A as a sample of students whose average performance on the state test is relatively low $\left(\bar{\hat{\theta}}_A\right)$, and group B as a sample of students whose average performance on the same state test is at the proficiency threshold $\left(\bar{\hat{\theta}}_B\right)$, where both group A and group B have equal outcome variances:

(4) $\hat{s}^2_{Y_A} = \hat{s}^2_{Y_B} = s^2$ and $\bar{\hat{\theta}}_A < \bar{\hat{\theta}}_B$

Given that the group level SEM^2 is the expectation of the individual CSEMs (from Equation 6 above) and that the CSEMs tend to be larger for scores in the tails of the distribution relative to the proficiency threshold, the lower average achievement level in sample A relative to sample B implies that:

(5) $SEM_A^2 > SEM_B^2$, or equivalently, $s^2_{e_A} > s^2_{e_B}$

Because we can decompose the observed variance of an outcome into explainable and unexplainable variance (given Equation [3]), it follows from Equation (5) that:

(6) $s^2_{T_A} + s^2_{e_A} = s^2_{T_B} + s^2_{e_B}$

And given Equations (5) and (6), and simple subtraction:

(7) $s^2_{T_A} < s^2_{T_B}$ and therefore $\dfrac{s^2_{T_A}}{s^2} < \dfrac{s^2_{T_B}}{s^2}$

That is, the amount (or proportion) of variance that is explainable in group A is less than the amount (or proportion) of variance that is explainable in group B. In other words, the location of the groups in the achievement distribution places an upper bound on the amount of variation that can be explained in the outcome. The squared value of the pretest-posttest correlation coefficient indicates the proportion of variance in the outcome that is explainable by the pretest and has as its maximum, or upper boundary, the amount of explainable variance in the outcome:

(8) $$\max\left(r^2_{pretest,\,posttest}\right) = \frac{s^2_T}{s^2}$$

From Equations (7) and (8), it follows that the maximum value of the squared pretest-posttest correlation coefficient in group A is less than the maximum value of the squared pretest-posttest correlation coefficient in group B:

(9) $$\max\left(r^2_{pretest_A,\,posttest_A}\right) < \max\left(r^2_{pretest_B,\,posttest_B}\right)$$

This illustrates how the maximum pretest-posttest correlation coefficient for samples of students drawn from the tails will be less than the maximum pretest-posttest correlation coefficient for samples of students drawn from the center of the distribution. Given that the theoretical maximum correlation coefficient for low performers is less than the maximum correlation coefficient for samples of students drawn from the center of the distribution, it is possible or even probable that the observed/empirical correlation coefficients calculated for low performers will be less than the observed/empirical correlation coefficients for samples of students drawn from the center. Note that this demonstration focuses solely on the fact that measurement error in the posttest potentially limits the pretest-posttest correlation coefficient, but it is also true that additional measurement error in the pretest for low-performers can attenuate the coefficient as well. It is of interest for prospective study design, therefore, to examine empirically the extent to which pretest-posttest correlation coefficients are attenuated for samples of students drawn from the tails of states' achievement distributions.

B. RESTRICTION IN RANGE (SAMPLE HOMOGENEITY)

Restricting the range or variability of observed test scores also tends to reduce the observed pretest-posttest correlation coefficient (Feldt and Brennan 1989; Hunter and Schmidt 2004;

Thompson and Vacha-Haase 2000). In education research, there are two ways in which the range of observed scores on an assessment can be restricted: (1) through selection of a homogeneous subgroup or (2) through use of assessments that are too difficult or too easy, leading to clustering of scores at the extremes for low and high achieving subgroups of students.

Pretest-posttest correlation coefficients are positively related to the total variability in test scores and hence to the heterogeneity of students in the sample (Thompson and Vacha-Haase 2000). In a relatively homogeneous subgroup, small changes in student performance (such as changes due to random/measurement error) can result in large changes in relative position, or rank, within the sample. Alternatively, in a relatively heterogeneous subgroup, small changes in student performance will have less of an influence on the relative rank of individuals within the sample. Given that pretest-posttest correlation coefficients are an indication of relative consistency of ranks in a sample, when a study focuses on a relatively homogeneous subgroup, correlations will be attenuated.

Another way in which the variability of scores on a given assessment can be restricted is if the test is too hard or too easy for the students in a sample. For example, if many students answer all of the questions incorrectly and obtain the minimum test score (a floor effect), the variability of scores is attenuated, reducing potential pretest-posttest correlation coefficients. Pretest-posttest correlation coefficients can also be reduced when many students answer all items on a test correctly (a ceiling effect). In essence, ceiling and floor effects create an artificial restriction in range by truncating the observed scores and making it more difficult to capture changes in performance over time. Ceiling and floor effects in the population can also create a situation in which there will be truncation in the data for samples drawn from the low and high ends of the achievement distribution.

Given that pretest-posttest correlation coefficients are affected by restriction in range, it should also be of interest to education evaluation researchers to further understand the unique contribution of sample homogeneity to pretest-posttest correlation coefficients. Pretest-posttest correlation coefficients calculated for different heterogeneity levels can provide insight into the extent to which the sample homogeneity issue will influence statistical power for education interventions.

C. OTHER FACTORS

The achievement level of the sample and the heterogeneity of the sample both influence correlation coefficients, and therefore, these two factors must be considered together in any analysis of correlation coefficients. For example, the motivating demonstration of how sample achievement levels can influence correlation coefficients appropriately accounted for sample homogeneity (Equation [4] fixes the posttest variance of the samples in the example). In the methods described in the following chapter, both of these critical factors will be examined simultaneously in all analyses to appropriately reflect their interdependent relationships with correlation coefficients.

In addition to the mean achievement level and the homogeneity of a study's student sample, there are other characteristics of state assessments that may be related to pretest-posttest correlation coefficients and can be particularly relevant for evaluations of certain types of educational interventions. Pretest-posttest correlations may vary significantly, for example, according to the subject matter of the assessment, the grade level of students being assessed, or (since state assessments are redesigned periodically) the academic year when the tests were administered. A stronger understanding of how pretest-posttest correlation coefficients vary along these factors should also be of value for education evaluation researchers.

D. PRETEST-POSTTEST CORRELATION COEFFICIENTS AND IMPLICATIONS FOR RESEARCH DESIGN

Information about pretest-posttest correlation coefficients can be used to help inform the design of education evaluations. Pretest-posttest correlation coefficients are commonly used in education measurement applications, for example, to demonstrate the stability of student test scores between two time points or to illustrate that this student score stability decreases with the length of time between assessments. They also can inform evaluation design, especially in regards to prospective power analysis. This latter application is motivated by the May et al. (2009) discussion paper, which highlights the need to understand how power can be attenuated when interventions are focused on particularly high- or low-performing students (p. 14).

Power analyses can be used for two alternative purposes. The analysis can determine the sample size necessary to detect impacts (if they exist) with a specified degree of confidence (Cohen 1988; Bloom 2006). Alternatively, when available resources define the size of a study sample, the power analysis can determine the MDES that can be discovered given that sample size and assumptions for Type I and II error rates. To perform these calculations, researchers commonly use the standard assumptions of 80 percent power (Type II error = β = 0.20) and a two-tailed hypothesis test (with Type I error = α = 0.05).

In order to calculate an MDES that is scale/metric free, a standardized effect size (ES) can be used. The standardized effect size for the difference between two groups can be calculated as the difference between the average performance of the treatment and control groups $\left(\overline{Y}_T \text{ and } \overline{Y}_C\right)$, divided by a common standard deviation unit $\left(s_Y\right)$:

$$(10) \qquad ES = \frac{\overline{Y}_T - \overline{Y}_C}{s_Y}$$

The standardized effect size can therefore be interpreted as the number of standard deviation units that separate the treatment and control group averages. Please note that in Equation (10), we have operationalized the common standard deviation unit as the standard deviation of the sample. An alternate specification could use the population standard deviation as the metric for standardized effect size. Given that samples tend to have smaller variances than populations, the interpretation of an effect size unit depends on the metric in which it is standardized (i.e., an effect size difference of 0.5 sample standard deviation units might only represent a difference of 0.2 population standard deviation units). Education researchers typically present results in terms of sample standard deviation units, and we will use the sample standard deviation unit as the standardizing metric for our analyses in this paper.

In the following sections, we present the MDES calculations for posttest-only designs and then show how the MDES can be reduced by including a pretest as a baseline covariate.

MDES for Posttest-Only RCT Designs

Assuming that the sample treatment and control group means have equal variance, the MDES (in standard deviation units of the outcome measure) can be calculated as follows:

$$(11) \qquad MDES\left(\bar{Y}_T - \bar{Y}_C\right) = M_{n-2} * \sqrt{\frac{1}{n * p * (1-p)}}$$

where M_{n-2} is a multiplier calculated as the sum of the t-values corresponding to the Type I and Type II error assumptions ($M_{n-2} = t_{\alpha/2} + t_{1-\beta}$), n is the total sample size in the study, and p corresponds to the proportion of the sample that is randomly assigned to the treatment condition (Bloom 2006).[2] This formula, therefore, provides the smallest average difference (in standard

[2] M_{n-2} is approximately 2.8 for large samples, reflecting a two-tailed alpha level (z = 1.96) and 80 percent power (z = 0.84). The *n-2* subscript refers to the number of degrees of freedom available for comparing the treatment and control means.

deviation units) that can be detected after the intervention given the stated assumptions for *n*, *P*, α*,* and β.[3]

MDES for Pretest-Posttest RCT Designs

Education researchers often include pretest assessments to improve power (or reduce MDES) to detect impacts in evaluation research (see, for example, Puma et al. 2005; Klein et al. 2008; Davidson et al. 2009). Inclusion of a pretest as a covariate in impact analyses helps to explain (error) variance in the posttest and improves the likelihood of uncovering program impacts by reducing the standard error of the impact estimate. Borrowing again from Bloom (2006), the impacts of a program can be estimated net of any baseline characteristics, including a pretest, using a regression framework:

$$(12) \qquad Y_i = \alpha + \beta_0 T_i + \sum_{j=1}^{k} \beta_j X_{ji} + \varepsilon_i$$

where T_i is an indicator for treatment status of person *i;* X_{ji} is a set of baseline covariates (that could include a pretest); and ε_i is a stochastic error term. The estimated impact of the hypothetical program is $\hat{\beta}_0$, which indicates the difference in the average outcome for the treatment and control groups, after adjusting for differences in the sample at baseline. Note that this analysis is identical to an analysis of covariance (ANCOVA) where the dependent variable is the posttest and the pretest is used as a covariate.[4]

[3] The MDES calculated in equation (11) provides effect size estimates in terms of *sample* standard deviation units of the outcome measure, a commonly used metric. As a result, the heterogeneity of the sample cannot directly affect the MDES calculations in these units, since both homogeneous and heterogeneous samples would be standardized to represent a standard deviation of one in the calculations. One contribution of this paper is to examine whether sample *homogeneity* attenuates pretest-posttest correlation coefficients in pretest-posttest designs, which *can* affect MDES calculations (see Table 1 below).

[4] Note also that the model presented in Equation (12) is *not* designed to explain changes that occurred between pretest and posttest (that is, achievement growth). Instead, the model is designed to detect differences in the average posttests of the treatment and control groups. This estimate of the intervention's impact is made more precise by

It is possible to estimate an MDES for a study, after adjusting for the inclusion of a pretest, using the following:

$$(13) \qquad MDES\left(\hat{\beta}_0\right) = M_{n-k-2} * \sqrt{\frac{1-R_A^2}{n*p*(1-p)}}$$

where $R_A{}^2$ is the proportion of variance in the outcome that is explained by covariates included in the analysis (Bloom 2006). Note that reduction in the MDES for a pretest-posttest design relative to a posttest-only design is directly related to the proportion of posttest variance explained by the pretest:

$$(14) \qquad MDES_{pretest\,\&\,posttest} = MDES_{posttest} * \sqrt{1-R_A^2}$$

When a pretest is the only covariate used in the analysis, the square of the pretest-posttest correlation coefficient for the sample of interest is equivalent to $R_A{}^2$ in Equation (13). For the purposes of this report, we focus on the contribution of the pretest as a sole covariate used to explain variance in the outcome, as it is expected to be the baseline variable with the largest bivariate correlation with the outcome.[5]

Based on these relationships, one can see how the inclusion of a pretest in a study design can reduce the MDES, improving precision or reducing the sample needed to detect impacts of a specific size. Consider a hypothetical, balanced randomized trial, where state assessment data are selected as the outcome measure and pretest data are available to improve power to detect

(continued)
including covariates that reduce the residual variance in the outcome. Additional discussion of the differences between these two approaches to estimating program impacts—that is, modeling growth versus modeling post-intervention outcomes—can be found in May et al. (2009).

[5] Schochet (2008) notes that in his analysis of three experiments, at least 50 percent of the within-classroom (as well as between-school) variance was explained by a student-level pretest. Schochet also notes that similar findings were observed in Gargani and Cook (2005) and Bloom et al. (1999).

effects. Under these assumptions and potential samples sizes of 250 and 500 students (selected for illustration), it is possible to calculate a variety of MDES estimates depending on the pretest-posttest correlation coefficient for a given sample. As Table 1 shows, for a given sample size, increasing the correlation between the pretest and the posttest decreases the MDES.[6]

TABLE 1

MDES AS A FUNCTION OF PRETEST-POSTTEST CORRELATION COEFFICIENT AND SAMPLE SIZE

Pretest-Posttest Correlation Coefficient (r)	Proportion of variance in outcome explained by pretest (r^2)	MDES when n=250	MDES when n=500
0.0	0.00	0.354	0.250
0.1	0.01	0.352	0.249
0.2	0.04	0.347	0.245
0.3	0.09	0.338	0.239
0.4	0.16	0.325	0.230
0.5	0.25	0.307	0.217
0.6	0.36	0.283	0.200
0.7	0.49	0.253	0.179
0.8	0.64	0.213	0.150
0.9	0.81	0.154	0.109

E. RESEARCH QUESTIONS

In an effort to quantify the extent to which the aforementioned factors are associated with pretest-posttest correlations coefficients and how this information might inform prospective power and MDES calculations, the present study addresses seven research questions (RQs):

RQ 1: *Are pretest-posttest correlation coefficients lower for samples of low-performing students (that is, students selected in the tails of the distribution) than for students selected from the center of the distribution?*

[6] Note that a pretest-posttest correlation coefficient of 0 is equivalent to a posttest-only RCT design, where there is no added benefit to the inclusion of a pretest as a covariate.

RQ 2: *Are pretest-posttest correlation coefficients of state assessments lower for homogeneous samples of students than for heterogeneous samples of students?*

RQ 3: *Do pretest-posttest correlation coefficients differ by subject matter (ELA versus Mathematics)?*

RQ 4: *Do pretest-posttest correlation coefficients differ by grade level (early elementary versus late elementary/early middle school versus middle school)?*

RQ 5: *Do pretest-posttest correlation coefficients differ by state?*

RQ 6: *Do pretest-posttest correlation coefficients differ over time?*

RQ 7: *How does attenuation in pretest-posttest correlation coefficients affect the MDES for experiments focused on low-performing students?*

III. DATA AND METHODS

The goal of our empirical analyses is to assess the variability in pretest-posttest correlation coefficients for different subgroups within a student population and to identify factors related to the size of these correlation coefficients. For this study, we used the three most recent years of available student population scale scores on ELA and Mathematics assessments from four states and two large districts. For the remainder of the paper, we refer to the four state and two large district data sets as our "population data sets" or "state data sets," using these terms interchangeably.[7] We use student records from grades 3 through 8 in these analyses, for two reasons: (1) these are the grades that are commonly assessed for Adequate Yearly Progress purposes; and (2) these grades allow for a one-year period between pretest and posttest, which is a commonly used interval for the length of an education experiment.[8] For more information about the use of state tests in pretest-posttest designs in other combinations (aside from using a state assessment as both the pretest and posttest with a one-year interval between assessments), please see Olsen et al. (2010) and Zhu et al. (2010).

The remainder of this chapter describes the methods used and our assumptions for the empirical analyses. First, we describe our data sources and how we selected our samples. We then describe our statistical methods and the way in which we aggregated our results to answer the research questions.

[7] This reflects (1) that, even in the two districts, we are examining the properties of *state* proficiency assessments and (2) that the correlation coefficients that are the focus of our analysis are calculated based on equally sized samples from the states and districts included in our analysis.

[8] The No Child Left Behind requirement for high school assessment is a single test (for both Mathematics and ELA) sometime during grades 10 and 12. As a result, a comparison of pretest-posttest correlations for high school data would require a longer interval between pretest (presumably 8th grade) and posttest (sometime between 10th and 12th grades).

A. DATA SOURCES AND SAMPLES

We obtained population data sets from three separate, ongoing Mathematica research studies.[9] Permission from states and sponsors of affiliated research studies was obtained prior to the use of the population data for this report. To maintain the anonymity of the participating states and districts, we do not reference the studies by name, nor do we include information on the years in which the assessments occurred or the size of the population of the states. Table 1 provides descriptive information about the population data sets included in our analyses.

TABLE 2

POPULATION ADMINISTRATIVE DATA SETS USED

	State A	State B	State C	State D	State E	State F*
Year 1 Assessment Grades	3, 4, 5, 6, 7	3, 4, 5	3, 6	5, 6	3, 4, 5, 6, 7	3, 4, 6, 7
Year 2 Assessment Grades	3, 4, 5, 6, 7, 8	3, 4, 5, 6	3, 4, 5, 6, 7	5, 6, 7	3, 4, 5, 6, 7, 8	3, 4, 5, 6, 7, 8
Year 3 Assessment Grades	4, 5, 6, 7, 8	4, 5, 6, 7	4, 5, 6, 7, 8	6, 7	4, 5, 6, 7, 8	4, 5, 6, 7, 8
Year 1, 2 Pretest-Posttest Analysis (Pretest Grades)	3, 4, 5, 6, 7	3, 4, 5	3, 6	5, 6	3, 4, 5, 6, 7	3, 4, 6, 7
Year 2, 3 Pretest-Posttest Analysis (Pretest Grades)	3, 4, 5, 6, 7	3, 4, 5, 6	3, 4, 5, 6, 7,	5, 6	3, 4, 5, 6, 7	3, 4, 6, 7
Subjects Assessed	Reading, Mathematics	Literacy, Mathematics	Language Arts, Mathematics	Reading, Mathematics	Reading, Mathematics	Reading, Mathematics

* Pretest-posttest analysis did not include grade 5 as a pretest year, due to the small population size for this state.

The goal of this study was to examine the variability in pretest-posttest correlation coefficients associated with the factors mentioned in RQs 1–6, and to contextualize these findings through MDES calculations to answer RQ 7. The availability of achievement test records for three consecutive years made it possible to examine two sets of pretest-posttest correlations within each state (that is, Year 1 with Year 2, and Year 2 with Year 3) while

[9] The three research studies were conducted on behalf of two private organizations and a large school district; none of the studies has generated restricted or public-use data files.

maintaining a one-year interval between the pretest and posttest periods. Individual students were included in our analyses (for a particular year/test combination) if they had nonmissing scale scores and grade indicators for two consecutive years. Student records that did not have appropriate test or grade data for a given year were eliminated from our analysis data sets.[10]

State tests are typically administered only in grades 3 through 8, which restricted our effective pretest samples to students in grades 3 through 7. Grade 8 students were eliminated from all pretest data sets because the only way they would have been observed in the relevant posttest year was if they were retained in grade 8. Students who were retained in other grades were included in all analyses.[11]

We created descriptive tables for each state and combination of pretest-posttest years and assessments (grades) examined. The eligible population size, together with the proportion of students receiving the minimum and maximum scores, is presented for each state as tables in the appendix.[12]

B. METHODS

In order to examine the variability in pretest-posttest correlation coefficients associated with the factors identified in RQs 1–6, we employed a multi-step procedure. First, we defined the factors of interest in RQs 1-6 (e.g., “low performing students” or “homogeneous samples”).

[10] Across all of the population data sets, on average, 1.03 percent of the student records were dropped because pretest-posttest data were unavailable in adjacent years in grades in which state assessments were administered.

[11] Across all of our population data sets, on average, 2.26 percent of students were retained in adjacent years in the grades in which state assessments were administered.

[12] We were able to confirm from state technical manuals that at least three of the assessments examined were vertically scaled. Note, however, that a vertical scale is not necessary for pretest-posttest RCT designs employing an ANCOVA-type impact model, as specified in Equation (3). Any variable can be included as a covariate to explain variance in the outcome, regardless of whether or not it is on the same interval scale as the outcome.

Next, we selected equal sized samples in each state using common procedures consistent with our factor definitions and calculated pretest-posttest correlation coefficients for each of those student samples. Last, we decomposed the variance in the observed pretest-posttest correlation coefficients by the factors of interest in an Analysis of Variance (ANOVA) model. This section describes our methods in more detail.

The first step involved making analytic decisions about those factors lacking a "natural" definition for our analyses. This included three of the factors—(1) exemplar achievement levels, (2) levels of sample variability, and (3) grade groupings—for which we planned to examine pretest-posttest correlation coefficients.

- **Sample average pretest achievement levels.** Three average pretest achievement levels were used to illustrate variability in sample achievement levels. We selected these achievement levels to represent segments of the overall student population that different educational interventions might target. The levels selected included 1.3 standard deviations below the population mean ($\bar{z}$ = -1.3), 0.7 standard deviations below the mean ($\bar{z}$ = -0.7), and the population average ($\bar{z}$ = 0). Figure 1 illustrates how samples of students can be identified from the population with a sample average achievement level at one of the three exemplar points, and a distribution of achievement in the sample around that midpoint.

  The lowest average achievement level ($\bar{z}$ = -1.3) represents an approximate threshold for students who are in the lowest decile of performance.[13] We chose this threshold as indicative of the type of performance that might be expected of students at a very low-performing school. Consider a school-based intervention that is intended to improve reading achievement. In a low-performing school (i.e., a school with average reading performance 1.3 standard deviations below the state population mean), there will be a distribution of student performance within that school around that mean. If we were to randomly assign students in that school to treatment and control conditions, we would create an experiment that has a sample that mimics the representation of the "Low Performers" sample in Figure 1. Note that this sample is

[13] This assumes that the achievement levels for a given assessment (in a given grade, state, and year) are normally distributed. The fact that these tests are imperfect measures of a particular ability level (and that student scores may regress toward the mean of the assessment) implies that the observed *z*-scores may underestimate the "true" level for samples of students whose achievement is below the population average. It may be the case that the students who are chosen into the lowest-scoring group have the largest negative error scores and thus, will be theoretically expected to have diminished pretest-posttest correlations. However, this situation mirrors reality for selection into interventions targeting low-performing students.

not defined as those students whose performance is below a certain threshold, as our framework has identified that sample variability is a critical component to control for in a comparison of correlation coefficients. As such, we have identified the low-performing group by a sample average and a distribution of scores around that average (see “Sample pretest variability levels” below).

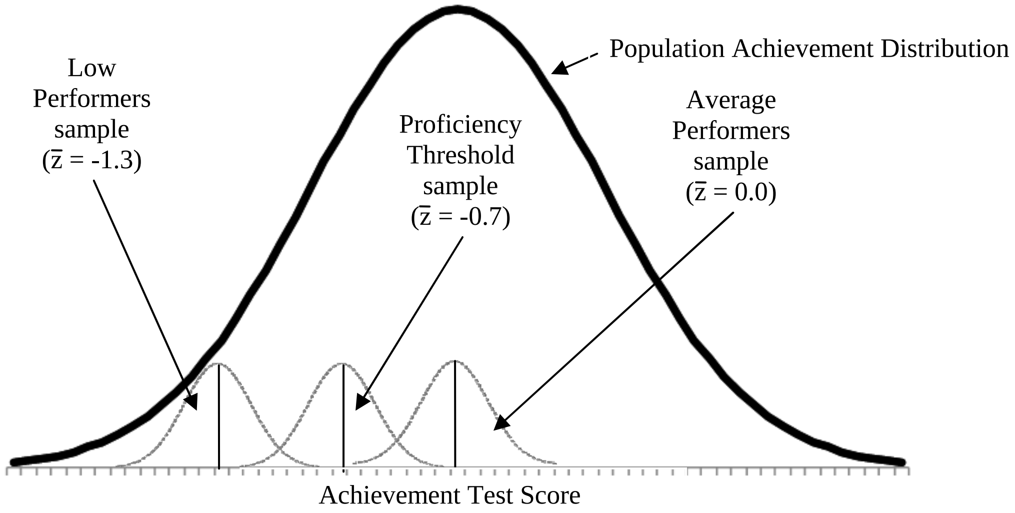
The second average achievement level ($\bar{z}$ = -0.7) represents the average achievement level for the proficiency threshold for four of the six states included in this study.[14] This threshold approximately represents the 24th percentile. Given that the CSEMs for state assessments tend to be lower at the proficiency threshold than at the tails of the distribution, examining the correlation coefficients at this achievement level provides an intuitive comparison for the correlation coefficients observed for low performers. According to the theoretical framework presented above, we should expect to see lower pretest-posttest correlation coefficients for low-performing students than for students closer to the proficiency threshold in a given state.

The final sample average achievement level of interest was average performance in the state ($\bar{z}$ = 0), which represents the 50th percentile of achievement. Similar to the proficiency cutpoint, the average achievement in the population serves as a strong comparison point against which to compare the correlation coefficients of low-performing students. Because the population average is another point in the “center” of the achievement distribution, we should expect to see lower pretest-posttest correlation coefficients in the low-performers sample than in the average-performers sample, based on the theoretical framework above.

[14] We were unable to identify the cut-scores of state assessments relative to the population average and standard deviation for the other two states. There was variability noted in the proficiency threshold across subject areas and states, and over time (standard deviation in proficiency *z*-scores was approximately 0.23). The chosen *z*-score of -0.7 was selected to represent the average “cusp” of proficiency as a threshold.

FIGURE 1

EXEMPLAR ACHIEVEMENT LEVEL SAMPLES, RELATIVE TO THE POPULATION


Population Achievement Distribution
Low Performers sample ($\bar{z}$ = -1.3)
Proficiency Threshold sample ($\bar{z}$ = -0.7)
Average Performers sample ($\bar{z}$ = 0.0)
Achievement Test Score

- **Sample pretest variability levels.** Our study focuses on pretest-posttest correlation coefficients for samples of students, which we expect to exhibit some degree of homogeneity relative to their corresponding student populations. To reflect this homogeneity, we elected to examine pretest-posttest correlations for samples in which the pretest standard deviation was approximately 50 percent of the population standard deviation. To provide a contrast that would enable us to isolate the influence of sample variability on pretest-posttest correlations, our analysis includes one additional sample at the average performance level ($\bar{z} = 0$) with a sample standard deviation that is 90 percent of the population standard deviation—that is, a relatively heterogeneous, average performance sample. The choice of these exemplar heterogeneity levels is based on the ranges of values of study sample standard deviations (as ratios of population standard deviations) observed by Zhu et al. (2010).

  The heterogeneous sample is selected only from the "average" of the population, not from the lower ends of the distribution, because the population distributions cannot support broad sampling from such a sparse area of the data. When a sample is to be selected from the lowest-achieving decile (that is, $\bar{z} = -1.3$), it is unlikely that the observed distribution of the sample will be so variable as to represent 90 percent of the standard deviation of the population.

- **Grade groupings.** We grouped students by their grade level during the pretest year and defined three grade groupings for which to examine differences in pretest-posttest coefficients. The grade groupings used were intended to reflect the groupings that are commonly used in evaluations of education interventions.[15] Students in grades 3 and 4 during the pretest year were combined into an "Early Elementary" group. Students in grade 5 during the pretest year were labeled "Late Elementary/Early Middle School," given that all nonretained students would progress into grade 6 during the posttest (that is, the intervention) year. Finally, students in grades 6 and 7 during the pretest year were considered members of the "Middle School" group.

Having formalized these analytic decisions, we performed the steps below in each state (A–F) for each test (ELA and Mathematics) for both available pretest years.

1. **Pretest standardization.** Within each grade, we standardized student pretest scores into z-scores (state population mean = 0, state population standard deviation = 1). The pretest standardization procedure allowed for comparisons of sample achievement levels across states (and grades) to be made in terms of state standard deviation units and simplified the sample selection procedure described below.

2. **Selection of students for study samples.** For each exemplar average achievement level ($\bar{z}$ = -1.3, $\bar{z}$ = -0.7, and $\bar{z}$ = 0*)* and grade group (Early Elementary, Late Elementary, and Middle School), a random sample of 500 students with sample standard deviation equal to 50 percent of the population standard deviation was selected from the eligible population.[16] We selected samples of 500 students because this is a reasonably sized study for an education experiment. A posttest-only RCT of 500 students is adequately powered to detect effect sizes of 0.25 or greater, without any covariates.

   In the random sampling procedure, we assigned each student a probability of being selected equal to the value of a Normal probability density function with a mean and standard deviation as described above (i.e., N($\bar{z}$, ½σ)) evaluated at the student's pretest score. In other words, the selection probabilities mirrored the intended distribution of pretest scores for each study subsample. Students were more likely to be selected into a given sample when their assigned probability of inclusion was relatively high, indicating that their pretest achievement level was similar to the desired pretest achievement level. In fact, the probability of selection was maximized when the student's pretest achievement score was equal to the mean desired pretest

[15] In education studies, it is sometimes necessary to combine results across grades to obtain a large enough sample to have sufficient statistical power to evaluate an intervention (May et al. 2009), or to provide evidence of broad impacts on performance in the context of interventions that span multiple grades and for which there are no theoretical reasons to analyze the grades separately. Researchers have used this strategy in a number of recent studies (see, for example, Glazerman et al. 2006; Constantine et al. 2009; and Clark et al. 2008).

[16] As described earlier, a sample standard deviation equal to 90 percent of the population standard deviation was also used as a contrast but only for the average achievement level (z = 0).

achievement level. Each sample was drawn independently; therefore, while each student appears no more than once in a given sample, a student may be included in more than one sample.

The random selection procedure produced normally distributed samples of 500 students with the desired pretest averages and standard deviations (i.e., the samples mimicked the presentation of Figure 1). Each random sample for a given grade group contained students from the included grades tested in the pretest year (i.e., the "Early Elementary" group contained students in both grades 3 and 4). Although the pretest means and standard deviations were fixed by design, the posttest means and standard deviations were free to vary. The posttest standard deviations did not differ significantly across the three achievement samples ($p > 0.20$), providing assurance that the comparison of pretest-posttest correlation coefficients has effectively controlled for sample variability as a key factor (see additional assurance of the control for sample variance levels in the robustness analysis below).

3. **Calculation of pretest-posttest correlation coefficients.** After we created the samples for a particular assessment in a given year, we calculated a Pearson-correlation coefficient and a 95 percent confidence interval for each sample using the unstandardized scale scores from the pretest and posttest years. These pretest-posttest correlation coefficients and 95 percent confidence intervals for all samples (all states, years, and assessment content areas) are presented in the appendix (see appendix tables 3, 6, 9, 12, 15, and 18). To provide context for these correlations, we also calculated the full population correlation coefficients for each assessment/year/state in a grade group and also present these in the appendix.

The next step involved aggregating our results across the multiple states, assessments, grade groups, and years examined. We used the results from this last step to answer the study's research questions.

- **Cross-cutting aggregation.** Within a given state, it was possible to estimate up to 12 different pretest-posttest correlation coefficients (three different grade groupings * two different content assessments * two separate year-over-year analyses) for a sample at a particular achievement level. Across the six states where population data were available, we calculated a total of 60 pretest-posttest correlation coefficients for each achievement level of interest.[17] The cross-cutting analysis examined the central tendency and variability of these pretest-posttest correlation coefficients. Because the population data sets available for this study constitute a convenience sample, results are presented as descriptive statistics (means, standard deviations, and minimum and

[17] We estimated a total of 60, not 72, pretest-posttest coefficients because some states did not provide data in all grades in all years. For example, in State B, there were no middle school students tested in year 1, so it was not possible to estimate a correlation coefficient for that grade group on either assessment in that year.

maximum values). These results provide a general indication of the extent to which there is variability in pretest-posttest correlation coefficients within our convenience sample of states.

We also performed an ANOVA on the Fisher z transformed pretest-posttest correlation coefficients to assess the significance of our findings.[18] Fixed effects for each of the following factors were included in the analysis: pretest achievement level (three categories), state (six categories), year (two categories), grade cluster (three categories), subject (two categories), and homogeneity level (two categories). Post hoc contrasts used Tukey's (1953) Honestly Significant Difference procedure to adjust for multiple comparisons.

- **Analysis of Variance.** The ANOVA was performed on a total of 240 pretest-posttest correlation coefficients. We calculated 60 pretest-posttest correlation coefficients for the homogeneous samples at *each* of the three exemplar achievement levels of interest (lowest performance, proficiency threshold, and average performance), for a total of 180 pretest-posttest correlation coefficients. Another 60 pretest-posttest correlation coefficients were calculated for the heterogeneous sample of students whose performance was centered around the population average.

To test the robustness of our findings, we also tested a model that contained the pretest and posttest standard deviations of the samples as covariates in the model, to see if variability in the samples (above and beyond the homogeneity factor) contributed to the correlation coefficients. As noted above, it is critical to account simultaneously for sample achievement level *and* sample homogeneity when comparing pretest-posttest correlation coefficients across samples, and this robustness check provides an additional check on the validity of the results.

[18] The Fisher *z* transformation of a correlation coefficient creates a transformed variable that is unbounded (the Pearson *r* is bounded on [-1, 1]), with constant variance for all population *p* values. This transformed variable is appropriate for hypothesis tests of correlation coefficients. Results were substantively the same (direction and significance) when the raw (nontransformed) correlation coefficients were used as the dependent variable.

Implications for Power Analysis. To answer RQ 7, the observed pretest-posttest correlation coefficients obtained from RQ 1 were used for prospective MDES calculations. The various MDES obtained for samples with different achievement levels provide an illustration of how correlation coefficient attenuation can be a concern for future research designs using state tests as pretest and outcome measures.

IV. RESULTS

In this chapter, we discuss our key findings for each research question, using tables and graphs to provide descriptive evidence for each question. Following this discussion, we revisit the implications of our findings for power analysis for prospective pretest-posttest RCT designs, using the observed data from several research questions to inform illustrative examples.

A. MAIN FINDINGS

The findings for RQs 1-6 are each presented in a similar format in this section. First, the research question is restated. Second, the empirical results for each research question are presented in prose. Following this prose, a table summarizing the descriptive results, including the estimated pretest-posttest correlation coefficients for each research question of interest is shown along with a figure for a visual representation of the data shown in the table. Research Question 7 is answered in section B.

The barometer used to answer RQs 1-6 was the *F*-test results for each factor of the ANOVA described in Chapter III. For RQ 1, the factor in question was the achievement level of the sample. For RQs 2-6, the factors of interest were the heterogeneity of the sample, the subject of the assessment, the grade grouping, and the state, respectively. The results of each ANOVA *F*-test (and *p*-values) are presented with the response for a given question.

The analysis of main effects (without interactions) explained 90.8 percent of the variance in Fisher *Z* pretest-posttest correlations and was selected as the most parsimonious model for explaining variation in the outcome of interest.[19] Including the pretest and posttest standard deviations as predictors of the correlation coefficients did not change the direction or significance of any of the findings (see RQ2 for the one minor exception). The results for the ANOVA analysis of the main effects are presented in the text, and are therefore robust to the threats of invalidity associated with having samples that differ in their homogeneity.

The tables and figures included in this chapter summarize our findings for each research question and follow a common structure. To facilitate the interpretation of these graphics, next we guide readers through their interpretation.

Tables. The average, standard deviation, and minimum and maximum values in each table are calculated from the pretest-posttest correlations estimated for a given subgroup of interest. For example, the population average correlation coefficient shown in Table 3 is calculated across all six states and all three years, in both ELA and Mathematics for all three grade groups. In this table, 60 observed pretest-posttest correlations contribute to each average. Similarly, the standard deviation and minimum and maximum columns in this table represent the variability and extreme values observed in pretest-posttest correlations for a particular subgroup of interest. For example, the minimum value for the lowest-performing subgroup (r = 0.37) occurred in State E, for the correlation between Year 1 and Year 2 scores for the middle school subgroup of students on the Mathematics assessment. Depending on the needs of the researcher, the average, minimum, or maximum pretest-posttest correlation coefficient could serve as a useful empirical reference point for prospective power calculations.

[19] The reported F-statistics for the analyses are therefore based on 227 denominator degrees of freedom. This error term consists of all of the interaction terms not included in the model.

Descriptive statistics are presented for both the tables and the figures (described below), rather than model based means for the correlation coefficients, for two reasons. First, our parsimonious main effects model did not include any interaction terms, and as such, the model based averages would not be able to highlight the effect of achievement level differences within other factors (e.g., achievement level differences in correlation coefficients within reading or mathematics). We believe that these descriptive results provide the most appropriate information for researchers planning interventions focused on particular samples of students (e.g., math interventions targeted at low-performing early elementary students). Second, this choice best reflects that our goal was to illustrate variability in correlation coefficients, not to identify a point estimate from our convenience sample of states.

Figures. To provide easily interpretable visual representations of the data, box and whisker plots of the observed correlation coefficients are included. In these graphs, the mean, median, quartiles, minimum and maximum scores for each exemplar achievement level are shown, stratified by the factor examined in RQs 1-5.[20]

RQ 1. Are pretest-posttest correlation coefficients lower for samples of low-performing students (that is, students selected in the tails of the distribution) than for students selected from the center of the distribution?

There were statistically significant differences in pretest-posttest correlation coefficients across achievement levels, F(2,227) = 20.03, $p < 0.0001$. When averaged across all subjects, states, years, and grade groupings, the "lowest performer" student samples—that is, samples of

[20] In the box and whisker plots, the lowest value on the whisker represents the minimum correlation coefficient for a given sample. The bottom line of the box represents the 25th percentile of correlation coefficients, the middle line of the box represents the median, and the top of the box represents the 75th percentile. The highest value on the whisker represents the maximum correlation coefficient for the sample, and the dot indicates the mean.

students selected at 1.3 standard deviations below the mean, or the threshold for the lowest decile of student achievement—had the lowest average pretest-posttest correlation coefficients with average correlation coefficients approximately .05 units lower than the other two conditions (and both differences were statistically significant at the $p < 0.0001$ level). The average pretest-posttest correlation coefficient (± standard deviation) for the lowest performers was $r = 0.60 \pm 0.13$ (See Table 3 and Figure 2). We found somewhat larger pretest-posttest correlation coefficients for the students at the proficiency threshold and at the population average achievement level ($r = 0.65$ for both). The average pretest-posttest correlation coefficients were not significantly different between the proficiency threshold and the average performers groups ($p = 0.30$).

As noted above, the *F*-test result for testing the differences in correlation coefficients across achievement levels was still significant after including both pretest and posttest standard deviations as covariates in the model. This result provides an assurance that the significant differences that are noted here are not attributable to a simple range-restriction effect. The fact that the three achievement level-defined samples had significant differences in their correlation coefficients, even after accounting for pretest and posttest standard deviations (which were not significantly different from each other), provides evidence of the role that measurement error might play in attenuating correlation coefficients for low-performing students.

TABLE 3

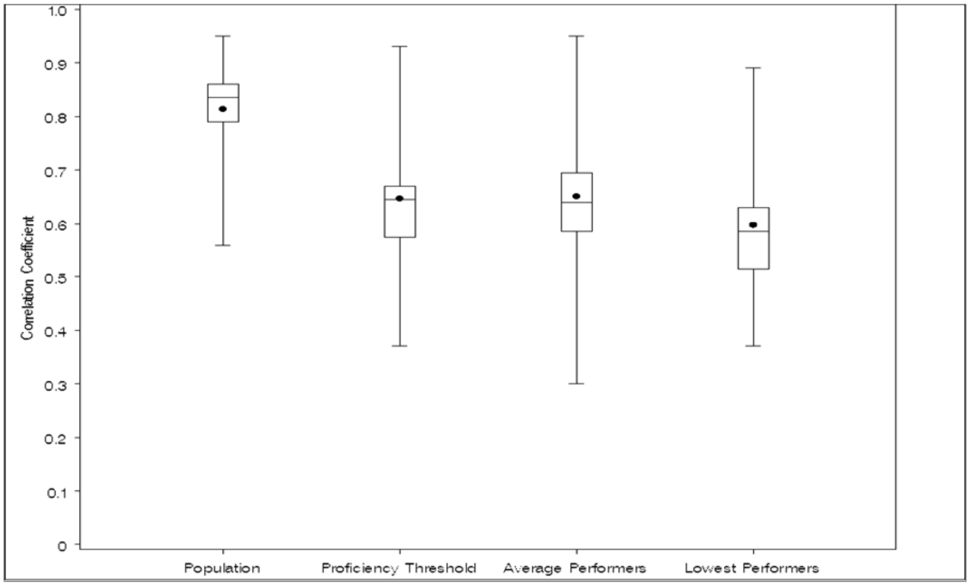
DESCRIPTIVE STATISTICS FOR AVERAGE PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY ACHIEVEMENT LEVEL

	Average	Standard Deviation	Minimum	Maximum
Population	0.81	0.09	0.56	0.95
Proficiency Threshold	0.65	0.12	0.37	0.93
Average Performers	0.65	0.13	0.30	0.95
Lowest Performers	0.60	0.13	0.37	0.89

Note: *n* = 60 pretest-posttest correlation coefficients, calculated for each achievement-level-defined sample (or population). The 60 pretest-posttest correlation coefficients include results aggregated across all six states, in both subject areas (ELA and Mathematics), for all grade groups, and over both pretest years.

FIGURE 2

PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY ACHIEVEMENT LEVEL



Note: The boxplots are ordered to reflect the descending average pretest-posttest correlations calculated for the population, proficiency threshold, average performers, and lowest performer samples of students.

RQ 2: Are pretest-posttest correlation coefficients of state assessments lower for homogeneous samples of students than for heterogeneous samples of students?

There were statistically significant differences in pretest-posttest correlation coefficients across heterogeneity levels, F(1,227) = 240.67, *p* < 0.0001. More heterogeneous samples of students (standard deviation SD = 90 percent of the population standard deviation) had pretest posttest correlation coefficients that were on average .14 units larger than homogenous samples (*p* < .0001). In the vast majority (97 percent) of comparisons of pretest-posttest correlations between heterogeneous samples and homogeneous samples (sample standard deviation = 50 percent of the population standard deviation) of "average performers," the correlation coefficient for the heterogeneous sample was larger (see Table 4 and Figure 3).

When the pretest and posttest standard deviations were included as covariates in the analytic model, the pretest standard deviation became a statistically significant predictor of correlation coefficients (F(1,225) = 8.24, *p* < .01) and the heterogeneity factor became insignificant (F(1,225) = 1.70, *p* < .19). Given that these variables were highly correlated (*r* = 0.96), this is not a surprising result and reflects the finding that is substantively similar to that of the main ANOVA model. This was the only research question in which the inclusion of the pretest and posttest standard deviations changed the significance of the *F*-test of a factor of interest. For the remainder of the paper, we do not report any additional findings from this sensitivity analysis.

TABLE 4

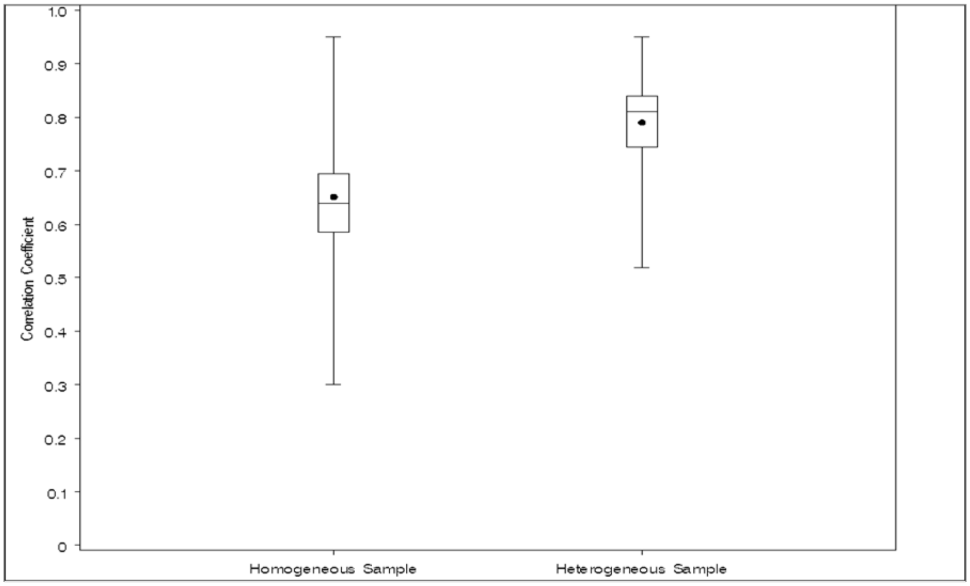
DESCRIPTIVE STATISTICS FOR AVERAGE PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY HETEROGENEITY LEVEL

	Average	Standard Deviation	Minimum	Maximum
Homogeneous Sample	0.65	0.13	0.30	0.95
Heterogeneous Sample	0.79	0.10	0.52	0.95

Note: *n* = 60 pretest-posttest correlation coefficients, calculated for homogenous (SD = 50 percent of population SD) and heterogeneous (SD = 90 percent of population SD) samples whose average performance was at the population average. The 60 pretest-posttest correlation coefficients include results aggregated across all six states, both subject areas (ELA and Mathematics), and across all grade groups.

FIGURE 3

PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY HETEROGENEITY LEVEL



RQ 3: Do pretest-posttest correlation coefficients differ by subject matter (ELA versus Mathematics)?

The pretest-posttest correlation coefficients were similar across the subjects of ELA and Mathematics (Table 5 and Figure 4). There was not a statistically significant difference in pretest-posttest correlation coefficients between the two subject areas, F(1,227) = 0.56, *p* = 0.46.

TABLE 5

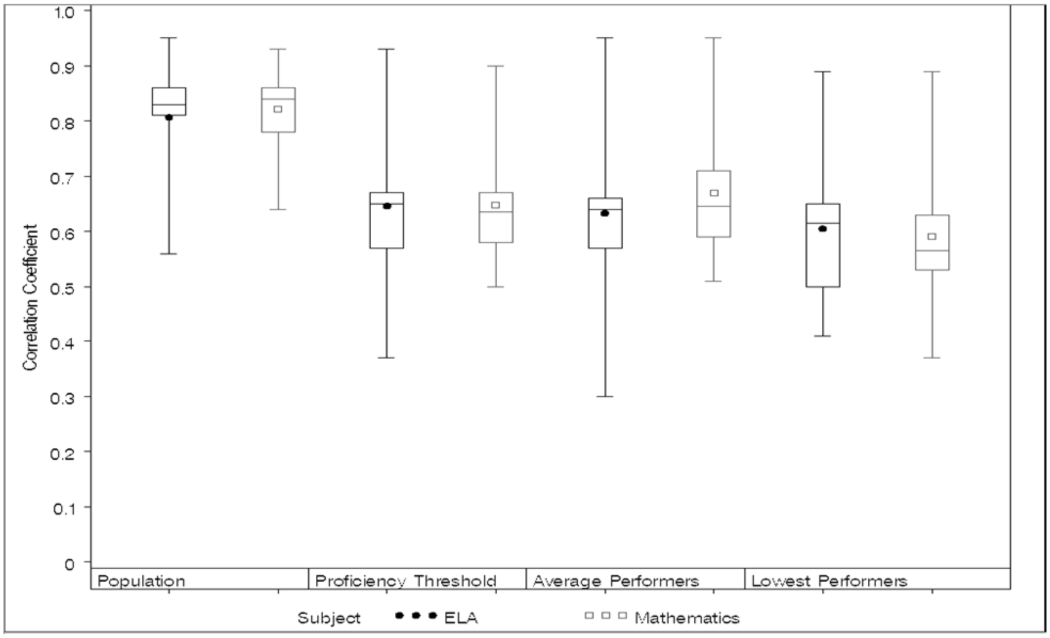
DESCRIPTIVE STATISTICS FOR AVERAGE PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY SUBJECT

	Subject	Average	Standard Deviation	Minimum	Maximum
Population	ELA	0.81	0.10	0.56	0.95
	Mathematics	0.82	0.07	0.64	0.93
Proficiency Threshold	ELA	0.65	0.13	0.37	0.93
	Mathematics	0.65	0.11	0.50	0.90
Average Performers	ELA	0.63	0.15	0.30	0.95
	Mathematics	0.67	0.12	0.51	0.95
Lowest Performers	ELA	0.61	0.13	0.41	0.89
	Mathematics	0.59	0.13	0.37	0.89

Note: *n* = 30 pretest-posttest correlation coefficients, calculated for each achievement-level-defined sample (or population). The 30 pretest-posttest correlation coefficients include results aggregated across all six states, in three grade groups, and over both pretest years.

FIGURE 4

PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY SUBJECT



Note: The boxplots are ordered to reflect the descending average pretest-posttest correlations calculated for the population, proficiency threshold, average performers, and lowest performer samples of students.

RQ 4: Do pretest-posttest correlation coefficients differ by grade level (early elementary versus late elementary/early middle school versus middle school)?

In the ANOVA results, there was not a statistically significant difference in pretest-posttest correlation coefficients across grade groups, F(2,227) = 0.55, *p* = 0.58. One descriptive finding from Table 6 and Figure 5 also merits attention. The standard deviation for the late elementary/early middle school group was less than half the size of the standard deviation for the other two grade groups (early elementary and middle school) in all three achievement levels of interest. Our descriptive data suggest that pretest-posttest correlation coefficients may be less variable (more stable) for students in these grades.[21]

[21] An alternate explanation might be that because the late elementary/early middle school grade group included only students in 5th grade, the smaller variability in correlation coefficients is due to composition differences in the grade group samples.

TABLE 6

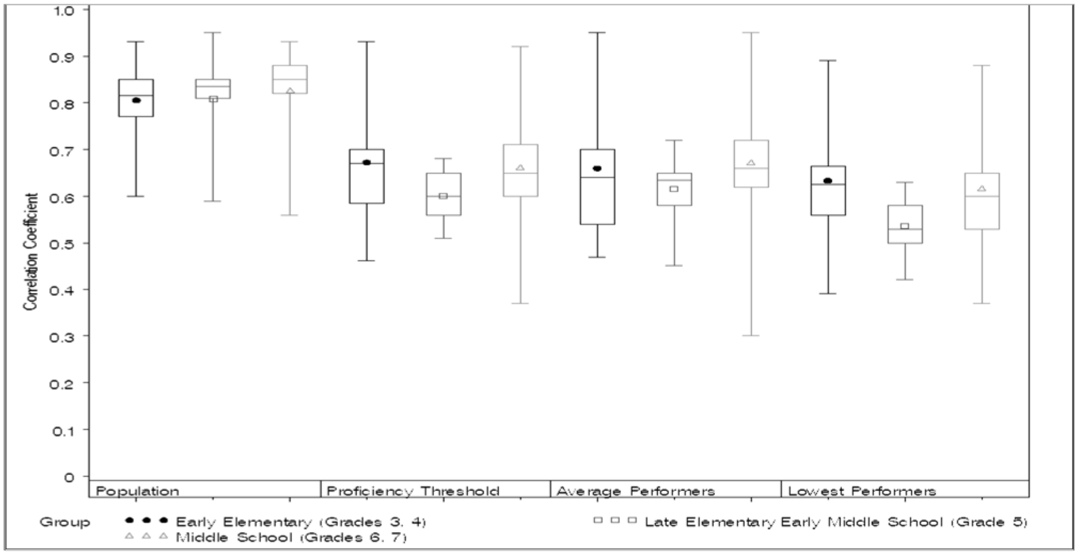
DESCRIPTIVE STATISTICS FOR AVERAGE PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY GRADE GROUP

	Grade Group	Average	Standard Deviation	Minimum	Maximum
Population	Early Elementary	0.81	0.09	0.60	0.93
	Late Elementary/Early Middle School	0.81	0.09	0.59	0.95
	Middle School	0.83	0.09	0.56	0.93
Proficiency Threshold	Early Elementary	0.67	0.13	0.46	0.93
	Late Elementary/Early Middle School	0.60	0.05	0.51	0.68
	Middle School	0.66	0.14	0.37	0.92
Average Performers	Early Elementary	0.66	0.15	0.47	0.95
	Late Elementary/Early Middle School	0.62	0.07	0.45	0.72
	Middle School	0.67	0.16	0.30	0.95
Lowest Performers	Early Elementary	0.63	0.15	0.39	0.89
	Late Elementary/Early Middle School	0.54	0.06	0.42	0.63
	Middle School	0.62	0.14	0.37	0.88

Note: *n* = 20, 18, and 22 pretest-posttest correlation coefficients, calculated for each achievement-level-defined sample (or population), for early elementary, late elementary/early middle school, and middle school, respectively. These pretest-posttest correlation coefficients include results aggregated across all six states, for both ELA and Mathematics assessments, and over both pretest years. Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

FIGURE 5

PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY GRADE GROUP



Note: The boxplots are ordered to reflect the descending average pretest-posttest correlations calculated for the population, proficiency threshold, average performers, and lowest performer samples of students.

RQ 5: Do pretest-posttest correlation coefficients differ by state?

Pretest-posttest correlation coefficients across states varied greatly, with state factor explaining more than 65 percent of the variance in the correlation coefficients ($\eta^2 = 0.65$). There were statistically significant differences in pretest-posttest correlation coefficients across states, F(5,227) = 322.25, $p < 0.0001$. Notably, the lowest correlations occurred where floor or ceiling effects were most prevalent. States E and F represent the extremes for pretest-posttest correlation differences across states (Table 7 and Figure 6). State E, which had the largest proportion of students scoring at the minimum and maximum scores across all six states (see Appendix Tables 13 and 14 for the ceiling and floor effects in State E), had the lowest average pretest-posttest correlation coefficients of all states, across all achievement levels.[22] In State E, the lowest-performing achievement sample had an average pretest-posttest correlation coefficient of r = 0.46. In State F, the lowest-performing achievement sample had an average pretest-posttest correlation coefficient of r = 0.87, a value higher than the population pretest-posttest correlation coefficient in State E (r = 0.67). Notably, while we find that sample achievement level helps explain variation in pretest-posttest correlation coefficients *within* states, the state factor explains more variance and can dominate the effect of achievement level when looking *across* states. The practical implication from this finding is that the pretest-posttest correlation coefficient observed in one state might not always serve as an appropriate estimate of the pretest-posttest correlation coefficient in a different state. Our analyses (using a convenience sample of states) suggest that correlation coefficients of samples of students with similar profiles do indeed vary across states, and the variation across states has implications for prospective research designs (see “Power Analysis” section below).

[22] The State E average was significantly lower than those of the other states (p < 0.0001) in all pairwise comparisons, and the State F average was significantly higher than those of the other states (p < 0.0001) in all comparisons, after multiple comparison adjustments.

TABLE 7

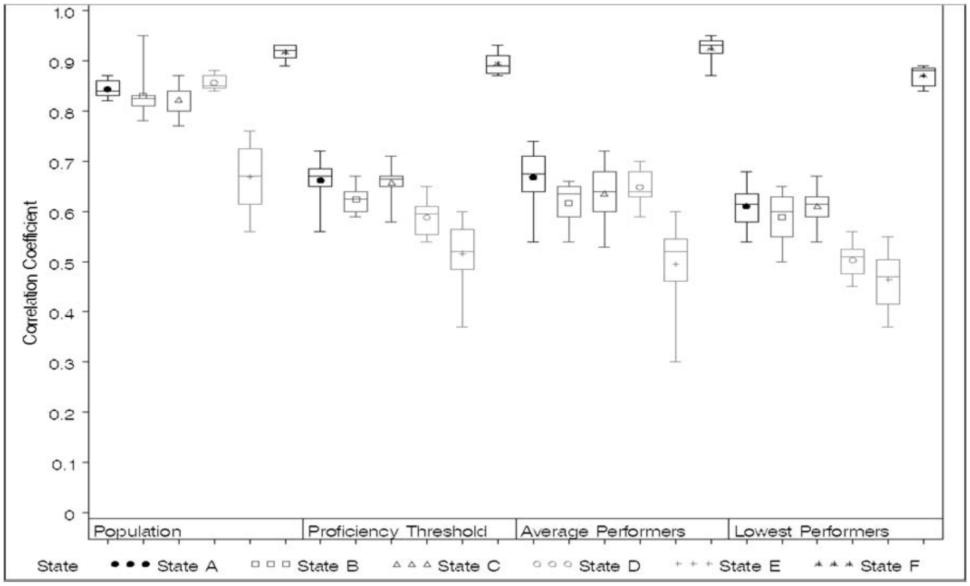
DESCRIPTIVE STATISTICS FOR AVERAGE PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY STATE

	State	Average	Standard Deviation	Minimum	Maximum
Population	A	0.84	0.02	0.82	0.87
	B	0.83	0.05	0.78	0.95
	C	0.82	0.03	0.77	0.87
	D	0.86	0.02	0.84	0.88
	E	0.67	0.07	0.56	0.76
	F	0.92	0.02	0.89	0.93
Proficiency Threshold	A	0.66	0.04	0.56	0.72
	B	0.63	0.03	0.59	0.67
	C	0.66	0.03	0.58	0.71
	D	0.59	0.04	0.54	0.65
	E	0.52	0.06	0.37	0.60
	F	0.89	0.02	0.87	0.93
Average Performers	A	0.67	0.05	0.54	0.74
	B	0.62	0.04	0.54	0.66
	C	0.64	0.05	0.53	0.72
	D	0.65	0.04	0.59	0.70
	E	0.50	0.08	0.30	0.60
	F	0.92	0.03	0.87	0.95
Lowest Performers	A	0.61	0.04	0.54	0.68
	B	0.59	0.05	0.50	0.65
	C	0.61	0.04	0.54	0.67
	D	0.50	0.04	0.45	0.56
	E	0.46	0.06	0.37	0.55
	F	0.87	0.02	0.84	0.89

Note: *n* = 12, 10, 10, 8, 12, and 8 pretest-posttest correlation coefficients, calculated for each achievement-level-defined sample (or population) for states A, B, C, D, E, and F, respectively. These pretest-posttest correlation coefficients include results aggregated across both ELA and Mathematics assessments, for all grade groups, and over both pretest years.

FIGURE 6

PRETEST-POSTTEST CORRELATION COEFFICIENTS, BY STATE



Note: The boxplots are ordered to reflect the descending average pretest-posttest correlations calculated for the population, proficiency threshold, average performers, and lowest performer samples of students.

RQ 6: Do pretest-posttest correlation coefficients differ over time?

Average pretest-posttest correlation coefficients were stable during the three years of analysis across the achievement distribution. There was not a statistically significant difference in pretest-posttest correlation coefficients across years, F(1,227) = 3.49, *p* = 0.06. The average change in absolute pretest-posttest correlation coefficients from one year to the next was between 0.03 and 0.05 across all grade groups, assessments, and states (Table 8).

TABLE 8

DESCRIPTIVE STATISTICS FOR ABSOLUTE CHANGE IN PRETEST-POSTTEST CORRELATION COEFFICIENTS OVER TIME

	Average	Standard Deviation	Minimum	Maximum
Population	0.03	0.04	0.00	0.14
Proficiency Threshold	0.05	0.03	0.00	0.10
Average Performers	0.04	0.03	0.00	0.10
Lowest Performers	0.04	0.04	0.00	0.17

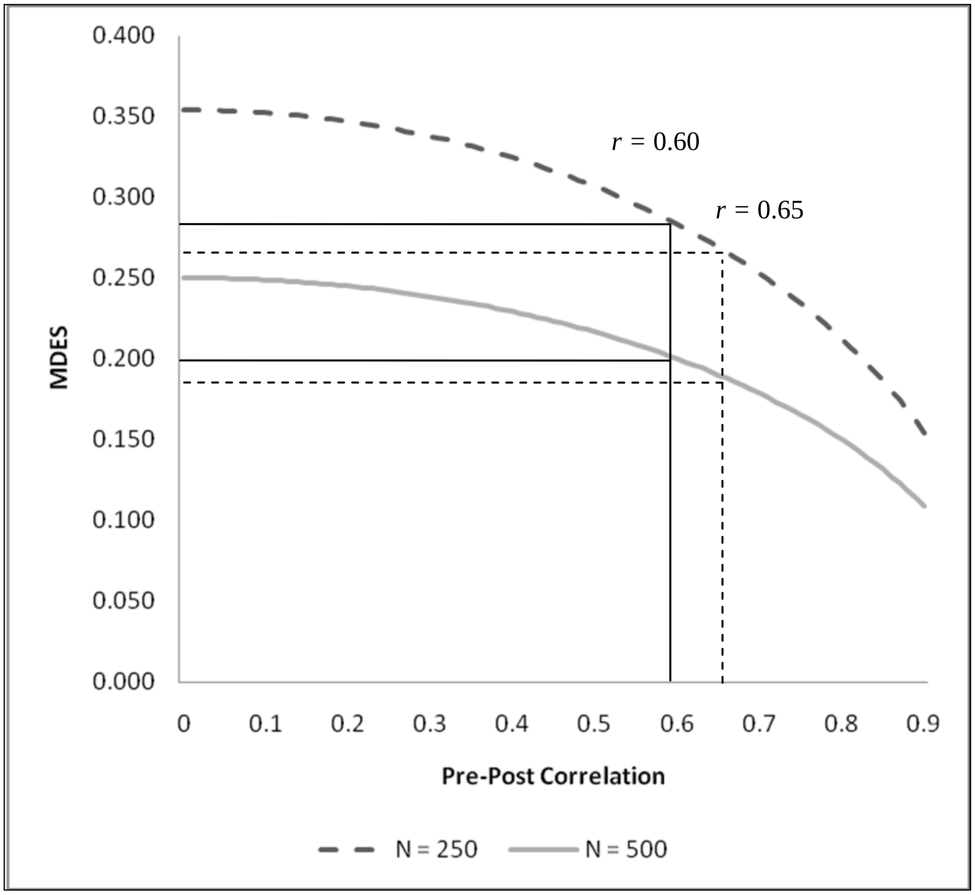
Note: *n* = 28 absolute differences in pretest-posttest correlation coefficients. Each of the 28 observations represents the absolute difference between the Year 1, Year 2 correlation coefficient and the Year 2, Year 3 correlation coefficient for a given grade group, assessment (ELA and Mathematics), and state combination (that is, $|r_{Year1,\ Year2} - r_{Year2,\ Year3}|$).

B. POWER ANALYSIS

Next, we examine the MDES that could be detected given the pretest-posttest correlation coefficients seen in our study. For example, Table 3 indicates that, across all six states, the average pretest-posttest correlation coefficient was $r = 0.65$ for “average performer” students, $r = 0.65$ for “proficiency threshold” students, and $r = 0.60$ for “lowest-performer” students. To estimate MDES given these correlations, we assume total study samples of 500 or 250 students, balanced assignment to treatment and control conditions, 80 percent power, and a two-tailed test with $\alpha = 0.05$ (see Equation 4 in Chapter II). Figure 7 relates pretest-posttest correlation coefficients to their corresponding MDES assuming that a pretest is to be included as a covariate.

FIGURE 7

MDES AS A FUNCTION OF PRETEST-POSTTEST CORRELATIONS, USING AVERAGE CORRELATIONS CALCULATED FOR ABILITY-DEFINED SUBGROUPS

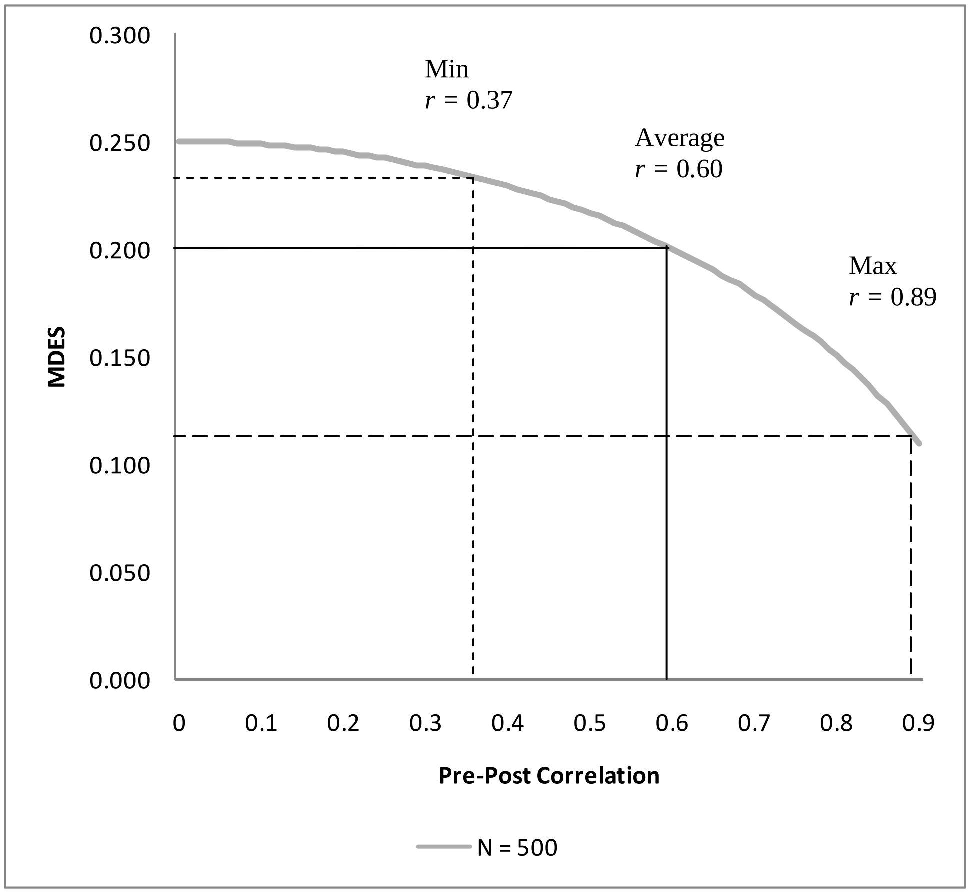


This figure suggests that, although attenuation in pretest-posttest correlation coefficients is found when samples of low performers are compared to samples of students with higher achievement, the effect of this attenuation for prospective power analysis may be modest. Using the squared pretest-posttest correlation coefficients to estimate the proportion of variance in the outcome that can be explained by the pretest, given a study sample of 500 students and other assumptions stated earlier, the MDES would be 0.19 for samples of either “average performing” or “proficiency threshold” students (r = 0.65) and 0.20 for samples of “lowest-performing” students (r = 0.60). These are relatively modest differences in MDES (5 percent), suggesting that the attenuation in pretest-posttest correlation coefficients that results from sampling students at different ability levels may not be a major concern for future RCT evaluations that use state proficiency assessments as outcome and pretest measures. Notably, this increase in MDES (of approximately 5 percent) will occur regardless of sample size, if the pretest-posttest correlation coefficient is reduced from $r = 0.65$ to $r = 0.60$. For example, in a sample of 250 students, the MDES would increase from approximately 0.27 to approximately 0.29.

The differences in MDES among *average* pretest-posttest correlation coefficients for achievement-defined samples obscure an important factor, however—the notable variability in correlation coefficients observed *across individual states*. Consider a power analysis for an evaluation of an intervention targeting very low-performing students—that is, the average achievement level for the study sample is approximately 1.3 population standard deviations below the state population mean. The observed pretest-posttest correlation coefficients for low performers in our study ranged from $r = 0.37$ to $r = 0.89$ (see minimum and maximum values for "lowest performers" in Table 3). We examined how changes in the correlation coefficient within this range can affect the power of an experiment and sample size requirements. Our results are displayed in Figure 8 below for a study with $n = 500$ students. In this figure, average ($r = 0.60$), minimum ($r = 0.37$), and maximum ($r = 0.89$) correlation levels are highlighted as pretest-posttest correlation levels of interest.

FIGURE 8

MDES AS A FUNCTION OF PRETEST-POSTTEST CORRELATION COEFFICIENT AND SAMPLE SIZE, WITH EMPIRICAL REFERENCE POINTS



With a pretest-posttest correlation of 0.89, the proportion of variance in the outcome explained by the pretest would be 0.89^2 or 79 percent. In this situation, the MDES for a sample of 500 students is 0.11 standard deviation units. However, if the pretest-posttest correlation were $r = 0.37$, then the MDES would be 0.23, an MDES more than twice as large.[23] If the hypothetical intervention of 500 students was expected to improve achievement levels by approximately 0.15 standard deviations, then impacts could be reliably detected only when state assessment have the largest pretest-posttest correlation (0.89) and not the two smaller values (0.60 or 0.37).

[23] As noted previously, if the pretest-posttest correlation coefficient were 0.60, then the MDES would be 0.20.

Furthermore, if a researcher naively used the population pretest-posttest correlation coefficient (*r* = 0.81) to estimate the proportion of variance explained in the outcome when planning a study, then he or she might incorrectly conclude that the study is well powered (because the incorrectly estimated MDES would be 0.15).

Readers should keep in mind that these estimates represent the extremes in variability in pretest-posttest correlation coefficients observed across the population data sets we were able to examine, because they pertain to the lowest-performing students. Given that the state factor in the ANOVA model explained more than 65 percent of the variance in correlation coefficients, these results highlight the differences in MDES that can occur given the differences in state proficiency assessments. The range in pretest-posttest correlation coefficients observed across our convenience sample of states suggests that cross-state differences, rather than achievement-level differences, may be of greatest concern for researchers planning to use state proficiency assessments in their experiments. As noted above, the implication for this finding is that researchers might consider exhibiting caution when assuming that the pretest-posttest correlation coefficient observed for a sample in one state is an appropriate estimate of the pretest-posttest correlation for an alternate state.

V. DISCUSSION

This report provided an empirical investigation of the extent to which pretest-posttest correlation coefficients differed across samples with varying mean achievement levels (and other factors). As an application, we examined empirically the extent to which the pretest-posttest correlation of state proficiency assessments is attenuated when low-performing students are selected as the sample of interest for an RCT evaluation of an educational intervention. As noted by May et al (2009), understanding how power is attenuated in study designs focused on these students is an important consideration, especially given the increased use of state assessments in education research. These instruments are seen as desirable outcome measures for education research because access to state assessment databases is inexpensive, data are widely available, and scores on these assessments have clear policy relevance (May et al. 2009).

The scaling of these state assessments tends to reduce the conditional standard error of measurement (CSEM) of scale scores at the center of the achievement distribution, at the expense of larger CSEM for scores at the low end of the achievement spectrum. This paper provided a theoretical motivation that the increased measurement error for low-performing students had the potential to reduce the pretest-posttest correlation coefficients for samples containing these students, relative to samples containing students drawn from the center of the achievement distribution.

In our analyses, pretest-posttest correlation coefficients were observed to be attenuated for low-performing students, relative to students whose scores were at the proficiency threshold or at the population average. Our analysis provided empirical reference points for the implications of this attenuation, which can provide insight for the planning of education experiments focused on low-performing students. Our analyses also suggest that attenuation in pretest-posttest correlations for the study's sample in a given state might be less of a concern for researchers than the considerable amount of between-state variability in pretest-posttest correlation coefficients.

In many prospective power analyses, researchers use the assumption that 50 percent of the variance in the outcome can be explained by a pretest, which corresponds to r = 0.71 (Schochet 2008; Bloom et al. 2005). Our results indicated that average pretest-posttest correlation coefficients for samples of homogenous students ranged between r = 0.60 and r = 0.65 for students of differing achievement levels, which suggests that this commonly used assumption could lead to underpowered studies when state assessments are used as outcome measures. However, this result is based on achievement data obtained from a convenience sample of states and large districts. This average result might not be generalizable outside of this sample.

The decision to use a state test is typically made during the planning stages of an intervention. In some circumstances, it may be useful to gather information on the pretest-posttest correlation in a sample of students similar to the group of interest for the interest. Our results suggest that pretest-posttest correlation coefficients are relatively stable from one year to the next. If an intervention is planned for a homogeneous sample of low-performing students, researchers could examine the pretest-posttest correlation coefficients for similar students in previous years on the intended outcome measure in the state(s) of interest. Using available administrative data will provide the most appropriate indicator of the expected correlation coefficients for the intended sample and will offer the best information for prospective power analyses.

This study focused solely on the power implications of pretest-posttest correlation coefficients for individual-level randomized designs. Of course, this is only one class of experimental designs that are of interest to education researchers. An important area for future research is to examine the extent to which the attenuation in correlation coefficients identified for low-performing students extends to attenuation in correlation coefficients for low-performing school averages. If there is an attenuation in the correlation coefficient for the aggregated, school-level averages in pretest-posttest correlation coefficients for low-performing schools, this will reduce power (and correspondingly increase MDES) for cluster-level assignment studies using state assessments as pretest and outcome measures.

Although the application of the analysis of pretest-posttest correlation coefficients in this paper focused on implications for MDES calculations, there are other ways in which pretest-posttest correlation coefficients can provide important information to education researchers. These correlation coefficients signal the extent to which student test scores (in a sample) are stable over the period between the pretest and the posttest. Education researchers focused on the stability of student performance (relative to their peers in the sample) might utilize the results from this paper to inform their own research in an application different from MDES calculations. The MDES illustrations in this paper are only one way that the pretest-posttest information from this paper can be used in future education research.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 2002.

Bloom H, I. Bos, and S. Lee. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for Evaluation of Education Programs." *Evaluation Review*, vol. 23, no. 4, 1999, pp. 445–469.

Bloom, H., L. Hayes, and A. Black. "Using Covariates to Improve Precision." New York: MDRC, 2005.

Bloom, H.S. "The Core Analytics of Randomized Experiments for Social Research." Working paper. New York: MDRC, 2006.

Boruch, R.F. "Randomized Experiments for Planning and Evaluation: A Practical Guide." *Applied Social Research Methods Series*, vol. 44. Thousand Oaks, CA: Sage Publications, 1997.

Chambers, B., P. Abrami, B. Tucker, R.E. Slavin, N.A. Madden, A. Cheung, and R. Gifford. "Computer-Assisted Tutoring in Success for All: Reading Outcomes for First Graders." *Journal of Research on Educational Effectiveness*, vol. 1, no. 2, 2008, pp. 120–137.

Clark, Melissa A., Sheena M. McConnell, Kristin Hallgren, Daniel W. Player, and Alison Wellington. "Evaluating Highly Selective Programs that Provide Alternative Routes to Teacher Certification: Feasibility and Design Issues." Final report submitted to the U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Princeton, NJ: Mathematica Policy Research, March 2008.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum, 1988.

Constantine, Jill M., Daniel W. Player, Timothy W. Silva, Kristin Hallgren, Mary Grider, and John G. Deke. "An Evaluation of Teachers Trained Through Different Routes to Certification." Final report submitted to the U.S. Department of Education, Institute of Education Sciences. Princeton, NJ: Mathematica Policy Research, February 2009.

Davidson, M.R., M.K. Fields, and J. Yang. "A Randomized Trial Study of a Preschool Literacy Curriculum: The Importance of Implementation." *Journal of Research on Educational Effectiveness*, vol. 2, no. 3, 2009, pp. 177–208.

Du Toit, M. (ed.). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc., 2003.

Feldt, L.S., and R.L. Brennan. "Reliability." In *Educational Measurement* (3rd ed.), edited by R. L. Linn. New York: American Council on Education and MacMillan, 1989.

Gargani J., and T. Cook. "How Many Schools? Limits of the Conventional Wisdom About Sample Size Requirements for Cluster Randomized Trials." Working paper. Berkeley, CA: University of California, 2005.

Glazerman, Steven M., Daniel P. Mayer, and Paul T. Decker. "Alternative Routes to Teaching: The Impacts of Teach For America on Student Achievement and Other Outcomes." *Journal of Policy Analysis and Management*, vol. 25, no. 1, winter 2006, pp. 75–96.

Hambleton, R.K., H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press, 1991.

Hunter, J.E., and F.L. Schmidt. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (2nd ed.). Thousand Oaks, CA: Sage, 2004.

Klein, Alice, Prentice Starkey, Douglas Clements, Julie Sarama, and Roopa Iyer. "Effects of a Pre-Kindergarten Mathematics Intervention: A Randomized Experiment." *Journal of Research on Educational Effectiveness*, vol. 1, no. 3, 2008, pp. 155–178.

Lee, W-C., R. L. Brennan, and M. J. Kolen. "Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study." *Journal of Educational Measurement,* vol. 37, 2000, pp. 1–20.

Lord, F.M., and M.R. Novick. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company, 1968.

May, Henry, Irma Perez-Johnson, Joshua Haimson, Samina Sattar, and Phil Gleason. *Using State Tests in Education Experiments: A Discussion of the Issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Olsen, R, and Unlu, F. "Using State Or Study-Administered Achievement Tests in Impact Evaluations" Paper presented at the Society for Research in Educational Effectiveness conference. Washington, D.C., 2010.

Petersen, N.S., M.J. Kolen, and H.D. Hoover. "Scaling, Norming, and Equating." In *Educational Measurement* (3rd ed.), edited by R.L. Linn. New York: American Council on Education and Macmillan, 1989.

Puma, M., Stephen Bell, Ronna Cook, Camilla Heid, Michel Lopez, Nicholas Zill, Gary Shapiro, Pam Proene, Debra Mekos, Monica Rohacek, Liz Quinn, Gina Adams, Janet Friedman, and Haidee Bernstein. "Head Start Impact Findings: First Year Findings." Final Report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, 2005.

Raju, N.S., L.R. Price, T.C. Oshima, and M.L. Nering. "Standardized Conditional SEM: A Case for Conditional Reliability." *Applied Psychological Measurement*, vol. 31, no. 3. 2007, pp. 169-180.

Schochet, Peter Z. "The Late Pretest Problem in Randomized Control Trials of Education Interventions." NCEE report 2009-4033. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

Spearman, C., "The Proof and Measurement of Association Between Two Things." *The American Journal of Psychology,* vol. 15, no. 1, 1904, pp. 72–101.

Spybrook, J., and S.W. Raudenbush. "An Examination of the Precision and Technical Accuracy of the First Wave of Group Randomized Trials Funded by the Institute of Education Sciences." *Educational Evaluation and Policy Analysis*, vol. 31, no. 3, 2009, pp. 298–318.

Thompson, B., and T. Vacha-Haase. "Psychometrics Is Datametrics: The Test Is Not Reliable." *Educational and Psychological Measurement,* vol. 60, 2000, pp. 174–195.

Tong, Y., and M. Kolen. "IRT Proficiency Estimators and Their Impact." Paper presented at the annual conference of the National Council on Measurement in Education. Denver, CO, 2010.

Tukey, J.W. "The Problem of Multiple Comparisons." Unpublished manuscript, 1953.

Vaughn, Sharon, Leticia R. Martinez, Sylvia Linan-Thompson, Colleen K. Reutebuch, Colleen D. Carlson, and David J. Francis. "Enhancing Social Studies Vocabulary and Comprehension for Seventh-Grade English Language Learners: Findings From Two Experimental Studies." *Journal of Research on Educational Effectiveness*, vol. 2, no. 4, 2009, pp. 297–324.

Zhu, P., A.M. Somers, and E. Wong. "How and When to Use State Tests to Measure Student Achievement: An Empirical Assessment Based on Four Recent Randomized Evaluations of Educational Interventions." Paper presented at the Society for Research in Educational Effectiveness. Washington, D.C., 2010

APPENDIX A

This appendix includes population descriptive statistics for ELA and Mathematics assessments for all eligible students in available years. In addition, we present the pretest-posttest correlation coefficients for the population and for three achievement levels. These appendix tables are grouped by state.

Appendix Tables 1 and 2 present the population descriptive statistics for the ELA assessment in State A for the paired years 1 and 2, and years 2 and 3, respectively. In these tables, we report only the population proportion of students who scored at the floor or at the ceiling of the state assessment, to protect the anonymity of the states. Given that observed scores for state tests are on different scales, providing information on population sizes, means, standard deviations, or scale minimum and maximum scores could potentially be used to identify states. Each row of the table presents data for a given grade. For example, in grade 3 approximately 0.16 percent of students obtained the minimum observed floor score for this assessment during the year 1 pretest and .01 percent of students scored at the ceiling during the pretest.

Appendix Table 3 presents the results of the pretest-posttest correlation analyses. Each row indicates correlations for a given pretest-posttest year combination for a given assessment in a particular grade grouping. The first row of Appendix Table 3 indicates pretest-posttest correlation coefficients for ELA assessments calculated across year 1 and year 2 for students in early elementary school (grades 3 and 4 during year 1). The population pretest-posttest correlation coefficient (r = 0.83) is calculated across all eligible students in this grade group. The next six columns present the pretest-posttest correlation coefficient and 95 percent confidence interval, calculated for the sample of 500 students defined for a given achievement level. In these six columns, all samples of 500 students were constructed to be homogeneous (standard deviation of the sample = 50 percent of the population standard deviation). Again, using the first row as an example, the pretest-posttest correlation coefficient (and 95 percent confidence interval) for the homogeneous, lowest-performing sample is 0.62 (0.56, 0.67). The final two columns of the table present the pretest-posttest correlation coefficients and confidence intervals for the heterogeneous (average achievement level) samples.

Similar results are presented for states B to F in Appendix Tables 4 to 15.

POPULATION DESCRIPTION

In all states except E, increases in average scale scores for a particular test corresponded with the higher grade in which the assessment was administered (the data are not shown to protect the anonymity of the states). The increase in average scale scores corresponding with increases in the grade assessed is a common feature of vertically equated tests. This was not the case in State E, where the mean score appeared relatively stable across all grades, and the assessment manual confirmed that this was not a vertically equated examination.

Across all assessments in all years except for States E and F, very few students scored at the floor or ceiling for any of the assessments. In States A, B, C, and D, less than one percent of students in a grade scored at either the floor or the ceiling (the minimum or maximum observed score, respectively) on a given assessment. Given the infrequent floor and ceiling effects in these data, this issue was not considered important as a component of pretest-posttest correlation coefficients for States A–D. However, in State E, there were considerable numbers of students with test scores at the floor or ceiling. On average, more than 1 percent of reading scores were at the floor and 4 percent were at the ceiling. In this state, the prevalence of floor and ceiling effects is notable. In State F, during the pretest year, approximately 1 percent of student scores (across

ELA and Mathematics) were at the floor. Floor effects were less prevalent during the posttest year (approximately 0.06 percent of students scored at the floor), and ceiling effects were relatively infrequent in both pretest (0.26 percent) and posttest (0.04 percent) years.

APPENDIX TABLE 1

STATE A ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.16	0.01	0.02	0.02
4	0.00	0.02	0.01	0.00
5	0.00	0.01	0.00	0.00
6	0.00	0.04	0.01	0.02
7	0.00	0.02	0.01	0.02

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.17	0.02	0.04	0.04
4	0.00	0.02	0.02	0.06
5	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.01
7	0.01	0.03	0.01	0.08

Source: State A student population administrative data.

APPENDIX TABLE 2

STATE A ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		Posttest (Year 2)	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.08	0.04	0.00	0.04
4	0.00	0.04	0.00	0.15
5	0.00	0.09	0.00	0.00
6	0.00	0.04	0.00	0.11
7	0.00	0.11	0.00	0.06

	Pretest (Year 2)		Posttest (Year 3)	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.05	0.11	0.00	0.05
4	0.00	0.05	0.00	0.09
5	0.01	0.14	0.00	0.13
6	0.00	0.11	0.00	0.11
7	0.01	0.13	0.00	0.08

Source: State A student population administrative data.

APPENDIX TABLE 3

STATE A PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Early Elementary	.83	.62	(.56, .67)	.67	(.62, .72)	.65	(.60, .70)	.81	(.77, .83)
	Year2–Year3	Early Elementary	.83	.61	(.55, .66)	.68	(.63, .72)	.69	(.64, .73)	.80	(.77, .83)
	Year1–Year2	Late Elementary	.83	.58	(.52, .64)	.56	(.49, .61)	.54	(.47, .60)	.81	(.78, .84)
	Year2–Year3	Late Elementary	.83	.58	(.52, .64)	.65	(.60, .70)	.63	(.57, .68)	.81	(.78, .84)
	Year1–Year2	Middle School	.82	.65	(.59, .70)	.71	(.66, .75)	.66	(.61, .71)	.80	(.77, .83)
	Year2–Year3	Middle School	.82	.63	(.57, .68)	.65	(.60, .70)	.62	(.57, .67)	.78	(.74, .81)
Mathematics	Year1–Year2	Early Elementary	.85	.68	(.63, .73)	.67	(.62, .72)	.66	(.60, .70)	.83	(.80, .86)
	Year2–Year3	Early Elementary	.85	.64	(.59, .69)	.69	(.65, .74)	.71	(.66, .75)	.83	(.80, .86)
	Year1–Year2	Late Elementary	.86	.56	(.49, .61)	.61	(.55, .66)	.69	(.65, .74)	.84	(.81, .86)
	Year2–Year3	Late Elementary	.86	.54	(.47, .60)	.68	(.63, .72)	.72	(.68, .76)	.83	(.80, .85)
	Year1–Year2	Middle School	.87	.61	(.55, .66)	.72	(.67, .76)	.74	(.70, .78)	.86	(.83, .88)
	Year2–Year3	Middle School	.87	.63	(.58, .68)	.66	(.60, .70)	.71	(.66, .75)	.84	(.81, .86)

Source: State A student population administrative data.
Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

APPENDIX TABLE 4
STATE B ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.01	0.02	0.00	0.01
4	0.00	0.03	0.00	0.05
5	0.01	0.02	0.04	0.02

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of population with Minimum Score	Percentage of Population with Maximum Score
3	0.15	0.01	0.02	0.02
4	0.13	0.01	0.01	0.03
5	0.12	0.05	0.03	0.03
6	0.11	0.01	0.03	0.02

Source: State B student population administrative data.

APPENDIX TABLE 5

STATE B ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.35	0.01	0.01	0.02
4	0.29	0.03	0.03	0.04
5	0.33	0.02	0.04	0.14

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.02	0.01	0.03	0.20
4	0.00	0.02	0.02	0.02
5	0.03	0.04	0.04	0.11
6	0.02	0.15	0.04	0.01

Source: State B student population administrative data.

APPENDIX TABLE 6

STATE B PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Early Elementary	.83	.63	(.58, .68)	.60	(.54, .65)	.65	(.60, .70)	.78	(.74, .81)
	Year2–Year3	Early Elementary	.82	.54	(.48, .60)	.59	(.53, .64)	.65	(.59, .70)	.79	(.76, .82)
	Year1–Year2	Late Elementary	.81	.62	(.57, .67)	.60	(.54, .65)	.57	(.51, .63)	.79	(.75, .82)
	Year2–Year3	Late Elementary	.85	.63	(.57, .68)	.67	(.62, .72)	.66	(.61, .71)	.83	(.80, .86)
	Year2-Year3	Middle School	.83	.65	(.60, .70)	.65	(.60, .70)	.59	(.53, .64)	.79	(.75, .82)
Mathematics	Year1–Year2	Early Elementary	.78	.57	(.51, .63)	.64	(.58, .69)	.59	(.53, .64)	.75	(.70, .78)
	Year2–Year3	Early Elementary	.80	.63	(.57, .68)	.64	(.58, .69)	.54	(.48, .60)	.74	(.70, .78)
	Year1–Year2	Late Elementary	.81	.50	(.43, .56)	.61	(.55, .66)	.65	(.59, .70)	.77	(.73, .80)
	Year2–Year3	Late Elementary	.83	.58	(.52, .63)	.63	(.57, .68)	.64	(.58, .69)	.80	(.76, .83)
	Year2-Year3	Middle School	.84	.55	(.49, .61)	.62	(.56, .67)	.63	(.57, .68)	.83	(.80, .86)

Source: State B student population administrative data.

Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

APPENDIX TABLE 7

STATE C ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.05	0.10	0.00	0.05
6	0.01	0.01	0.03	0.07

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.06	0.30	0.00	0.12
4	0.03	0.04	0.02	0.01
5	0.03	0.02	0.01	0.01
6	0.02	0.01	0.00	0.05
7	0.03	0.07	0.00	0.03

Source: State C student population administrative data.

APPENDIX TABLE 8

STATE C ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.04	0.76	0.15	0.14
6	0.30	0.03	0.00	0.03

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.05	0.64	0.00	0.13
4	0.16	0.14	0.00	0.23
5	0.27	0.07	0.00	0.13
6	0.00	0.06	0.00	0.07
7	0.70	0.02	0.01	0.03

Source: State C student population administrative data.

APPENDIX TABLE 9

STATE C PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Early Elementary	.77	.62	(.56, .67)	.71	(.67, .75)	.60	(.54, .65)	.73	(.68, .76)
	Year2–Year3	Early Elementary	.81	.65	(.59, .70)	.67	(.62, .72)	.59	(.53, .65)	.80	(.77, .83)
	Year2-Year3	Late Elementary	.81	.61	(.55, .66)	.67	(.62, .72)	.65	(.60, .70)	.84	(.81, .86)
	Year1–Year2	Middle School	.84	.63	(.57, .68)	.66	(.61, .71)	.64	(.58, .69)	.79	(.76, .82)
	Year2–Year3	Middle School	.87	.67	(.62, .72)	.67	(.62, .72)	.64	(.58, .69)	.79	(.76, .82)
Mathematics	Year1–Year2	Early Elementary	.77	.59	(.53, .64)	.58	(.52, .63)	.53	(.46, .59)	.73	(.68, .77)
	Year2–Year3	Early Elementary	.80	.63	(.57, .68)	.67	(.62, .72)	.63	(.57, .68)	.78	(.74, .81)
	Year2-Year3	Late Elementary	.84	.54	(.48, .60)	.65	(.59, .69)	.68	(.63, .72)	.82	(.79, .85)
	Year1–Year2	Middle School	.85	.59	(.53, .64)	.65	(.60, .70)	.69	(.64, .73)	.82	(.79, .85)
	Year2–Year3	Middle School	.87	.59	(.53, .64)	.65	(.60, .70)	.72	(.67, .76)	.86	(.84, .88)

Source: State C student population administrative data.

Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

APPENDIX TABLE 10

STATE D ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
5	0.00	0.02	0.01	0.02
6	0.00	0.03	0.02	0.03

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
5	0.00	0.00	0.04	0.03
6	0.02	0.02	0.00	0.03

Source: State D student population administrative data.

APPENDIX TABLE 11

STATE D ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
5	0.00	0.02	0.00	0.08
6	0.01	0.05	0.00	0.03

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
5	0.03	0.03	0.01	0.08
6	0.02	0.08	0.00	0.03

Source: State D student population administrative data.

APPENDIX TABLE 12

STATE D PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Late Elementary	.86	.50	(.43, .56)	.54	(.48, .60)	.64	(.58, .69)	.84	(.81, .86)
	Year2–Year3	Late Elementary	.85	.49	(.42, .55)	.59	(.52, .64)	.59	(.54, .65)	.82	(.79, .85)
	Year1–Year2	Middle School	.88	.45	(.38, .52)	.65	(.59, .70)	.70	(.65, .74)	.84	(.81, .86)
	Year2–Year3	Middle School	.88	.46	(.38, .52)	.62	(.56, .67)	.70	(.65, .74)	.86	(.83, .88)
Mathematics	Year1–Year2	Late Elementary	.84	.52	(.45, .58)	.55	(.49, .61)	.63	(.57, .68)	.82	(.79, .84)
	Year2–Year3	Late Elementary	.84	.52	(.45, .58)	.60	(.54, .65)	.64	(.58, .68)	.84	(.81, .87)
	Year1–Year2	Middle School	.85	.53	(.47, .59)	.60	(.54, .65)	.66	(.60, .70)	.84	(.81, .86)
	Year2–Year3	Middle School	.85	.56	(.50, .62)	.56	(.50, .62)	.63	(.58, .68)	.81	(.77, .84)

Source: State D student population administrative data.

Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

APPENDIX TABLE 13

STATE E ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.87	15.31	0.00	0.00
4	4.11	1.76	0.01	0.00
5	3.97	2.43	2.25	4.85
6	3.11	3.71	0.04	0.01
7	3.32	2.13	1.40	3.72

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.96	11.68	0.27	0.00
4	1.50	4.36	0.00	0.00
5	3.66	2.49	0.28	7.04
6	1.46	4.79	0.02	0.01
7	1.53	1.53	0.00	8.04

Source: State E student population administrative data.

APPENDIX TABLE 14

STATE E ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.01	4.13	0.01	0.00
4	0.02	0.00	0.00	0.00
5	0.02	0.00	0.14	4.18
6	1.86	2.79	0.00	0.00
7	1.90	1.14	0.03	0.00

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.01	0.00	0.00	0.00
4	0.01	6.77	0.00	0.00
5	0.01	0.00	0.01	0.00
6	0.79	4.16	0.00	0.00
7	0.85	1.88	0.03	0.00

Source: State E student population administrative data.

APPENDIX TABLE 15

STATE E PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Early Elementary	.68	.41	(.34, .48)	.56	(.50, .62)	.47	(.40, .54)	.58	(.52, .64)
	Year2–Year3	Early Elementary	.60	.43	(.36, .50)	.46	(.39, .53)	.48	(.40, .54)	.62	(.56, .67)
	Year1–Year2	Late Elementary	.71	.50	(.43, .56)	.52	(.46, .58)	.45	(.38, .52)	.72	(.68, .76)
	Year2–Year3	Late Elementary	.59	.51	(.45, .58)	.57	(.51, .63)	.53	(.46, .59)	.54	(.48, .60)
	Year1–Year2	Middle School	.63	.50	(.43, .56)	.47	(.39, .53)	.40	(.33, .47)	.54	(.47, .60)
	Year2–Year3	Middle School	.56	.49	(.42, .56)	.37	(.30, .45)	.30	(.22, .38)	.49	(.42, .56)
Mathematics	Year1–Year2	Early Elementary	.74	.39	(.32, .46)	.50	(.43, .57)	.54	(.48, .60)	.70	(.66, .75)
	Year2–Year3	Early Elementary	.66	.55	(.49, .61)	.55	(.49, .61)	.51	(.44, .57)	.67	(.62, .71)
	Year1–Year2	Late Elementary	.76	.42	(.35, .49)	.51	(.44, .57)	.60	(.54, .65)	.75	(.71, .78)
	Year2–Year3	Late Elementary	.64	.45	(.37, .52)	.60	(.54, .65)	.58	(.51, .63)	.64	(.58, .69)
	Year1–Year2	Middle School	.76	.37	(.29, .44)	.57	(.51, .63)	.55	(.49, .61)	.75	(.70, .78)
	Year2–Year3	Middle School	.70	.54	(.48, .60)	.52	(.45, .58)	.54	(.47, .60)	.67	(.61, .71)

Source: State E student population administrative data.

Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

APPENDIX TABLE 16

STATE F ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (ELA)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	1.19	0.09	0.06	0.03
4	2.49	0.03	0.07	0.03
5	0.03	0.10	0.03	0.03
6	0.03	0.03	0.03	0.03
7	0.07	0.13	0.07	0.03

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.86	0.11	0.03	0.03
4	1.08	0.15	0.06	0.03
5	1.21	0.10	0.03	0.03
6	1.49	0.06	0.03	0.03
7	0.79	0.06	0.03	0.24

Source: State F student population administrative data.

APPENDIX TABLE 17

STATE F ELIGIBLE POPULATION DESCRIPTIVE STATISTICS (MATHEMATICS)

	Pretest (Year 1)		**Posttest (Year 2)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	0.93	0.41	0.06	0.03
4	1.74	0.03	0.07	0.03
5	0.03	0.13	0.06	0.03
6	0.03	0.06	0.03	0.06
7	0.03	0.10	0.23	0.03

	Pretest (Year 2)		**Posttest (Year 3)**	
Grade	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score	Percentage of Population with Minimum Score	Percentage of Population with Maximum Score
3	1.14	1.58	0.03	0.03
4	0.88	0.55	0.03	0.03
5	1.18	1.08	0.03	0.03
6	3.63	0.10	0.16	0.03
7	2.44	0.32	0.03	0.03

Source: State F student population administrative data.

APPENDIX TABLE 18

STATE F PRETEST-POSTTEST CORRELATION COEFFICIENTS

				Homogeneous Sample (Sample standard deviation = 50% of population standard deviation)						Heterogeneous Sample (Sample standard deviation = 90% of population standard deviation)	
				Lowest Performers		Proficiency Threshold		Average Performance		Average Performance	
Subject	Year	Grade Level	Population Correlation Coefficient	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval	Correlation Coefficient	95% Confidence Interval
Reading	Year1–Year2	Early Elementary	.91	.85	(.82, .87)	.89	(.88, .91)	.92	(.91, .94)	.95	(.94, .96)
	Year2–Year3	Early Elementary	.93	.89	(.86, .90)	.93	(.91, .94)	.91	(.90, .93)	.95	(.94, .96)
	Year1–Year2	Middle School	.90	.88	(.85, .89)	.89	(.87, .91)	.93	(.92, .94)	.93	(.92, .94)
	Year2–Year3	Middle School	.93	.88	(.85, .90)	.92	(.91, .93)	.95	(.94, .96)	.93	(.92, .94)
Mathematics	Year1–Year2	Early Elementary	.92	.85	(.82, .87)	.88	(.86, .90)	.93	(.92, .94)	.93	(.92, .94)
	Year2–Year3	Early Elementary	.93	.89	(.87, .90)	.87	(.85, .89)	.95	(.94, .96)	.89	(.87, .91)
	Year1–Year2	Middle School	.89	.84	(.81, .86)	.87	(.85, .89)	.87	(.85, .89)	.92	(.90, .93)
	Year2–Year3	Middle School	.92	.88	(.86, .90)	.90	(.88, .91)	.93	(.92, .94)	.94	(.93, .95)

Source: State F student population administrative data.

Note: Early Elementary = grades 3 and 4 during pretest year; Late Elementary/Early Middle School = grade 5 during pretest year; Middle School = grades 6 and 6 during pretest year.

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA

www.ed.gov ies.ed.gov