Discreteness Causes Bias in Percentage-Based Comparisons: A Case Study From Educational Testing

Citation

Yee, Darrick S. and Andrew D. Ho. 2015. Discreteness causes bias in percentage-based comparisons: A case study from educational testing. *American Statistician*, 69: 174-181.

Published Version

doi:10.1080/00031305.2015.1031828

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:27471534>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Discreteness causes bias in percentage-based comparisons:

A case study from educational testing

Darrick Yee and Andrew Ho

Harvard Graduate School of Education

January 22, 2015

Discretizing continuous distributions can lead to bias in parameter estimates. We present a case study from educational testing that illustrates dramatic consequences of discreteness when discretizing partitions differ across distributions. The percentage of test-takers who score above a certain cutoff score (percent above cutoff, or “*PAC*”) often describes overall performance on a test. Year-over-year changes in *PAC*, or ΔPAC , have gained prominence under recent U.S. education policies, with public schools facing sanctions if they fail to meet *PAC* targets. In this paper, we describe how test score distributions act as continuous distributions that are discretized inconsistently over time. We show that this can propagate considerable bias to *PAC* trends, where positive ΔPAC s appear negative, and vice versa, for a substantial number of actual tests. A simple model shows that this bias applies to any comparison of *PAC* statistics in which values for one distribution are discretized differently from values for the other.

Keywords: Education, Estimation, Testing

Darrick S. Yee is a doctoral candidate, Harvard Graduate School of Education, Cambridge, MA 02138 (e-mail: dsy783@mail.harvard.edu). Andrew D. Ho is Professor of Education, 455 Gutman Library, 6 Appian Way, Cambridge, MA 02138 (e-mail: andrew_ho@harvard.edu). This research was supported in part by a grant from the Institute of Education Sciences (R305D110018). The authors thank Sean Reardon, Judith Singer, and Richard Murnane for their helpful feedback. We claim responsibility for any errors.

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*).

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 ΔPAC , 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 ΔPAC metric that, to our knowledge, has never been addressed formally. Previous authors have observed that the relationship between ΔPAC 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 ΔPAC 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.

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 ΔPAC 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 *PAC*-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 PAC_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 ΔPAC , is then

$$\begin{aligned}\Delta PAC &= PAC_1 - PAC_0 = (1 - F_1(k^*)) - (1 - F_0(k^*)) \\ &= F_0(k^*) - F_1(k^*)\end{aligned}\tag{1}$$

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

$$\begin{aligned}\Delta PAC^{DISC} &= PAC_1^{DISC} - PAC_0^{DISC} = (1 - G_1(k^*)) - (1 - G_0(k^*)) \\ &= (1 - F_1(h_1^*)) - (1 - F_0(h_0^*)) \\ &= F_0(h_0^*) - F_1(h_1^*)\end{aligned}\tag{2}$$

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 ΔPAC 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 ΔPAC of 0, but the ΔPAC^{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).

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 ΔPAC^{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 ΔPAC^{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 ΔPAC^{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 ΔPAC^{DISC} at 400 and 401. Figure 1 shows that the observed ΔPAC^{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 ΔPAC Estimator*

We formalize the results above to show that, for any sample of students, the observed ΔPAC^{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

$$\widehat{PAC}_i^{DISC} = \frac{1}{N_i} \sum_{j=1}^{N_i} I(X_{ij} > h_i^*), \quad (3)$$

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

$\widehat{PAC}_i^{DISC}$ is therefore binomially distributed with count parameter N_i and probability parameter $1 -$

$F_i(h_i^*)$. Using $\hat{\theta} = \widehat{PAC}_1^{DISC} - \widehat{PAC}_0^{DISC}$ as an estimator for θ produces a bias of

$$\begin{aligned} E(\hat{\theta} - \theta) &= E(\widehat{PAC}_1^{DISC} - \widehat{PAC}_0^{DISC}) - (F_0(k^*) - F_1(k^*)) \\ &= (F_0(h_0^*) - F_1(h_1^*)) - (F_0(k^*) - F_1(k^*)) \\ &= \left(\int_{-\infty}^{h_0^*} f_0(t) dt - \int_{-\infty}^{h_1^*} f_1(t) dt \right) - \left(\int_{-\infty}^{k^*} f_0(t) dt - \int_{-\infty}^{k^*} f_1(t) dt \right) \end{aligned}$$

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

Similarly, we can show that $\hat{\theta}$ is an inconsistent estimator of θ . 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 σ^2 is a constant that does not depend on N . Recall that the numerator of $\widehat{PAC}_i^{DISC}$ is binomially distributed; then, by a straightforward central-limit argument, $\widehat{PAC}_i^{DISC}$ is asymptotically normally distributed with mean $1 - F_i(h_i^*)$. Thus, $\hat{\theta} = \widehat{PAC}_1^{DISC} - \widehat{PAC}_0^{DISC}$ is asymptotically normal with mean $\nu + \theta$. For large N ,

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

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

$$\begin{aligned} P(|\hat{\theta} - \theta| \geq \varepsilon) &= P(\hat{\theta} - \theta \geq \varepsilon) + P(\hat{\theta} - \theta \leq -\varepsilon) \\ &= P\left(\frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta - \nu) \geq \frac{\sqrt{N}}{\sigma}(\varepsilon - \nu)\right) + P\left(\frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta - \nu) \leq -\frac{\sqrt{N}}{\sigma}(\varepsilon + \nu)\right) \\ &\approx 1 - \Phi\left(\frac{\sqrt{N}}{\sigma}(\varepsilon - \nu)\right) + \Phi\left(-\frac{\sqrt{N}}{\sigma}(\varepsilon + \nu)\right). \end{aligned}$$

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

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

Thus, $\hat{\theta}$ is an inconsistent estimator of θ . This latter result implies that, in practice, ΔPAC 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 ΔPAC 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 ΔPAC estimator as ΔPAC^{DISC} .

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

$$\begin{aligned}\nu &= \int_{k^*}^{h_0^*} f_0(t)dt - \int_{k^*}^{h_1^*} f_1(t)dt \\ &= \int_{k^*}^{h_0^*} f_0(t)dt - \int_{k^*}^{h_0^*} f_1(t)dt - \int_{h_0^*}^{h_1^*} f_1(t)dt \\ &= \int_{k^*}^{h_0^*} (f_0(t) - f_1(t))dt + \int_{h_1^*}^{h_0^*} f_1(t)dt\end{aligned}\tag{5}$$

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 ΔPAC 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 ΔPAC 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 ΔPAC 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 ΔPAC^{DISC} at each cutoff score k^* is then equal to the continuous ΔPAC 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 ΔPAC evaluated at a score of 4.5. In effect, ΔPAC^{DISC} gives the continuous ΔPAC 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 ΔPAC^{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 ΔPAC^{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.

Figure 2. Change in percent above cutoff score (ΔPAC) under two discretization scenarios.

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 CAUSED BY DISCRETIZATION

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 ΔPAC and the empirical ΔPAC^{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.¹ 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

¹ 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.

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 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 ΔPAC using the official state Proficient cutoff score, k^* . We then find the lowest discrete score exceeding k^* (across both years), compute the ΔPAC at this score, and subtract the original ΔPAC at k^* . The result is the change in the ΔPAC 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.

<i>Table 1.</i> Summary of changes in year-over-year increases in percent-Proficient (ΔPAC) 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 (ΔPAC), and subtracting the observed ΔPAC at the original Proficient cutoff score from the result. A “sign change” indicates that the “incremented” ΔPAC values include at least one value whose sign is the opposite of the Proficient ΔPAC . <i>Observed</i> values are those actually reported by states. <i>Smoothed</i> 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 ΔPAC was -1.36 percentage points. The lowest discrete score exceeding 400 was 401 in 2010 and 404 in 2011; therefore, we compute the observed ΔPAC at 401, which we find to be -1.16 percentage points. 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 ΔPAC of 0.2 percentage

points. Meanwhile, the highest discrete score below 400 in either year was 396, at which the observed ΔPAC was 2.11 percentage points. Decrementing the cutoff score by 1 discrete score thus increases the ΔPAC by 3.5 percentage points. For this test, the maximum change in ΔPAC 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 ΔPAC was negative, while the ΔPAC at 396 was positive, the ΔPAC 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 ΔPAC 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 ΔPAC 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 ΔPAC 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 ΔPAC 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 ΔPAC 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.

Figure 3. Estimated bias in reported change in percent-Proficient (ΔPAC) from 2010 to 2011 attributable to changes in score discretizations for NCLB tests in 13 states.

5. DISCUSSION

We have shown above that bias in the observed ΔPAC 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 ΔPAC 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 ΔPAC from one year to the next. The observed PAC in each year, given in (3), follows a scaled binomial distribution, and thus the observed ΔPAC has standard error $\sqrt{\frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1}}$, where $\sigma_i^2 = F_i(h_i^*)(1 - F_i(h_i^*))$ 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 ΔPAC 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 ΔPAC . Using finer-grained partitions will reduce the bias caused by misalignment of discrete score cells, all else equal. States wishing to minimize ΔPAC 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 ΔPAC reporting. As a last resort, states or researchers could employ “smoothing” to estimate continuous ΔPAC 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 ΔPAC 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 ΔPAC 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.

References

- American Institute of CPAs (2013). "CPA Examination Passing Rates," retrieved from <http://www.aicpa.org/BECOMEACPA/CPAEXAM/PSYCHOMETRICSANDSCORING/PASSINGRATES/Pages/default.aspx>
- Bandeira de Mello, V. (2011), *Mapping State Proficiency Standards onto the NAEP Scales: Variation and Change in State Standards for English and Mathematics, 2005–2009* (NCES 2011-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC: Government Printing Office.
- Braun, H. & Mislevy, R. (2005), "Intuitive test theory," *Phi Delta Kappan*, 86, 489-497.
- Bryant, M. J., Hammond, K. A., Bocain, K. M., Rettig, M. F., Miller, C. A., & Cardullo, R. A. (2008), "School performance will fail to meet legislated benchmarks," *Science*, 321, 1781-1782.
- College Board, The (2013), *The 9th Annual AP Report to the Nation*, New York, NY: The College Board.
- Dempster, A. P. and Rubin, D.B. (1983), "Rounding error in regression: The appropriateness of Sheppard's corrections," *Journal of the Royal Statistical Society, Series B (Methodological)*, 45(1), 51-59.
- Flanagan, J. C. (1951), "Units, scores, and norms," in E. F. Lindquist (ed.), *Educational Measurement*, Washington, DC: American Council on Education, pp. 695-763.
- Fulton, M. (2006), "Minimum subgroup size for Adequate Yearly Progress (AYP): State trends and highlights," Denver, CO: Education Commission for the States.
- Glass, G. V., McGaw, B., and Smith, M. L. (1981), *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage.
- Hedges, L. V., and Hedberg, E. C. (2007), "Intraclass correlation values for planning group-randomized trials in education," *Educational Evaluation and Policy Analysis*, 29(1), 60-87.
- Heijtan, D. F. (1989), "Inference from grouped continuous data: A review," *Statistical Science*, 4(2), 164-179.
- Heijtan, D. F. and Rubin, D.B. (1991), "Ignorability and coarse data," *Annals of Statistics*, 19(4), 2244-2253.
- Ho, A.D. (2008), "The problem with 'proficiency': Limitations of statistics and policy under No Child Left Behind," *Educational Researcher*, 37(6), 351-360.
- Holland, P. (2002), "Two measures of change in the gaps between the CDFs of test-score distributions," *Journal of Educational and Behavioral Statistics*, 34, 201-228.
- Holland, P. W., and Dorans, N. J. (2006), "Linking and equating," in R. Brennan (ed.), *Educational Measurement* (4th ed.), Westport, CT: American Council on Education / Praeger Publishers, pp. 187-220.
- Horton, N.J., Lipsitz, S.R., and Parzen, M. (2003), "A potential for bias when rounding in multiple imputation," *The American Statistician*, 57(4), 229-232.
- Kolen, M.J., and Brennan, R.L. (2004)., *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.), New York: Springer-Verlag.
- Lord, F. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, New Jersey: Erlbaum.

- McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., and Miles, J. A. (2013), "Evidence-based standard setting: Establishing a validity framework for cutoff scores," *Educational Researcher*, 42, 78-88.
- National Conference of Bar Examiners (2012), "2011 statistics," *The Bar Examiner*, 81(1), 6-41.
- Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: University of Chicago Press (1981).
- Schneeweiss, H., Komlos, J., and Ahmad, A.S. (2010), "Symmetric and asymmetric rounding: A review and some new results," *Advances in Statistical Analysis*, 94(3), 247-271.
- Sheppard, W.F. (1897), "On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale," *Proceedings of the London Mathematical Society*, 1(1), 353-380.
- U.S. Department of Education (n.d.), *NAEP Data Explorer*, Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences.
- U.S. Department of Education (2012), *ESEA Flexibility*. Available at <http://www.ed.gov/esea/flexibility/documents/esea-flexibility-acc.doc>
- Wallis, W., and Steptoe, S. (2007), "How to fix No Child Left Behind," *Newsweek*. 169(23), 34-41.
- Yen, W. M., and Fitzpatrick, A. R. (2006), "Item response theory," in R. Brennan (ed.), *Educational Measurement* (4th ed.), Westport, CT: American Council on Education / Praeger Publishers, pp. 111-154.