A Three-Study Examination of Test-Based Accountability Metrics

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:33052855

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
A Three-Study Examination of Test-Based Accountability Metrics

Darrick Shen-wei Yee

Andrew D. Ho, Professor of Education (Chair)
Daniel M. Koretz, Henry Lee Shattuck Professor of Education
Luke W. Miratrix, Assistant Professor of Education

A Thesis Presented to the Faculty
of the Graduate School of Education of Harvard University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

2017
Acknowledgements

I would like to express my deepest appreciation and gratitude to the following individuals and organizations, without whom none of this work would have been possible:

Harvard University

Andrew Ho, Daniel Koretz, Luke Miratrix, Thomas Hehir, John Willett, Richard Murnane, Carol Yu, and Joseph Blitzstein

Center for Education Policy Research at Harvard

Margaret Nipson, Timothy Brennan, Lindsay Page, and William Marinell

Delaware Department of Education

Atnre Alleyne and Liru Zhang

Stanford University

Sean Reardon and Demetra Kalogrides
Table of Contents

Abstract .......................................................................................................................................... iv

I. Introduction .................................................................................................................................. 1

II. Discreteness Causes Bias in Percentage-Based Comparisons: A Case Study from 
    Educational Testing ................................................................................................................ 3

   1. Introduction ........................................................................................................................ 3
   2. Test Score Discretization and Linking ............................................................................. 5
   3. Model ........................................................................................................................................ 7
   4. Estimation of $\Delta PAC$ Bias in State Tests ................................................................. 17
   5. Discussion .......................................................................................................................... 21
   A. Addendum ....................................................................................................................... 25

    State-wide Natural Experiment ............................................................................................. 38

   1. Introduction ....................................................................................................................... 38
   2. Background and Research Agenda ................................................................................. 39
   3. Research Design ............................................................................................................... 43
   4. Findings ............................................................................................................................ 52
   5. Threats to Validity ............................................................................................................ 62
   6. Effects of Content Changes .............................................................................................. 69
   7. Discussion .......................................................................................................................... 74

IV. The Sensitivity of Gain-Based Growth Metrics to Transformations of Scale ...................... 78

   1. Introduction and background .......................................................................................... 78
   2. Motivation and Research Agenda ................................................................................... 80
3. Framework ............................................................................................................................. 83
4. Evaluating School Growth Rating Sensitivity .................................................................. 97
5. Results .................................................................................................................................. 105
6. Summary and Conclusion ............................................................................................... 112

V. Conclusion .......................................................................................................................... 115

References ................................................................................................................................ 116
Abstract

Recent state and federal policy initiatives have led to the development of a multitude of statistics intended to measure school performance. Of these, statistics constructed from student test scores number among both the most widely-used and most controversial. In many cases, researchers and policymakers alike are not fully aware of the ways in which these statistics may lead to unjustified inferences regarding school effectiveness. A substantial amount of recent research has attempted to remedy this, although much remains unknown.

This thesis seeks to contribute to these research efforts via three papers, each examining how a commonly-employed accountability statistic may be influenced by factors unrelated to student proficiency or school effectiveness. The first paper demonstrates how the discrete nature of test scores leads to biased estimates of changes in the percentage of “proficient” students between any two given years and examines estimators that provide better recovery of this parameter. The second paper makes use of a state-wide natural experiment to show that a change in testing program, from paper-and-pencil to computer-adaptive, may cause apparent changes in achievement gaps even when relative student proficiencies have remained constant. The third paper examines “growth-based” accountability metrics based on vertically-scaled assessments, showing that certain types of metrics based on gain scores can be modeled via nonlinear transformations of the underlying vertical scale. It then makes use of this result to investigate the potential magnitude of impacts of such transformations on growth-based school accountability ratings.
I. Introduction

Among the many changes in educational policy and practice precipitated by the No Child Left Behind Act of 2001, two notable impacts stand out. First, nearly all public school students in the nation were required to take annual assessments in mathematics and reading between grades 3 and 8 and at least once between grades 9 and 12. Second, a new emphasis on “accountability” for schools (and, later, teachers) led to the use of assessment results as a basis for evaluating the effectiveness of educators and educational institutions, with consequences for those that were determined to perform poorly.

While measuring student achievement via standardized tests was by no means new, the use of test scores to construct statistical measures of educational effectiveness meant that assessment results were, at an unprecedented scale, being employed for purposes for which they were not originally designed. Perhaps because of this, the properties of many test-based statistics employed in accountability policy are often poorly understood or not clearly documented. Although a large and growing body of research has emerged to address these gaps in understanding, the proliferation of novel assessment-based accountability metrics across states and school districts generally continues to outpace research-based efforts to document their unintended consequences. The papers I present below all seek to contribute to this research by investigating validity-related issues for three widely-used accountability indicators.

The first paper, “Discreteness Causes Bias in Percentage-Based Comparisons: A Case Study from Educational Testing,” demonstrates that simple changes in “percent proficient” often lead to incorrect inferences regarding student progress when estimated
from discrete score scales employed by many states; its addendum shows that simple linear interpolation generally provides substantially better estimates.

The second paper, “Does Computer-Adaptive Testing Affect Score Gaps? Evidence from a State-wide Natural Experiment,” analyzes a state-wide transition from paper-and-pencil to computer-adaptive testing (CAT) and finds that the transition caused the apparent score gap between male and female students in Reading to narrow by nearly half. Although the tests were not designed as equivalent, the results suggest that at least some of this narrowing was attributable to proficiency-irrelevant differences between the tests. A conclusion of this paper is that policymakers should guard against drawing unjustified achievement-gap comparisons when some schools are tested using CATs and others are not.

The third paper, “The Sensitivity of Gain-Based Growth Metrics to Transformations of Scale,” presents a basic framework for analysis of school accountability ratings that intend to measure student growth, showing that a broad range of metrics based on “gain scores” on vertically-scaled assessments can be modeled by subjecting vertical scale scores to nonlinear transformations and computing gain scores from the resulting transformed scores. Using simulated and empirical data, it then investigates the sensitivity of accountability results to plausible variations in scale score and growth model characteristics, in terms of their effects on school rankings.

My hope is that these papers will help policymakers make better-informed decisions – or, at least, avoid making unjustified inferences – when interpreting accountability statistics derived from test scores, especially when those statistics may carry high stakes for educators.
II. Discreteness Causes Bias in Percentage-Based Comparisons: A Case Study from Educational Testing

1 Introduction

Discretization of continuous distributions is ubiquitous in practice. Previous studies have explored discretization in a variety of forms, including rounding, where data are discretized evenly to integers or decimal places; heaping, when some data are discretized coarsely and others finely; and interval censoring, when data are known to exist within some interval and are assigned the value of the interval endpoint (Heitjan, 1989; Heitjan and Rubin, 1991). The terms binning, grouping, and coarsening are less specific and refer broadly to a process that sacrifices precision for simplicity by assuming similar observations are equal in value. Generally, we desire that estimates derived from discretized data will recover parameters of the continuous data.

The consequences of discretization depend upon the nature of the discretization and the target parameter. Sheppard’s correction (1898) is a well-known adjustment for bias in moments of an evenly discretized normal distribution, with implications that extend to parameters from least squares regression models (e.g., Dempster and Rubin, 1983; Schneeweiss, Komlos, and Ahmad, 2010). Horton, Lipsitz, and Parzen (2003) describe how rounding to prevent implausible values in multiple imputation procedures can impart bias to results. In this paper, we show that discreteness can cause considerable bias when we compare two or more distributions by a particular summary statistic: the cumulative proportion, or its complement, the “percent above cutoff” (PAC).

1 This paper was coauthored with Andrew D. Ho, Professor of Education, Harvard Graduate School of Education.
The $PAC$-based comparison is a staple of descriptive reporting. The poverty rate, for example, is a $PAC$ statistic that uses an income-based cutoff, while the obesity rate is a $PAC$ statistic with a cutoff based on the body-mass index. Comparisons of poverty and obesity rates – for example, across regions, subpopulations, or time periods – rely on the assumption that the cutoff is the same for each comparison group. We focus on a case study in education where discretization leads to a consequential violation of this assumption. In this context, the $PAC$ often represents a passing rate, or a percentage of students considered “Proficient” in a particular academic subject, such as mathematics or English. Examples include licensure and certification tests (e.g., American Institute of CPAs, 2013; National Conference of Bar Examiners, 2012), Advanced Placement exams (The College Board, 2013), and the U.S. Department of Education’s National Assessment of Educational Progress (U.S. Department of Education, n.d.). The change in $PAC$, or $\Delta PAC$, is thus a measure of educational progress or improvement.

In 2002, this metric gained newfound importance with the signing of the No Child Left Behind (NCLB) Act, an ambitious piece of U.S. federal legislation that set the goal of 100% student “Proficiency” by 2014. The policy required U.S. states to administer standardized tests in multiple subjects to the vast majority of public school students; set cutoff scores on the tests such that students achieving or exceeding the cutoff would be considered “Proficient” in the tested subject; calculate percentages of Proficient students at various levels of aggregation; and increase these percentages to 100% by 2014. Schools with insufficient percentages of Proficient students faced sanctions, including possible school restructuring and closure. Although federal policies have since allowed some flexibility, in particular for the 100% goal in 2014 that no state ultimately met,
percentages remain a central metric for reporting and incentivizing educational progress (U.S. Department of Education, 2012).

We model and address a significant source of bias associated with the $\Delta PAC$ metric that, to our knowledge, has never been addressed formally. Previous authors have observed that the relationship between $\Delta PAC$ and changes in average test scores, $\Delta \bar{X}$, is nonlinear and determined by the shape of the distributions and the magnitude of the initial and final percentages (Holland, 2002; Ho, 2008). This relationship is smooth and generally predictable. In contrast, we examine a source of bias attributable to unpredictable changes in discretization over time. We model this process by varying the “discretizing partitions” applied to each comparison group. We show that year-over-year changes in these partitions impart severe volatility to the $\Delta PAC$ metric that threatens trend interpretations, leads to sign reversals, and overshadows conventional sampling variability for most large-scale (district- and state-level) applications.

2. Test Score Discretization and Linking

First, we review the two steps in educational test score construction that ultimately impart this unpredictable bias to $PAC$-based trends: discretization and linking. Later, we will describe a general model and show how it may arise in other situations.

2.1. Discretization

Although classroom intuition holds that test scores are simple counts of correctly answered questions (“number-correct scores”), large-scale testing programs generally convert these counts to “scale scores,” which span ranges such as the 200-800 SAT scale and the 1-36 ACT scale used in many U.S. college admissions decisions. The state of the art for scale score construction is Item Response Theory (IRT; Lord, 1980; Yen and
Fitzpatrick, 2006), a modeling framework that allows test “items” (the formal term for test questions) to differ across examinees and over time while still providing comparable scores on a continuous latent scale. For the purpose of this presentation, IRT is only important in that it reflects a commonsense intuition about a continuous score scale for academic proficiency – one on which reported scores may be restricted to integers, but that nonetheless allows the theoretical possibility of intermediate scores and, importantly, intermediate cutoff scores. In principle, this is analogous to situations in which, for example, weights may be reported in kilograms or heights in inches, but finer-grained differences exist between individuals who are reported as having the same discrete weight or height. Test score reporting similarly involves, in part, the discretization of a theoretically continuous distribution of academic proficiency among test-takers.

1. Longstanding tenets of score reporting for individual test-takers maintain that no more than 30 to 60 score points should distinguish among them (Flanagan, 1951; Kolen and Brennan, 2004). These rules of thumb are based on the standard errors of individual scores and were developed to discourage distinctions among individuals that the precision of scores could not support. Adhering to this rule of thumb can result in the “binning” of multiple number-correct scores to the same scale score. In practice, then, scale scores are typically number-correct scores that are transformed, binned, and rounded to integers for individual score reports. This discretization is one of the two elements that imparts unpredictable bias to $\Delta PAC$ statistics.

2.2. Linking

Operational testing programs generally require replacement of test items to discourage cheating and sensitization. The resulting tests are not identical and will
naturally differ in difficulty to a degree that cannot (and need not) be eliminated (Holland and Dorans, 2006). Linking functions transform scores from one test to be comparable to scores from the other, relying on common items, common examinees, or randomly equivalent groups across tests (Kolen and Brennan, 2004). If items are more difficult, the linking function will map the same number-correct scores to higher scale scores; if items are less difficult, the linking function will map the same number-correct scores to lower scale scores.

The key observation that supports the remainder of this paper is that, unless the items appearing on two tests are identical, the linking function from number-correct scores to scale scores will be different for each test. This difference in linking functions effectively produces a different discretization of continuous scores for each test. In the next section, we model this changing discretization formally and illustrate the consequences for $P\!A\!C$-based comparisons.

3. Model

We start by examining student proficiencies in a baseline year 0 and comparison year 1. We assume latent proficiencies in year $i$ can be represented by continuous (real-valued) scores, with probability density function $f_i(\cdot)$ and cumulative distribution function $F_i(\cdot)$. A student is considered “Proficient” or “above the cutoff” if her continuous score exceeds a cutoff score, $k^*$, which is common to both years. The quantity $1 - F_i(k^*)$ is then the proportion (or, trivially, the percentage/100) of students who are Proficient in year $i$. We call this the “percentage above cutoff in year $i$, “ or $P\!A\!C_i$. 
We model discretization as a two-step process. First, the continuous scale is partitioned into a finite number of intervals, or “cells.” Second, a single “discrete score” is assigned to each cell. For example, in order to discretize a continuous scale ranging from 0 to 3, one might partition the interval \([0, 3]\) into four cells:
\([0, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 3]\). Assigning the discrete scores 0, 1, 2, and 3 to the first through fourth cells, respectively, would produce a discretization equivalent to rounding to the nearest integer. We will refer to this as “integer-rounding.” Alternatively, one might discretize the same continuous scale by partitioning it into cells \([0, 1), [1, 2), [2, 3]\) and assigning discrete scores 0.5, 1.5, and 2.5 to the corresponding cells. This rounds to the nearest “midpoint” between integers; we will call this “midpoint-rounding.”

Note that, in any discretization, each cell has an upper and lower bound. Additionally, for any discretization of continuous scores, there exists a single discrete score that is the minimum discrete score exceeding the cutoff score, \(k^*\). We use \(h_i^*\) to denote the lower bound of the cell associated with this discrete score under the discretization in year \(i\). To expand on the previous example, suppose integer-rounding is applied in year 0, while midpoint-rounding is applied in year 1. Suppose that the Proficient cutoff score in both years is \(k^* = 1.2\). In year 0, the lowest discrete score that exceeds \(k^*\) is 2. The cell associated with this discrete score is the interval \([1.5, 2.5]\), for which the lower bound is 1.5. Therefore, \(h_0^* = 1.5\). Similarly, in year 1, the lowest discrete score exceeding \(k^*\) is 1.5, and its corresponding cell is \([1, 2]\); thus, \(h_1^* = 1\).
For any pair of continuous score distributions \( F_0(\cdot) \) and \( F_1(\cdot) \), the proportion above cutoff (\( PAC \)) in year 0 is \( PAC_0 = 1 - F_0(k^*) \), while in year 1 it is \( PAC_1 = 1 - F_1(k^*) \). The year-over-year difference in \( PAC \), or \( \Delta PAC \), is then

\[
\Delta PAC = PAC_1 - PAC_0 = (1 - F_1(k^*)) - (1 - F_0(k^*))
\]

\[
= F_0(k^*) - F_1(k^*)
\]

(1)

However, if discrete scores are used to compute the \( PAC \) in each year, then we have \( PAC_0^{DISC} = 1 - F_0(h_0^*) \) and \( PAC_1^{DISC} = 1 - F_1(h_1^*) \). To see why, consider integer-rounding. In year 0, the set of discrete scores is \{0, 1, 2, 3\}. Since the cutoff score is \( k^* = 1.2 \), students with discrete scores of 0 and 1 are considered non-Proficient, while students with a discrete score of 2 (the lowest discrete score exceeding 1.2) or higher are considered Proficient. The “discrete \( PAC \)” in year 0 is therefore the proportion of students who receive discrete scores of 2 and higher. Since continuous scores in the interval \([1.5, 2.5)\) are all assigned a discrete score of 2, all continuous scores in this interval (and higher) are considered “above the cutoff,” while all scores below this interval are “below the cutoff.” The discrete \( PAC \) for year 0 is therefore \( PAC_0^{DISC} = 1 - F_0(1.5) \). Similarly, the discrete \( PAC \) for year 1 is \( PAC_1^{DISC} = 1 - F_1(1) \).

More formally, any discretization of continuous scores with PDF \( f_i(\cdot) \) and CDF \( F_i(\cdot) \) implies a probability mass function \( g_i(\cdot) \) and corresponding CDF \( G_i(\cdot) \), in which \( G_i(k^*) = F_i(h_i^*) \). The year-over-year change in \( PAC \) using discrete scores is then

\[
\Delta PAC^{DISC} = PAC_1^{DISC} - PAC_0^{DISC} = \left(1 - G_1(k^*)\right) - \left(1 - G_0(k^*)\right)
\]

\[
= \left(1 - F_1(h_1^*)\right) - \left(1 - F_0(h_0^*)\right)
\]

\[
= F_0(h_0^*) - F_1(h_1^*)
\]

(2)
To continue our above example, suppose that continuous scores in both years were normally distributed with mean 1.5 and variance 1; recall that scores range from 0 to 3, and integer-rounding is applied in year 0, while midpoint-rounding is applied in year 1. The continuous and discrete $\Delta PAC$s are then given by the expressions in (1) and (2), respectively:

$$\Delta PAC = PAC_1 - PAC_0 = F_0(1.2) - F_1(1.2) = \Phi(-0.3) - \Phi(-0.3) = 0$$

$$\Delta PAC^{DISC} = PAC_1^{DISC} - PAC_0^{DISC} = F_0(1.5) - F_1(1) = \Phi(0) - \Phi(-0.5) = 0.19,$$

where $\Phi(\cdot)$ is the standard normal CDF. In this example, continuous scores would produce a $\Delta PAC$ of 0, but the $\Delta PAC^{DISC}$ resulting from the use of discrete scores would suggest a 19 percentage-point increase in Proficient students, from 31% to 50%. In effect, when the discretization pattern changes in this way, students in year 0 are subjected to a higher cutoff score than students in year 1. On the other hand, if the discretizations were reversed, then the opposite would hold, resulting in $\Delta PAC^{DISC} = F_0(1) - F_1(1.5) = -0.19$. In general, this “misalignment” of discrete score partitions, where $h_0^* \neq h_1^*$, causes students in one year to be subjected to a different cutoff score from students in the other.

The critical issue for policy is that test administrators cannot control the discrete score partitions that are applied to continuous scores. As noted in Section 2.2, these partitions depend primarily on the properties of the items that appear on each test (and, to a lesser extent, the error involved in estimation of item parameters). Thus, in any pair of years, $h_0^* \neq h_1^*$ almost surely, and students in one year face a different cutoff score from students in the other. Figure 1 illustrates the consequences of this problem.
Figure 1. Observed and estimated change in the proportion of students scoring above cutoff from 2010 to 2011, Washington state Grade 8 Mathematics (N = 150,875).

\[ \Delta PAC \] denotes the observed change in proportion of students scoring above the cutoff score. For any cutoff score on the horizontal axis, the solid line indicates the observed change in the proportion of students scoring above that cutoff score from 2010 to 2011. Dotted vertical line indicates the “Proficient” cutoff score, for which the actual reported year-over-year change was -1.36 percentage points. Curve represents \[ \Delta PAC \] estimated using smoothed continuous CDFs; see Section 4 for details.


In Figure 1, given any cutoff score on the horizontal axis, the plot’s value on the vertical axis indicates the observed year-over-year change (that is, the sample estimate of \[ \Delta PAC^{DISC} \]) in the proportion of students scoring above that cutoff score. For instance, at the Proficient cutoff score of 400 (indicated by the vertical dashed line), the observed change between 2010 and 2011 was −1.36 percentage points. Similarly, at a cutoff score of 401, the observed \[ \Delta PAC^{DISC} \] was −1.16 percentage points. Qualitatively, these represent small-to-moderate declines in the percentage of Proficient students.

However, at a cutoff score of 402, the observed \[ \Delta PAC^{DISC} \] was 2.85 percentage points. In other words, if one were to assess year-over-year progress using a cutoff score only 2 points higher than the “official” Proficient score of 400, one would conclude that
there had been a moderate increase in the percentage Proficient students – the opposite of the conclusion when observing $\Delta PAC^{DISC}$ at 400 and 401. Figure 1 shows that the observed $\Delta PAC^{DISC}$ would swing wildly back and forth if the cutoff score were changed. Such a result would not occur if scores were measured on a continuous scale. However, this is exactly the behavior one would expect if changes in discretization caused students in different years to be subjected to different “effective” cutoff scores, $h_0^*$ and $h_1^*$.

3.1. Bias and Inconsistency of the $\Delta PAC$ Estimator

We formalize the results above to show that, for any sample of students, the observed $\Delta PAC^{DISC}$ is a biased and inconsistent estimator of the change in the percentage of students with continuous scores above a given cutoff score.

Our parameter of interest is the change in proportion of students whose continuous scores exceed the cutoff score, $k^*$:

$$\theta = \Delta PAC = F_0(k^*) - F_1(k^*),$$

where $F_0(\cdot)$ and $F_1(\cdot)$ are the continuous CDFs of student scores in years 0 and 1, respectively.

Given a sample size of $N_i$ students in year $i$, the discrete-score estimator for $F_i(k^*)$ is

$$\bar{PAC}_i^{DISC} = \frac{1}{N_i} \sum_{j=1}^{N_i} I(X_{ij} > h_i^*),$$

(3)

where $I(\cdot)$ is an indicator function, $X_{ij}$ is the continuous score for student $j$ in year $i$, and $h_i^*$ is the lower bound of the partition cell corresponding to the lowest discrete score above $k^*$, as defined previously. Continuous student scores in year $i$ are distributed according to $F_i(\cdot)$; thus, the probability that any given student will be observed as having
a score above \( k^* \) is \( P(X_{ij} > h_i^*) = 1 - F_i(h_i^*) \). The numerator of \( \overline{PAC}_i^{DISC} \) is therefore binomially distributed with count parameter \( N_i \) and probability parameter \( 1 - F_i(h_i^*) \).

Using \( \hat{\theta} = \overline{PAC}_1^{DISC} - \overline{PAC}_0^{DISC} \) as an estimator for \( \theta \) produces a bias of

\[
E(\hat{\theta} - \theta) = E(\overline{PAC}_1^{DISC} - \overline{PAC}_0^{DISC}) - (F_0(k^*) - F_1(k^*))
\]

\[
= (F_0(h_0^*) - F_1(h_1^*)) - (F_0(k^*) - F_1(k^*))
\]

\[
= \left( \int_{-\infty}^{h_0^*} f_0(t) dt - \int_{-\infty}^{h_1^*} f_1(t) dt \right) - \left( \int_{-\infty}^{k^*} f_0(t) dt - \int_{-\infty}^{k^*} f_1(t) dt \right)
\]

\[
= \int_{k^*}^{h_0^*} f_0(t) dt - \int_{k^*}^{h_1^*} f_1(t) dt \neq 0 \tag{4}
\]

Similarly, we can show that \( \hat{\theta} \) is an inconsistent estimator of \( \theta \). For simplicity, we assume that sample sizes in both years are equal.

Let \( \nu = E(\hat{\theta} - \theta) \) be the bias of \( \hat{\theta} \), and let \( \frac{\sigma^2}{N} \) be the variance of \( \hat{\theta} \), where \( N \) is the sample size in each year, and \( \sigma^2 \) is a constant that does not depend on \( N \). Recall that the numerator of \( \overline{PAC}_i^{DISC} \) is binomially distributed; then, by a straightforward central-limit argument, \( \overline{PAC}_i^{DISC} \) is asymptotically normally distributed with mean \( 1 - F_i(h_i^*) \). Thus,

\( \hat{\theta} = \overline{PAC}_1^{DISC} - \overline{PAC}_0^{DISC} \) is asymptotically normal with mean \( \nu + \theta \). For large \( N \),

\( \hat{\theta} - \theta \sim N\left( \nu, \frac{\sigma^2}{N} \right) \)

and therefore \( \frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta - \nu) \) has a standard normal distribution. Then, for any \( \varepsilon > 0 \),

\[
P(|\hat{\theta} - \theta| \geq \varepsilon) = P(\hat{\theta} - \theta \geq \varepsilon) + P(\hat{\theta} - \theta \leq -\varepsilon)
\]

\[
= P\left( \frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta - \nu) \geq \frac{\sqrt{N}}{\sigma}(\varepsilon - \nu) \right) + P\left( \frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta - \nu) \leq -\frac{\sqrt{N}}{\sigma}(\varepsilon + \nu) \right)
\]

13
\approx 1 - \Phi \left( \frac{\sqrt{N}}{\sigma} (\varepsilon - \nu) \right) + \Phi \left( -\frac{\sqrt{N}}{\sigma} (\varepsilon + \nu) \right).

Choosing any \( \varepsilon \in (0, |\nu|) \) then produces

\[
\lim_{N \to \infty} P\left( |\hat{\theta} - \theta| \geq \varepsilon \right) \approx \lim_{N \to \infty} \left( 1 - \Phi \left( \frac{\sqrt{N}}{\sigma} (\varepsilon - \nu) \right) + \Phi \left( -\frac{\sqrt{N}}{\sigma} (\varepsilon + \nu) \right) \right) = 1
\]

Thus, \( \hat{\theta} \) is an inconsistent estimator of \( \theta \). This latter result implies that, in practice, \( \Delta PAC \) estimates are likely to be incorrect even with very large sample sizes, perhaps contrary to typical intuition.

3.2. **Partition Misalignment and Volatility of the \( \Delta PAC \) Estimator**

Next, we show that “misalignment” of discrete score partitions is primarily responsible for the volatility, or “sawtooth” pattern, observed in Figure 1. We refer to the expected value of the biased, discrete-score \( \Delta PAC \) estimator as \( \Delta PAC^{DISC} \).

Recall that \( \Delta PAC^{DISC} = \theta + \nu \), where \( \theta \) is the value of the continuous \( \Delta PAC \) and \( \nu \) is the bias of the estimator. The bias given in (4) can be written

\[
\nu = \int_{k^*}^{h_0^*} f_0(t) dt - \int_{k^*}^{h_1^*} f_1(t) dt
\]

\[
= \int_{k^*}^{h_0^*} f_0(t) dt - \int_{k^*}^{h_0^*} f_1(t) dt - \int_{h_0^*}^{h_1^*} f_1(t) dt
\]

\[
= \int_{k^*}^{h_0^*} (f_0(t) - f_1(t)) dt + \int_{h_0^*}^{h_1^*} f_1(t) dt
\]

When the same discretization is applied to continuous scores in both years, \( h_0^* = h_1^* \), and the second term of (5) evaluates to zero. The remaining term then imparts a small bias to the discrete \( \Delta PAC \) estimator, causing it to be evaluated at \( h_0^* = h_1^* \), rather than \( k^* \).
We refer to this component of the bias as *rounding bias*, since it is commonly described as “rounding error”: its primary effect is to produce a discrete, step-function version of the continuous $\Delta PAC$ curve.

On the other hand, when different discretizations are applied to scores in each year, $h_0^* \neq h_1^*$, and the second term of (5) is nonzero. Furthermore, because its integrand consists only of a single density function, rather than a difference between densities, its magnitude is relatively large. We refer to this component as *misalignment bias*, since it results from misalignment of the discrete-score partition cells in each year. We present visual examples of each bias component in Figure 2.

In Figure 2, scores range from 0 to 10, and the dotted lines depict the $\Delta PAC$ values for continuous scores. In panel A, scores are normally distributed with $\mu = 5, \sigma^2 = 1$ in the first year and $\mu = 5.2, \sigma^2 = 2$ in the second year. Integer-rounding is applied to continuous scores in both years; thus, $h_0^* = h_1^*$ for all cutoff scores, and misalignment bias is zero. The $\Delta PAC^{DISC}$ at each cutoff score $k^*$ is then equal to the continuous $\Delta PAC$ evaluated at the lower bound of the discrete-score partition cell associated with the lowest integer equal to or exceeding $k^*$. For example, the lower bound of the cell associated with a discrete score of 5 is $h_0^* = h_1^* = 4.5$; thus, for any $k^* \in (4, 5]$, we have $\Delta PAC^{DISC} = F_0(4.5) - F_1(4.5)$, the continuous $\Delta PAC$ evaluated at a score of 4.5. In effect, $\Delta PAC^{DISC}$ gives the continuous $\Delta PAC$ based on scores rounded to the nearest integer.

Panel B, on the other hand, isolates the effect of misalignment bias. Continuous scores in both years have the same distribution ($\mu = 5, \sigma^2 = 1$), and therefore $f_0(t) = f_1(t)$ for all $t$, so that rounding bias is zero. However, integer-rounding is applied in the
first year, while midpoint-rounding is applied in the second; thus, $h_0^* \neq h_1^*$ for all cutoff scores, and $\Delta PAC^{DISC}$ suffers from misalignment bias. The “sawtooth” volatility observed in Figure 1 is clearly visible in panel B but absent from panel A, confirming the fact that it results from the application of different discretizations in each year. For any cutoff score $k^* \in (4.5, 5]$, the next-highest discrete score is 5 in the first year and 5.5 in the second year, giving $h_0^* = 4.5$ and $h_1^* = 5$. As shown in (5), the $\Delta PAC^{DISC}$ thus has a bias of $\int_5^{4.5} f(t) dt = F(4.5) - F(5)$, effectively subjecting scores in the second year to a higher cutoff score than in the first.
In short, the use of different discretizing partitions introduces bias beyond what would be expected when discretizations are identical. Below, we estimate this bias using state test data and show that, where state testing is concerned, its magnitude is likely to be substantial in many cases.

4. **Estimation of ΔPAC Bias in State Tests**

**NOTE:** The original, published version of this section appears below. Please see the Addendum following the main paper for an analysis of alternative estimators that may be more appropriate than the methods we use below.

Our model suggests a number of methods for reducing the effect of bias in the ΔPAC estimator in large samples. We employ a relatively simple method via OLS regression, in which we estimate continuous CDF values in each year by fitting polynomials of increasing degree to discrete scores near each test’s official Proficient
cutoff score, stopping when the estimated \( PAC \) at the cutoff score changes by less than 1.5 percentage points (varying the stopping criterion does not substantially change results). We then use these “smoothed” estimates to produce estimates of the continuous \( \Delta PAC \) and the empirical \( \Delta PAC^{DISC} \) bias across 107 state test score trends.

We acknowledge that more sophisticated methods – such as those that incorporate weighting, explicit modeling of the error term, or improved stopping criteria, among many others – should produce more robust results than this procedure. However, this relatively simple continuizing procedure is sufficient to illustrate the bias that is our interest.

4.1. Dataset

Our data consist of frequency distributions for individual student scale scores in 2010 and 2011 for Math and Reading/English Language Arts (ELA) tests, all gathered from publicly available technical reports for 13 states, with one “state” consisting of a four-state consortium using a common testing program.2 For each testing program in our dataset, discrete scale scores and the number of students achieving each score were compiled, where available, for students in grades 3 through 8. For the analyses below, we exclude tests for which data appear incomplete, as well as tests that were not comparable from 2010 to 2011 due to large-scale changes in test design, state policy, or similar factors. Our final dataset includes 60 Math and 47 Reading/ELA test score distributions, representing approximately 22 million student scores. Sample sizes for each test range from about 9,000 to more than 300,000 students per year, with a mean

\[2\text{ States included in the final dataset were Alaska, Arizona, Idaho, Maine, Nebraska, New Hampshire, New Jersey, New York, North Carolina, Oklahoma, Pennsylvania, Rhode Island, South Dakota, Texas, Vermont, and Washington.} \]
slightly over 100,000; thus, the impact of sampling error on our results is likely to be negligible.

4.2. Empirical Results

We summarize the scope of this problem, as well as the results of our smoothing algorithm, in Table 1. For each test, we compute the $\Delta PAC$ using the official state Proficient cutoff score, $k^*$. We then find the lowest discrete score exceeding $k^*$ (across both years), compute the $\Delta PAC$ at this score, and subtract the original $\Delta PAC$ at $k^*$. The result is the change in the $\Delta PAC$ attributable to increasing the cutoff score by one discrete score increment. We repeat this with the highest discrete score lower than $k^*$ to produce the change attributable to a -1 score increment.

| Table 1. Summary of changes in year-over-year increases in percent-Proficient ($\Delta PAC$) from 2010 to 2011 attributable to changing the Proficient cutoff score to the next-highest or next-lowest discrete score for state NCLB tests in 13 states. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | Math ($N = 60$) |                 | Reading/English Language Arts ($N = 47$) |                 |
|                 | Std. dev.       | Min    | Max    | No. of sign changes | Std. dev. | Min    | Max    | No. of sign changes |
| Observed        | 1.5             | -4.6   | 4.0    | 14 (23%)            | 1.8       | -4.7   | 4.3    | 4 (9%)             |
| Smoothed        | 0.2             | -0.9   | 0.5    | 0 (0%)              | 0.3       | -1.0   | 0.9    | 2 (4%)             |

NOTE: All decimal values in percentage points. Values were computed by increasing and decreasing the Proficient cutoff score by one scaled score increment, computing the observed change in percent above the new cutoff score ($\Delta PAC$), and subtracting the observed $\Delta PAC$ at the original Proficient cutoff score from the result. A “sign change” indicates that the “incremented” $\Delta PAC$ values include at least one value whose sign is the opposite of the Proficient $\Delta PAC$. Observed values are those actually reported by states. Smoothed values were computed using polynomial smoothing of CDFs in each year. Test data were compiled from state technical reports.

For example, in Washington state, the official Proficient cutoff score was 400, at which the observed $\Delta PAC$ was -1.36 percentage points. The lowest discrete score exceeding 400 was 401 in 2010 and 404 in 2011; therefore, we compute the observed
incrementing the cutoff score by 1 discrete score (that is, increasing the cutoff score to the next-highest discrete score in either year) thus produces a change in the observed $\Delta PAC$ of 0.2 percentage points. Meanwhile, the highest discrete score below 400 in either year was 396, at which the observed $\Delta PAC$ was 2.11 percentage points. Decrementing the cutoff score by 1 discrete score thus increases the $\Delta PAC$ by 3.5 percentage points. For this test, the maximum change in $\Delta PAC$ attributable to a one-increment change in cutoff score is therefore 3.5 percentage points, while the minimum is 0.2 percentage points. Additionally, since the official $\Delta PAC$ was negative, while the $\Delta PAC$ at 396 was positive, the $\Delta PAC$ for this distribution changed sign.

We repeat this for all tests in our dataset, using both observed and “smoothed” values, and summarize the results in Table 1. The “Observed” row in Table 1 reports results using officially reported data and shows that the pattern in Figure 1 is not unique to Grade 8 Math in Washington state. Small changes in the cutoff score produce swings in $\Delta PAC$ of between -4.6 and 4.0 percentage points in Math and between -4.7 and 4.3 percentage points for Reading/ELA. More disturbingly, a sign change is observed in roughly 1 out of every 6 tests. For these tests, a small change in the cutoff score would produce a year-over-year “improvement” result that was qualitatively the opposite of what was officially reported.

The “Smoothed” row presents results after application of the smoothing algorithm to the CDFs in each year; a visual example for Washington state appears in Figure 1. Across all tests in our sample, the estimated $\Delta PAC$s exhibit far greater stability across cutoffs when distributions are smoothed. This is unsurprising under our model and
suggests that smoothing reduces the effect of misalignment bias. We use the empirical variance of the differences between observed and smoothed $\Delta PAC$s as an approximation of the true variance of the discrete-score bias values.

In Figure 3, we present the empirical distributions of these estimated bias values. The average absolute values of the observed $\Delta PAC$s in our sample were 1.67 for Math and 2.93 for Reading/ELA. We estimate standard deviations of 1.11 and 1.28 percentage points for the Math and Reading/ELA $\Delta PAC$ bias values, respectively. We consider these standard deviations to be large. In short, eliminating the bias caused by discretization would likely change interpretations of the magnitude of year-over-year improvement for a large proportion of the tests in our sample.

<table>
<thead>
<tr>
<th>Figure 3. Estimated bias in reported change in percent-Proficient ($\Delta PAC$) from 2010 to 2011 attributable to changes in score discretizations for NCLB tests in 13 states.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Math (N = 60)</strong></td>
</tr>
<tr>
<td><strong>Reading/ELA (N = 47)</strong></td>
</tr>
</tbody>
</table>

Bias estimated by using observed (discrete) cumulative distributions for scores in 2010 and 2011 to estimate continuous CDF values and $\Delta PAC$s for each test. Standard deviations of estimated bias values are 1.11 for Math and 1.28 for Reading/English Language Arts (ELA). Average absolute magnitudes of reported $\Delta PAC$s were 1.67 for Math and 2.93 for Reading/ELA. Test data compiled from state technical reports.

5. **Discussion**

We have shown above that bias in the observed $\Delta PAC$ caused by discretization may lead to incorrect substantive conclusions regarding year-over-year educational
progress. Changes in discretizing partitions effectively subject students to different
cutoff scores over time. In this section, we discuss four additional issues: bias at the
district and school levels, the relative contribution of sampling variability, solutions
implementable by state testing programs, and implications of this case study for the
general problem of comparing discretized distributions.

First, at the district and school levels, where similarities among students typically
cause scores to be more concentrated (non-zero intraclass correlations; see Hedges and
Hedberg, 2007), bias is likely to be larger, particularly when the state cutoff score
happens to be close to the modal score for the district or school. To see why, recall that
the second term of (5) depends only on the continuous-score density in a single year.
Because test score distributions are generally unimodal, and since the values of $h_0^*$ and $h_1^*$
apply to all schools in the state, smaller variances in scores imply larger values of
$\int_{h_1^*}^{h_0^*} f_1(t) dt$ for values of $h_0^*$ and $h_1^*$ near the mode of a district or school’s score
distribution. Our model thus implies that $\Delta PAC$ bias is almost certainly larger than
estimated in Section 4 in schools or districts in which large proportions of students score
near the cutoff score.

Second, at the school and subgroup level, this bias is likely to be overshadowed
by error. In smaller samples, such as the “minimum subgroup size” that ranges from 5 to
100 across states (Fulton, 2006), sampling variability will typically overwhelm any useful
information that might be gleaned from the observed $\Delta PAC$ from one year to the next.
The observed $PAC$ in each year, given in (3), follows a scaled binomial distribution, and
thus the observed $\Delta PAC$ has standard error $\sqrt{\frac{\sigma_0^2}{N_0} + \frac{\sigma_i^2}{N_i}}$, where $\sigma_i^2 = F_i(h_i^*)\left(1 - F_i(h_i^*)\right)$
and $N_i$ is the sample size in year $i$. With $F_0(h_0^*) = F_1(h_1^*) = 0.5$ and $N_0 = N_1 = 20$, for example, this results in a standard error of nearly 16 percentage points – far too large to draw reliable inferences regarding year-over-year improvement using only a single $\Delta PAC$ observation. At samples of 5,000 students per year, it declines to one percentage point, commensurate with the magnitudes of bias that we report here. Although we have demonstrated that this bias will always be a factor, it will be most salient at the level of reporting for states and large districts.

Third, our results have practical implications for state testing programs that report trends using $\Delta PAC$. Using finer-grained partitions will reduce the bias caused by misalignment of discrete score cells, all else equal. States wishing to minimize $\Delta PAC$ volatility should estimate the statistic using the most fine-grained data available at each time point (e.g., through the use of scale score estimates from a two- or three-parameter IRT model). If concerns remain about continuous or fine-grained scores imparting a false sense of precision, states could use coarse values when comparing individual observations, while retaining finer values for the purposes of $\Delta PAC$ reporting. As a last resort, states or researchers could employ “smoothing” to estimate continuous $\Delta PAC$s. A broader recommendation, following Ho (2008), is to dispense with $PAC$-based metrics in favor of average-based metrics. Average-based trend metrics are not cutoff-score dependent and are more robust to discretization and changes in discretizing partitions.

Finally, although we focus on $\Delta PAC$ statistics in educational testing, our model applies generally to any percentage- or proportion-based comparison in which discretizing partitions differ between groups. Such comparisons appear in a wide variety of fields of study, and discretizations of the underlying data (including height, weight,
temperature, and currency) often differ depending on their source. Consider the problem of comparing poverty rates in the U.K. and the U.S., where incomes in each country are rounded to the nearest thousand pounds or dollars, respectively. Applying an exchange rate reveals that discretizing partitions differ across countries, effectively holding U.K. residents to a different poverty standard than U.S. residents. Similarly, rate comparisons in which data for one group are rounded to British units and the other rounded to metric, such as low-birth-weight comparisons, can be expected to suffer from the same bias. Discreteness, different partitions, and the $\Delta P A C$ metric interact to produce the volatile pattern in Figure 1, and addressing any one of these (by “continuizing” distributions, using aligned partitions, or using average-based metrics, for example) will reduce potential bias in percentage-based comparisons.
A. Addendum

In this Addendum, I explore alternative estimators intended to mitigate the effects of “misalignment” on parameter estimates, particularly the “sawtooth” volatility across cutoff scores that may substantially impact educational accountability results. Specifically, I aim to answer the question: Given the available discrete score data, are there estimators that reduce the bias caused by misalignment without producing an offsetting increase in variance?

In the original paper, we showed that, if \( \theta \) is the \( \Delta PAC \) parameter based on continuous scores, and its estimator \( \hat{\theta} \) is based on discrete scores, then the estimator’s bias, \( E(\hat{\theta} - \theta) \neq 0 \), is especially consequential when the \( PAC \) estimates used to compute \( \hat{\theta} \) are based on scores that are discretized differently. More formally, if \( \mathcal{H}_0 \) is a discrete-score partition on one set of scores, and \( \mathcal{H}_1 \) is a partition on the other, then when \( \hat{\theta} \) is computed using the discrete scores produced by \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \),

\[
E(\hat{\theta} - \theta \mid \mathcal{H}_0, \mathcal{H}_1) \neq 0
\]

almost surely, with \( \hat{\theta} \) exhibiting a “sawtooth” pattern when \( \mathcal{H}_0 \neq \mathcal{H}_1 \). This expression emphasizes the fact that the expectation above depends on the partitions, \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \), and is properly described as “bias” in cases where the partitions are constant (e.g., when \( \mathcal{H}_0 \) rounds values to centimeters, and \( \mathcal{H}_1 \) rounds them to inches).

As we noted in section 2, however, the score discretizations produced by educational tests are typically random – depending, among other things, on the characteristics of randomly-selected items on each test. In theory, across all possible pairs of discrete-score partitions, the differences between \( \hat{\theta} \) and \( \theta \) are likely to “average out” to zero – that is, it is likely that \( E \left( E(\hat{\theta} - \theta \mid \mathcal{H}_0, \mathcal{H}_1) \right) = 0 \). For this reason, it may
be more technically appropriate to refer to the expectation above as “error,” rather than “bias.”

We avoid modeling such variation, due to the additional assumptions required – regarding, for example, the distributions of partition cells across years and the parameters of these distributions, which are likely to vary from test to test – and because it is not central to our findings. Therefore, we continue to employ the term “bias” below, in order to keep terminology consistent with the original paper and to more clearly distinguish between an estimator’s “conditional bias” \( E(\hat{\theta} - \theta|\mathcal{H}_0, \mathcal{H}_1) \) and its “conditional error” \( (\hat{\theta} - E(\hat{\theta} - \theta|\mathcal{H}_0, \mathcal{H}_1)) \). Although we omit conditional notation in what follows for clarity’s sake, the terms “bias,” “error,” and “variance” should be understood to be conditional on exogenously-defined partitions, \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \).

We begin by developing and examining expressions for comparing the linear-interpolation estimator, \( \hat{\theta}^{LIN} \), to the usual discrete-score estimator, \( \hat{\theta}^{DISC} \).

5.1. Linear Interpolation

The linear-interpolation estimator, defined formally below, produces \( PAC \) estimates for each year by interpolating between observed \( PAC \) values for discrete scores immediately below and above the cutoff score, rather than simply using the latter \( PAC \) value as an estimate of the continuous \( PAC \). The interpolated \( PAC \) values are then differenced to produce an estimate of the \( \Delta PAC \).

As a function of the cutoff score, the expected value of the linear-interpolation \( PAC \) estimator, \( \overline{PAC}_i^{LIN} \), is piecewise linear, unlike the step function produced by the discrete-score estimator, \( \overline{PAC}_i^{DISC} \). The linear-interpolation \( \Delta PAC \) estimator therefore also has a piecewise-linear expected value, as illustrated in Figure A1.
Figure A1. Example of expected values for proportion above cutoff (PAC) and difference in proportion above cutoff (ΔPAC) estimators.

**PAC estimators**

Top panel: Continuous scores have standard normal distribution. Discrete scores are rounded down to the nearest standard deviation (-3, -2, -1, …).

Bottom panel: Scores in year 0 are distributed and rounded as in top panel. Scores in year 1 are normally distributed with mean 0.1 and standard deviation 1.1 and rounded down to the nearest half-standard deviation (-2.5, -1.5, -0.5, …). ΔPAC value is year 1 PAC minus year 0 PAC.
Because the expected value of the linear-interpolation estimator is continuous with respect to cutoff score, it does not suffer from the “sawtooth” volatility of the discrete-score estimator, so that small perturbations in the cutoff score do not produce dramatic leaps in ΔPAC estimates. However, under certain conditions, the linear-interpolation estimator may produce larger bias and variance than the discrete-score estimator. We explore these conditions in sections A1.1 and A1.2, employing the definitions below.

As above, let \( k \) be a cutoff score common to tests in both years; \( F_i(\cdot) \) be the CDF for continuous scores in year \( i \), with \( f_i(\cdot) \) its associated PDF; and \( h_i \) be the lower bound of the discrete-score partition cell associated with the lowest discrete score equal to or exceeding \( k \). For the results below, we assume that continuous scores are rounded down to the nearest discrete score; other discretization schemes produce similar results where misalignment is concerned, but with an “offset” applied to cell bounds.

A few additional definitions are needed for the linear interpolation estimator. First, let \( g_i \) be the lower bound of the partition cell immediately below \( k \) in year \( i \), and let

\[
\alpha_i = \frac{h_i - k}{h_i - g_i}
\]

be the “weighting factor” for the linear interpolation estimator.

Additionally, let \( T_i(x) \) be the number of students who score above \( x \) in year \( i \).

The discrete-score \( PAC \) estimator for year \( i \) is then

\[
PAC_i^{DISC} = \frac{T_i(h_i)}{N_i},
\]

where \( N_i \) is the number of students tested. The linear-interpolation estimator, which interpolates between discrete scores adjacent to \( k \), is defined as

\[
PAC_i^{LIN} = \alpha_i \frac{T_i(g_i)}{N_i} + (1 - \alpha_i) \frac{T_i(h_i)}{N_i}.
\]
For both estimators, the estimand is \(1 - F_i(k)\), the proportion of students with proficiencies exceeding \(k\).

Finally, the discrete-score and linear-interpolation \(\Delta_PAC\) estimators are \(\hat{\theta}_\text{DISC} = \overline{PAC}_i \text{DISC} - \overline{PAC}_0 \text{DISC}\) and \(\hat{\theta}_\text{LIN} = \overline{PAC}_i \text{LIN} - \overline{PAC}_0 \text{LIN}\), respectively, with the estimand \(\theta = F_0(k) - F_1(k)\), the change or difference in proportion of students with proficiencies exceeding \(k\).

5.1.a. Bias

As we show in the main paper, the bias of \(\hat{\theta}_\text{DISC}\) is:

\[
E(\hat{\theta}_\text{DISC} - \theta) = \int_k^{h_0} f_0(t) - f_1(t)dt + \int_{h_1}^{h_0} f_1(t)dt. \tag{A1}
\]

\[
= \int_k^{h_0} f_0(t)dt - \int_{k}^{h_1} f_1(t)dt. \tag{A2}
\]

For the linear interpolation estimator, the bias of \(\overline{PAC}_i \text{LIN}\) can be written:

\[
E(\overline{PAC}_i \text{LIN} - (1 - F_i(k))) = \alpha_i(1 - F_i(g_i)) + (1 - \alpha_i)(1 - F_i(h_i)) - (1 - F_i(k))
\]

\[
= F_i(k) - F_i(h_i) + \alpha_i(F_i(h_i) - F_i(g_i))
\]

\[
= - \int_k^{h_1} f_i(t)dt + \alpha_i \int_{g_i}^{h_i} f_i(t)dt.
\]

The bias of \(\hat{\theta}_\text{LIN}\) is then:

\[
E(\overline{PAC}_i \text{LIN} - \overline{PAC}_0 \text{LIN} - \theta) = \int_k^{h_0} f_0(t)dt - \int_k^{h_1} f_1(t)dt - (\alpha_0 \int_k^{h_0} f_0(t)dt - \alpha_1 \int_{g_1}^{h}\int_{g_1}^{h_1} f_1(t)dt). \tag{A3}
\]

Comparing (A1) and (A2), we see that the bias of \(\hat{\theta}_\text{LIN}\) exceeds that of \(\hat{\theta}_\text{DISC}\) when

\[
\left| \int_k^{h_0} f_0(t)dt - \int_{k}^{h_1} f_1(t)dt - (\alpha_0 \int_{g_0}^{h_0} f_0(t)dt - \alpha_1 \int_{g_1}^{h_1} f_1(t)dt) \right| > \left| \int_k^{h_0} f_0(t)dt - \int_{k}^{h_1} f_1(t)dt \right|
\]

This expression reveals that, given some fixed cutoff score, it is possible for the linear-interpolation estimator to suffer larger bias than the discrete-score estimator.

However, in section A1.3, we show that, under plausible conditions, the linear-
interpolation estimator has smaller average bias and, at worst, is only slightly more biased than the discrete-score estimator.

5.1.b. Variance

We next examine the relative variances of the two estimators. To simplify notation, let \( p_{hi} = 1 - F_i(h_i) \) and \( p_{gi} = 1 - F_i(g_i) \). Then \( T_i(g_i) \) and \( T_i(h_i) \) are binomially distributed with probability parameters \( p_{gi} \) and \( p_{hi} \), respectively, and count parameter equal to sample size, \( N_i \).

The variance of \( \bar{PAC}_i^{LIN} \) is then

\[
\left( \frac{1}{N_i^2} \right) \text{Var}(\alpha_i T_i(g_i) + (1 - \alpha_i) T_i(h_i))
\]

\[
= \left( \frac{1}{N_i^2} \right) \left( \alpha_i^2 \text{Var}(T_i(g_i)) + (1 - \alpha_i)^2 \text{Var}(T_i(h_i)) + 2\alpha_i(1 - \alpha_i) \text{Cov}(T_i(g_i), T_i(h_i)) \right),
\]

with

\[
\text{Cov}(T_i(g_i), T_i(h_i)) = E(T_i(g_i) \cdot T_i(h_i)) - E(T_i(g_i)) E(T_i(h_i)).
\]

To evaluate the left-hand expectation, let \( G_j = I \left( X_{ij} \geq 1 - F(g_i) \right) \) be an indicator for student \( j \) scoring above \( g_i \), so that \( T_i(g_i) = \sum_{j=1}^{N_i} G_j \). Define \( H_j \) similarly, so

\[
T_i(h_i) = \sum_{j=1}^{N_i} H_j.
\]

Then, by binomial expansion,

\[
T_i(g_i) T_i(h_i) = \left( \sum_{j=1}^{N_i} G_j \right) \left( \sum_{j=1}^{N_i} H_j \right)
\]

\[
= \sum_{j=1}^{N_i} G_j H_j + \sum_{j=k}^{N_i} G_j H_k + \sum_{j<k}^{N_i} G_k H_j.
\]

For each student, \( G_j H_j = 1 \) only if her score exceeds \( h_i \); thus, \( E\left( \sum_{j=1}^{N_i} G_j H_j \right) = E\left( T_i(h_i) \right) \).
For $G_j H_k$ with $j \neq k$, by independence, $E(G_j H_k) = E(G_j)E(H_k) = p_g i p_{hi}$. Since there are $2 \binom{N_i}{2}$ such terms, we have $E\left(\sum_{j<k}^N G_j H_k + \sum_{j<k}^N G_k H_j\right) = p_g i p_{hi} N_i (N_i - 1)$.

Therefore,

$$E(T_i(g_i) T_i(h_i)) = p_{hi} N_i + p_g i p_{hi} N_i (N_i - 1)$$

Finally, since $E(T_i(g_i)) E(T_i(h_i)) = p_{gi} p_{hi} N_i^2$, we have

$$Cov(T_i(g_i), T_i(h_i)) = p_{hi} N_i + p_{gi} p_{hi} N_i (N_i - 1) - p_{gi} p_{hi} N_i^2$$

$$= N_i p_{hi} (1 - p_{gi}).$$

For the variance of $\hat{PAC}_i^{LIN}$ to be larger than the variance of $\hat{PAC}_i^{DISC}$, we must have

$$\alpha_i^2 Var(T_i(g_i)) + 2\alpha_i (1 - \alpha_i) Cov(T_i(g_i), T_i(h_i)) + (1 - \alpha_i)^2 Var(T_i(h_i)) > Var(T_i(h_i)).$$

After some algebra, this inequality can be written as

$$\alpha_i > 2 \left(\frac{p_{hi}}{1 - (p_{gi} - p_{hi})}\right).$$

In general, this condition requires i) the proportion of students achieving the highest discrete score below the cutoff \(\frac{T_i(g_i)}{N_i}\) to receive large weight in the linear-interpolation estimator, and ii) $p_{gi}$ to be closer to 0.5 than $p_{hi}$ (so that $T_i(g_i)$ has larger variance than $T_i(h_i)$).

5.1.c. **Computational Comparisons**

In order to evaluate the relative accuracy and precision of $\hat{\theta}^{LIN}$ compared to $\hat{\theta}^{DISC}$, we employ the following general algorithm:

1. Partition scores in each year into $M$ cells, ensuring that partitions are misaligned.

2. Choose a minimum cutoff score, $k$. 

31
3. Compute the bias and standard error of \( \hat{\theta}^{LIN} \) and \( \hat{\theta}^{DISC} \) at \( k \) and save the results.

4. Increment the cutoff score by a small amount.

5. Repeat steps 3 and 4 until a maximum cutoff score is reached.

6. Increase \( M \), the number of partition cells.

7. Repeat steps 1 through 6 until a maximum number of cells is reached.

For each \( M \), we compare \( \hat{\theta}^{LIN} \) and \( \hat{\theta}^{DISC} \) according to two criteria. First, we compute the average absolute bias across cutoff scores for each estimator, and second, the average root mean squared error (RMSE). These serve as relative measures of overall accuracy and efficiency, respectively, for each estimator when continuous scores are discretized into \( M \) discrete scores.

We begin with \( M = 20 \) cells and increase this to 100 in increments of 5. Scores in each year range from -3 to 3, and \( k \) is evaluated from -2.5 to 2.5 in increments of 0.001. Continuous scores in year 0 are standard normal, while in year 1 they are normally distributed with mean 0.1 and a standard deviation of 1.1; these values were chosen to reflect plausible variation in test scores between consecutive years. For standard error and RMSE calculations, we set a sample size of 10,000 students for each test, commensurate with single-grade enrollment in a small state or large district.

Year 0 scores are partitioned into \( M \) equally-sized cells, resulting in cells of width \( w = \frac{6}{M} \). The partition in year 1 is similar to the year 0 partition, but “shifted” in the positive direction by \( 0.1w \) to produce a small amount of misalignment. This has the effect of setting \( h_0 \) close to \( h_1 \) across about 90% of cutoff scores. Combined with the
fact that \( f_0(\cdot) \) is similar to \( f_1(\cdot) \), these parameter choices are highly favorable to \( \hat{\theta}^{DISC} \) (see expression (A1)).

Finally, we emulate “midpoint rounding” by computing bias values at \( k - 0.5w \), rather than \( k \). For example, suppose \( w = 1 \) and partition cell bounds are located at \( H_i = \{ ..., -1.5, -0.5, 0.5, 1.5, ... \} \). Evaluating the discrete-score estimator at \( k \) by defining its estimator function as

\[
\hat{PAC}_i^{DISC}(k) = 1 - F_i \left( \min_{h_i \in H_i} \{ h_i: h_i \geq k \} \right)
\]

would then reflect “rounding down,” since \( \hat{PAC}_i^{DISC}(k) = 1 - F_i(-0.5) \) for all \( k \in (-1.5, -0.5] \); \( 1 - F_i(0.5) \) for \( k \in (-0.5, 0.5] \); and so on. However, setting \( \hat{PAC}_i^{DISC}(k) = 1 - F \left( \min_{h_i \in H_i} \{ h_i: h_i \geq k - 0.5w \} \right) \) results in \( 1 - F_i(-0.5) \) for \( k \in (-1, 0] \), \( 1 - F_i(0.5) \) for \( k \in (0, 1] \), etc., equivalent to rounding to the nearest integer. Overall, this reduces the bias of \( \hat{PAC}_i^{DISC} \) in each year, while increasing the bias of \( \hat{\theta}^{LIN} \) at nearly all cutoff scores, further favoring \( \hat{\theta}^{DISC} \).

Results are displayed in Figure A2. Dashed lines indicate values for \( \hat{\theta}^{DISC} \), while solid lines indicate \( \hat{\theta}^{LIN} \); average absolute bias is shown in black and average RMSE in gray. On both measures, \( \hat{\theta}^{LIN} \) is superior to \( \hat{\theta}^{DISC} \), regardless of the number of discrete scores, with the differences between estimators decreasing as the number of discrete scores grows larger (implying “finer” partitions). Although not pictured here, \( \hat{\theta}^{LIN} \) remains superior to \( \hat{\theta}^{DISC} \) on both measures even at implausibly large numbers of discrete scores (1,000 or more). In short, Figure A3 shows that, even under conditions highly favorable to the discrete-score estimator, the linear-interpolator is less biased and closer to the continuous \( \Delta PAC \) value, on average, across a large range of plausible cutoff scores.
Figure A2. Comparison of absolute bias and RMSE for linear-interpolation and discrete-score ΔPAC estimators.

Number of discrete scores is the number of partition cells in year 0, which have equal width. Partition cells in year 1 are identical to year 0, but moved in the positive direction by one-tenth of cell width. Continuous scores in year 0 and year 1 are normally distributed with $\mu = 0, \sigma = 1$ and $\mu = 0.1, \sigma = 1.1$, respectively. Values for each “Number of discrete scores” were computed at each cutoff score between -2.5 to 2.5 at increments of 0.001.

While bias and RMSE averages may be helpful as high-level measures for comparing the estimators, they do not provide information on specific conditions that may be relevant in practice. The results from sections A1.1 and A1.2 show that it is possible, under certain conditions, for $\hat{\theta}^{LIN}$ to have larger bias and variance than $\hat{\theta}^{DISC}$. If a cutoff score happens to be located at a score that meets these conditions, then $\hat{\theta}^{LIN}$ will in fact be worse than $\hat{\theta}^{DISC}$.

Given this, it may be helpful to ask, “How much worse than $\hat{\theta}^{DISC}$ can $\hat{\theta}^{LIN}$ possibly be, and vice versa?” We address this question by computing the maximum differences in absolute bias and RMSE between $\hat{\theta}^{LIN}$ and $\hat{\theta}^{DISC}$, using the same parameter
combinations used for Figure A2, to illustrate the “worst-case scenarios” for each estimator. We present the results in Figure A3.

**Figure A3.** Comparison of maximum difference in absolute bias and RMSE for linear-interpolation and discrete-score ΔPAC estimators.

Number of discrete scores is the number of partition cells in year 0, which have equal width. Partition cells in year 1 are identical to year 0, but moved in the positive direction by one-tenth of cell width. Continuous scores in year 0 and year 1 are normally distributed with $\mu = 0, \sigma = 1$ and $\mu = 0.1, \sigma = 1.1$, respectively.

Figure A3 shows that, in terms of both absolute bias and RMSE, the worst-case scenarios for $\hat{\theta}^{LIN}$ relative to $\hat{\theta}^{DISC}$ are fairly benign. For example, with only 20 discrete scores, the bias of $\hat{\theta}^{LIN}$ is at most about 0.6 percentage points larger than that of $\hat{\theta}^{DISC}$; the maximum RMSE difference is even smaller (less than 0.3 percentage points).

However, the worst-case scenario for $\hat{\theta}^{DISC}$ relative to $\hat{\theta}^{LIN}$ is far more extreme: its bias and RMSE may be more than 9 percentage points larger than those of $\hat{\theta}^{LIN}$.
Although the gap between $\hat{\theta}^{LIN}$ and $\hat{\theta}^{DISC}$ diminishes as the number of discrete scores grows larger, the linear-interpolator remains superior to the discrete-score estimator on this metric even when the score range is divided into hundreds of discrete scores. Additional results (not pictured here) reveal that it is possible for $\hat{\theta}^{LIN}$ to have strictly smaller bias than $\hat{\theta}^{DISC}$ over the entire score range; meanwhile, there are always at least some scores at which $\hat{\theta}^{DISC}$ suffers larger bias than $\hat{\theta}^{LIN}$ when the continuous-score distribution is “smooth.” In short, the worst-case scenario for $\hat{\theta}^{DISC}$ is generally far worse than for $\hat{\theta}^{LIN}$.

Overall, the analyses above point to the following conclusions:

1. Linear interpolation eliminates the “sawtooth” volatility across cutoff scores produced by the discrete-score estimator, providing a continuous $\Delta PAC$ function that does not result in large “leaps” when cutoff scores are slightly perturbed.

2. On average, across plausible cutoff score ranges, the linear-interpolation estimator is less biased and has smaller RMSE than the discrete-score estimator.

3. At worst, the linear-interpolation estimator may suffer slightly larger bias and RMSE than the discrete-score estimator at some cutoff scores.

Taken together, these results suggest that linear interpolation substantially alleviates $\Delta PAC$ estimation problems identified in the original paper.
5.2. **Polynomial Interpolation**

Rearranging terms, the linear PAC interpolator for \( k \in (g_i, h_i] \) can be written as:

\[
\overline{PAC}^{LIN} = \frac{h_i T_i(g_i) - g_i T_i(h_i)}{N_i(h_i - g_i)} + \left( \frac{T_i(h_i) - T_i(g_i)}{N_i(h_i - g_i)} \right) k = \beta_0 + \beta_1 k,
\]

where \( \beta_0 = \frac{h_i T_i(g_i) - g_i T_i(h_i)}{N_i(h_i - g_i)} \) and \( \beta_1 = \frac{T_i(h_i) - T_i(g_i)}{N_i(h_i - g_i)} \). The polynomial estimator,

\[
\overline{PAC}^{POLY} = \beta_0 + \beta_1 k + \beta_2 k^2 + \cdots + \beta_n k^n
\]

is a natural extension of this, where the coefficients \( \beta_0, \ldots, \beta_n \) are computed based on scores immediately below \( g_i \) and above \( h_i \). Avoiding under-specification of the coefficient-estimation system for a polynomial of degree \( n \) requires at least \( n + 1 \) points of data. Using the \( n + 1 \) score/value pairs closest to \( k \) to derive coefficients for a polynomial of degree \( n \) provides an exact solution and is straightforward to compute via a homogeneous system of linear equations; the linear-interpolation estimator is a special case of this, with \( n = 1 \).

Analysis of polynomial estimators of arbitrary degree is cumbersome and not provided here. However, any polynomial of reasonably low degree improves upon the discrete-score estimator according to the criteria outlined in section A1.3. When polynomial and linear-interpolation estimators differ substantially, it is generally not possible to determine which estimator has greater bias, as this depends on the shape of the (unobservable) continuous-score distribution. In practice, though, the difference between polynomial and linear-interpolation estimators is typically extremely small, as cutoff scores tend to be located where CDF values are close to linear with respect to discrete scores. Thus, linear interpolation should suffice for most test comparisons.
III. Does Computer-Adaptive Testing Affect Achievement Gaps?

Evidence from a State-wide Natural Experiment

1. Introduction

Achievement gaps between student subgroups have long been of interest to educational policymakers as indicators of equity in educational opportunity. At the federal and state levels, gaps as measured by scores on state-wide standardized tests began to receive special attention under the No Child Left Behind Act of 2001, whose subtitle read, “An Act to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind” (U.S. Department of Education, 2001). More recently, spurred by the federal Race to the Top initiative, a growing number of states has begun to incorporate score “gap closure” among the “annual measurable objectives” by which schools are evaluated for accountability purposes (Center for American Progress, 2014). Thus, measurement of gaps in a manner that plausibly reflects actual differences in student proficiency has become increasingly important.

At the same time, state accountability assessments have been changing rapidly. As of this writing, at least one-third of states have transitioned or declared plans to transition from traditional to computer-adaptive tests (CATs), at least in English language arts and mathematics (Gewertz, 2016). Although CATs offer many theoretical benefits over traditional paper-and-pencil tests – including increased precision in estimation of proficiency, decreased scoring and reporting time, and discouragement of certain types of cheating, such as memorizing response strings or “leaking” of test forms prior to administration (Linacre, 2000) – few large-scale studies have investigated their potential impact on the relative test performance of student subgroups. Furthermore, most studies
investigating differences between CAT and paper-and-pencil performance have, understandably, tended to focus on the broad question of whether or not CATs in general produce different results from traditional tests, rather than the more specific question of if, and by how much, the use of CAT-based testing programs in lieu of paper-and-pencil tests may affect accountability-relevant student outcomes in operational contexts.

This study focuses on the latter question, making use of a state-wide natural experiment to investigate the potential magnitudes of changes in observed achievement gaps between male versus female, black versus white, and low-income versus non-low-income students resulting from a state-wide transition to CATs. Compared to most experimental studies, the large and representative state-wide sample, combined with the use of operational state tests, enables estimation of gap effects with greater precision and in a context that is directly relevant to accountability outcomes, while at the same time limiting inferences regarding which test characteristics may be most responsible for gap changes.

2. Background and Research Agenda

Recent years have seen an increase in the number of studies employing randomized-trial or controlled designs to isolate the effects of item and test mode on test outcomes on the relative performance of student subgroups. Gallagher et al. (2002), in a meta-analysis of large-sample, experimental and quasi-experimental computer-adaptive studies focusing on the SAT, GRE, GMAT, and Praxis exams, found small gender-gap impacts on both mathematics and verbal scores, with computer-adaptive forms favoring males, and sizable impacts by race and ethnicity, with computer-adaptive tests having a positive impact for African-Americans compared to white examinees. Horne (2007), in a
set of experimental studies of middle- and high-school-aged UK students, found that girls significantly outperformed boys on paper-and-pencil literacy tests, while finding no significant gender differences on computer-adaptive forms. In contrast, Sandene et al. (2005) and Horkay et al. (2006) investigated gender gaps on computer-based and paper-and-pencil forms of a writing assessment from the National Assessment of Educational Progress among a nationally-representative sample of eighth-grade students and found no significant difference between the two test forms. Various smaller studies have also failed to find consistent gap differences between computer-based tests and their paper-based equivalents in literacy (Singleton, 2001; Higgins et al., 2005) and other subjects (Poggio et al., 2005; Wallace & Clariana, 2005).

Controlled experiments clearly provide critical benefits. If students are successfully randomized and subjected to identical conditions aside from the assessments assigned to each group, then mean scores for students taking the paper-and-pencil test provide an unbiased counterfactual for those taking the computer-based test – that is, an unbiased estimate of how the computer-based test-takers would have scored if they had taken the paper-and-pencil test instead – and differences in means between the groups can credibly be attributed to differences between the tests. Furthermore, if the computer-based tests are equivalent to the paper-and-pencil tests except in their mode of delivery, then differences in means can be attributed to mode effects. Experiment-based studies are thus well-suited to isolating the effects of delivery mode on student performance.

At the same time, because such studies are typically not designed to examine impacts on accountability metrics, generalizing their results to accountability-relevant contexts can prove difficult. First, voluntary samples may introduce selection bias, as
students who are likely to volunteer for such studies may differ in unobserved ways from the general population of K-12 students; this may partially explain the diversity of findings, and it raises the possibility that findings may not predict outcomes well for public K-12 populations. Second, the limited sample sizes typical of these studies often prevent them from detecting effects that may be small but consequential with respect to accountability outcomes.

Moreover, mode effects represent only one of a number of factors that may lead to differences in relative performance but are unrelated to differences in subject proficiency. While fixed-form tests are typically constructed according to one (or at most, a small number) of blueprint specifications, CAT blueprints often vary from student to student by construction, leading to potentially substantial differences in test content (and, therefore, changes in the proficiency targeted by the test). Paper-and-pencil tests are also generally easier to administer to nearly all students at once, while limitations in technological infrastructure (e.g., the number of available computers in a school) often lead states to administer CATs across multiple sessions or to smaller groups of students at a time. Administrative procedures for CATs can also be substantially more complex than for paper-and-pencil tests for administrators and students, leading to confusion and anxiety that may affect test performance (e.g., California Department of Education, 2017a & 2017b). To the extent that such factors produce differences in performance that are unrelated to student proficiency on a common domain, inferences regarding school contributions to achievement gap closure may be sensitive to the choice of testing format.
For the purpose of examining the potential accountability impacts of a change from a paper-and-pencil to CAT assessment program, then, an ideal experiment would involve thousands of broadly representative public school students, each assigned randomly to a paper-and-pencil test or CAT equivalent of the actual accountability test and taking their assigned test under conditions that closely mimic the conditions under which “live” assessments would be administered.

Natural experiments provide distinct advantages over randomized experiments in this regard, while suffering a number of disadvantages. The use of observational data for all available students ensures that the sample is representative of the population of interest and maximizes statistical power, while the use of observed scores on state tests – exactly the same tests on which accountability outcomes are based – ensures that findings are generalizable to operational tests and uninfluenced by “laboratory” conditions.

On the other hand, a critical limitation is that, because students are no longer assigned to tests by researchers, correlations between unobserved student characteristics and test format may produce bias in effect estimates if the former are causally related to test performance. For example, if students choose between paper-and-pencil and CAT formats based on their level of comfort and familiarity with technology, then mean performance on the former will likely produce a biased estimate of how CAT test takers would have performed on the paper-and-pencil test. Careful investigation of the assumptions required to identify the causal effects of the test change is therefore needed to produce plausible estimates of the impact of test format (Angrist & Pischke, 2008; Murnane & Willett, 2010). Additionally, unless the CAT was successfully developed as an equivalent to the paper-and-pencil test, the effects of changes to test format may be
confounded with changes in the target proficiency measured by the test. Although any change in test gaps between the two test formats may arguably be consequential for accountability, changes attributable to the latter may limit the generalizability of findings beyond the two specific tests in question.

This study makes use of an abrupt, state-wide assessment change in Delaware, from the paper-and-pencil Delaware Student Testing Program (DSTP) test to the computer-adaptive Delaware Comprehensive Assessment System (DCAS) test, as a source of arguably exogenous test assignment in a natural experiment intended to answer the question, “How much might a change from paper-and-pencil to computer-adaptive testing affect achievement gaps in an operational testing system, and for which groups?” Furthermore, given the lack of equivalency between the two tests, I also investigate the question, “To what extent might gap changes be attributable to changes in test content?”

3. Research Design

3.1. Data and Sample

The data for this study were provided by the Delaware Department of Education (DDOE) to the Center for Education Policy Research at Harvard (CEPR) and were assembled from three sources. The first contains annual test scores in Mathematics and Reading for the academic years ending 2007 through 2014. The second, covering the academic years ending in 2007 through 2012, includes student demographic and school assignment information and was produced using DDOE’s primary, “transactional” student records systems. The third, also containing demographic and school assignment information, was produced using DDOE’s recently-implemented longitudinal data system and includes records only from 2009-2014. All three datasets include unique student
identifiers, enabling observations to be merged into a single dataset, with one observation
per student per year, spanning academic years 2007-2014. Among students who appear
in both demographic datasets, approximately 95% matched exactly on all characteristics
used in this study. Where conflicts existed, values from the more recent 2009-2014
dataset were used.

I restrict the dataset to student-years in grades 3 through 10 and exclude rare
observations (0.4%) in which the student did not take a test matching his or her enrolled
grade level. Among the remaining students, 37 exhibited a change in gender between
years, primarily due to dataset discrepancies. An additional 581 exhibited a change in
race, mostly due to apparent reporting inconsistencies (e.g., “black” or “white” changing
to “multiple races”). For the sake of within-student consistency, and because such
changes were uncommon, I exclude these students from the analysis sample, resulting in
3,281 dropped observations (0.5%).

The final analysis dataset comprises 600,848 student-year observations associated
with 165,840 unique students tested during academic years 2007 through 2014, with
72,569 to 77,643 students per year.

3.2. Research Variables

3.2.a. Test Scores

To avoid public confusion and unwanted comparisons between scale scores,
DCAS scores were reported on a different scale from those on the DSTP. As a result,
neither means nor gaps can be meaningfully compared between the two tests using scale
scores. For the main analyses, I therefore follow common practice by standardizing scale
scores within grades and tests to their means and standard deviations in the year closest to
the test transition (2010 for DSTP scores and 2011 for DCAS).

Standardization may be inappropriate or produce misleading results when score
distributions exhibit extreme characteristics, such as strong ceiling or floor effects or
large changes in shape between years. To help evaluate whether such characteristics
might be present, I examine grade-year-level scale score histograms for each test, of
which I present a subsample below. DSTP histograms are presented as discrete, while
DCAS scores are binned, due to the large number of distinct scores (approximately 400)
on the CAT. While histograms reveal variation in skewness across grade-years in both
tests, there do not appear to be large, systematic differences between the two tests, and no
noticeable ceiling or floor effects are present in any grade or year.

| Figure 1. Scale score histograms for selected Delaware state tests. |
|-----------------------------|-----------------------------|-----------------------------|
| Grade 6                     | Grade 7                     | Grade 8                     |
| DSTP, 2010                  |                             |                             |
|                             |                             |                             |
| DCAS, 2011                  |                             |                             |
| Source: Delaware Department of Education. DSTP scores are presented as discrete. DCAS scores treated as continuous, due to the large number of score values (approx. 400 per grade-year). |

DSTP and DCAS tests for the same subject and grade were designed to align with
the Delaware Content Standards, but often differed in content (DDOE, 2009; AIR,
2011b). Because the DCAS, like most computer-adaptive tests, required a very large
item bank, the DSTP item bank was greatly expanded to produce the item bank for the DCAS. Furthermore, test blueprints – specifications for the number of items in each strand or standard that appear on a test – were fixed on the DSTP, but varied both overall and on a per-student basis on the DCAS, due to its “adaptive” nature. Additionally, the DSTP included human-graded short-answer and open-response items, while the DCAS was entirely machine-scored and included interactive, machine-scored constructed-response items.

Technical documentation, available for the 2008 DSTP and 2011 DCAS tests, provides information on the distributions of items by standard for Mathematics for both tests, which are presented in Table 1, below.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Standard 1:</th>
<th>Standard 2:</th>
<th>Standard 3:</th>
<th>Standard 4:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Numeric Reasoning</td>
<td>Algebraic Reasoning</td>
<td>Geometric Reasoning</td>
<td>Quantitative Reasoning</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>15</td>
<td>8</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>11</td>
<td>14</td>
<td>9</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>15</td>
<td>21</td>
<td>8</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>58</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>25</td>
<td>15</td>
<td>11</td>
<td>61</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>22</td>
<td>16</td>
<td>12</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>DCAS (Ranges)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30-33</td>
<td>8-9</td>
<td>9-10</td>
<td>1-3</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>24-27</td>
<td>8-9</td>
<td>12-14</td>
<td>4-5</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>22-24</td>
<td>8-10</td>
<td>8-10</td>
<td>8-9</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>24-27</td>
<td>8-9</td>
<td>8-9</td>
<td>8-9</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>16-18</td>
<td>15-17</td>
<td>9-10</td>
<td>8-10</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>9-10</td>
<td>23-25</td>
<td>8-9</td>
<td>8-9</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>2-3</td>
<td>35-38</td>
<td>2-3</td>
<td>9-10</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>2-3</td>
<td>18-22</td>
<td>20-21</td>
<td>8-9</td>
<td>50</td>
</tr>
</tbody>
</table>

DCAS ranges are as specified in 2010-2011 technical report.
Compared to the DSTP, the DCAS tests generally included substantially more Numeric Reasoning items in the lower grades and fewer in the higher grades, a large spike in Algebraic items and decline in Geometric items in grades 8 and 9, and fewer Quantitative Reasoning items in grades 3 and 4.

In Reading, changes to technical documentation prevent direct comparison of item distributions between tests. Both tests were restricted to items aligned to two standards, “Reading Comprehension” and “Literary Text.” DCAS tests in grades 3-10 contained 30-40 items in the first category and 10-20 in the other (out of 50 total), but item distributions among standards were unavailable in the DSTP documentation. Items were further classified according to passage type, “Informational” or “Literary,” but DCAS documentation provides distributions in terms of item frequencies, while DSTP documentation provides only passage frequencies. The available information by passage type is presented in Table 2. With the exception of grade 5, the proportion of DCAS items in each type overlaps the proportion of corresponding passages on the DSTP, although the number of items per passage is not known.

<table>
<thead>
<tr>
<th>Grade</th>
<th>DSTP - # of passages</th>
<th>DCAS - # of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

DSTP figures reflect frequencies on 2008 test. DCAS figures indicate ranges specified in 2010-11 technical report.
The differences between tests suggest that content changes between the two tests may be primarily responsible for any observed changes in gaps. Although this cannot be entirely ruled out with the available data, I explore this possibility in Section 6.

3.2.b. Student Characteristics

For subgroup identification, I employ gender indicators as reported in the source dataset and follow U.S. Office of Management and Budget guidelines in coding students as Hispanic if they were reported as Hispanic of any race (U.S. OMB, 1995); I then assign non-Hispanic students to mutually-exclusive “black,” “white,” or “other” categories. As noted above, any students whose race or gender changed during the observation window were dropped from the sample; race and gender are therefore time-invariant within students in the sample.

For time-varying characteristics, I code student-years as English Language Learners (ELL) if they were assigned any ELL status in that year and as disabled if they were assigned any disability status that year. The state’s definition of “low-income” was in flux during the period covered by the observation window, with a transition from “Free or Reduced-Price Lunch” (FRPL) in 2007-2010 to a broader “Economically Disadvantaged” indicator in 2011-2013, with a further redefinition of the latter in 2014 (DDOE, n.d.). However, the “transactional” demographic dataset includes student FRPL status indicators for each year from 2007 to 2012. Since this is the only available low-income indicator that remains consistent across the DSTP-to-DCAS transition, I use FRPL status as a low-income indicator, coding students as FPRL if they were identified as eligible for either free or reduced-price lunch. As a consequence, models that include FRPL as a predictor are restricted to academic years 2007-2012.
3.2.c. School Assignment

Finally, as in most states, school assignment accounts for a considerable proportion of the variation in student scores, and such variation often depends on within-school grade levels. More importantly, gap-based accountability outcomes are typically most consequential for schools, rather than districts or states. For these reasons, I include school-by-grade fixed effects in the main analyses, so that results are based only on variation within school-grades.

3.3. Analytic Strategy

As with any causal study, identification of the effects of the DCAS transition depends critically on obtaining unbiased estimates of the counterfactual DSTP gaps during the years in which students took the DCAS (Rubin, 1974). The key identifying assumption underlying this analysis is that the population of students who took the DSTP did not differ from the DCAS population in ways that would affect their relative performance. If this assumption holds true, then DSTP gaps observed during the 2007-2010 period would have persisted during the period 2011-2014 if students had taken the DSTP, rather than the DCAS, in those years. Gaps observed during the DSTP years would then provide a valid counterfactual for gaps during the DCAS years; any significant differences between the counterfactual and observed DCAS gaps would therefore be attributable to the DCAS transition.

I employ a difference-in-differences approach (Angrist & Pischke, 2009; Murnane & Willett, 2010; Dynarski, 2003), in which first differences are defined as the mean gaps on each of the two tests, and the difference-in-differences is the gap change between
tests, to identify the effect of the DCAS transition, modeling the standardized score for student $i$ in school $s$, grade $g$, and year $t$ as

$$Y_{sgt} = \alpha_{sg} + \delta \cdot DCAS_{ti} + \gamma \cdot MALE_{i} + \beta \cdot (DCAS_{ti} \times MALE_{i}) + \theta'X_{ti} + \varepsilon_{ti} \quad (A)$$

where $DCAS_{ti} = 1$ if the test was a DCAS test and 0 if DSTP, $MALE_{i}$ is an indicator for the student being male, and $X_{ti}$ is an optional column vector of (possibly time-varying) student characteristics. The school-by-grade fixed effect, $\alpha_{sg}$, accounts for any unobserved, time-invariant differences between schools and grades, so that all other parameters are estimated using only variation within school-grade units.

When covariates $X_{ti}$ are excluded from the model, the parameter $\delta$ captures the average within-school-grade score change between the DSTP and DCAS for girls, while $\gamma$ captures the average DSTP score gap between boys and girls. Under “parallel trends” assumption underlying this difference-in-differences model, the counterfactual DCAS gap – that is, the gap that would have been observed if students had taken the DSTP in years when they actually took the DCAS – is identical to the observed DSTP gap, $\gamma$. The primary estimand of interest, $\beta$, is the average difference between this counterfactual and the actual gap on the DCAS; a value of $\beta = 0$ implies that gaps did not change between tests.

For each subgroup, the difference-in-differences approach effectively relies on a single, “pooled” score estimate for the DSTP period, and another for the DCAS period. Over the eight-year observation window employed here, this approach may produce biased estimates if test scores exhibit secular score trends that differ by subgroup. For example, suppose male-female score gaps were unaffected by the DCAS transition but were gradually decreasing over the entire observation period, 2007-2014. The average
observed gap over the DSTP period, 2007-2010, would then be larger than the gap over
the DCAS period, 2011-2014. The difference-in-differences estimate would therefore
lead to the conclusion that the DCAS caused performance gaps to shrink, when in
actuality, the “shrinkage” would have occurred regardless of which test was used.

I address this potential problem in two ways. First, I refit the model in (A) using
only the years immediately before and after the DCAS transition, 2010 and 2011. This
employs the narrowest possible “window” with which a DCAS effect can be estimated; if
\( \beta \) estimates differ substantially between this restricted sample and the full sample, then
longer-term gap trends may be biasing results based on the full-sample. Compared to the
full-sample estimation, the primary drawback of this model is a decrease in statistical
power, due to the reduced number of observations.

Second, I estimate \( \beta \) using a trend-adjusted model, in which I estimate a distinct
linear trend for each of the four possible subgroup-test type combinations:

\[
Y_{sgti} = \alpha_{sg} + \tau_1 \cdot T_t + \gamma \cdot MALE_i + \tau_2 \cdot (T_t \times MALE_i) + \delta \cdot DCAS_{ti} + \tau_3
\]

\[
\cdot (T_t \times DCAS_{ti}) + \beta \cdot (MALE_i \times DCAS_{ti}) + \tau_4 \cdot (T_t \times MALE_i \times DCAS_{ti}) + \theta'X_{ti} + \epsilon_{ti} \tag{B}
\]

The equation above is similar to (A), except for the introduction of the linear school-year
term, \( T_t \), and its interactions with \( MALE_i, DCAS_{ti} \), and \( MALE_i \times DCAS_{ti} \). The
parameters associated with these terms, \( \tau_1, ..., \tau_4 \), allow scores to change linearly across
time within each of the four subgroup-by-period combinations. I set \( T_t \) to be equal to the
observation’s school year minus 2011; thus, \( T_t = 0 \) for tests taken in the first year of the
DCAS. The equation is then identical to (A) for observations in 2011, so that \( \beta \) again
captures the average difference between the observed and counterfactual gaps. The
primary difference between the two models is that the counterfactual in (A) is defined as the average observed DSTP gap, while in (B), the counterfactual gap is predicted based on the assumption that linear score trends observed over the DSTP years would continue into 2011.

4. Findings

4.1. Male-Female gaps

Results for male-female gaps in Reading and Math are presented in Tables 3 and 4, respectively. For each model, the DSTP Gap row contains the estimated DSTP gap between boys and girls, with positive values indicating males scored higher than females. For models suffixed with “(a),” this is simply the parameter estimate associated with Male; for covariate-adjusted models suffixed with “(b),” the estimate is computed using within-group covariate means for students in the sample in 2011, the first year of the DCAS. The latter is required in order to estimate the counterfactual gap in that year (i.e., the gap that would have been observed among students in 2011 if they had taken the DSTP). For all models, the DCASxMale row contains the estimated effect ($\hat{\beta}$) of the DCAS transition on score gaps.

Differences-in-differences models labeled as “All years” are fit using all available data, while those labeled “2010-2011” are limited to student-years just before and after the DCAS transition. “Trend-adjusted” models are fit using all available data. Covariate-adjusted models include indicators for race (Black, Hispanic, and Other), free or reduced-price lunch status (FRPL), English language learner status (ELL), and whether or not the student was identified with a disability. Additionally, I include interactions between each of these indicators and both Male and DCAS indicators, in
order to avoid confounding gender-specific effects with those attributable to other factors. For example, suppose disabled students performed significantly better in Reading on the DCAS, compared to the DSTP. Failing to allow the effect of disability to vary between periods would cause this change to be captured by the time-dependent DCAS and DCASxMale parameter estimates, and since disabled students are disproportionately male, this would produce a positive bias in the latter. Finally, because FRPL indicators are unavailable in 2013 and 2014, covariate-adjusted “All years” models are limited to 2007-2012.
In Reading, all models point to a significant narrowing of the gap between boys and girls after the DCAS transition. Prior to the transition, boys scored an average of .242 to .269 standard deviation (SD) lower than girls in the same school and grade, depending on the model. After the transition, boys exhibited an estimated relative score increase of .082 to .116 SD, narrowing the gap by 34-44%. The estimates from the full-sample models, 1(a) and 1(b), are somewhat smaller than the corresponding restricted-sample models, 2(a) and 2(b), suggesting that the results may be sensitive to the width of the
observation window. As can be observed in Figure 1 below, this is likely due to the fact that the gap in Reading was slowly widening over the last four years of the DSTP.

Importantly, effect estimates from the covariate-adjusted models, 2(b) and 3(b), are not significantly different from the corresponding models without covariates, 2(a) and 3(a). Adding covariates to model 2(a) changes the DCAS effect estimate from .105 to .095 SD, while standard errors for the estimates are .009 and .008, respectively. Meanwhile, the covariates greatly improve the explanatory power of the model, increasing its R-squared from .167 to .345. The stability of the estimate after the addition of highly explanatory covariates provides support for the model’s identifying assumptions: if the DCAS effect estimate observed in model 2(a) were attributable to unobserved factors correlated with the transition and scores, then adding such factors to the model should change the estimate. The trend-adjusted estimates exhibit similar robustness to the inclusion of covariates: adding covariates to model 3(a) changes the estimate from .116 (SE = .010) to .107 (SE = .009) SD, while increasing the R-squared from .165 to .339. Overall, the results strongly suggest that the DCAS transition’s effects on gaps are not attributable to endogeneity, though I explore this further in Section 5.
Table 4. Estimated effect of DCAS transition on standardized score gaps between Male and Female students: Math

<table>
<thead>
<tr>
<th>Model</th>
<th>Difference-in-differences</th>
<th>Trend-adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All years</td>
<td>2010-2011</td>
</tr>
<tr>
<td></td>
<td>4(a) 4(b)</td>
<td>5(a) 5(b)</td>
</tr>
<tr>
<td>DCASxMale</td>
<td>0.011*</td>
<td>0.017 0.010</td>
</tr>
<tr>
<td></td>
<td>(0.005) (0.005)</td>
<td>(0.009) (0.008)</td>
</tr>
<tr>
<td>DSTP Gap</td>
<td>0.037***</td>
<td>0.027*** 0.023***</td>
</tr>
<tr>
<td></td>
<td>(0.003) (0.003)</td>
<td>(0.007) (0.006)</td>
</tr>
<tr>
<td>Covariates</td>
<td>X X</td>
<td>X X</td>
</tr>
<tr>
<td>School-grade effects</td>
<td>X X X X</td>
<td></td>
</tr>
<tr>
<td>N (student-years)</td>
<td>599,768 446,217</td>
<td>150,392 150,392</td>
</tr>
<tr>
<td>R²</td>
<td>0.170 0.338</td>
<td>0.18 0.344</td>
</tr>
<tr>
<td>School-grade groups</td>
<td>944 882</td>
<td>714 714</td>
</tr>
</tbody>
</table>

*p < 0.05, **p < 0.01, ***p < 0.001. Standard errors in parentheses. Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. DCASxMale is estimated effect of DCAS transition on 2011 score gaps between male and female students. DSTP Gap is estimated counterfactual gap in 2011. Covariates comprise indicators for race (black, Hispanic, and “other”), free or reduced-price lunch status (FRPL), English language learner status, and disability status, and their two-way interactions with Male and DCAS. “All years” includes 2007-2014 for basic models (a) and 2007-2012 for covariate-adjusted models (b); latter models exclude years in which FRPL indicators are unavailable.

Results in Math are more sensitive to model specification. On average, boys performed slightly better than girls on the DSTP, but this gap was generally narrowing over the 2007-2010 period. The basic difference-in-differences models thus indicate a small but significant advantage for boys on the DSTP – .023 to .027 SD, according to the restricted-sample models 5(a) and 5(b) – while estimating that the effect of the DCAS was positive, but either small or insignificant. On the other hand, the trend-adjusted models, which extrapolate the narrowing of the DSTP gap into the first DCAS year, estimate a smaller or insignificant counterfactual DSTP gap in 2011 (.012-.016 SD), while suggesting that the DCAS caused the gap to widen by .021 to .031 SD, a
statistically significant departure from the DSTP trends. As with the results for Reading, estimated effects within each model type are highly robust to the addition of covariates that explain a large amount of score variation, with coefficients between the (a) and (b) models differing by only about one standard error.

The appropriateness of the basic model compared to the trend-adjusted model depends on the extent to which gap trends observed over 2007-2010 would have extended into the first year of the DCAS if students had taken the DSTP instead. I provide an illustration of these trends for both subjects in Figure 1, along with year-by-year, within-school-grade gap means. The points corresponding to Reading and Math in 2010 indicate the \textit{DSTP Gap} point estimates in models 2(a) and 5(a), respectively; points in 2011 are equal to the sum of the \textit{DSTP Gap} and \textit{DCASxMale} estimates. Fitted trend lines are based on the results of models 3(a) and 6(a), with the difference between trend lines in 2011 equal to the \textit{DCASxMale} estimate from the corresponding model.

Figure 2 illustrates that models 3(a) and 5(a) estimate a slightly widening gap between boys and girls in Reading and a narrowing gap in Math, respectively, over the DSTP years, while gaps in the DCAS years appear to be roughly constant in both subjects. Overall, however, the results for Math are substantively similar for all models: boys performed slightly better than girls on the DSTP, and the DCAS may have caused this gap to widen slightly.
Figure 2. Means and fitted trend lines for standardized score gaps between male and female students (positive favoring males) within school and grade, by year.

Students took the DSTP test in 2007-2010 and the DCAS test in 2011-2014. Scores standardized within grade to 2010 for DSTP and 2011 for DCAS.

4.2. Black-White and FRPL Gaps

To investigate additional gaps, I refit the restricted-sample difference-in-differences and trend-adjusted models, this time modifying predictors to allow \( \hat{\beta} \) to estimate the effect of the DCAS transition on gaps between i) black and white students, and ii) FRPL and non-FRPL students.

For black-white gaps, I limit the sample to only black and white students, who together comprise approximately 5/6ths of all observations. Results for both subjects are presented in Table 5. In Reading, unlike for male and female students, I find little or no evidence of a gap change caused by the DCAS transition. Effect estimates are close to zero and barely significant at most, ranging from .002 to .022 standard deviation. In contrast, I find a small, significant, positive effect in Math, with estimates ranging from...
.039 to .063 SD. However, the effects are small when compared to the overall DSTP gap of approximately 0.6 SD. At most, the results would imply that the DCAS transition caused the black-white gap in Math to narrow by about one-tenth. As with the male-female results, adding covariates increases the fit of each model without substantially changing estimates of the effect.

Results for FRPL gaps are presented in Table 6. As with black-white gaps, the score gaps between FRPL and non-FRPL students are large in both subjects, at approximately one-half of a standard deviation. While point estimates of the DCAS transition effect are negative, suggesting the DCAS may have exacerbated performance gaps between FRPL and non-FRPL students, they are close to zero and marginally significant, at best ($p > 0.01$ in all except one case). Overall, the results suggest that the DCAS transition had little, if any, effect on score gaps between FRPL and non-FRPL students.
Table 5. Estimated effect of DCAS transition on standardized score gaps between Black and White students.

<table>
<thead>
<tr>
<th>Model</th>
<th>Reading</th>
<th></th>
<th></th>
<th>Math</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diff-in-diffs.</td>
<td>Trend-adjusted</td>
<td></td>
<td>Diff-in-diffs.</td>
<td>Trend-adjusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7(a) 7(b)</td>
<td>(a) (b)</td>
<td></td>
<td>9(a) 9(b)</td>
<td>(a) (b)</td>
<td></td>
</tr>
<tr>
<td>DCASxBlack</td>
<td>0.004 0.017 (0.010) (0.010)</td>
<td>0.022* 0.002 (0.010) (0.011)</td>
<td>0.052*** 0.063*** (0.010) (0.010)</td>
<td>0.039*** 0.049*** (0.010) (0.011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSTP Gap</td>
<td>-0.497*** -0.535*** (0.008) (0.007)</td>
<td>-0.503*** -0.538*** (0.009) (0.008)</td>
<td>-0.580*** -0.613*** (0.007) (0.007)</td>
<td>-0.569*** -0.605*** (0.009) (0.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariates</td>
<td>x x x x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School-grade effects</td>
<td>x x x x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (student-years)</td>
<td>125,448 125,448 499,700 376,209</td>
<td>125,623 125,623 500,287 376,762</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.200 0.338 0.200 0.333</td>
<td>0.236 0.339 0.230 0.337</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School-grade groups</td>
<td>702 702 932 866</td>
<td>704 704 936 870</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. Samples include only White and Black students. DCASxBlack is estimated effect of DCAS transition on 2011 score gaps between Black and White students. DSTP Gap is estimated counterfactual gap in 2011; negative values indicate Black students scored lower than White students. Covariates comprise indicators for gender, free or reduced-price lunch status (FRPL), English language learner status, and disability status, and their two-way interactions with Black and DCAS. Diff-in-diffs. models are simple difference-in-differences models fitted to observations in 2010 and 2011. Trend-adjusted models employ observations from 2007-2014 in non-covariate-adjusted models and 2007-2012 in covariate-adjusted models (due to unavailability of FRPL indicators).
Table 6. Estimated effect of DCAS transition on standardized score gaps between FRPL and non-FRPL students.

<table>
<thead>
<tr>
<th>Model</th>
<th>Reading</th>
<th>Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diff-in-diffs.</td>
<td>Trend-adjusted</td>
</tr>
<tr>
<td></td>
<td>11(a)</td>
<td>11(b)</td>
</tr>
<tr>
<td>DCASxFRPL</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.013</td>
<td>-0.020*</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>DSTP Gap</td>
<td>-0.482***</td>
<td>-0.522***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Covariates</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>School-grade effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>N (student-years)</td>
<td>149,949</td>
<td>149,949</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.206</td>
<td>0.346</td>
</tr>
<tr>
<td>School-grade groups</td>
<td>712</td>
<td>712</td>
</tr>
</tbody>
</table>

*p < 0.05, **p < 0.01, ***p < 0.001. Standard errors in parentheses.

Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. $DCASxFRPL$ is estimated effect of DCAS transition on 2011 score gaps between students with free or reduced-price lunch status (FRPL) and those without. $DSTP$ is estimated counterfactual gap in 2011; negative values indicate FRPL students scored lower than non-FRPL students. Covariates comprise indicators for race (Black, Hispanic, or Other), gender, English language learner status, and disability status, and their two-way interactions with FRPL and DCAS. Diff-in-diffs. models are restricted to 2010 and 2011. Trend-adjusted models are based on observations from 2007 through 2012.
From the perspective of school accountability, the male-female gap change in Reading is likely to be the most consequential. I therefore focus primarily on this gap change in what follows.

5. **Threats to Validity**

5.1. **Selection Bias**

As noted above, the key identifying assumption underlying these findings is that the gaps observed on the DSTP can be used to estimate an unbiased counterfactual gap in the first year of the DCAS. This assumption would be violated if students within the same school and grade who took the DCAS differed from those who took the DSTP in ways that would affect the relative performance of subgroups. One plausible means by which this assumption would be violated would involve students manipulating their test participation – in this case, by exiting or entering the sample in anticipation of the test transition – in a manner that significantly affected achievement gaps. Although the robustness of the above estimates to the addition of covariates suggests that this did not occur, there remain many unobserved factors that may be correlated with both the DCAS transition and test scores and affect relative performance, but are not captured by the covariates. If, for example, boys who anticipated doing worse on the DCAS than the DSTP were more likely than their female peers to exit the sample in anticipation of the transition, then the results for male-female gaps above may be attributable to selection bias.

To investigate possible violations of this identifying assumption, I code an indicator for voluntary sample exit, ExitSample, for observations associated with students in grades 3 through 9 in years prior to 2014, with a value of 1 if the student was not
present in the sample the following year and 0 otherwise. For grade 10 or year 2014 observations, the indicator is undefined, since all such students exited the sample the next year.

I then make use of this indicator in two ways. First, I estimate the probability of voluntary sample exit as a function of being present in the sample in 2010, allowing parameter estimates to vary between boys and girls, in order to determine if exit rates differed significantly between boys and girls just prior to the DCAS transition. Because exit rates are highly dependent on grade level, I also include grade-level indicators and their interactions with Male in this estimation. The resulting model can be written as:

\[ \Lambda^{-1}(P(ExitSample_{it} = 1)) = \alpha + \delta \cdot MALE_i + \gamma \cdot Year2010_t + \beta \cdot (MALE_i \times Year2010_t) + \theta' X_{it}, \]

where \( \Lambda^{-1}(\cdot) \) is the standard logit function \( \left( \Lambda^{-1}(p) = \ln \frac{p}{1-p} \right) \); \( P(ExitSample_{it} = 1) \) is the probability of voluntary exit for student \( i \) in year \( t \); \( MALE_i \) is an indicator for a male student; \( Year2010_t \) is an indicator for academic year 2010; and \( X_{it} \) is a vector of grade-level indicators and their interactions with \( MALE_i \).

Second, for students present in the sample in 2010, I estimate standardized scores in each subject, using \( ExitSample \) and its interaction with Male as key predictors, to determine if within-school-grade, male-female gaps among students who exited the sample before the DCAS differed significantly from gaps among students who stayed. The resulting model is similar to model (A), but with \( ExitSample_{it} \) in place of \( DCAS_{ti} \).

I also code an indicator for voluntary sample entry, \( EnterSample \), for students in grades 4 through 10 in years 2008-2014, which takes a value of 1 if the student was not in
the sample the previous year and 0 otherwise. I then conduct analyses similar to those for sample exit, but with 2011 as the focal year, rather than 2010. These analyses, analogous to the two described above, intend to determine if i) boys were more likely than girls to enter the sample in reaction to the DCAS transition, and ii) if male-female score gaps for students new to the sample in the first DCAS year were significantly different from gaps among students who were not new.

Results from the exit analyses are presented in Table 7. In model (mf3), the positive, significant coefficient on Year2010 (.105) indicates that students in general were more likely to exit the sample after 2010, compared to 2007-2009 and 2011-2013. The positive coefficient on Male indicates that boys were more likely than girls to exit the sample in any given year. However, this difference in exit probability between boys and girls was not significantly higher or lower in 2010 than the overall, long-term average, as indicated by the non-significant coefficient on Male x Year2010.
Table 7. DSTP score gaps in 2010 and differences in sample exit probability between Male and Female students.

<table>
<thead>
<tr>
<th></th>
<th>Standardized Scores</th>
<th>Log-odds of Sample Exit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading (mf1)</td>
<td>Math (mf2)</td>
</tr>
<tr>
<td>ExitSample x Male</td>
<td>-0.041 (0.025)</td>
<td>-0.038 (0.025)</td>
</tr>
<tr>
<td>ExitSample</td>
<td>-0.300*** (0.019)</td>
<td>-0.301*** (0.019)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.255*** (0.007)</td>
<td>0.026*** (0.007)</td>
</tr>
<tr>
<td>School-grade effects</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

R² 0.200 0.200  0.021
N (students) 66,774 67,062 N (student-years) 463,415
School-grade groups 610 611

*p < 0.05, **p < 0.01, ***p < 0.001. Standard errors in parentheses. Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. ExitSample is an indicator for voluntary sample exit, coded as 1 if the student was absent from the sample the following year and 0 otherwise. Male indicates a male student. Year2010 is an indicator for academic year 2010, the final year of DSTP administration. Grade effects include main effects and their interactions with Male. Samples exclude observations in grade 10 and the final year of the dataset (2014). Standardized score models restricted to year 2010.

Models (mf1) and (mf2) compare score gaps between students who exited the sample just before the DCAS transition to those among students who remained to take the DCAS. While students who exited the sample performed significantly worse than those who remained (.3 standard deviation lower in both subjects), the gap between boys and girls among those who exited was not significantly different from those who stayed. Taken together, the results provide additional evidence that endogenous sample exit is not responsible for the DCAS gap-narrowing effect.
Table 8. DCAS score gaps in 2011 and differences in sample entry probability between Male and Female students.

<table>
<thead>
<tr>
<th></th>
<th>Standardized Scores</th>
<th>Log-odds of Sample Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading (mf4)</td>
<td>Math (mf5)</td>
</tr>
<tr>
<td>EnterSample x Male</td>
<td>-0.026</td>
<td>-0.048</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>EnterSample</td>
<td>-0.033</td>
<td>-0.142***</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.153***</td>
<td>0.042***</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>School-grade effects</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

|               | N (students) 64,871 | N (student-years) 461,528 |
| R²            | 0.161          | 0.188                  |
| School-grade groups | 523              | 522                    |

* p < 0.05, ** p < 0.01, *** p < 0.001. Standard errors in parentheses.
Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. EnterSample is an indicator for voluntary sample entry, coded as 1 if the student was absent from the sample the previous year and 0 otherwise. Male indicates a male student. Year2011 is an indicator for academic year 2011, the first year of DCAS administration. Grade effects include main effects and their interactions with Male. Samples exclude observations in grade 3 and the first year of the dataset (2007). Standardized score models restricted to year 2011.

Similarly, results for sample entry, presented in Table 8, indicate that sample entry cannot account for the DCAS transition effect. Although students were significantly less likely to enter the sample in 2011 compared to other years (log-odds = −.306), no significant difference exists between boys and girls in this respect. Additionally, those who entered the sample in 2011 exhibit score gaps that were not significantly different from students who were present in the prior year.

Overall, these results imply that the DCAS transition effect on score gaps between boys and girls is unlikely to be attributable to self-selection of students into and out of the sample.
5.2. Comparison with NAEP Gaps

While selection bias represents one mechanism that could introduce bias, a more general assumption underlying the restricted-sample, difference-in-differences models above is the “parallel trends” assumption that gaps in 2011 would have remained the same as in 2010 if students had taken the DSTP instead of the DCAS. This assumption might be violated if, for example, a large-scale program or intervention coincided with the DCAS transition and affected Reading proficiency differently for boys than for girls.

Unlike Delaware’s state tests, the federally-administered National Assessment of Educational Progress (NAEP) did not change in format during the DSTP-to-DCAS transition. Additionally, as with this study’s sample, the state-level NAEP sample for Delaware was designed to be representative of the state’s public school students. With respect to the main analyses above, the main drawbacks of NAEP are that i) the NAEP was administered in 2009 and 2011, rather than 2010 and 2011; ii) NAEP grade levels overlap with Delaware’s tested grades in only grades 4 and 8; and iii) the NAEP tests content and standards that differ from the Delaware Content Standards. Nevertheless, NAEP results may serve as an informative reference for investigating if the DSTP-to-DCAS gap changes may be attributable to proficiency changes in the tested population.

To provide for appropriate comparisons with NAEP results, I fit difference-in-differences models for each subgroup comparison without covariates and with years restricted to 2009 and 2011. I further restrict each model to a single grade (4 or 8), replacing school-grade fixed effects with school fixed effects. The resulting models produce, for each subgroup, a comparison of the 2011 DCAS gap to the 2009 DSTP gap, in terms of within-grade standard deviations, for grades 4 and 8. I then compare the
differences in these gaps to the corresponding gap changes on the NAEP, obtained via the NAEP Data Explorer (US DOE, n.d.).

Table 9. Estimated gap changes on NAEP and Delaware state tests, 2009 to 2011.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th></th>
<th>Math</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSTP-DCAS</td>
<td>NAEP</td>
<td>DSTP-DCAS</td>
<td>NAEP</td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-Female</td>
<td>0.037</td>
<td>0.003</td>
<td>0.049</td>
<td>-0.058</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.057)</td>
<td>(0.027)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>Black-White</td>
<td>0.102***</td>
<td>0.076</td>
<td>0.050</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.068)</td>
<td>(0.029)</td>
<td>(0.056)</td>
</tr>
<tr>
<td>FRPL-Non FRPL</td>
<td>0.041</td>
<td>-0.119*</td>
<td>0.023</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.055)</td>
<td>(0.026)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>Grade 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-Female</td>
<td>0.071**</td>
<td>-0.031</td>
<td>-0.056*</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.060)</td>
<td>(0.027)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>Black-White</td>
<td>0.030</td>
<td>0.003</td>
<td>0.069*</td>
<td>-0.030</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.062)</td>
<td>(0.028)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>FRPL-Non FRPL</td>
<td>-0.044</td>
<td>0.042</td>
<td>-0.031</td>
<td>-0.068</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.065)</td>
<td>(0.026)</td>
<td>(0.050)</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001. Standard errors in parentheses.
NAEP results standardized within grade to 2011 values.

The results, presented in Table 9, provide for a number of important conclusions. First, in nearly all cases, NAEP gap changes are not statistically significant, lending confirmatory support for the parallel trends assumption. (The lone exception, for the FRPL gap change in grade 4 Reading, suggests that results from models 11(a) and 11(b)
may be attributable to population changes, rather than the test transition.) Second, the key finding that the DCAS produced an increase in performance in Reading for boys relative to girls cannot be explained by the NAEP results. The point estimate of the NAEP gap change in grade 4 is close to zero, while the estimate in grade 8 is negative, indicating a widening of the gap; thus, if the NAEP differences suggest a violation of the parallel trends assumption, the bias from this violation must be causing the DCAS effect to be underestimated.

6. Effects of Content Changes

Content changes between the two tests imply that the proficiencies measured by the two tests may differ, potentially explaining the male-female gap change in Reading. This explanation cannot be entirely ruled out without additional, item-level data. Rather, the objective of this section is to determine the extent to which content changes, compared to factors unrelated to student proficiency, might account for the change in gap.

I approach this problem by investigating gap heterogeneity by grade. The logic behind this is as follows: DSTP assessments in different grades share the same format and administration procedures but are explicitly designed to test different content and standards. Content differences therefore contribute to within-test, between-grade heterogeneity, but test format differences do not. On the other hand, the DCAS and DSTP tests in the same grade and subject include overlapping items, but differ in format and administration. Thus, test format differences contribute to between-test, within grade variation, while the effect of content differences should be smaller than within tests, between grades, due to similar standards and at least some degree of overlapping content.
I examine gap heterogeneity by grade by fitting the restricted-sample, difference-in-differences models 2(a) and 5(a), with the addition *GRADE*, a vector of grade-level indicators for grades 4 through 10. To allow the effect of the DCAS transition to vary by grade, I include the two sets of two-way interactions *DCASxGRADE* and *MALExGRADE*, as well as the set of three-way interactions, *DCASxMALExGRADE*. The results are presented visually in Figure 3.
Figure 3. Estimated within-school gender gaps (positive favoring males) on 2010 DSTP and 2011 DCAS tests, by grade level.

Scores standardized within grade to 2010 for DSTP and 2011 for DCAS. Black lines indicate 95% confidence intervals.
The three-way interactions, $DCAS \times MALE \times GRADE$, collectively explain a statistically significant amount of the variation in scores, indicating that the effect of the DCAS on performance gaps varies significantly by grade level in both subjects. Within tests, both DCAS and DSTP gaps also vary significantly across grades. These results lend support to the hypothesis that male-female gaps may be sensitive to changes in content, as content differences are likely to contribute to this between-grade variation.

Of particular note is the fact that DSTP gaps vary a great deal from grade to grade and are especially large in grades 6 and 7, while DCAS gaps are smaller than average in these grades and more consistent overall; as a consequence, between-test gap changes are very large in grades 6 and 7. While content changes could certainly account for this, these large gap changes may also be attributable to proficiency-irrelevant changes between the two tests. Familiarity with technology may play a role if many reading passages and their associated questions no longer fit on the same computer screen for tests in these grades. Differences in handwriting may be more consequential in these grades if short-answer and open-response questions abruptly become more complex and require longer answers at the middle-school level (Bolger & Kellaghan, 1990), or differences in grader expectations between boys and girls may be larger at the middle school level, compared to the elementary and high school levels (Chase, 1986). These factors would explain the relatively large gaps in grades 6 and 7 on the DSTP, where graders influence scores, and the smaller grade-to-grade variation on the DCAS, which was entirely machine-scored. Further analyses, such as comparisons of score trends across grades for boys and girls and item-level performance differences between the two groups, could shed further light on these hypotheses.
To investigate the magnitude of the potential contributions of content changes to the overall DSTP-DCAS gap change in Reading, I compare between-grade gap changes on the DSTP to within-grade, between-test gap changes on the DSTP and DCAS, which I present in Table 10.

<table>
<thead>
<tr>
<th>Grade</th>
<th>DSTP prior grade</th>
<th>DSTP to DCAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.131</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.025</td>
<td>0.062</td>
</tr>
<tr>
<td>5</td>
<td>0.021</td>
<td>0.037</td>
</tr>
<tr>
<td>6</td>
<td>0.014</td>
<td>0.174</td>
</tr>
<tr>
<td>7</td>
<td>0.065</td>
<td>0.205</td>
</tr>
<tr>
<td>8</td>
<td>0.050</td>
<td>0.112</td>
</tr>
<tr>
<td>9</td>
<td>0.054</td>
<td>0.065</td>
</tr>
<tr>
<td>10</td>
<td>0.035</td>
<td>0.064</td>
</tr>
<tr>
<td>Average</td>
<td>0.038</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Scores are standardized within grade and year. “DSTP prior grade” is the absolute difference in gaps between the indicated grade and the previous grade. “DSTP to DCAS” is the gap difference between DSTP and DCAS. Figures computed from 2010 DSTP and 2011 DCAS results.

The results reveal that between-grade gap changes on the DSTP Reading test are smaller, both on average and for each grade, than the within-grade gap changes between the DSTP and DCAS. To more directly investigate within- vs. between-test variation, I compare the standard deviation of gap point estimates across grades on the DSTP and compare this to a within-grade, between-test pooled standard deviation obtained by taking the square root of the average the variances of within-grade, between-test gap point estimates, providing a rough estimate of the effects of grade compared to the effects of test type. The between-grade standard deviation of DSTP gap estimates is .046, while the within-grade, between-test estimate is .078.
While not conclusive, these results suggest that the DCAS gap effect cannot be attributed solely to changes in content. If content differences indeed contribute to gap differences, then the effect of content changes should be larger across grades within the DSTP than within grades across the two tests, since same-grade content across both tests is more similar, on average, than between-grade content on the DSTP. Put differently, for content changes to account entirely for the main results, content differences between tests sharing common items and aligned to the same content standards (i.e., DSTP and DCAS tests for the same grade) must have a stronger effect on gaps than differences between tests composed of entirely different items and aligned to different standards (DSTP tests for adjacent grades).

7. Discussion

This paper contributes to a growing body of literature investigating the potential effects of computer-adaptive vs. paper-and-pencil testing programs on observed score gaps between student subgroups, finding that the transition from a paper-and-pencil to computer-adaptive state-wide test caused gender gaps in Reading to decrease significantly among Delaware public school students, while gender gaps in Math, as well as black-white and FRPL-non FRPL gaps in both subjects, experienced negligible or insignificant effects. The results are robust to alternative model specifications and the addition of highly explanatory covariates, and the results for the gender gap in Reading cannot be explained by self-selection or secular changes in proficiency as measured by NAEP. Taken together, these findings support the notion that the change in the observed gender gap in Reading was indeed attributable to the test change, rather than plausible alternative factors.
Unlike traditional, controlled studies, however, this study is unable to disentangle the effects of changes in the proficiency measured by the tests from the effects of delivery mode, changes in administrative procedures, and other factors unrelated to student proficiency. While a comparison of within-test, between-grade gap variation to within-grade, between-test variation suggests that the gap change cannot be attributed solely to differences in content, a more rigorous and detailed analysis of item responses and testing-environment data would be needed to produce conclusive findings in this area.

On the other hand, the observational data employed in this study provide a number of key benefits for applying findings to accountability systems that evaluate schools based on performance gaps between students. First, the large sample sizes enable effects to be estimated with high precision, with standard errors for gap-change estimates on the order of one-hundredth of a standard deviation. Because of this, effects that are small but have a large potential impact on accountability ratings are more likely to be detected, and zero or near-zero estimates can be considered substantively negligible. Second, the sample comprises nearly all public school students in the state, so that findings are likely to be more generalizable to accountability-relevant populations than the voluntary, self-selected samples common to most controlled designs. Finally, the findings are based on outcomes from operational state tests, rather than tests designed for special populations (e.g., aspiring college or graduate students) or for purposes not directly related to measuring proficiency on state standards, and are thus directly applicable to school accountability outcomes. From a methodological perspective, this
study thus serves to exemplify both the benefits and limits of a natural-experimental approach to investigating the effects of a change in test type.

At the same time, because CATs, and computer-based tests in general, vary greatly in terms of user interface (keyboard-only, keyboard-and-mouse, touchscreen), item type (multiple-choice, machine-scored constructed-response, open- or extended-response, etc.), and administration (multiple vs. single-session, in-school vs. external location), gap effects from one assessment study may not generalize well to other tests or populations. Rapid developments in human interface design, along with their speedy proliferation among school-age children, may also lead to substantial changes in how tested populations perform on computer-based assessments, as children adapt to new technologies. Credibly evaluating gap effects caused by differences in test types may therefore require test-specific analyses, performed at regular intervals (Wang & Kolen, 2001; Wang et al., 2007).

An increasing adoption of CATs for state accountability tests, combined with increased use of “achievement gap closure” as an accountability metric, has underscored the importance of large-scale analyses investigating the potential for differential performance impacts between student subpopulations. States in which both paper-based and CAT forms are used for accountability – a group that includes nearly all states using the Smarter Balanced Assessment Consortium assessment, among others – represent the most obvious examples, since high-stakes consequences cannot be applied equitably if gaps are not comparable between tests, and large samples will likely be required to detect potential differences with adequate precision. Even states such as Delaware, where effectively all students take CATs, should benefit from investigations of gap effects, as
instructional efforts aimed at “gap closure” as measured by one type of test may be misplaced if the gap is smaller or nonexistent on the state test, potentially drawing resources away from other activities. The expanding use of CATs alongside paper-and-pencil equivalents may also provide opportunities for large-scale studies aimed at isolating mode effects on accountability tests from effects attributable to other, CAT-related factors, enabling policymakers to better identify the factors that contribute most strongly to differential test performance and to focus more narrowly on reducing their impact.
IV. The Sensitivity of Gain-Based Growth Metrics to Transformations of Scale

1. Introduction and background

Vertically-scaled assessments, in which test scores are theoretically comparable across students in adjacent grades, have seen increased use in recent years (Gewertz, 2016). Examples include the scales from the Measures of Academic Progress assessment (MAP) developed by the Northwest Evaluation Association (NWEA), the Common-Core-aligned Smarter Balanced Assessment Consortium assessment employed by roughly one-third of U.S. states, and a number of other assessments developed by state education agencies.

Vertical scales are often appealing because they support, at least nominally, intuitive notions about academic growth. Subtracting a student’s prior-grade score from her current-grade score produces a “gain score” that ostensibly measures growth on a scale that does not depend on prior-grade scores and can be meaningfully compared among students. In practice, however, these simple gain scores are rarely used directly to measure school contributions to academic growth. Instead, many states assign growth ratings to schools by weighting prior-grade and current-grade scores differentially. Some states use “value tables” that reward schools for students that achieve gains above some cutoff that varies by prior-grade score levels (Delaware Department of Education, 2010; Missouri Department of Elementary and Secondary Education, 2013; Illinois State Board of Education, 2013; South Carolina Education Oversight Committee, 2013; Florida Department of Education, 2016). Others effectively ignore the cross-grade, “vertical” properties of their test scales in favor of regression-based methods, such as “value-added” growth models and student growth percentiles (Sanders & Horn, 1994; McCaffrey et al.,...
2004; Betebenner, 2008; Colorado Department of Education, 2013; Ohio Department of Education, 2016; Delaware Department of Education, 2017). When applied to test scores reported on a vertical scale, these approaches can be viewed as an additional layer of computation that converts simple gains to growth values, which are then used to evaluate schools.

Mechanical relationships between gains and score levels, as well as other statistical limitations on the inferences that vertical scales support, have been widely discussed for some time (Rogosa & Willett, 1985; Tong & Kolen, 2007; Ballou, 2009). More recent research has focused on the effects of test design decisions on vertical scale properties, such as trends in means and variances between grades, and the sensitivity of accountability outcomes based on vertical score gains or value-added metrics to such properties (Briggs & Weeks, 2009; Briggs & Domingue, 2013).

With this paper, I seek to contribute to this body of knowledge in two ways. First, I provide a framework for examination of growth-based accountability models, which I define as any system that measures school contributions to student growth using current-grade scores and at least one prior-grade score. This framework reveals connections between value-table-based growth values, value-added school accountability metrics, between-grade trends in scale score parameters (mean gains, variance trends, skewness, and correlations), and transformations of scale scores in a unified manner. The framework also characterizes the effects of these factors on school growth outcomes relative to a simple mean-gain-score model. I use this framework to show that nonlinear transformations of scale scores are equivalent to perturbations or modifications of values that policymakers choose for value-tables. Nonlinear transformations are also equivalent
in impact to certain types of variation in between-grade score trends. Second, I examine the magnitudes of the changes that such transformations produce in gain-based school accountability outcomes, as well as the association between these changes and parameters of the underlying scale, including between-grade mean differences, variance trends, and score correlations.

2. Motivation and Research Agenda

Nearly all growth models employed for school accountability can be described as a process in which 1) each student is assigned a “growth value” that depends on their current-grade score and at least one prior-grade score, and 2) growth values are aggregated within schools to produce a “rating” used to evaluate school contributions to growth. Although growth models vary in their aggregation methods (Castellano & Ho, 2015), for simplicity’s sake, I focus on the first of these steps – the assignment of growth values to students – and restrict analyses to school-level means.

Growth values, like many educational metrics, are often sensitive to factors that have little to do with actual school contributions to student growth (McCaffrey et al., 2004; Castellano & Ho, 2013). In particular, in any growth model, any individual school (with some fixed set of student outcomes) may obtain very different growth ratings relative to others than it would have if the properties or parameters of the growth model had been set differently. As a simple example, a key feature of models that rate schools on a normative scale (including value-added and SGP models) is that an individual school’s rating depends on the performance of students who do not attend the school (i.e., all other students in the state). For simple gain-score models, a fixed set of student item responses may produce different gain scores, depending on how the test’s vertical scale is
constructed, thus leading to different school-level mean gains and school rankings (Briggs & Weeks, 2009).

Value tables, unlike gain-score models, do not require a vertical scale, but are similar in that they attempt to assess growth on an absolute scale rather than a normative one. Value tables represent a type of growth model in which the impact of policy decisions on growth values can be seen directly. An example, taken from Illinois’ growth model, is provided in Tables 1 and 2.

<table>
<thead>
<tr>
<th>Table 1. Performance level score ranges in Reading from Illinois value table growth model system.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance Level Score Ranges - Reading</strong></td>
</tr>
<tr>
<td><strong>Academic Warning</strong></td>
</tr>
<tr>
<td><strong>Grade</strong></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>


In the Illinois growth model, a student’s growth value is obtained by first determining her performance levels in the current and prior grade using the score ranges in Table 1, then finding the growth value corresponding to her current-grade, prior-grade performance level combination in Table 2. Growth values may then be averaged across all students in a school to obtain a school growth rating.
The example above illustrates the fact that, under a value-table system, growth values will be sensitive to at least two factors that have little to do with school or student performance: the values assigned to each cell in the value table, and the cutoff scores between performance level categories. Additionally, between-grade differences in the means, standard deviations, and shapes of score distributions, as well as changes in these parameters over time, will also affect growth value results. While this latter variation can certainly be caused by differences in actual school contributions to student growth, these parameters can also vary due to test construction decisions, error in linking and equating, and secular changes in student populations, so that growth ratings may change considerably even when school contributions to growth have remained the same.

The primary question this paper seeks to answer is, “How sensitive might school growth rankings be to variation in factors that are unrelated to school contributions to student growth?” As noted above, however, such variation can come from a large range of sources, making it impractical and cumbersome to model variation in each one.
individually. This motivates the other question addressed by this study: “How might we model at least some of this variation in a unified manner?” I address this question below, in Section 3, and turn to the former question in Section 4.

3. Framework

Generally speaking, growth-based school accountability models assign ratings to schools based on their students’ scores in the current grade and at least one prior grade. Formally, a two-grade growth-based accountability model may be described using a triplet, \( \{X, H, g\} \), whose elements are defined as:

1. A pair of sets, \( X = \{X_t, X_{t-1}\} \), of obtainable scale scores in the current and prior grades, \( t \) and \( t - 1 \);
2. A pair of grade-specific partitions on scale scores, \( H = \{H_t, H_{t-1}\} \), with \( H_t \subseteq X_t \) and \( H_{t-1} \subseteq X_{t-1} \); and
3. A “growth value function,” \( g: \{H_t \times H_{t-1}\} \to \mathbb{R} \), mapping cells in the Cartesian product of sets in \( H \) to “growth values,” which are then aggregated across students in a school to obtain the school’s overall growth rating.

Value tables, such as the one illustrated in Tables 1 and 2, can be easily described using these definitions, by defining \( H_t \) and \( H_{t-1} \) as the cutoffs (and overall bounds) of scale scores in grades \( t \) and \( t - 1 \), and defining \( g(h_t, h_{t-1}) \) according to the value table.

A system that assigns school growth ratings using an average of simple gain scores would be modeled by setting

- \( H_t = X_t, H_{t-1} = X_{t-1} \), (partition elements for every scale score in the current- and prior-grade, preserving all information), and
• \[ g(h_t, h_{t-1}) = h_t - h_{t-1} \] (equal and opposite-signed weights on the current and prior score respectively, resulting in growth values that are gain scores).

This framework is developed further via the examples below. First, I show that a simple value-table model in which growth values are commensurate with grade-to-grade changes in evenly-spaced proficiency categories is equivalent to a “coarsening” of a simple gain-score model, and similarly, the latter represents a “refinement” of the category-change model. Second, I extend this example to show that a) any value-table model can be approximated to arbitrary precision by a “refined” polynomial function, and b) certain classes of these approximation functions are equivalent to simple gain scores computed from scale scores that have been transformed. This latter result implies that many types of variation in factors affecting growth values – including perturbations in category cutoffs and value-table values, as well as perturbations in the shape and parameters of scale score distributions – can be modeled by transforming scale scores and computing gain scores from the transformed results. Finally, I show that this framework can also accommodate regression-based value-added models, but argue that their “growth to empirical expectation” property implies that they should be considered distinct from gain-based models.

3.1. Example 1: Category-change model

Consider a growth-based accountability model in which scale scores in adjacent grades are divided into evenly-spaced, ordinal proficiency categories (e.g., “Warning,” “Basic,” “Proficient,” and “Advanced”), and students’ growth values correspond to the number of categories they have moved between grades (positive for movement into
higher categories and negative for lower categories). Panel A of Figure 1 displays a plot illustrating such a system, with current-grade scale scores on the vertical axis, prior-grade on the horizontal axis, and growth values represented by shaded areas, with lighter shades indicating higher scores.

In Figure 1, Panel A, the union $X_t \cup X_{t-1}$ of obtainable scale scores across both grades consists of integers from 0 to 1000, which are divided into evenly-spaced categories according to partitions $H_t = H_{t-1} = \{0, 250, 500, 750, 1000\}$. Students are assigned 1 growth value point per category for grade-to-grade improvement, 0 for remaining in the same category, and -1 per category they decline. The growth values obtained at each current-grade, prior-grade pair, $(X_t, X_{t-1})$, are illustrated by shaded regions, with white indicating a score of 3, darkest gray a score of -3, and intermediate shades indicating integer scores from -2 to 2. With respect to measuring growth for school accountability purposes, the system effectively distinguishes among students using only four scores in each grade.

**Figure 1.** Refinement of hypothetical category-change growth metric into gain scores.

<table>
<thead>
<tr>
<th>A. Four categories</th>
<th>B. 20 categories</th>
<th>C. Gain scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Shaded regions indicate growth values under an accountability model in which students are awarded one point per proficiency category increment between years, with lighter shades indicating higher values." /></td>
<td><img src="image" alt="Axis values indicate scale score levels in current (Grade $t$) and prior (Grade $t-1$) years. Scale scores are partitioned into four categories in Panel A and 20 categories in Panel B. Simple gain scores are obtained when scores in each grade are “partitioned” into one cell per obtainable scale score. Panel C illustrates gain scores, with values indicated by shading and iso-gain curves (level sets) shown as dashed lines." /></td>
<td><img src="image" alt="Shaded regions indicate growth values under an accountability model in which students are awarded one point per proficiency category increment between years, with lighter shades indicating higher values." /></td>
</tr>
</tbody>
</table>

85
Finer-grained distinctions could be achieved by dividing the four existing categories into five subcategories each, producing $H'_t = H'_{t-1} = \{0, 50, 100, ..., 950, 1000\}$, resulting in $20 \times 20 = 400$ cell combinations, as shown in Figure 1, Panel B. Although the growth value magnitudes have changed from Panel A, the latter can be recovered by coarsening the new partitions to produce those in Panel A, aggregating over growth values in the coarsened cells (e.g., by choosing the minimum score), and multiplying the resulting values by a scaling constant. From this example, it is easy to see that, as $H_t \rightarrow X_t$ and $H_{t-1} \rightarrow X_{t-1}$, the growth values converge to simple gain scores, illustrated in Panel C. This category-change growth model is thus equivalent to a coarsening, up to a multiplicative constant, of a model in which gain scores are employed as growth values. Conversely, the gain-score model can be viewed as a refinement of the category-change model.

### Example 2: Category change with arbitrary partitions.

In practice, proficiency category partitions are almost never evenly-spaced or identical between grades; instead, they typically are set by policymakers and experts according to criteria that reflect expectations and goals for achievement. To explore the effect of this unequal spacing on the relationship between gain scores and value tables, I consider a value-table system in which scale scores are partitioned according to the four major Illinois performance level categories in grades 3 and 4 shown in Table 1 and values are assigned according to the category-change metric. A plot of this system is shown in Figure 2.

The growth system in Figure 2 can be modeled by defining $H_t$ according to the cutoff scores for Grade 4 and defining $H_{t-1}$ using the Grade 3 cutoffs, then setting
and so on. Figure 2 highlights the fact that growth-based accountability models serve as a means of assigning different degrees of “weight” or “credit,” depending on student score levels, for proficiency gains that are equal on the vertical score scale. In this example, a student achieving a gain of 60 scale score points would contribute a growth value of 0 to her school if she had scored in the top category (“Exceed Standards”) in grade 3, 1 if she had scored in the “Meet Standards” category, 2 if her grade 3 scale score had been in the upper part of the “Below Standards” category, 1 again if she had scored lower but still “Below Standards,” and so on, as shown by the iso-gain curve corresponding to a gain of 60. This apparently odd relationship between gain scores and growth values is attributable to the unequal widths of proficiency category partition cells in each grade and is not uncommon in operational accountability systems, where both partition locations and the growth values assigned to the cells they create may be set judgmentally or to satisfy policy goals.
Figure 2. Growth values and iso-gain curves from hypothetical category-change growth model based on Illinois proficiency categories in Reading, grades 3 and 4.

Hypothetical growth model assigns 1 point per category improvement and -1 per category decline between grades. Shaded regions indicate growth values, with darkest gray corresponding to -3 and white indicating 3.

A plot of “category-change” growth values for students with Grade 3 scores of $x_{t-1} = 140$ is presented in Figure 3. The black horizontal bars consist of densely-packed points that indicate growth values at each Grade 4 scale score value, with the growth value scale displayed on the left axis. These points are obtained by refining the grade 4 performance-level category partition to the scale score level, then assigning to each point the growth value corresponding to its partition cell. Gain scores are also plotted in Figure 3 as gray points with scale displayed on the right axis; these consist of points along a line with a slope of 1 that passes through the point corresponding to a grade 4 scale score of 140 and gain score of 0.
Figure 3 illustrates the discontinuous nature of growth values with respect to Grade 4 scale scores, contrasted against the linear nature of gain scores. A key point of this illustration is to show that value-table values and gain scores simply represent two alternative mappings of current-grade scores, $X_t$, to growth values for students with any particular prior-grade score (in this case, students with $X_{t-1} = 140$); that is, for students at any fixed $X_{t-1}$, both gain scores and growth values from any value table can be obtained using functions of $X_t$.

These examples show that gain scores can be seen as equivalent to growth values from a value-table system in which performance categories are as small as possible and values are determined by the vertical scale itself. While value tables enable policymakers and experts to set growth values for each $(X_t, X_{t-1})$ pair directly via category cutoffs and
cell values, gain scores represent growth values from a value-table system where category cutoffs and cell values are determined during vertical scale construction.

3.3. Growth values as transformations of scale

The preceding examples highlight the fact that gain scores for students with any fixed prior-grade score can be expressed as a simple linear function of current-grade scores. However, value-table growth values are less mathematically convenient, requiring “lookup values” based on arbitrary category partitions and having a discrete codomain. In order to provide a more unified framework under which variation in factors affecting both gain scores and value tables can be modeled in a similar way, it is necessary to characterize both types of growth models in a more coherent manner.

Below, I show that a) any value table growth function $g(h_t, h_{t-1})$ can be approximated to arbitrary precision by a polynomial function of current- and prior-grade scores $g^e(x_t, x_{t-1})$, and b) for any such approximation, if $\frac{\delta^2 g^e}{\delta x_t \delta x_{t-1}} = 0$, then growth values from any value table can be approximated to arbitrary precision by transforming scale scores and computing gains from the transformed scores.

Although the growth value function for any value table is discrete, so long as the set of obtainable scale scores, $X_t$, is finite, it is always possible to produce a continuous, piecewise-linear function that exactly reproduces the discrete, coarsened growth value values by interpolating linearly between the maximum score in one cell to the minimum in the next-highest cell. An illustration of the piecewise-linear function corresponding to the growth values from Figure 3 is shown below, in Figure 4.
Figure 4. Growth values on category-change metric and polynomial approximation among students with a Grade 3 score of 140.

Solid line indicates growth values that would be obtained under a growth-based accountability model that assigns students 1 point per category change between grades. Dashed curve is a polynomial approximation of growth value function, whose value is within 0.5 of the growth value at all Grade 4 scale scores.

Since scale scores are bounded, such a function can be approximated to arbitrary precision with a polynomial function of sufficient degree (Stone, 1948). The dashed curve in Figure 3 illustrates one such function, with values within 0.5 of the original, discrete growth values at all Grade 4 scale scores.

Extending this logic to the full growth value function, which takes a vector-valued input, it follows that any value table system can be approximated to arbitrary precision by a “smooth” polynomial function. Thus, for any growth-based accountability model \(\{X, H, g(h_t, h_{t-1})\}\), there exists at least one differentiable function \(g^\varepsilon: X \rightarrow \mathbb{R}\) that approximates student-level growth values \(g: H \rightarrow \mathbb{R}\) to within some arbitrary \(\varepsilon > 0\) and can be written directly as a function of scale scores \(X\).
One example of such an approximation is displayed in Panel A of Figure 5. In this figure, growth values are modeled as a continuous function, $g^e(x_t, x_{t-1})$, of scale scores in the current and prior grades. Level sets of $g^e$ are illustrated as shaded regions, with values corresponding to those from Figure 2, revealing that growth values from this function approximately mimic values from the “coarse” model in Figure 2. Hypothetical scale scores drawn from a bivariate normal distribution are also shown as points on this plot.

*Figure 5.* Polynomial approximation of category-change growth metric in scale score and transformed score coordinate spaces, with hypothetical student scale score data.

<table>
<thead>
<tr>
<th>A. Scale score space</th>
<th>B. Transformed score space</th>
</tr>
</thead>
</table>

![Figure 5](image)

Shading indicates growth values, with darkest shade indicating a score of -3 and white a score of 3. Panel A is a polynomial approximation of hypothetical category-change growth model with category cutoffs defined according to Illinois state tests in Reading, grades 3 and 4. Panel B displays Panel A growth values in coordinate system in which scale scores have been transformed such that growth values are equal to the difference between transformed scores.

Choosing some fixed prior-grade score, $c_{t-1} \in X_{t-1}$, results in a function similar to the polynomial in Figure 4, where growth values for students with a prior-grade score of $c_{t-1}$ can be approximated by a nonlinear transformation, $g^e(x_t, c_{t-1})$, of current-grade scores $x_t$. Now, consider some fixed score pair, $(c_t^*, c_{t-1}^*)$. For any growth-based
accountability model with growth value function $g$, if there exists a differentiable approximation $g^\varepsilon$ of $g$ such that $\frac{\delta^2 g^\varepsilon}{\delta x_t \delta x_{t-1}} = 0$, then $g^\varepsilon(x_t, x_{t-1}) = g^\varepsilon(x_t, c^*_t - 1) + C$ for some constant $C$, and $g^\varepsilon(x_t, x_{t-1}) = g^\varepsilon(c^*_t, x_{t-1}) + D$ for some constant $D$. That is, the curve defined by $g^\varepsilon(x_t, x_{t-1})$ at any chosen prior-grade score is parallel to the curve at every other prior-grade score, and similarly, the curve at any current-grade score is parallel to the curves at all other current-grade scores. Thus, $g^\varepsilon$ can be written

$$g^\varepsilon(x_t, x_{t-1}) = g^\varepsilon(x_t, c^*_t - 1) - g^\varepsilon(c^*_t, x_{t-1}).$$

When this condition holds, since $g^\varepsilon(x_t, c^*_t - 1)$ is simply a transformation of $x_t$, and $g^\varepsilon(c^*_t, x_{t-1})$ is a transformation of $x_{t-1}$, growth values can be obtained by applying these transformations to scale scores and differencing the results. In other words, growth values will be equivalent to gain scores on the transformed scale.

Panel B of Figure 5 shows a plot of the same growth value function and with the same student score distribution from Panel A, but in the space of transformed scores, $g^\varepsilon(x_t, c^*_t - 1)$ and $g^\varepsilon(c^*_t, x_{t-1})$. Iso-gain curves from Panel A are also included as a visual reference, with labels removed for clarity. Note that the transformations produce substantial changes in the marginal distributions of scale scores – in particular, the transformations cause scale scores in both grades to be more negatively skewed and slightly reduce the variance of current-grade scores relative to prior-grade scores, while “spreading out” scores near the means in each year. The shaded regions show that level sets of the growth value function have a slope of 1, reflecting the fact that growth values for each student are obtained by subtracting the student’s horizontal axis value from their value on the vertical axis; the growth model in Panel A is therefore identical to a gain-score model based on transformed scores, as illustrated in Panel B.
The equivalence of growth values between the plots in Panels A and B underscores two key points. First, gain scores computed from any nonlinear transformation of scale scores can be seen as equivalent to growth values from an accountability model in which the relationship between gain scores and growth values is nonlinear; hence, “perturbations” in growth values can be modeled by applying nonlinear transformations to scale scores. As can be seen in Figure 4, smaller category widths imply a vertical “stretching” of the relationship between scale scores and growth values. For current-grade scores, decreasing the width of a category cell relative to the one below it increases the rate of increase of growth values for higher scale scores; keeping the widths unchanged while increasing the growth value associated with the higher category has a similar effect. Both perturbations thus imply a convex transformation of the relationship between current-grade scale scores and growth values. Hence, a similar effect could be achieved by keeping growth values and category cells fixed and applying a concave transformation to scale scores.

Second, gain scores from tests where adjacent-grade score distributions are not bivariate normal can be viewed as equivalent to growth values from an accountability model in which scores are bivariate normal, but growth values are assigned such that the relationship between gains and growth is nonlinear. Perturbations in the distribution of scores in which student score rankings are preserved (within and between grades) are thus also equivalent to perturbations in growth values assigned to gain scores, and therefore also can be modeled by transforming scores nonlinearly. This logic underlies the sensitivity analyses presented below.
Not all value-table growth models can be modeled by computing gain scores solely from transformed current- and prior-grade scale scores. In particular, if any two rows in the value table contain scores such that the relationship between growth-value increments and columns differs to a sufficient degree, then any $g^e$ that approximates growth values sufficiently closely must have $\frac{(\delta g^e)^2}{\delta x_t \delta x_{t-1}} \neq 0$, implying that any transformation of current-grade scores must vary with prior-grade scores, and vice versa.

Nevertheless, an appropriate choice of $g^e$ may still provide a useful basis for analysis, especially with regard to the relative magnitude of “bonuses” or “penalties” that such accountability models effectively assign to students at different score levels and the potential incentives that these create for educators. Furthermore, although this study focuses on growth measured between two grades, the framework above may be extended to growth models incorporating scale scores from any number of prior grades, as well as non-score variables, such as student or school characteristics. Thus, nearly any growth system can be modeled and analyzed using a differentiable growth value function.

I leave these as subjects for future research; for the purposes of this paper, I primarily make use of the fact that nonlinear scale transformations are equivalent to modeling both perturbations in growth values and changes in scale score distribution parameters, such as skewness and relative variance.

3.4. Regression residuals

Residuals from an ordinary least-squares (OLS) regression of current-grade scale scores on prior-grade scores constitute perhaps the simplest form of “value-added” model, a class of accountability metrics in which growth values are measured as the difference between observed current-grade scores and an expected score estimated using
prior data. The “value-added” aspect of such models arises from the inference that the deviations between expected and observed score outcomes are caused by school and teacher contributions to student learning. In the current framework, any residual-based accountability model can be modeled by setting $H_t = X_t$ and $H_{t-1} = \hat{X}_{t-1}$, where $\hat{X}_{t-1}$ is the set of predicted current-grade scores and is tantamount to a (possibly covariate-dependent) transformation of at least one prior-grade score, and defining $g$ as a simple gain score function, $g(h_t, h_{t-1}) = h_t - h_{t-1}$.

For the purposes of this study, the key difference between regression-residual and value-table models is that, while the latter are typically static across years, regression models will adjust for year-to-year variation in test outcomes, including (to an extent, depending on the model) secular or idiosyncratic changes in vertical scale score properties. While gain-based value-table accountability models can be seen as addressing the question, “How much did a school’s students grow, in terms of targets determined by policymakers?”, residual-based models answer the question, “How much did a school’s students grow, relative to the observed average for similar students (and, depending on the model, in the average school)?” Put differently, the former measures growth on some predetermined, absolute scale, while the latter seeks to measure growth referenced to some empirically-determined expectation.

I therefore treat residual-based models as a separate class of growth-based accountability model that should be expected to be more robust than gain scores to transformations of scale given the additional parameter(s) in the model. For the purposes of this paper, a simple OLS regression of current-grade scores on a single prior-grade’s
scores suffices to illustrate the potential magnitudes of differences between residual- and gain-based growth models with regard to their sensitivity to scale score transformations.

4. **Evaluating School Growth Rating Sensitivity**

With this framework in mind, the remainder of this paper focuses on the question, “To what extent can growth-based school accountability outcomes be expected to change in response to transformations of the underlying vertical scale?” In light of the results above, this can be seen as equivalent to the question, “How much would school accountability ratings change under different political or judgmental choices of growth values?” I also show that the impact can be compared to the answer to the question, “How would changes in scale score characteristics, such as relative means, variances, and skewness between grades, affect gain-based school accountability results?”

From a practical perspective, the latter two questions may appear somewhat unrelated; growth values and their attendant category partitions are directly set by policymakers, while variation in the parameters of bivariate scale score distributions is largely “empirical,” in the sense that it results from processes outside of policymakers’ direct control (although it is influenced by decisions made during test construction; see Lord, 1980; Kolen, 2006; Briggs and Domingue, 2013). The framework outlined above, however, shows that these two types of variation can be considered equivalent where growth values are concerned.

I first investigate this question using simulated scale scores, by generating scores and computing school-level mean gains and residuals from a simple OLS regression of current-grade scores on prior-grade scores as a baseline. I then subject scale scores to transformations of varying convexity and concavity and re-compute mean gains and
residuals from the transformed scale scores. Finally, I evaluate the magnitude of the effects of the transformation by examining rank correlations and percentile-rank changes between the transformed and untransformed results. The procedure and approach are described in more detail below.

4.1. Simulated Data

For each student $i$ in school $j$, I generate a pair of scores in adjacent grades, $t$ and $t - 1$, according to the data-generating process,

$$
\begin{bmatrix}
  x_{ijt} \\
  x_{ijt-1}
\end{bmatrix} = \begin{bmatrix}
  \mu \\
  0
\end{bmatrix} + \begin{bmatrix}
  \sigma & 0 \\
  0 & 1
\end{bmatrix} \left( \sqrt{\rho_{ICC}} \begin{bmatrix}
  v_{jt} \\
  v_{jt-1}
\end{bmatrix} + \sqrt{1 - \rho_{ICC}} \begin{bmatrix}
  \varepsilon_{it} \\
  \varepsilon_{it-1}
\end{bmatrix} \right),
$$

where

$$
\begin{bmatrix}
  v_{jt} \\
  v_{jt-1}
\end{bmatrix}, \begin{bmatrix}
  \varepsilon_{it} \\
  \varepsilon_{it-1}
\end{bmatrix} \sim N \left( \begin{bmatrix}
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  1 & \rho \\
  \rho & 1
\end{bmatrix} \right).
$$

This results in a bivariate normal scale score distribution,

$$
\begin{bmatrix}
  x_{t} \\
  x_{t-1}
\end{bmatrix} \sim N \left( \begin{bmatrix}
  \mu \\
  0
\end{bmatrix}, \begin{bmatrix}
  \sigma^2 & \rho \sigma \\
  \rho \sigma & 1
\end{bmatrix} \right),
$$

where prior-grade scores $x_{t-1}$ are standard normal, current-grade scores $x_t$ are normally distributed with mean $\mu$ and standard deviation $\sigma$, scores have a between-grade correlation of $\rho$, and the school-level intra-class correlation in each grade is $\rho_{ICC}$. The current-grade mean, $\mu$, is equal to the mean gain between grades expressed in terms of the standard deviation of prior-grade scores, and the current-grade standard deviation, $\sigma$, is expressed as a ratio relative to the prior-grade standard deviation. I set between-grade correlations for school effects and student effects to be equal, with value $\rho$.

For each simulation, I draw 200 schools with 100 students per school, then compute gain scores and residuals from a simple OLS regression of current-grade scores on prior-grade scores. I then transform scores and re-compute gain scores and residuals.
on the transformed scores. I execute 200 simulations for each unique combination of the following parameters:

\[ \mu: \{1.5, 1.0, 1.1\}, \sigma = \{0.9, 0.85, 0.90\}, \rho_{ICC} = \{0.15, 0.25\} \]

These values were chosen to reflect plausible ranges and common values in observed data (Castellano & Ho, 2015; Hedges & Hedberg, 2007) and in the NWEA data below. This results in a total of \((3 \times 3 \times 3 \times 2) = 54\) parameter combinations, for a total of 10,800 simulations.

4.2. Empirical data

To investigate the degree to which theoretical and simulated results extend to operational tests, I also examine the impact of transformations on gain-score and residual growth metrics using NWEA MAP assessment data in Math and Reading from four states, spanning the years 2007 through 2013. Tests are distinguished by their grade span (2-5 or 6-10), administering state, version number (two distinct versions of each test were administered during this time span), and subject. Longitudinal student identifiers enable student-level links between current- and prior-grade scale scores. After removing school-grade-years having fewer than 20 students with valid current- and prior-grade scores on the same test and grade-year-test cells with fewer than 10 schools or 2,000 students, the dataset contains 297 grade-year-test combinations for which gain scores and OLS residuals may be calculated, analogous to 297 “simulations” with varying combinations of scale score parameters.

For each grade-year-test, I standardize student scale scores to their prior-grade mean and standard deviation. Gains can then be understood, as with the simulated data, as effect sizes relative to prior-grade scores and current-grade standard deviations as
ratios, with values less than 1 indicating scale shrinkage and greater than 1 indicating scale expansion. Selected summary statistics for the resulting dataset are presented in Table 2.

Table 3. Summary statistics for NWEA MAP assessment data across 297 unique grade-year-test combinations.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>10th %ile</th>
<th>90th %ile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean gain</td>
<td>0.47</td>
<td>0.10</td>
<td>0.95</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.03</td>
<td>0.95</td>
<td>1.10</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.571</td>
<td>-0.950</td>
<td>-0.203</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.831</td>
<td>0.787</td>
<td>0.886</td>
</tr>
<tr>
<td>School ICC</td>
<td>0.168</td>
<td>0.114</td>
<td>0.232</td>
</tr>
<tr>
<td>Number of schools</td>
<td>182.4</td>
<td>33</td>
<td>435</td>
</tr>
<tr>
<td>Number of students</td>
<td>17,051.7</td>
<td>2,701</td>
<td>38,335</td>
</tr>
</tbody>
</table>

4.3. Scale score transformations

In order to limit the scope of these analyses, I seek transformations that reflect plausible perturbations in growth value values (or, equivalently, plausible variations in score skewness and mean and variance trends) and are easily understood in terms of their effects on growth values at higher and lower levels of the scale score distribution. The first condition requires a monotonic transformation applied to scale scores across both grades, so that student rankings are preserved both within and between grades. To satisfy the second condition, the analyses are limited to transformations that are strictly concave or convex; the former assign increased “weight” lower-scoring students relative to higher-scoring students (i.e., larger magnitudes for the same scale score gains), while the latter achieve the opposite. I therefore follow Reardon and Ho (2015) in transforming scores using the function,

\[ f(x) = -\frac{\text{sgn}(c)}{\sqrt{e^{c^2} - 1}} \left( 1 - e^{c x - \frac{c^2}{2}} \right).\]
The growth value function for each transformation is then $g(x_t, x_{t-1}) = f(x_t) - f(x_{t-1})$, where $x_t$ and $x_{t-1}$ are current- and prior-grade scale scores, respectively, on the original, untransformed scale. This transformation has the convenient property that, if $x$ has a standard normal distribution, then $f(x)$ is log-normally distributed, with a skewness of

$$-\text{sgn}(c)(e^{c^2} + 2)\sqrt{e^{c^2} - 1},$$

but translated and re-scaled to have mean 0 and variance 1. The parameter $c$ then controls the concavity or convexity of the transformation (equivalently, the change in skewness, variance ratio, and current-grade means of the resulting scores), with negative values producing concave and positive producing convex transformations.

I choose values of $c \in \{-0.3143, -0.0830, 0.0830, 0.3143\}$, corresponding to skewness changes of $-1$, $-0.25$, $0.25$, and $1$, respectively, to standard normal data. The negative values are commensurate with the range of skewness values spanned by most of the NWEA data. Thus, for the empirical results, the convex transformations applied to these scores can be seen as partially or fully “unskewing” scores, while concave transformations amplify the skewness.

4.4. Effects of variation in scale score parameters

Simulated data provide an opportunity to isolate the effects of scale score parameters on growth value sensitivity that may be difficult to disentangle in practice. Below, I review a few key parameters that affect the stability of growth values (gain scores and regression residuals) and, in operational tests, vary due to factors that are “empirical,” in the sense that they are determined by factors external to the accountability system.
Mean student gains. The magnitude of the trend in means between adjacent grades, as measured in prior-grade standard deviations, varies considerably across grade levels, tests, and subjects (Dadey & Briggs, 2012). Additionally, mean trends may vary across time within the same test, due to sampling variation or changes to student populations (perhaps influenced by instructional effectiveness).

Referring again to Figure 5, if mean gains were larger in a particular grade or year, all else equal, the observed scale score distribution would be shifted in the positive vertical direction, or equivalently, iso-growth and iso-gain curves would be shifted in the negative vertical direction. In either case, the effect is to increase the discrepancy between simple gain scores and growth values at the student level, so that larger mean gains can also be viewed as assignment of growth values with larger “disagreement” with gain scores at values close to a mean gain of zero. Transformations on scores with larger mean gains will then result in greater changes to school growth rankings.

Regression residuals are designed to accommodate changes in means, and school mean residual rankings would be entirely unaffected by changes in mean gains if scores were transformed linearly. Nonlinear transformations affect the relative rankings of residuals by inducing a nonlinear relationship between growth values and scale scores; the magnitude of a transformation’s effect on the relative ranking of residuals will then be commensurate with the “severity” of the nonlinearity of the growth value function, \( g \), across the range of observed scores, as can be measured by changes in the slope of \( g \) with respect to the iso-gain curves around which scale scores are clustered.

Since iso-gain curves are parallel to the vector \([1, 1] \), the slope of the growth value function along an iso-gain curve can be obtained by evaluating \( \nabla g(x_t, x_{t-1}) \).
\([\sqrt{2}, \sqrt{2}]^\top\) at \((x_{t-1} + y, x_{t-1})\), where \(\nabla g\) denotes the gradient of \(g\) and \(y\) is a gain value. Given two fixed, observed prior-grade scale scores \(a\) and \(b\), the ratio of slopes evaluated at the current-score, prior-score points \((b + y, b)\) and \((a + y, a)\) can be used as an indicator of the nonlinearity of the growth value function for students with a gain of \(y\) (as long as slopes are nonzero at these points). Regression residuals will then be sensitive to changes in mean gains if this ratio depends on \(y\), due to the vertical “shifting” of iso-gain curves.

For the present transformations, this slope ratio is \(e^{c(b-a)}\) at all values of \(y \neq 0\), and thus school rankings from the residual model are insensitive to variation in mean gains; this is done in order to provide a “stable” reference for the gain-based models when mean gains are varied. However, it should be noted that residuals in general will not be stable under changes in mean gains, and the framework above should provide an initial basis for estimating their sensitivity.

**Scale shrinkage or expansion.** Between-grade variance trends, where increasing variance implies “expansion” and decreasing implies “shrinkage” of the scale, can differ noticeably both between tests and within different administrations of the same test. All else equal, changes in variance trends directly affect relative student rankings on gain-score metrics, due to the fact that the expected correlation between prior-grade scores and gains is \(\rho - \frac{\sigma_{t-1}}{\sigma_t}\), where \(\rho\) is the between-grade score correlation, and \(\sigma_t\) and \(\sigma_{t-1}\) are the score standard deviations in the current and prior grades. Sufficient scale expansion leads to \(\frac{\sigma_{t-1}}{\sigma_t} < \rho\) and induces a positive correlation between prior-grade scores and gains, while \(\frac{\sigma_{t-1}}{\sigma_t} > \rho\) leads to a negative correlation, due to mean regression. Similarly, the slope
parameter for an OLS regression of current scores on prior scores is equal to \( \frac{\sigma_r}{\sigma_{t-1}} \rho \), so that changes in variance trends will change student residual rankings. These effects are attributable to the fact that modifying variance trends while holding all else fixed requires modifying between-grade student rankings; for example, increasing the variance of current-grade scores while keeping prior-grade scores fixed requires that students with the same score in both grades on the original scale be assigned lower current-grade scores on the increased-variance scale if they scored in the lower part of the current-grade distribution.

The effects of scale expansion or shrinkage caused by monotonic transformations of scale is less clear, but can be expected to be less severe, due to the preservation of score rankings across grades. For the simulations below, the magnitude of the change in variance trends induced by transformations is controlled by varying the initial standard deviation ratio (parameter \( \sigma \)) on the untransformed scale.

*Within-school score variances.* The dispersion of scale scores within schools can be expected to affect the sensitivity of school gains to transformations of scale, as a larger “spread” of students results in a wider range of distinct growth values contributing to the school mean. When these growth values change relative to each other, the impact of these changes on school mean growth values will affected by within-school score variances.

For the simulations below, this is controlled via two parameters: \( \rho_{ICC} \), the school-level intra-class correlation, and \( \rho \), the between-grade marginal score correlation. When marginal standard deviations are held constant, the former affects within-school score variances in both grades, while the latter affects the spread of scores conditional on prior-
grade scores without affecting the prior-grade score distribution. All else equal, variations in $\rho$ should therefore be expected to have a somewhat stronger impact on transformed-score school rankings than variations in $\rho_{ICC}$.

5. **Results**

I present Spearman rank correlations between school-mean growth values based on transformed vs. original vertical scale scores for both simulated and empirical data in Table 3. Note that, since the transformations are strictly monotonic, student-level scores have a rank correlation of 1 in all comparisons. For the simulated data, I initially limit results to the 200-replication sets whose parameters most closely match the overall means of the NWEA data ($\rho = 0.85$, $\rho_{ICC} = 0.15$, and $\sigma = 1$) and vary only with respect to their mean gains, $\mu \in \{0.15, 0.50, 0.85\}$. I present empirical data that correspond roughly to these simulated means by disaggregating results by mean gain tercile, with means and bounds for each tercile displayed in the leftmost column of the bottom panel. Each row of the bottom panel of Table 3 thus aggregates 99 grade-year-test observations.
Table 4. Spearman rank correlations between school-average growth metrics under monotonic, nonlinear transformations of vertical scale scores.

<table>
<thead>
<tr>
<th>Transformation coefficient:</th>
<th>Gain Scores</th>
<th>OLS Residuals</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-.3143</td>
<td>-.0830</td>
<td>.0830</td>
<td>.3143</td>
<td>-.3143</td>
</tr>
<tr>
<td>Simulated data</td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
<td>(d)</td>
<td>(e)</td>
</tr>
<tr>
<td>Mean gain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.995</td>
</tr>
<tr>
<td>0.150</td>
<td>0.990</td>
<td>0.999</td>
<td>0.999</td>
<td>0.990</td>
<td>0.995</td>
</tr>
<tr>
<td>0.500</td>
<td>0.954</td>
<td>0.996</td>
<td>0.996</td>
<td>0.954</td>
<td>0.995</td>
</tr>
<tr>
<td>0.850</td>
<td>0.895</td>
<td>0.990</td>
<td>0.990</td>
<td>0.893</td>
<td>0.995</td>
</tr>
<tr>
<td>Empirical data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.995</td>
</tr>
<tr>
<td>Mean gain (min-max) within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.974</td>
</tr>
<tr>
<td>tercile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td>0.157 (0.038-0.268)</td>
<td>0.955</td>
<td>0.995</td>
<td>0.995</td>
<td>0.954</td>
<td>0.974</td>
</tr>
<tr>
<td>0.387 (0.269-0.558)</td>
<td>0.910</td>
<td>0.991</td>
<td>0.991</td>
<td>0.910</td>
<td>0.975</td>
</tr>
<tr>
<td>0.851 (0.558-1.731)</td>
<td>0.837</td>
<td>0.984</td>
<td>0.983</td>
<td>0.840</td>
<td>0.979</td>
</tr>
</tbody>
</table>

Negative transformation coefficients imply concave transformations; positive imply convex. Gain scores are school-level averages of student gains. OLS residuals are school-level averages of residuals from OLS regression of current-grade scale scores on prior-grade scale scores. Simulated data statistics are based on 200 simulations of 200 schools each, with 100 students per school-grade, school intra-class correlation set to 0.15, between-grade scale score correlation to 0.85, and equal scale score variances in each grade. Empirical data are based on NWEA MAP assessment scores on 32 distinct assessments from four states, with 99 grade-year-tests per “Mean gain” category.
For gain scores, the decreasing correlations within each column illustrate the magnitudes of the sensitivity of school growth values to perturbations of the score scale as a function of the trend in means between grades. Larger differences in means between grades result in less consistency in school growth rankings. However, this relationship depends heavily on the strength of the transformation: in both the simulated and empirical data, the correlations for the largest mean-gain groups in columns (b) and (c) are just as strong, or stronger, than for the smallest mean-gain groups in columns (a) and (d). Thus, even small mean gains can result in substantial rank changes in response to moderately large skewness changes (or equivalently, growth values that “disagree” with gain scores to a moderate degree). In contrast, regression residuals are somewhat sensitive to scale transformations, primarily due to nonlinearity, but almost entirely insensitive to differences in mean gains, as expected.

Mean absolute changes in school percentile rankings, which provide a more intuitive sense of the impact of transformations, are presented in Table 4 below. In the most extreme cases, at large mean gains and transformation coefficients of -0.3143 and 0.3143, schools move 10 percentile ranks (PRs), on average, after transformation in the simulated data, and about 12 PRs in the NWEA data. While residuals are generally more stable, it should be noted that the magnitudes of their PR changes may be consequential for many schools, approaching or exceeding an average of 4 PRs in the empirical data.
Table 5. Average absolute change in percentile ranks of school growth metrics under monotonic, nonlinear transformations of vertical scale scores.

<table>
<thead>
<tr>
<th>Transformation coefficient:</th>
<th>Gain Scores</th>
<th>OLS Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-.3143</td>
<td>-.3143</td>
</tr>
<tr>
<td></td>
<td>-.0830</td>
<td>-.0830</td>
</tr>
<tr>
<td></td>
<td>.0830</td>
<td>.0830</td>
</tr>
<tr>
<td></td>
<td>.3143</td>
<td>.3143</td>
</tr>
<tr>
<td>Simulated data</td>
<td>(a)</td>
<td>(e)</td>
</tr>
<tr>
<td>Mean gain</td>
<td>(b)</td>
<td>(f)</td>
</tr>
<tr>
<td>0.150</td>
<td>(c)</td>
<td>(g)</td>
</tr>
<tr>
<td>2.92</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>0.500</td>
<td>6.35</td>
<td>1.84</td>
</tr>
<tr>
<td>0.850</td>
<td>9.84</td>
<td>2.91</td>
</tr>
<tr>
<td>Empirical data</td>
<td>(d)</td>
<td>(h)</td>
</tr>
<tr>
<td>Mean gain (min-max) within</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tercile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.157 (0.038-0.268)</td>
<td>5.96</td>
<td>1.79</td>
</tr>
<tr>
<td>0.387 (0.269-0.558)</td>
<td>8.82</td>
<td>2.59</td>
</tr>
<tr>
<td>0.851 (0.558-1.731)</td>
<td>12.19</td>
<td>3.69</td>
</tr>
</tbody>
</table>

Negative transformation coefficients imply concave transformations; positive imply convex. Gain scores are school-level averages of student gains. OLS residuals are school-level averages of residuals from OLS regression of current-grade scale scores on prior-grade scale scores. Simulated data statistics are based on 200 simulations of 200 schools each, with 100 students per school-grade, school intra-class correlation set to 0.25, between-grade scale score correlation to 0.85, and equal scale score variances in each grade. Empirical data are based on NWEA MAP assessment scores on 32 distinct assessments from four states, with 99 grade-year-tests per “Mean gain” category.
Scale shrinkage and expansion on the untransformed scale has a negligible impact on these results. In the most extreme cases, the rank correlations change by less than .003 when \( \sigma \) varies between 0.9 and 1.1, with mean PR changes exhibiting a range of less than 0.2 percentile ranks. While variance trends may theoretically affect school rankings under more extreme transformations, differences in scale score variance trends commensurate with those in observed data have nearly no impact under the monotonic transformations employed here.

As noted above, within-school variances, as simulated by variations in school ICC and between-grade student score correlations, should also affect the sensitivity of gain scores to scale transformations. To investigate the effect of ICC variations, I compare simulation results with \( \rho_{ICC} = 0.15 \) to those with \( \rho_{ICC} = 0.25 \), restricting results to those with \( \mu = 0.85 \) and \( \rho = 0.80 \), and pooling results across all three values of \( \sigma \) in order to maximize the differences. The results, displayed in Table 5, show that while ICC differences do affect results, the impacts are generally small.

<table>
<thead>
<tr>
<th>ICC</th>
<th>Transformation coefficient (c)</th>
<th>-.3143</th>
<th>-.0830</th>
<th>.0830</th>
<th>.3143</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>8.71</td>
<td>2.55</td>
<td>2.54</td>
<td>8.69</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>8.56</td>
<td>2.52</td>
<td>2.51</td>
<td>8.56</td>
<td></td>
</tr>
</tbody>
</table>

Negative transformation coefficients imply concave transformations; positive imply convex. Statistics based on 600 simulations of 200 schools each, with 100 students per school-grade, between-grade score correlations set to 0.80, mean gain set to 0.85, and current- to prior-grade standard deviation ratios ranging from 0.9 to 1.1.
Between-grade score correlations, on the other hand, have a more substantial impact on rank stability. Figure 6 depicts the relationship between scale score correlations and mean absolute changes in school percentile ranks for the transformation with $c = .3143$, disaggregated by mean gain levels. For gain scores, stronger between-grade scale score correlations produce larger changes in PR, with the magnitude of the effect dependent on mean gain levels. Correlations impact OLS regression residuals, as well, since stronger correlations magnify the “stretching” of school-level means among high-scoring schools and increase the slope of the regression line; however, the effects are small.

**Figure 6.** Effect of correlation changes on mean absolute percentile rank changes in growth values after nonlinear, monotonic scale score transformation ($c = .3143$).

Each bar displays result from 200 simulations of 200 schools with 100 students per school-grade. Untransformed scores have a school-level intra-class correlation of .15 and standard deviation of 1 in both current and prior grades. Untransformed mean gains among students displayed on horizontal axis.

Isolating the effect of between-grade scale score correlations conditional on mean gains in the NWEA data is difficult, due to the sparseness of the data and the inverse
relationship (across grade-years) between mean gains and score correlations; the former leads to insufficient mean-by-correlation “cells” for presenting disaggregated means, while the latter causes confounding between the effects of mean gains and correlations. To provide a rough estimate of the empirical effect of correlation changes conditional on mean gains, I fit linear regression models to the set of 297 grade-year-test combinations available in the dataset of the form,

\[ \Delta |PR|_i = \beta_0 + \beta_1(GAIN_i) + \beta_2(CORR_i) + \beta_3(CORR_i \times GAIN_i) + \epsilon_i, \]

where \( \Delta |PR|_i \) is the mean absolute change in percentile ranks among schools in grade-year-test \( i \) after transformation, \( GAIN_i \) is the mean student-level gain on the untransformed score scale, \( CORR_i \) is the between-grade student score correlation, and \( \epsilon_i \) is an error term. I fit one model per transformation (for a total of four models), with heteroskedasticity-adjusted standard errors, and use the resulting coefficients to estimate the effect of an increase in score correlation from .80 to .90, conditional on mean gains of .15, .50, and .85.

The results are presented in Figure 7, along with simulated results. Considering the limitations of the data and regression model, empirical results are remarkably similar to simulated results, with the effects of a .10 correlation increase growing larger with both mean gains and transformation convexity/concavity. In the NWEA data, the larger effect for \( c = -0.3143 \) compared to \( c = 0.3143 \) is attributable to the fact that untransformed scores tend to have negative skew, causing the former transformation to have an exaggerated effect, due to its exponential relationship with skewness values. While larger mean gains lead to greater sensitivity of gain scores to transformations, the
effect is moderated to a non-trivial degree by the strength of between-grade score correlations.

**Figure 7.** Simulated and estimated empirical relationship between change in absolute percentile ranks of school gain-based growth values caused by scale score transformations and scale score correlation increase from .80 to .90, conditional on mean student gains on untransformed scale.

**A. Simulated data**

Simulated results based on 200 simulations of 200 schools each, with 100 students per school, school intra-class correlation set to .15 and equal untransformed scale score standard deviations between grades. Empirical data estimated based on linear regression of absolute percentile rank change on (untransformed) mean gains, between-grade scale score correlations, and their interaction using 297 grade-year-test observations. Coefficient c parameterizes transformations, with positive values producing convex and negative producing concave transformations.

**B. Empirical data**

**6. Summary and Conclusion**

This paper has presented a basic framework for analysis of growth-based accountability rating systems that leads to a number of findings. First, any growth rating system based on value tables may be approximated to arbitrary precision by a polynomial function of current- and prior-grade scores. This suggests opportunities for future mathematical and empirical analysis of the effects of operational growth models on different populations of schools and students, such as unforeseen or unintended
discrepancies between growth value “weightings” and gains for target populations, or identification of perverse incentives certain models may produce.

Second, certain classes of growth-based accountability systems can be modeled as simple gain-score models applied to transformed scale scores, providing a unified view of growth metrics in which values assigned to gains, many aspects of variation in empirical score distributions, and transformations of the score scale can be considered equivalent. Given that the first of these factors is entirely controlled by policymakers, while the second results primarily from “nature,” their equivalence under this framework implies that growth values may be used to compensate for, or amplify, the effects of unusual or undesirable variation in scale score distributions.

Third, regression-based growth models, including most value-added models, can also be modeled in this framework, but are distinct from gain- or value-based models in that their “transformations” change from year to year in response to changes in observed data. As a consequence, they are generally more robust to transformations that preserve student rankings between grades.

Additionally, this study has shown that, while school growth ratings based on gains can be quite sensitive to transformations of scale, this sensitivity varies considerably depending on properties of the underlying score scale. Specifically, shifts in the mean gains obtained by students can have a dramatic effect on school rankings, and variation in between-grade score correlations may magnify or attenuate these effects. On the other hand, differences in scale shrinkage or expansion have a negligible effect on transformation sensitivity, and differences in school intra-class correlations have a fairly small impact at typical values. Regression residuals are generally far more robust to
scale transformations, since by design they adjust to perturbations in scale score distributions; however, the magnitude of changes in regression-based rankings may still be consequential if transformations are sufficiently nonlinear and the model fails to correct for this.

Where policy is concerned, this study highlights the need for value-table systems to choose both growth values and category cutoffs carefully, with a particular eye toward the degree of relative “credit” or “weight” assigned to students with the same gains but different prior-grade scores. Additionally, the sensitivity of school growth results to even minor fluctuations in values or score distributions suggests that values and category cutoffs be examined regularly, in light of empirical results, to ensure that they continue to reflect reasonable policy demands and are free from “pathological” outcomes.
V. Conclusion

Federal policy developments in recent years have provided states with an unprecedented degree of flexibility in how they evaluate school performance. As a result, test-based accountability metrics are more varied and complex than ever. While many such metrics may seem intuitively sensible and meaningful, their use of test scores as inputs can lead to systematic variation that is unrelated to actual school effectiveness. This collection of papers has investigated three examples of this phenomenon that appear in operational tests. Because many more are likely to exist, policymakers would be well-advised to examine new accountability metrics more closely, to ensure that such measures are both useful and defensible.
References


California Department of Education (2017a). *CAASP Online Test Administration Manual, Chapter 6: Administering the Summative Assessments to*


121


