Estimating Achievement Gaps from Test Scores Reported in Ordinal "Proficiency" Categories

Andrew D. Ho

Harvard Graduate School of Education

Sean F. Reardon

Stanford University

Author Note

This paper was supported in part by a grant from the Institute of Education Sciences (#R305A070377) and a fellowship from the National Academy of Education and the Spencer Foundation. The authors benefited from the research assistance of Demetra Kalogrides and Erica Greenberg of Stanford University and Katherine Furgol of the University of Iowa.

Abstract

Test scores are commonly reported in a small number of ordered categories. These contexts include state accountability testing, Advanced Placement tests, and English proficiency tests. This paper introduces and evaluates methods for estimating achievement gaps on a familiar standard-deviation-unit metric using data from these ordered categories alone. These methods hold two practical advantages over alternative achievement gap metrics. First, they require only categorical proficiency data, which are often available where means and standard deviations are not. Second, they result in gap estimates that are invariant to score scale transformations, providing a stronger basis for achievement gap comparisons over time and across jurisdictions. We find three candidate estimation methods that recover full-distribution gap estimates well when only censored data are available.

Keywords: achievement gaps, proficiency, nonparametric statistics, ordinal statistics

Estimating Achievement Gaps from Test Scores Reported in Ordinal "Proficiency" Categories

Achievement gaps are among the most visible large-scale educational statistics. Closing achievement gaps among traditionally advantaged and disadvantaged groups is an explicit goal of state and federal education policies, including current and proposed authorizations of the Elementary and Secondary Education Act (Miller & McKeon, 2007). Gaps and gap trends are thus a commonplace topic of national and state report cards, newspaper articles, scholarly articles, and major research reports (e.g., Education Week, 2010; Jencks & Phillips, 1998; Magnuson & Waldfogel, 2008; Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009).

Researchers selecting an achievement gap metric face three issues. First, average-based gaps—effect sizes or simple differences in averages—are variable under plausible transformations of the test score scale (Ho, 2007; Reardon, 2008a; Seltzer, Frank, & Bryk, 1994; Spencer, 1983). Second, gaps based on percentages above a cut score, such as differences in "proficiency" or passing rates, vary substantially under alternative cut scores (Ho, 2008; Holland, 2002). Third, researchers often face a practical challenge: Although they may wish to use an average-based gap metric, the necessary data may be unavailable.

This last situation has become common even as the reporting requirements of the No Child Left Behind Act (NCLB) have led to large amounts of easily accessible test score data. The emphasis of NCLB on measuring proficiency rates over average achievement has led states and districts to report "censored data": test score results in terms of categorical achievement levels, typically given labels like "below basic," "basic," "proficient," and "advanced." These censored data are often reported in lieu of traditional distributional statistics like means and standard deviations. A recent Center on Education Policy (2007) report noted that state-level black and white means and standard deviations required for estimating black-white achievement gaps were available in only 24 states for reading and 25 states for mathematics. Moreover, many

of these states only made these statistics available upon formal request. Without access to basic distributional statistics, much less full distributional information, research linking changes in policies and practices to changes in achievement gaps becomes substantially compromised in the absence of alternative methodological approaches.

This paper develops and evaluates a set of methods for estimating achievement gaps when standard distributional statistics are unavailable. The first half of this paper reviews traditional gap measures and their shortcomings and then presents alternative gap measures in an ordinal, or nonparametric framework. Links to a large literature in nonparametric statistics and signal detection theory are emphasized. This nonparametric approach generally assumes full information about the test score distributions of both groups. The second half of the paper introduces and evaluates methods for estimating achievement gaps using censored, unanchored data. This describes most readily available state testing data under NCLB, where only a small number of categories are defined, and the cut scores delineating categories are either unknown or not locatable on an interval scale. The contribution of the paper is a toolbox of transformation-invariant gap estimation methods that either overcomes or circumvents the aforementioned theoretical and practical challenges: transformation-dependence, cut-score-dependence, and the scarcity of standard distributional statistics.

Traditional Achievement Gap Measures and Their Shortcomings

A test score gap is a comparative statistic describing the difference between two distributions. Typically, the target of inference is the difference between central tendencies. Three "traditional" gap metrics dominate this practice of gap reporting. The first is the test's score scale, where gaps are most often expressed as a difference in group averages. For a student

test score, X, a typically higher scoring reference group, a, and a typically lower scoring focal group, b, the difference in averages, d^{avg}, follows:

$$d^{avg} = \bar{X}_a - \bar{X}_b. \quad (1)$$

The second traditional metric expresses the gap in terms of standard deviation units. This metric allows for standardized interpretations when the test score scale is not well established and affords aggregation and comparison across tests with differing score scales (Hedges & Olkin, 1985). Sometimes described as Cohen's d, this effect size expresses d^{avg} in terms of a quadratic average of both groups' standard deviations, s_a and s_b, as follows:

$$d^{coh} = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2 + s_b^2}{2}}}. \quad (2)$$

The third traditional metric, the percentage-above-cut (PAC) metric, has become particularly widespread under NCLB, which mandates state selection of cut scores delineating "proficiency." Schools with insufficient percentages of proficient students face the threat of sanctions. The relevance of the cut score and the mandated reporting of disaggregated proficiency percentages lead to a readily available gap statistic: the difference in percentages of proficient students. If PAC_a and PAC_b are the percentages of groups a and b above a given cut score, the PAC-based gap is

$$d^{pac} = PAC_a - PAC_b. \quad (3)$$

The PAC-based gap in Equation 3 is known to be dependent upon the location of the cut score (Ho, 2008; Holland, 2002). If the two distributions are normal and share a common variance, however, this cut-score dependence can be eliminated by a transformation of PACs onto the standard-deviation-unit metric using an inverse normal transformation. The resulting

gap estimate, denoted d^{tpac}, for the "difference in transformed percentages-above-cut" (DTPAC), is written:

$$d^{tpac} = \Phi^{-1}(PAC_a) - \Phi^{-1}(PAC_b). \quad (4)$$

This method implicitly assumes the test scores in groups a and b are both normally distributed with equal variance. The resulting gap can be interpreted in terms of standard deviation units. If the distributions meet this strict normal, equal-variance assumption, Equation 4 returns the same effect size regardless of cut-score location. Moreover, this common effect size will be equal to Cohen's d. More formal demonstrations of the logic of this transformation are widespread (e.g., Hedges and Olkin, 1985; Ho, 2009).

Table 1 displays these four gap measures as they describe the White-Black fourth-grade reading test score gap in Alabama.[1] Data are from the National Assessment of Educational Progress (NAEP) in 2005 and 2007. Gap trends are also shown as the simple gap change from 2005 to 2007. Row 1 of Table 1 shows that Alabama's White-Black gap decreased 6.41 scale score points: from 32.0 to 25.6 between 2005 and 2007. Row 2 of Table 1 shows that the d^{coh}-based gap for Alabama decreased from 0.88 standard deviation units in 2005 to 0.74 standard deviation units in 2007, a decrease of 0.14 standard deviation units. Rows 3-5 show d^{pac}-based gaps for Alabama on the three NAEP cut scores. The NAEP cut scores are the lower borders of three NAEP achievement levels: Basic, Proficient, and Advanced. The gaps in the percentages above these cut scores are denoted in Rows 3, 4, and 5 as d_1^{pac}, d_2^{pac}, and d_3^{pac} respectively.

The differences between gaps across the three cut scores are stark. Gaps appear to be much larger at Basic and Proficient cut scores than at the Advanced cut score. Gaps decrease

[1] We chose Alabama for this illustrative example for two reasons. First, Alabama was one of only 7 out of a possible 200 state-subject-grade combinations with statistically significant NAEP White-Black gap trends between 2005 and 2007 (Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009). Second, the White-Black gap in Alabama in 2007 was similar in magnitude to that of the U.S. as a whole.

5.52 percentage points at the Basic cut score but increase 1.45 and 1.34 percentage points at the Proficient and Advanced cut scores. These seeming contradictions cloud the overall finding of "a statistically significant gap closure" and encourage inferences about interactions between gap trends and particular regions of the score distributions. Finally, the DTPAC approach shown in Rows 6-8 yields results that differ substantially across cut scores, revealing violations of the normal, equal variance assumption in practice. For example, the DTPAC-based gap trend at the "Advanced" cut score shows twice the decline of the gap trend at other cut scores. Note that the relatively large gap trend for Alabama over other states makes sign-reversal—the most dramatic dependence—less likely across metrics. Alabama thus provides a somewhat conservative example of the degree to which interpretations may change under gap metric selection.

Each of the four metrics compared in Table 1 has shortcomings. The first two average-based metrics depend on the assumption that the test score scale has equal-interval properties (Reardon, 2008b). If equal-interval differences do not share the same meaning throughout all levels of the test score distribution, nonlinear transformations become permissible, and distortions of averages and Cohen-type effect sizes will result. In educational measurement, arguments for strict equal interval properties are difficult to support (Kolen & Brennan, 2004; Lord, 1980; Spencer, 1983; Zwick, 1992). And the magnitude of differences under plausible scale transformations can be of practical significance. Ho (2007) has shown that d^{coh} can vary by more than 0.10 from baseline values under plausible monotone transformations. Further, estimates of d^{coh} based on different reported test score metrics of the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K) reveal cross-metric gap trend differences as large as 0.10 (author calculations from Pollack, Narajian, Rock, Atkins-Burnett, & Hausken, 2005). This range is sufficient to call many gap comparisons and gap trends into question.

Gap inferences based on PAC-based metrics are subject to a different kind of distortion. Holland (2002) demonstrates that gap magnitude is confounded by the location of the cut score. With two normal distributions, for example, PAC-based gaps are maximized when the cut score is at the midpoint between the modes of the two distributions. In Table 1, this explains the greater magnitudes of the gaps in Row 3 over Rows 4 and 5, as the NAEP Basic cut score is generally more central than the higher Proficient and extremely high Advanced cut scores. As a corollary, when the scores of both groups are increasing, gap trends based on low cut scores will show gap closure, and gap trends based on high cut scores will show gap increases. Although it is tempting to interpret the gap trends in Rows 3, 4, and 5 as a story about gap trends increasing for high-scoring students but decreasing for low-scoring students, there is an alternative interpretation: proportionately more white students than black students crossed the Advanced NAEP cut score simply because more of them were near the cut score initially.

The inverse normal transformation of the percentages-above-cuts helps to address the predictable relationship between estimated gaps and the locations of cut scores, but it rests on the assumption of normality and equal variance, which is rarely satisfied in practice (Ho, 2008, 2009). As Table 1 illustrates, when this assumption is not met, estimated gaps and gap trends may be highly sensitive to the location of the cut score used to estimate them.

Taken together, these shortcomings raise serious concerns about the gap and gap trend metrics in Table 1. The first two rows assume not only equal-interval scale properties but also, for gap trends, the maintenance of equal-interval properties over time. The gaps and gap trends in the next three rows confound the movement of score distributions with the density of students adjacent to the cut score. Unadjusted d^{pac} and Δd^{pac} statistics will either appear contradictory across cut scores or encourage inaccurate inferences about gaps and gap trends at different

regions of the distribution. The gaps and trends in the last three rows illustrate the consequences of a violation of the strict normal, equal variance distributional assumption. These shortcomings motivate an alternative approach to achievement gap reporting.

An Ordinal Framework for Gap Trend Reporting

The literature on ordinal distributional comparisons contains attractive alternatives to traditional gap metrics. When the scale-dependence of gap statistics is a concern, gaps can be derived from transformation-invariant representations like the probability-probability (PP) plot (Ho, 2009; Livingston, 2006; Wilk & Gnanadesikan, 1968). The PP plot is best described by considering the two Cumulative Distribution Functions (CDFs), $F_a(x)$ and $F_b(x)$, which return the proportions of students (p_a and p_b) at or below a given score x in groups a and b, respectively. The left panel of Figure 1 shows the normal CDFs of Alabama's White and Black test score distributions on the 2005 Grade 4 NAEP Reading test as an example. These are labeled generically as a lower-scoring focal distribution, F_b, and the higher scoring reference distribution, F_a. The vertical axis expresses the proportion at or below a given NAEP scale score x. The left panel of Figure 1 shows that, for the basic cut score of 208, 33% of the reference group is at or below Basic, whereas 69% of the focal group is at or below Basic.

Due to the construction of the PP plot from paired cumulative proportions, all statistics generated from a PP plot are transformation-invariant. One useful statistic is the area under the PP curve, which is equal to $\Pr(X_a > X_b)$, denoted $P_{a>b}$ for short: the probability that a randomly drawn group *a* student has a greater score than a randomly drawn group *b* student. This statistic has a substantial background in the nonparametric and ordinal statistics literature (e.g., Cliff, 1993; McGraw & Wong, 1992; Vargha & Delaney, 2000). In signal detection theory and medical testing, a mathematically equivalent expression is known as the Area Under the Curve

(AUC) of the Receiver Operating Characteristic (ROC) Curve, where the ROC curve is a PP plot with a particular interpretation. In this literature, the two distributions are usually those of healthy and sick populations along some test criterion, and the interpretation of AUC is as a summary measure of the diagnostic capability of the criterion (Swets & Pickett, 1982). The use of these approaches for expressing achievement gaps in education is fairly limited (exceptions include Livingston, 2006; Neal, 2006; Reardon, 2008b).

Although the interpretation of $P_{a>b}$ may be appealing, a Cohen-like effect size that avoids the proportion metric allows for better equal-interval properties and interpretation in terms of standard deviation units. For this purpose, Ho and Haertel (2006) and Ho (2009) propose the V statistic, a nonlinear monotonic transformation of $P_{a>b}$:

$$V = \sqrt{2}\Phi^{-1}(P_{a>b}). \tag{5}$$

The V statistic has several useful properties. First, V is equal to the Cohen effect size when the two test score distributions are normal, even if they have unequal variances.[2] However, even in these circumstances, Cohen's d will vary under scale transformations whereas V will not. The implicit condition equating V to d^{coh} is *respective normality* (Ho, 2009). That is, the two distributions need not be normal in the metric in which they are observed, but there must be a common transformation of that metric that would render both distributions normal. This is a more flexible assumption than assuming distributions are normal on their extant common scale. It is, in fact, departures from *respective* equivariant normality, not just

[2] The V statistic arises from this relationship between the parameters of two normal distributions and $P_{a>b}$: the area under the PP curve for the two normal distributions. When both distributions are normal with mean and variance parameters μ_a, μ_b, σ_a^2, and σ_b^2, the relationship follows (Downton, 1973):

$$P_{a>b} = \Phi\left(\frac{d^{coh}}{\sqrt{2}}\right).$$

Solving for d^{coh} yields V. Equivalent expressions to V have proposed in the ROC literature (e.g., Simpson and Fitter, 1973), where it is commonly known as d_a. However, in the context of medical tests, AUC-type measures are most commonly used (Pepe, 2003).

equivariant normality, that lead to disagreements between d^{tpac} gaps estimated from different cut scores. In general, distributional assumptions in an ordinal framework are best described as respective or transformation-inducible. In the ROC literature, where the concern is sensitivity and specificity of diagnostic tests, the transformation-inducible normality assumption has been described as "binormal" (Swets & Pickett, 1982). In the context of gaps between test score distributions, we retain the descriptor "respective normal" to make more explicit that the distributions need not be normal and need not be restricted to two in number.

The V statistic can be understood as the difference in mean test scores between two groups, both with normal test score distributions, that would correspond to a PP plot with an area under the curve of $P_{a>b}$. As shown in Equation 5, V can be computed directly from the area under the PP curve. It is thus broadly interpretable as a transformation-invariant analogue of Cohen's d even when distributions are not respectively normal.

When the full CDFs are known for both groups, the calculation of nonparametric gap statistics like V and $P_{a>b}$ follows in straightforward fashion from the PP plot. When only censored, PAC-type data are available, however, these statistics cannot be calculated exactly. The single-cut-score statistics d^{pac} and d^{tpac} are estimable, but, as Table 1 shows, they can vary widely across alternative cut scores. The next section describes the use of PAC data as observed points to estimate a PP curve. These curves allow for nonparametric gap estimates from ordered categorical data alone.

Estimating Ordinal Gaps from Censored Data

In order to estimate the gap measure V using censored data, we apply the PP framework described in Figure 1. Extending previous notation, we begin with an interest in the gap between two groups, a and b. The ith student score in group g is given by X_{ig}. Suppose there are $K+1$

distinct achievement levels common to both groups. These are delineated by K cut scores $x_1 < x_2 < \cdots < x_K$, so that a student i from group g achieves Level 1 if $X_{ig} < x_1$; Level k if $x_{k-1} \leq X_{ig} < x_k$ for $k \in \{2 \ldots K\}$; and Level $K+1$ if $x_k \leq X_{ig}$. The CDF F_g returns the cumulative proportion of students in group g at or below cut score k, denoted $p_g^k = F_g(x_k)$. Note that these proportions are simply the complements of the PAC statistics described above: $PAC_g^k = 1 - p_g^k = 1 - F_g(x_k)$.

If we had the full data from the test score distributions (that is, if we knew F_a and F_b), we would be able to compute any gap measure we like. We could plot the full PP curve (Figure 1) describing the proportion of members of group b with scores below given percentiles of group a:

$$G(p_a) = F_b\left(F_a^{-1}(p_a)\right). \quad (6)$$

The problem arises when we do not know F_a or F_b but instead have access only to the proportions of each group above cut scores. That is, we know only PAC_a^k and PAC_b^k (and, of course, the associated p_g^k, because $p_g^k = 1 - PAC_g^k$) for some small number of cut scores K. Usefully, the representation of the PP plot, G, allows for the possibility of an estimate of the PP plot, $\hat{G}$, from the PAC data. In fact, the K points, $(p_b^1, p_w^1), (p_b^2, p_w^2), \ldots, (p_b^K, p_w^K)$, fall on the curve described by G, by definition. The points $(0,0)$ and $(1,1)$ can be added given the logic that some score exists below all observed score points, and some score exists above all observed score points. Figure 2 shows these $K+2$ points for the previously used example of Alabama Grade 4 NAEP Reading in 2005. The point defined by the NAEP Basic cut score is highlighted, where 33% of reference group is below Basic and 69% of the focal group is below Basic. The other two points are defined by cumulative proportions for the Proficient and Advanced cut scores respectively.

Our strategy will be to use these $K + 2$ points to estimate the function G within the unit square. If these points provide enough information to estimate G reliably, then we can obtain reliable estimates of $P_{a>b}$, as the area under $\hat{G}$, and reliable estimates of V from $\hat{P}_{a>b}$. We denote this version of V, estimated from censored data alone, as $V_{cen} = \sqrt{2}\Phi^{-1}(\hat{P}_{a>b})$. The contrasting target statistic, estimated from the full distributions, is $V_{full} = \sqrt{2}\Phi^{-1}(P_{a>b})$. In the next section, we describe nine candidate methods that attempt to minimize the distance between V_{cen} and V_{full}: obtaining usable gap statistics from censored data alone.

The criteria for evaluation of these methods have both theoretical and statistical motivations. First, symmetry is a desirable property. Logically, the distance between groups a and b should be the same, whether the expression is "group a over group b" or "group b under group a." Under symmetry, the following expression will hold: $P_{a>b} = 1 - P_{b>a}$. As a corollary, following Equation 5, a V statistic calculated using $P_{a>b}$ will have the opposite sign of a V statistic calculated using $P_{b>a}$. Second, the function, $\hat{G}$, should be monotonically nondecreasing on the unit interval, following the theoretical restrictions on PP curves. Third, the estimate of V_{full} should be unbiased over a range of realistic situations: the average distance $V_{cen} - V_{full}$ should be zero. Finally, the magnitude of the distance between the estimate and the target should be as small as possible over a range of realistic situations. This will be evaluated using the root mean square deviation (RMSD) between V_{cen} and V_{full}. The nine candidate methods follow.

Piecewise Linear Interpolation (PLI)

A graphically simple approach is to fit a linear spline function to the $K + 2$ points, essentially "connecting the dots" to estimate G. Computing $\hat{P}_{a>b}$, the integral of $\hat{G}$ over the unit interval, is then a straightforward sum of areas of rectangles and triangles:

$$\hat{P}_{a>b}^{PLI} = \sum_{k=1}^{K+1} \left[\left(p_a^{k-1} \cdot \left(p_b^k - p_b^{k-1} \right) \right) + \frac{1}{2} (p_a^k - p_a^{k-1}) \left(p_b^k - p_b^{k-1} \right) \right], \quad (7)$$

where $p_g^0 = 0$ and $p_g^{K+1} = 1$. The PLI approach is also notable because of its equivalence to the so-called midrank convention (Conover, 1973), a conventional nonparametric approach to adjusting $P_{a>b}$ when a pair of full distributions has ties. Ties result in unconnected PP points on a PP plot: the same problem addressed by this paper. The midrank convention adjusts $P_{a>b}$ as follows: $P_{a>b}^{midrank} = P(a > b) + \frac{P(a=b)}{2}$; this is equivalent to Equation 7 if the censored distributions are treated as the full distributions of interest.

Although this method has the advantage of being relatively simple, the linear spline function is unlikely to accurately describe the underlying distributional shape. In particular, the integral will be biased toward 0.5 if the true function G has a regular shape—concave up everywhere or concave down everywhere—because the linear spline will tend to truncate portions of the area between G and the 45-degree line.

Polynomial Fitted Curve (PFC)

A second straightforward approach is to fit a J^{th}-order polynomial (where $J \leq K + 1$) to the $K + 2$ points: $p_b = \sum_{j=1}^{J} \beta_j p_a^j$. To obtain the area under the curve, the appropriate integral over the unit interval follows:

$$\hat{P}_{a>b}^{PFC} = \int_0^1 \left[\sum_{j=0}^{J} \hat{\beta}_j q^j \right] dq = \sum_{j=0}^{J} \frac{\hat{\beta}_j}{j+1}. \quad (8)$$

When the number of points is small, as in the common testing scenario where $K = 3$, one simple approach is to fit a cubic polynomial using ordinary least-squares. Symmetry can be achieved by refitting the function after reversing p_a and p_b, then averaging as follows,

$$V_{cen}^{PFC} = \frac{\Phi^{-1}\left(\hat{P}_{a>b}^{PFC}\right) - \Phi^{-1}(\hat{P}_{b>a}^{PFC})}{\sqrt{2}}.$$

A variation of this approach incorporates the known variance of the cumulative proportions, $\frac{p(1-p)}{n}$, into a weighted least-squares regression. The equation can be further constrained by ensuring that the function passes through (0,0) and (1,1), points that should exist on a PP curve regardless of sampling or measurement error. Algebraic reconfiguration of the regression equations under these constraints allows for estimation of regression parameters through standard statistical software programs. We evaluate both free and weighted-constrained polynomial fitted curves and designate them PFCf and PFCwc respectively. One anticipated drawback of the PFC approaches is that theoretically impossible PP curves can result, including those with negative slopes and those that break the vertical boundaries of the unit square.

Monotone Cubic Interpolation (MCI)

A third approach that combines some of the advantages of the PLI and PFC approaches fits a piecewise cubic spline through the data using the Fritsch-Carlson (1980) method. The Fritsch-Carlson method guarantees a function that is monotonic, differentiable everywhere, and passes through each data point. For the purpose of fitting PP curves, this affords three primary advantages. First, the estimated curve, $\hat{G}$, passes through each of the K+2 points, like PLI but unlike PFC. Second, the function is monotonic, resolving the problem of negative slopes and unbounded PP curves that can arise under the PFC approach. Third, the curve is smooth everywhere on the unit interval, potentially resolving the bias that may arise with PLI. Given the MCI-fitted curve $\hat{G}$ and coefficients returned by the Fritsch-Carlson algorithm, we can compute:

$$\hat{P}_{a>b}^{MCI} = \sum_{k}^{K+1}\left[\int_{p_b^{k-1}}^{p_b^{k}} \left(\hat{\alpha}_k + \hat{\beta}_k x + \hat{\gamma}_k x^2 + \hat{\delta}_k x^3\right) dx\right] \quad (9)$$

A drawback of the MCI approach is asymmetry. Like the PLI method, MCI will return asymmetrical gaps when groups a and b are switched on the axes. As with the PFC approaches, we resolve this by averaging.

Fitted Curve to Transformed Data

An alternative to fitting the PP points directly is to transform the two axes and fit the transformed data points. We can then transform the fitted line back into the original metric and integrate in order to compute $\hat{P}_{a>b}$. If the transformation results in a more familiar or easily estimable functional relationship between the variables, such as a line, then we can obtain more accurate estimates of $P_{a>b}$ and V_{full}. We call this method by the mnemonic acronym TFIT (Transform-Fit-Inverse Transform). In particular, we investigate two potential transformations: one using the probit (inverse-normal) function that is denoted PTFIT and one using the logit function that is denoted LTFIT. Both functions are monotonic mappings of the domain $(0,1)$ to the range $(-\infty, +\infty)$. Due to the infinite mappings of (0,0) and (1,1), we exclude these two theoretical points and apply TFIT only to the K observed data points.

Probit transform-fit-inverse transform (PTFIT).

By transforming each p_a^k and p_b^k using the probit function, we can fit a curve to the K transformed points. That is, we can fit a J^{th}-order polynomial of the form

$$\Phi^{-1}(p_a) = \sum_{j=0}^{J} \beta_j \left[\Phi^{-1}(p_b)\right]^j , \tag{10}$$

where $J < K$. Moreover, J should be odd such that the fitted curve goes toward $(-\infty, -\infty)$ and (∞, ∞). Such a curve will approach $(0,0)$ and $(1,1)$ when inverse-transformed back to PP space. When $K = 3$, as is standard in NAEP and common in many state accountability systems, the

linear fit is the only option. This estimated line can be transformed back into PP space and evaluated numerically as the following integral:

$$\hat{P}_{a>b}^{PTFIT} = \int_0^1 \Phi\left(\hat{\beta}_0 + \hat{\beta}_1\left[\Phi^{-1}(x)\right]\right)dx.$$

Symmetry may be obtained by fitting a principal axis regression line and obtaining $\hat{\beta}_0$ and $\hat{\beta}_1$. However, preliminary results showed marked improvement with a weighted least squares approach. Each PP point may be weighted by the inverse of the variance in the transformed space. When plotting group b on group a, as in a typical PP plot, an estimate of the standard error of each point in the transformed space is given by the delta method:

$$\hat{\sigma}(\hat{p}_b) = \frac{\sqrt{\hat{p}_b(1-\hat{p}_b)/N_b}}{\phi\left(\Phi^{-1}(\hat{p}_b)\right)}. \quad (11)$$

Here, ϕ is the normal density function, Φ^{-1} is the probit function, and $1/\phi\left(\Phi^{-1}(\hat{p}_b)\right)$ is the slope of the probit function at $\hat{p}_b$. Fitting Equation 10 while weighting each point by the inverse of the square of Equation 11, we obtain a weighted least squares estimate of the slope and intercept. Due to the asymmetry of the approach, we can achieve an average by repeating the process and plotting group a on group b. A geometric average of the slopes provides the appropriate estimate of $\hat{\beta}_1$, and $\hat{\beta}_0$ can be estimated knowing that the line must pass through the intersections of each individual line. These may be transformed back into PP space for numeric integration as described before.

When PTFIT is constrained to a linear fit, a more theoretically appealing estimate of V_{full} follows from the fact that a line in probit-transformed (normal-normal) space defines two respectively normal distributions. The averaged $\hat{\beta}_1$ and $\hat{\beta}_0$ can be then used to obtain V_{cen} directly. It is relatively straightforward to show that, if two distributions a and b are respectively

normal and can be transformed to have normal parameters $\mu_a,\ \mu_b,\ \sigma_a,$ and σ_b, the probit-transformed PP plot will be a line with slope $m = \frac{\sigma_a}{\sigma_b}$ and intercept $n = \frac{\mu_a - \mu_b}{\sigma_b}$ (e.g., Pepe, 2003). Thus, we can express V as a function of m and n in a quasi-Cohen expression:

$$V_{cen}^{PTFIT} = \frac{\mu_a - \mu_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}} = \frac{n}{\sqrt{\frac{m^2 + 1}{2}}}. \quad (12)$$

Fitting a line through the probit-transformed PP points therefore implicitly assumes that the two distributions are respectively normal. With enough cut scores (at least 4), one could fit a higher-order odd polynomial through the PP points. Such a procedure would not imply respective normality, and numerical integration procedures would be required.

Logit transform-fit-inverse transform (LTFIT).

A similar procedure uses the logit function in place of the probit function. The logit and probit functions are very similar, but the logistic function has thicker tails and might be more stable when fitting extreme proportions, particularly when cut scores are very high or low for certain groups. In this case, let a standard logistic function be $\Lambda(x) = \frac{1}{1+e^{-x}}$, and the inverse-logistic, the logit, follows: $\Lambda^{-1}(p) = ln\left(\frac{p}{1-p}\right)$. Following an analogous estimate of variance as that shown in Equation 11, we can transform each proportion using the logit function, then weight each point in a linear regression by the following factor:

$$W = \frac{\lambda\left(\Lambda^{-1}(\hat{p}_b)\right)}{\hat{p}_b(1 - \hat{p}_b)}. \quad (13)$$

Here, λ is a standard logistic density function given by $\lambda(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$. The geometric average slope and appropriate intercept are calculated as before, then numerical

integration obtains $\hat{P}_{a>b}^{LTFIT}$ after the line is transformed back into PP space. Note that no direct computation of $V_{a>b}^{LTFIT}$ is available as it was in the linearly fit PTFIT case.

Average Normal Shift (ANS)

The Normal Shift (NS) approach was introduced by Furgol, Ho, and Zimmerman (2010) as a method of estimating V from censored data. The authors adapt a maximum-likelihood-based algorithm from Wolynetz (1979) that estimates a mean and variance from censored data with known cut scores assuming an underlying normal distribution. With cut scores for state tests, cut scores are either unavailable or lack strong equal-interval properties for meaningful use. Therefore, the authors established cut scores by assuming the reference distribution, F_a, was standard normal, Φ, leading to K cut scores defined by $\Phi^{-1}(p_a^k)$ for $k = 1 \ldots K$. These cut scores anchor the cumulative proportions for the focal group, p_b^k, and these are used to estimate the mean and variance, $\hat{\mu}_b$ and $\hat{\sigma}_b^2$, via the Wolynetz algorithm. Given the assumed standard normal parameters of the reference distribution, the appropriate effect size estimate is simply

$$V_{cen}^{ANS} = \frac{-\hat{\mu}_b}{\sqrt{(1+\hat{\sigma}_b^2)/2}}. \quad (14)$$

A weakness of the NS model is that, like the PFC and MCI approaches, gap estimates are not symmetric under the choice of reference group. We resolve this by averaging V_{cen}^{ANS} with its negative when the groups are reversed, and we contrast this approach with the Furgol, Ho, and Zimmerman approach by describing this as the Average Normal Shift (ANS) approach. Both approaches assume respective normality but allow for variances to differ across the groups. It is similar to the linear PTFIT approach in its assumptions but uses a maximum likelihood approach on the CDFs instead of a weighted regression on transformed cumulative percentages.

Receiver Operating Characteristic Fit (ROCFIT)

A previous section described the interpretation of a PP plot as a ROC curve in signal detection theory. Within this literature, maximum likelihood estimates of the parameters for the ROC curve have been developed by Dorfman and Alf (1969) under the binormal or respectively normal assumption. We use the algorithm as implemented in the Stata command, "rocfit"; it is also available in the R package "pROC." In Stata, the area under the curve, or $\hat{P}_{a>b}^{ROCFIT}$, is stored as the scalar *e(area)*, and the V statistic, V_{cen}^{ROCFIT}, is stored as the scalar, *e(da)*.

The ROCFIT approach can be considered a more formal version of the linearly constrained PTFIT. It fits the K probit-transformed PP points in normal-normal space using a maximum likelihood approach. It enjoys the property of symmetry. A similar maximum likelihood extension of linearly constrained LTFIT in the ROC context was proposed by Ogilvie and Creelman (1968); we do not evaluate this bilogistic approach due to the lack of available software and the lack of promise of results to be shown.

Average Difference in Transformed Percents-Above-Cut (ADTPAC)

A previous section described d^{tpac}, an achievement gap measure that expresses the difference between groups a and b by taking the difference of probit-transformed PACs. Table 1 showed that this measure of the gap may differ across the K cut scores due to departures from respective normality. If the two test score distributions are normal with equal standard deviations, however, this measure will be the same regardless of which score is used as the cut score. Moreover, it will be equal to Cohen's d and, in this special case, to V. Given that deviations from normality are expected, a simple method for obtaining a gap estimate is to average across the K estimates of d_k^{tpac}.

We use an improved approach that takes advantage of the same weighting principles introduced for PFCwc and the two TFIT approaches. Equation 11 can be used to obtain an

approximation of the variance of the difference in transformed PACs: $\hat{\sigma}^2\left(\hat{d}^{tpac}\right) = \hat{\sigma}^2(\hat{p}_a) + \hat{\sigma}^2(\hat{p}_b)$. The inverse of these variances can be used as weights, w_k, to obtain a weighted average difference of transformed PACs:

$$V_{cen}^{ADTPAC} = \sum_{k=1}^{K} \frac{w_k}{W} d_k^{tpac}. \quad (15)$$

Here, $w_k = \frac{1}{\sigma^2\left(\hat{d}_k^{tpac}\right)}$ and $W = \sum_{k=1}^{K} w_k$. Due to the inverse normal transformations in the calculation of each d_k^{tpac}, this approach implicitly assumes that the two distributions are respectively normal with equal standard deviations. The variation in the d_k^{tpac} gaps over k is assumed to be sampling variation. The average is thus an estimate of V obtained without directly estimating the PP curve, G.

Table 2 summarizes the nine proposed methods of estimating V_{full} from the observed censored data. Note that most of the methods guarantee monotonicity, and few of them are inherently symmetric. For the asymmetric methods, we find some approach to taking an average of gaps estimated both ways in order to avoid the arbitrariness of the choice. Due to the ordinal framework, the implied distributional assumptions are not traditional but respective. The respective normal assumption implies that some shared transformation can render both distributions normal. Likewise, the respective logistic assumption implies that some shared transformation can shape both distributions to the derivative of the logistic function: the shape of a Hubbert curve. The respective distributional assumptions for the TFIT approaches apply only for linear models in the transformed space: when $J = 1$. The TFIT and PFC approaches are thus a much larger family of approaches when greater numbers of cut scores allow for higher-order polynomial fits.

Evaluating Approaches to Ordinal Gap Estimation

This section uses simulated and real data to compare approaches as they attempt to recover full-distribution gap estimates, V_{full}, using censored data alone. As Table 2 describes, there are strong a priori reasons to discount seemingly straightforward approaches, such as the anticipated bias of the PLI approach and the possibility of unbounded or decreasing functions with PFC approaches. The first subsection compares the performance of different approaches across simulated scenarios. The second subsection compares recovery of V_{full} in the real data context of NAEP White-Black achievement gaps in 2003, 2005, and 2007.

Recovery of V_{full} in Controlled Scenarios

The initial goal is to identify the best gap estimation approaches in controlled scenarios where respective distributional shapes are known. This section presents three simulation scenarios: an equivariant normal scenario, an unequal-variance normal scenario, and a skewed scenario using lognormal distributions. For each of these, we 1) draw two samples from generating distributions with known parameters, 2) record V_{full} using these two full samples, 3) define a set of centered, plausible cut scores, 4) apply these cut scores to the two samples to obtain cumulative proportions and PACs, 5) apply each approach in Table 2 to these cumulative proportions to obtain V_{cen} values, and 6) repeat this many times to evaluate bias and variance under sampling. We add the gap between the generating distributions as an additional factor to understand how the magnitudes of bias and variance vary for gaps between 0 and 1.5 standard deviation units in size.

For these scenarios, we draw 2000 students for the reference group a and 500 students for the focal group b, approximating the median sample sizes used for state NAEP. Following the NAEP design and the designs of many state testing programs, we censor the data using three cut scores. Larger numbers of cut scores will increase the similarity between the censored and

full distributions and dampen the differences between estimation approaches. To establish generic cut score locations, we use a symmetric approach with respect to both distributions: The three cut scores result in unweighted cross-group averages of PACs as follows: 80% above Basic, 50% above Proficient, and 20% Advanced (cumulative proportions of 0.2, 0.5, and 0.8).

The cut scores are obtained through an approach akin to mixture modeling that results in the appropriate PACs (or cumulative proportions) for the mixture of both distributions. For example, when two normal distributions with unit variance are centered on 0 and 1 respectively, the cut scores -0.45, 0.50, and 1.45 result in 20%, 50%, and 80% PACs for the unweighted mixture of the two distributions. This approach also has the advantage of keeping the amount of cut-score information constant—in the sense that the combined cumulative proportions are always the same—even as the gap between distributions shifts from 0 to 1.5. Figure 3 shows the generating distributions mapped into PP space along with the PP points that would be generated in the population.

These cut scores censor the distributions, leaving three pairs of cumulative proportions. Following the example in the previous paragraph, for a Basic cut score of -0.45, the normal CDF shows that 32.6% of a low-scoring, $N(0,1)$ distribution scores below Basic, and 7.4% of a high-scoring, $N(1,1)$ distribution scores below Basic. Note that these proportions average to 0.20, as expected by the cut score approach detailed in the previous paragraph. Clearly, these percentages may change in any given sample due to sampling. The sampled PP point will thus vary around the point $(.074,\ .326)$ on a PP plot; this point can be found in Figure 3.

The other two cut scores define two more PP points, and these, with the two theoretical points at the origin and $(1,1)$ where appropriate, are the data that the methods in the previous section use to obtain V_{cen}. The second, unequal variance scenario increases the variance of the

generating distribution for the low-scoring group by 50%, and the third, skewed scenario uses the lognormal distribution to impart respective positive skew. These are also shown in Figure 3 and are described in greater detail in the next subsections.

The criteria for the recovery of V_{full} are bias, the average of $V_{cen} - V_{full}$ over all replications, and the root mean squared deviation (RMSD), the square root of the average of $\left(V_{cen} - V_{full}\right)^2$ over all replications. We use 5000 replications for each distance between generating distributions, drawing 2000 for the reference group and 500 for the focal group for every replication. The distance between the generating distributions is varied between 0 and 1.5 at intervals of .02. This allows comparison of approaches across a range of plausible gap magnitudes and across distributional scenarios likely to arise in practice. Note that V_{full} is the appropriate criterion over d_{coh}, because d_{coh} is a transformation-dependent statistic that cannot be fully specified within an ordinal framework. Although d_{coh} happens to be equal to V_{full} in the two normal scenarios that follow, this does not change the fact that a transformation can distort d_{coh} but not V_{full}.

The normal, equivariant scenario.

The most straightforward model for test scores is the normal model, and the equal-variance assumption is an appropriate baseline assumption in the absence of other information. Figure 3 displays the population PP curves that result from the normal, equivariant model when the mean difference is 0, 0.5, 1.0, and 1.5 standard deviation units. The figure displays the population normal, equivariant PP curves as black, solid lines above the diagonal, and the "observed" points from the cut score algorithm in the population are also shown. As expected, the curves bulge from the diagonal as the mean difference increases. The observed points, however, stay on a line with slope -1, as expected from the cut score algorithm that keeps the

cumulative proportion of the mixture of distributions constant over mean differences. The goal of each of the nine proposed approaches is to approximate the full curve using the five observed points alone.

The top half of Figure 4 shows the bias—the average of $V_{cen} - V_{full}$ over 5000 replications—over the range of mean differences and for each approach. When the mean difference is zero, the PP curve is the line $y = x$, and all nine approaches estimate this easily. As the mean difference increases, the PLI approach is the most biased, underestimating the full gap by almost 10%. This is not surprising given that the linear approach truncates area under any concave curve. The PFCf approach also underestimates V_{full}, and its weighted, constrained counterpart, PFCwc, is positively biased. The MCI and LTFIT approaches show slight negative bias when gaps are very large, and PTFIT, ANS, ROCFIT, and ADTPAC perform very well in a scenario that matches their assumptions perfectly.

The bottom half of Figure 4 shows the RMSD, and there is a clear distinction between the PLI/PFC approaches and the others. The weighted, constrained PFC approach can be seen to improve upon the free PFC approach when gaps are small. However, when gaps are large, one or both groups may have cumulative proportions close to 0 or 1. In these cases, the PFCwc approach places heavy weight on these extreme points while also forcing the curve to go through the origin and $(1, 1)$. It is clear that this results in poor performance. Although there is a substantive and statistical justification to weight points and force the curve through the origin and $(1, 1)$, there appears to be little advantage to doing so.

The normal, unequal-variance scenario.

Returning back to Figure 3, the unequal variance scenario can be seen mapped into PP space and reversed over the diagonal to unclutter the figure. The generating distribution for the

low-scoring focal group has a variance of 1.5 compared to the unit variance of the reference group's generating distribution. This difference in variance, equivalent to increasing the standard deviation by 22.5%, represents a fairly high variance difference in practice, but differences in observed variances are not uncommon. For example, the average absolute White-Black variance ratio for 2009 NAEP was 1.15 across 172 state-subject-grade combinations, and 4 state-subject-grade combinations exceeded an absolute ratio of 1.5. As expected of the cut-score-selection algorithm, comparing the "observed" PP points across the population PP curves reveals alignment on a line with slope of -1.

The top half of Figure 5 shows the bias plotted on the standardized mean difference as defined by V_{full} in the population. The results are very similar to Figure 4 in spite of the notable variance differences. The PLI approach remains negatively biased to a similar degree, and the weighted, constrained PFC approach begins performing more poorly than its free counterpart regardless of gap size. The ROCFIT, PTFIT, and ANS account for variance differences explicitly and therefore perform without bias. A notable difference from Figure 4 is that ADTPAC begins to show negative bias when gaps are large. This is a reminder that ADTPAC assumes respective normality with equal variances, and its performance worsens when this assumption is not met. The bottom half of Figure 5 shows the RMSD for the normal, unequal-variance scenario. It is worth noting that the RMSD for V_{full} recovery by the six best approaches hovers around .015, a fairly small amount of variability for the estimation of gaps when the only three paired cumulative proportions are available.

The lognormal, skewed scenario.

To challenge the assumptions of respective normal approaches like ANS, ROCFIT, and PTFIT, which account for respective normality and unequal variances, we use respectively

skewed lognormal distributions. We define a random variable whose log is distributed $N(0,0.3)$. Such a distribution has mean 1.05, a standard deviation of 0.32, and positive skew of 0.95. To generate a gap, we shift one distribution above another such that V_{full} varies from 0 to 1.5 standard deviation units. Unlike the previous two scenarios, this is not equivalent to shifting d_{coh} from 0 to 1.5, as $d_{coh} = V$ only when distributions are normal. Cut scores are generated as before. These PP curves are also plotted back in Figure 3.

A useful conceptual point is that two respectively lognormal distributions are not equivalent to two shifted normal distributions on the same scale that are transformed by the exponential function. This latter construction is ordinally equivalent to the respective normal distributions presented in the first scenario. Respectively lognormal distributions cannot be transformed to normal with a single transformation unless their CDFs completely overlap.

The top half of Figure 6 shows the bias plotted on the V_{full} metric as before. The performances of the PFCwc and ADTPAC approaches are notably worse. It is clear at this point that the PLI and PFC approaches are flawed under even the most typical scenarios; they will not be considered further. Taking Figures 5 and 6 together, the poor performance of ADTPAC under variance differences and skewness indicates its inability to adequately estimate the full distribution through weighted averaging of transformed PACs. In contrast, although ANS, ROCFIT, and PTFIT show biased recovery of V_{full}, their bias in this case is very small at around .004 for the largest gaps. In this scenario, the ANS approach outperforms ROCFIT and PTFIT by very small amounts. The LTFIT and MCI approaches have a larger amount of negative bias at around -.015; this is still less than 1% of the largest gaps.

The bottom half of Figure 6 shows the RMSD, where the decline in ADTPAC performance is also apparent. The efficiency of recovery of the five best approaches continues to

hover at around .015 and increases to just over .022 when population gaps are very large. These approaches outperform seemingly attractive alternatives like PLI by a considerable margin and suggest that gap recovery is possible even when respective normal assumptions are not met.

Recovery of V_{full} in Real Data Scenarios

This subsection assesses the performance of these approaches in real data scenarios. We use the full distributions of plausible values from NAEP state distributions, averaging over the five available sets of plausible values as recommended by Mislevy, Johnson, and Muraki (1992) to obtain V_{cen} and V_{full}. The state distributions correspond to White and Black students in 2003, 2005, and 2007, for Reading and Mathematics in Grades 4 and 8. Out of 600 possible state-subject-grade-year combinations (50 states by 2 subjects by 2 grades by 3 years), 490 have sufficient sampling of Black students to allow for the reporting of achievement gaps. We calculate the nonparametric gap measure, V_{full}, for these 490 White-Black gaps; these are the targets that the approaches attempt to recover under censored data scenarios. The criteria are bias and RMSD averaging over these 490 trials.

The full distributions clearly cannot have their standardized mean differences, variances, or skew manipulated as in Figures 4-6, as these distributions are real. Their properties remain the same as those actually reported. However, the factor of cut score location can be usefully introduced into this analysis, as recovery of gaps is expected to depend on the location of cut scores in the distributions. We vary cut score location along two dimensions, the breadth of the cut scores and the stringency of the cut scores. The cut scores are indexed by the average cumulative percentages as before, except instead of fixing the average cumulative percentages at 20%, 50%, and 80%, they are varied systematically. The breadth dimension has average cumulative percentages varying from 5%, 50%, and 95% (broad cut scores) to 45%, 50%, and

55% (narrow cut scores). The stringency dimension has average cumulative percentages varying from 5%, 30%, and 55% (low cut scores leading to low cumulative percentages and high PACs) to 45%, 70%, and 95% (high cut scores leading to high cumulative percentages and low PACs).

Unlike NAEP reporting, where there are common cut scores for each subject-grade combination, this approach allows each pair of distributions to have its own trio of cut scores. This is done to ensure that the interpretation of "broad" or "stringent" is consistent across pairs of distributions. If a common set of cut scores were used, broad or stringent cut scores for one pair of distributions would be less broad or stringent for another. Note also that some approximation of the results from the actual NAEP cut scores is located reasonably high along the stringency dimension, where the unweighted average cumulative proportions between White and Black students approach 45%, 70%, and 95% (55% basic and above, 30% proficient and above, 5% advanced) across state-subject-grade combinations.

Recovery of gaps depending on cut score breadth.

The top half of Figure 7 shows the bias of the 6 best-performing metrics in their recovery of real-data gaps across broad and narrow cut scores. As noted previously, the PLI and PFC approaches perform poorly in common scenarios and are not considered further. The top half of Figure 7 shows that the overall bias of these six candidate methods can be very low. The LTFIT and MCI approaches do not perform well when cut scores are narrow. However, the four best metrics, ADTPAC, PTFIT, ANS, and ROCFIT have bias less than .02. Focusing on these four approaches, the ADTPAC approach performs relatively poorly, and PTFIT does not perform as well as ANS or ROCFIT particularly when cut scores are broadly spaced. The lowest bias across all methods occurs close to cumulative proportions of 20%, 50%, and 80%. This suggests that the bias scenarios in Figures 4-6 are optimistic. However, for ANS and ROCFIT in particular,

the bias ranges from -.007 to +.013, a very small bias given that the median White-Black NAEP gaps are generally about .75 standard deviations in size.

The bottom half of Figure 7 shows the RMSD across cut score breadth. As before, the MCI and LTFIT approaches perform poorly when cut scores are more narrowly spaced. Within the top four approaches, the ADTPAC and PTFIT approaches perform relatively poorly when cut scores are extreme. The poor ADTPAC performance is consistent with the findings in Figures 5 and 6. The performance of ANS and ROCFIT are indistinguishable along the RMSD criterion. The overall efficiency when cut scores are neither broad nor narrow is quite good, with RMSDs bottoming out at around .009, a small percentage of White-Black gaps in practice.

Recovery of gaps depending on cut score stringency.

The top half of Figure 8 shows the bias in recovery across cut score stringency. The symmetry of these curves suggests that methods perform best when cut scores are central with respect to the unweighted mixture of both distributions. The MCI and LTFIT approaches continue to perform worse than their counterparts, with negative bias. The absolute bias of the PTFIT, ANS, and ROCFIT approaches are similar, and ADTPAC bias is negative when cut scores are low.

The bottom half of Figure 8 shows the RMSD of the approaches and results in similar conclusions. The LTFIT and MCI approaches perform relatively poorly. The ANS and ROCFIT approaches perform the best, with slightly better efficiency than PTFIT. ADTPAC does not perform as well outside of the region where it happens to show no bias. Focusing on the right-hand portion of the graph, where cut scores are closer to their real-world NAEP counterparts, the RMSDs are between .025 and .029. This may still be considered surprisingly low given how little information exists about the lower half of the respective distributions. When the basic cut

score is lower, as it often is in practice, Figure 7 suggests that performance will improve. Further, because state cut scores are usually lower or much lower than NAEP cut scores, the RMSDs are likely to be closer to those seen towards the center of Figure 8.

Discussion

These results suggest three promising candidates for the estimation of gaps under censored data scenarios. The two best approaches are ROCFIT, implemented by Stata in a command motivated by signal detection theory, and ANS, a simple adaptation of a maximum likelihood estimation procedure developed by Furgol, Ho, and Zimmerman (2010). Both result in very small amounts of bias and RMSD across a range of simulated and real-data scenarios. The ROCFIT approach is symmetrical and estimates a PP curve directly, a comparative theoretical advantage over ANS, which is asymmetrical and estimates normal CDFs. In addition, the ANS implementation in R does not have documentation and is not widely available. Both packages also allow for the estimation of standard errors; the ROC approaches to standard error estimation are reviewed by Pepe (2003).

For those who do not have access to ROCFIT approaches in Stata, the PTFIT approach is intuitive, easy to implement with standard routines in statistical packages, and shows little loss in performance across scenarios. There may be greater possibilities for PTFIT when more cut scores are available, and higher-order polynomials can be fit to data on the probit-transformed axes. The magnitudes of the bias and RMSD for all three of these methods are rarely over .02 and are usually much less, an impressive result under the real-data and lognormal scenarios, where the respective normal assumption is threatened or violated outright. These results suggest that the estimation approach is robust to deviations from respective normality across a range of cut score locations.

The applicability of these approaches extends beyond gap estimation for censored state testing data. Tests reported on score scales with few ordinal categories, such as Advanced Placement exams, which report scores on a 1-5 integer scale, and some exams for English Learners are also natural applications for these gap estimation approaches. In these cases, the data are treated as censored even if the grain-size of the data is the finest available. The argument in favor of the use of this framework is that some continuous scale underlies the observed scale. Similarly, a case where ceiling or floor effects compress a theoretically distinguishable score range into a single undifferentiated score point—this is a censored data problem. These are cases where an ANS-, PTFIT-, or ROCFIT-estimated V statistic may in fact be preferable to effect sizes calculated from means and standard deviations on the established score scale.

A small number of technical issues remain. The effects of sample size, sample size ratio across groups, and the overall number of cut scores are of interest. We do not spend time on them here because the findings will be straightforward: more is better. Increasing cut scores and sample size beyond the levels here will also mute the differences between methods that were our primary interest. The adequate recovery of V_{full} when there are only three cut scores suggests that a higher benchmark for the minimum number of cut scores is not necessary.

When sample sizes are smaller, cut scores are extreme, group differences are large, or some combination of these instances, there is an increased likelihood that the highest or lowest score category will have no student representation from one group or another. In these situations, a number of the methods proposed here will fail, including PTFIT and ANS, which would both attempt to take an inverse-normal transformation of 0 or 1. A simple correction involves adding a student or a fraction of a student to the highest or lowest score bin.

Measurement error is known to attenuate Cohen-type effect sizes by inflating standard deviations. The same issues arise in PP plots, as measurement error will attenuate PP curves toward the main diagonal. The NAEP examples are adjusted for measurement error through the plausible values methodology (Mislevy, Johnson, & Muraki, 1992), however, gap comparison across tests, times, or groups with different degrees of measurement error must acknowledge or adjust for attenuation. An ad hoc disattenuation approach treats V statistics like their d_{coh} counterparts and divides by the square root of the reliability estimate; this is discussed briefly in Ho (2009).

Finally, it may seem straightforward to extend these analyses from gaps to trends. After all, any two distributions on the same scale can be expressed as a PP plot; it may not seem to matter whether they are Groups a and b or Times 1 and 2. However, we recommend caution in using these methods for descriptive trend analyses for a few reasons. First, if cross-sectional, within-grade trends are the target of inference, these are much smaller in magnitude, and the degree of bias and variance reported here will have a greater impact on substantive interpretations. Second, trends rely on the year-to-year linking of score scales, a source of error that this ordinal framework does not currently address. Third, discrete score scales are typically aligned imperfectly from year to year due to linking functions. A score bin in one year may not exist in the next, or two bins may be combined that once were separate. This has little effect on the means of the full distributions, but a small shift in the scale can lead to a large percentage of students crossing one of a small number of cut scores. This is not an issue with within-year gap measures, where score points align, but this issue may severely distort PP-based trend measures.

Interestingly, this latter problem with trends does not necessarily generalize to a problem with gap trends. As Ho (2009) has noted, one can express a gap trend as a “change in gap” or a

"difference in changes." These are equivalent in an average-based framework but not in an ordinal framework. A "difference in changes" formulation subjects a gap trend to not one but two opportunities for the distortions noted in the previous paragraph. However, a change-in-gap formulation, where gaps are estimated within each year and then subtracted from each other, manages to avoid the problems of year-to-year equating and misaligned score bins. This is the recommended approach to tracking gaps over time.

With widespread reporting of test scores in ordinal achievement levels, researchers interested in achievement gaps are increasingly faced with censored data scenarios. This paper evaluates ordinal approaches for estimating achievement gaps using censored data alone and finds three approaches—ROCFIT, ANS, and PTFIT—whose performance justifies recommendation. These estimates represent dramatic improvement over any gap estimate derived from a single cut score. They also recover gaps well over a range of scenarios, in both an absolute sense and relative to alternative ordinal approaches. The resulting estimates are interpretable on a familiar Cohen-type metric and are transformation-invariant: particularly useful properties for gap comparisons across different tests, times, grades, and jurisdictions.

References

Center on Education Policy. (2007). Answering the question that matters most: Has student achievement increased since No Child Left Behind? Retrieved November 1, 2008, from http://www.cep-dc.org/index.cfm?fuseaction=document.showDocumentByID&nodeID=1&DocumentID=200

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

Conover, W. J. (1973). Rank tests for one sample, two sample, and *k* samples without the assumption of a continuous distribution function. *The Annals of Statistics, 1,* 1106-1125.

Dorfman, D. D., & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology, 6*, 487-496.

Downton, F. (1973). The estimation of Pr(Y>X) in the normal case. *Technometrics, 15*, 551-558.

Education Week. (2010, January 14). State of the states: Sources and notes. *Education Week*, *29*(17), 49-50. Retrieved June 1, 2010, from http://www.edweek.org/ew/articles/2010/01/14/17sources.h29.html

Fritsch, F. N., & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *Society for Industrial and Applied Mathematics: Journal on Numerical Analysis, 17*, 238-246.

Furgol, K. E., Ho, A. D., & Zimmerman, D. L. (2010). Estimating trends from censored assessment data under No Child Left Behind. *Educational and Psychological Measurement, 70*(5), 760-776.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Ho, A. D. (2007). *Describing the pliability of growth statistics under transformations of the vertical scale*. Paper presented at the 2007 annual meeting of the National Council on Measurement in Education.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351-360.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*, 201-228.

Ho, A. D., & Haertel, E. H. (2006). *Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples* (CSE Report No. 665). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3-17.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White Test Score Gap*. Washington D.C.: Brookings Institution Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer-Verlag.

Livingston, S. A. (2006). Double P-P plots for comparing differences between two groups. *Journal of Educational and Behavioral Statistics, 31*, 431-435.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Magnuson, K., & Waldfogel, J. (Eds.). (2008). *Steady Gains and Stalled Progress: Inequality and the Black-White Test Score Gap*. New York: Russell Sage Foundation.

Massachusetts Department of Education. (2002). 2001 MCAS Technical Report. Retrieved October 10, 2009, from http://www.doe.mass.edu/mcas/tech/01techrpt.pdf

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

Miller, G., & McKeon, H. (2007, August 28). Miller-McKeon Discussion Draft: Amendments to Title I. Retrieved June 1, 2010, from http://republicans.edlabor.house.gov/Media/File/PDFs/Discussion_Draft_Title_I.pdf

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131-154.

Neal, D. A. (2006). Why has Black-White skill convergence stopped? In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 1, pp. 511-576): Elsevier.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5*, 377-391.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

Pollack, J. M., Narajian, M., Rock, D. A., Atkins-Burnett, S., & Hausken, E. G. (2005). *Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Fifth Grade* (NCES Report No. 2006-036). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Reardon, S. F. (2008a). *Differential Growth in the Black-White Achievement Gap During Elementary School Among Initially High- and Low-Scoring Students*. Stanford, CA:

Working Paper Series, Institute for Research on Educational Policy and Practice, Stanford University.

Reardon, S. F. (2008b). *Thirteen Ways of Looking at the Black-White Test Score Gap*. Stanford, CA: Working Paper Series, Institute for Research on Educational Policy and Practice, Stanford University.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*(1), 41-49.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80*, 481-488.

Spencer, B. D. (1983). Test scores as social statistics: Comparing distributions. *Journal of Educational Statistics, 8*(4), 249-269.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress, (NCES 2009-455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved June 9, 2010, from http://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf

Vargha, A., & Delaney, H. D. (2000). A critique and modification of the common language effect size measure of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*, 101-132.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55*, 1-17.

Wolynetz, M. S. (1979). Algorithm AS 138: Maximum likelihood estimation from confined and censored normal data. *Applied Statistics, 28*, 185-195.

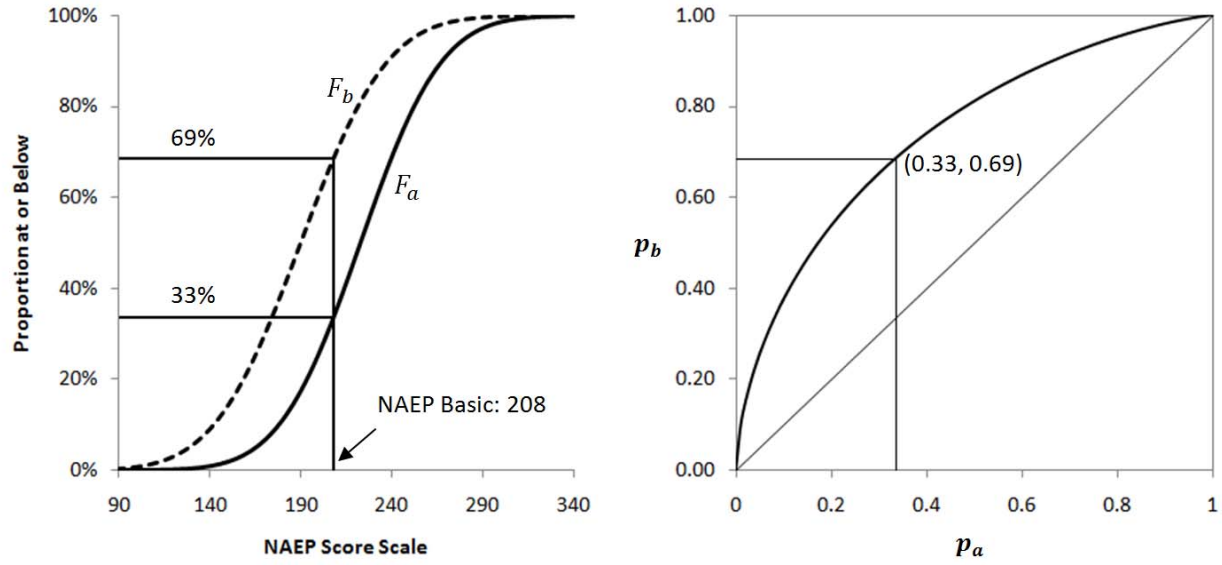
Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Measurement, 20*, 299-326.

TABLE 1. *White-Black gaps and gap trends on four different metrics; Alabama, Grade 4, 2005-2007.*

	2005	**2007**	**Change,**
d^{avg}	32.00	25.59	-6.41
d^{coh}	0.88	0.74	-0.14
d_1^{pac}	35.36	29.84	-5.52
d_2^{pac}	23.85	25.31	+1.45
d_3^{pac}	6.09	7.43	+1.34
d_1^{tpac}	0.92	0.79	-0.13
d_2^{tpac}	0.95	0.82	-0.13
d_3^{tpac}	1.01	0.76	-0.26

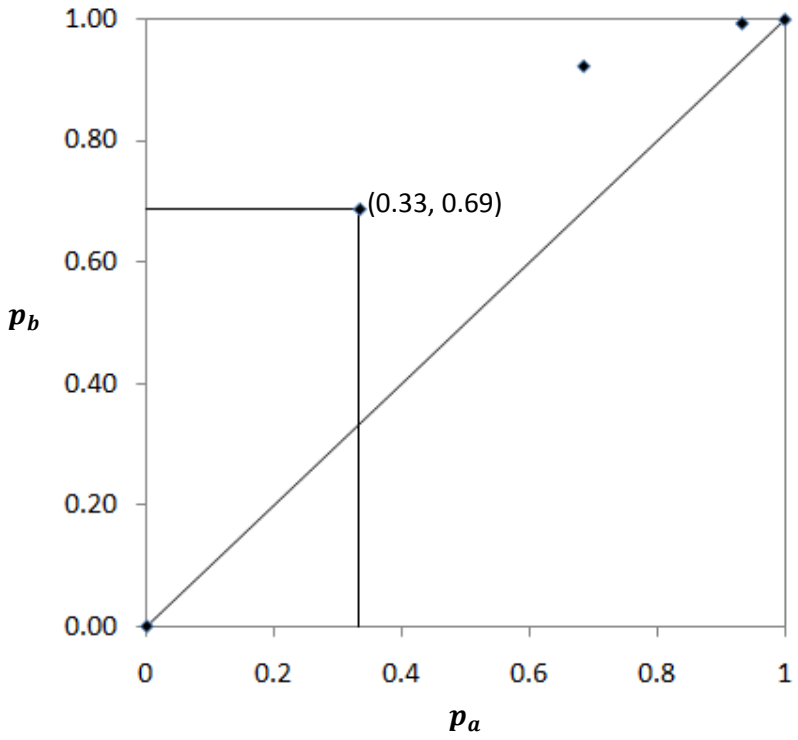
Note: Authors' calculations from data from Alabama's Grade 4 Reading administration of the National Assessment of Educational Progress, 2005-2007, obtained at http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx.
d^{avg} denotes the average-based metric; d^{coh} denotes Cohen's d; d^{pac} denotes the percentage-above-cut metric; d^{tpac} denotes the transformed percentage-above-cut metric.

FIGURE 1. *Illustrating the construction of a Probability-Probability (PP) plot from the paired cumulative proportions of distributions.*



Note: Data from idealized normal distributions using parameters from Alabama's Grade 4 Reading administration of the National Assessment of Educational Progress, 2005.

FIGURE 2. *The 5 points on a PP plot from paired cumulative proportions defined by 3 cut scores.*



Note: Data from Alabama's Grade 4 Reading administration of the National Assessment of Educational Progress, 2005. The PP pair for the Basic cut score is referenced.

TABLE 2. *Characteristics of Proposed Methods of Estimating V*

Method	Properties		Respective Distributional Assumptions	Notes
	Monotonicity	Symmetry		
PLI	✓	✓		Probable bias toward zero gap.
PFCf PFCc				Lack of monotonicity may lead to substantial bias and variance.
MCI	✓			Implemented by Matlab's "pchip" spline option.
PTFIT	✓		Normal when $J = 1$	$K < 4$ requires linear constraint.
LTFIT	✓		Logistic when $J = 1$	$K < 4$ requires linear constraint.
ANS	✓		Normal	Maximum Likelihood. Not readily available.
ROCFIT	✓	✓	Normal	Maximum Likelihood. Implemented by Stata's "rocfit" command.
ADTPAC	✓	✓	Normal, Equal Variance	Simple to implement.

FIGURE 3. *Generating distributions and the observed PP points in the population for the equivariant normal, unequal variance (reversed over the diagonal for visual clarity), and lognormal scenarios.*

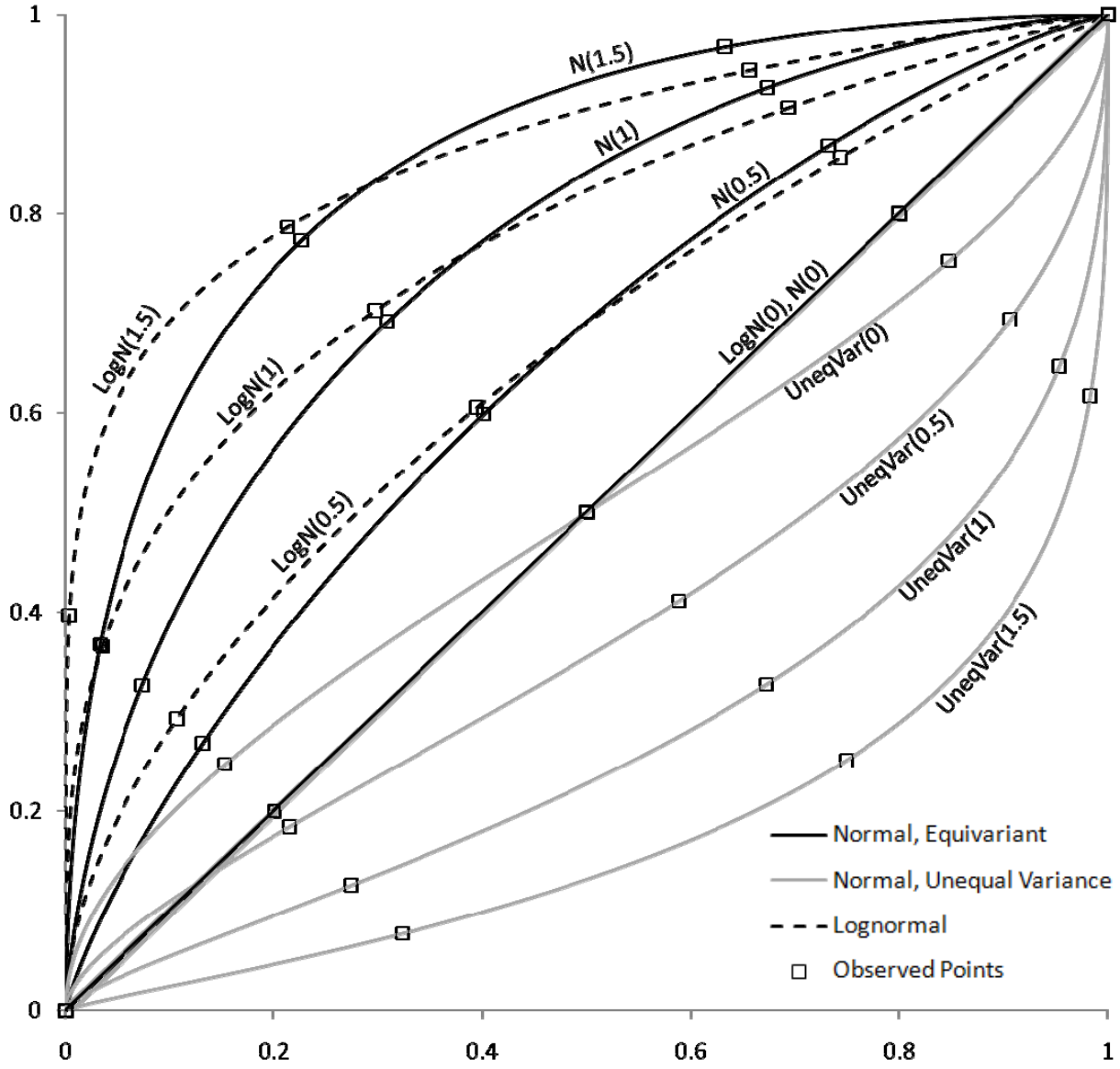


FIGURE 4. *Bias and Root Mean Squared Deviation of gap estimation approaches plotted on the size of the true gap in a normal, equivariance scenario.*

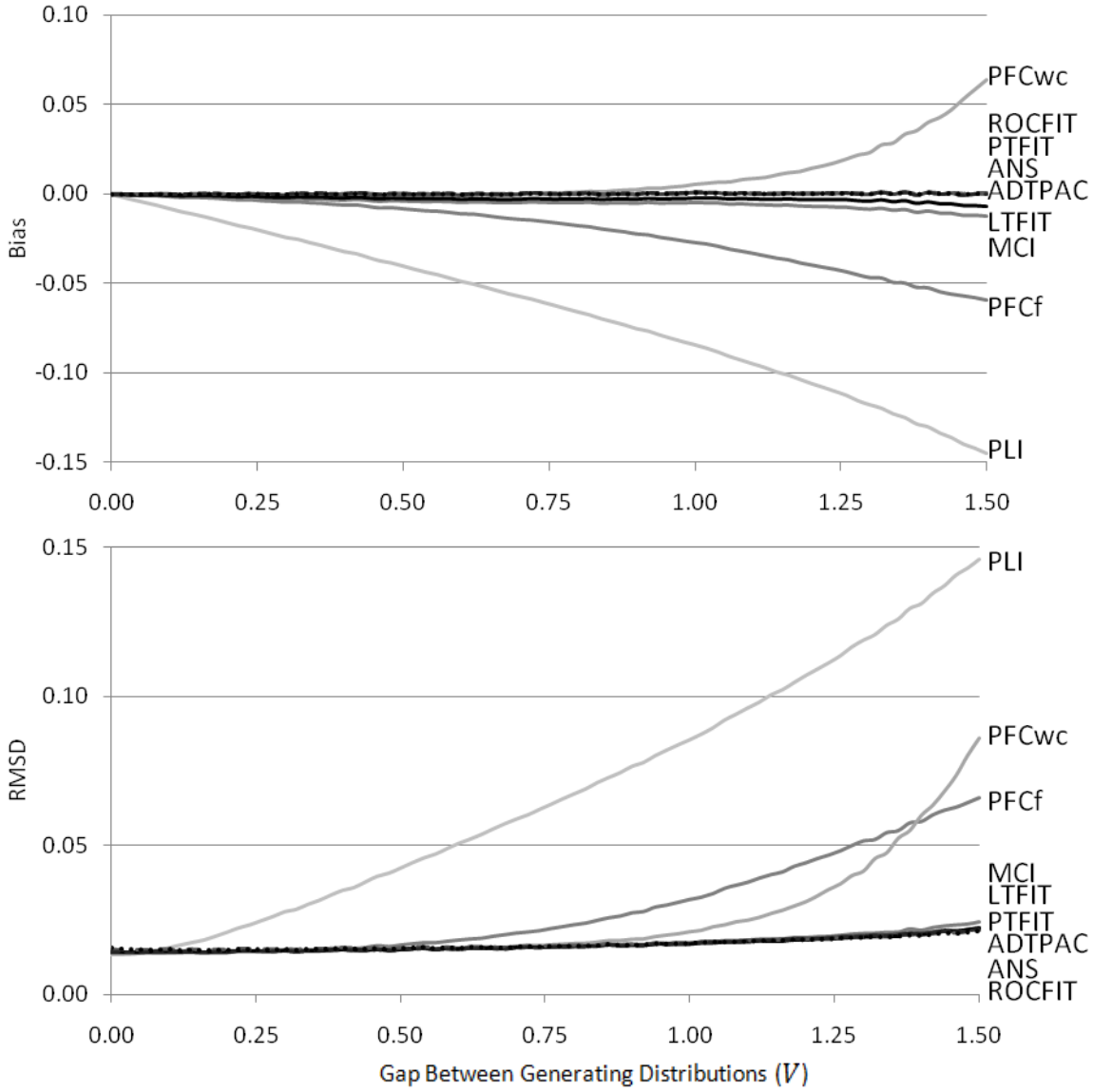


FIGURE 5. *Bias and Root Mean Squared Deviation of gap estimation approaches plotted on the size of the true gap in a normal, unequal-variance scenario.*

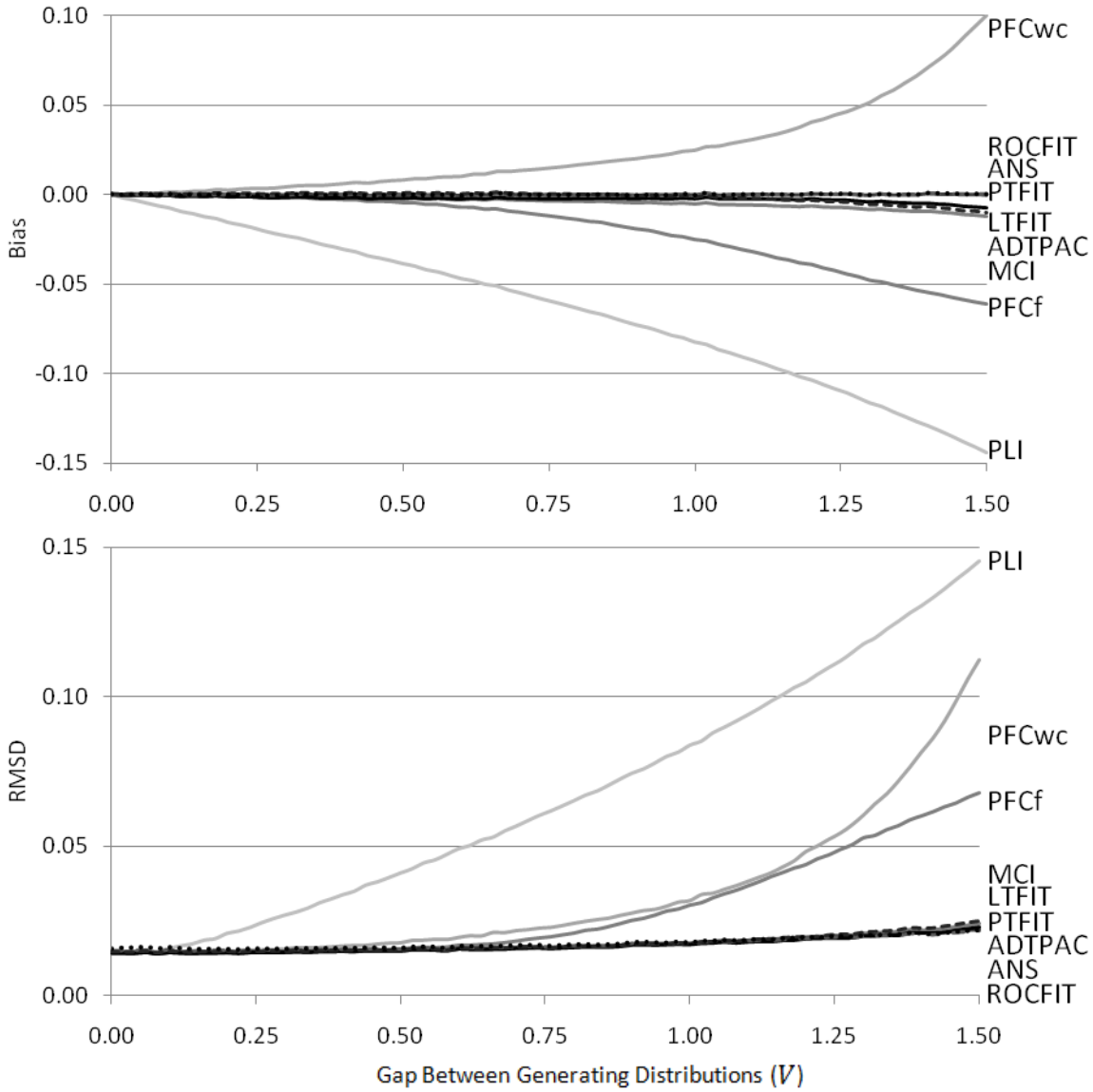


FIGURE 6. *Bias and Root Mean Squared Deviation of gap estimation approaches plotted on the size of the true gap in a lognormal scenario.*

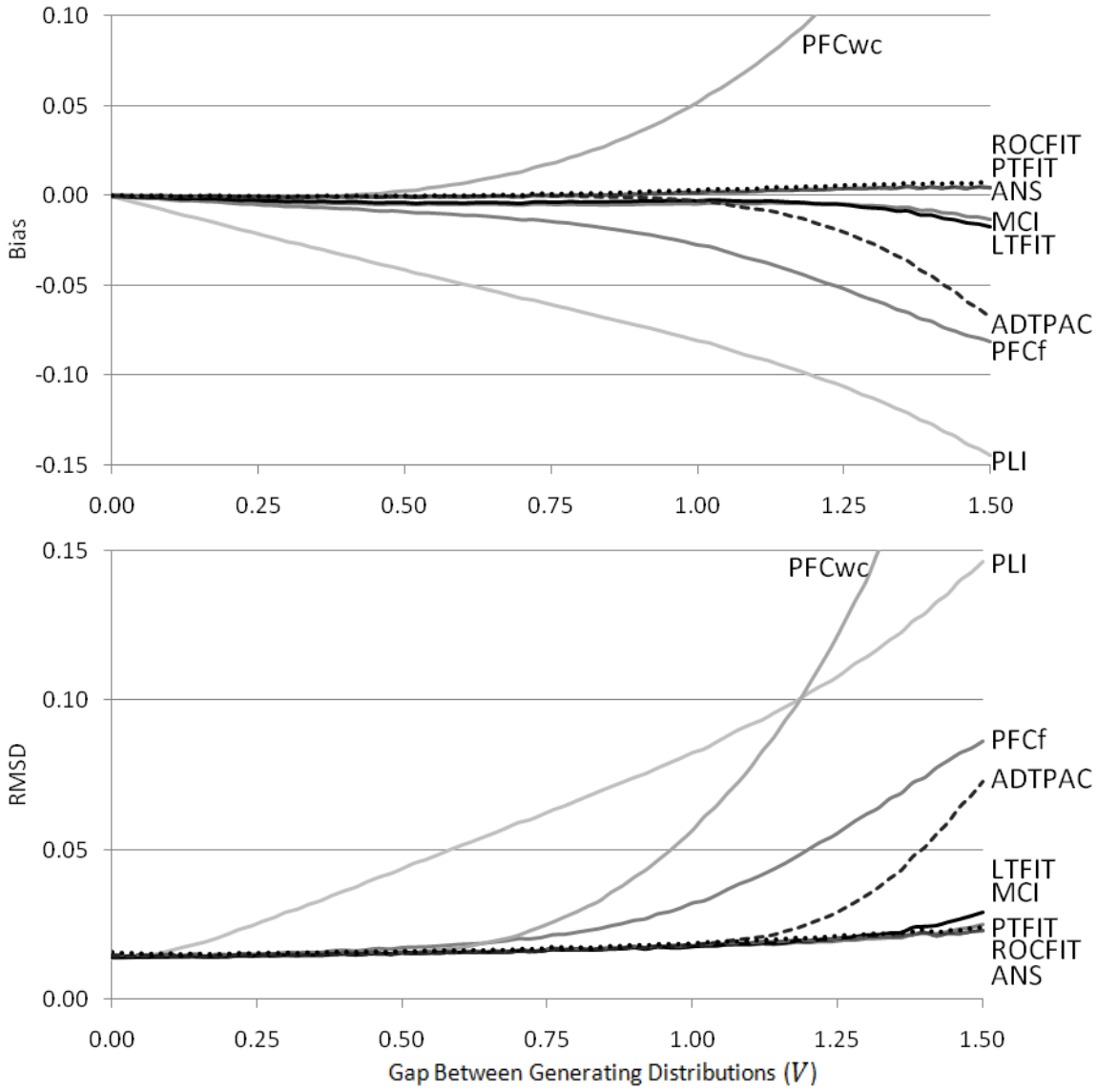


FIGURE 7. *Bias and Root Mean Squared Deviation of gap estimation approaches for White-Black state achievement gaps on the National Assessment of Educational Progress plotted on the breadth of three cut scores.*

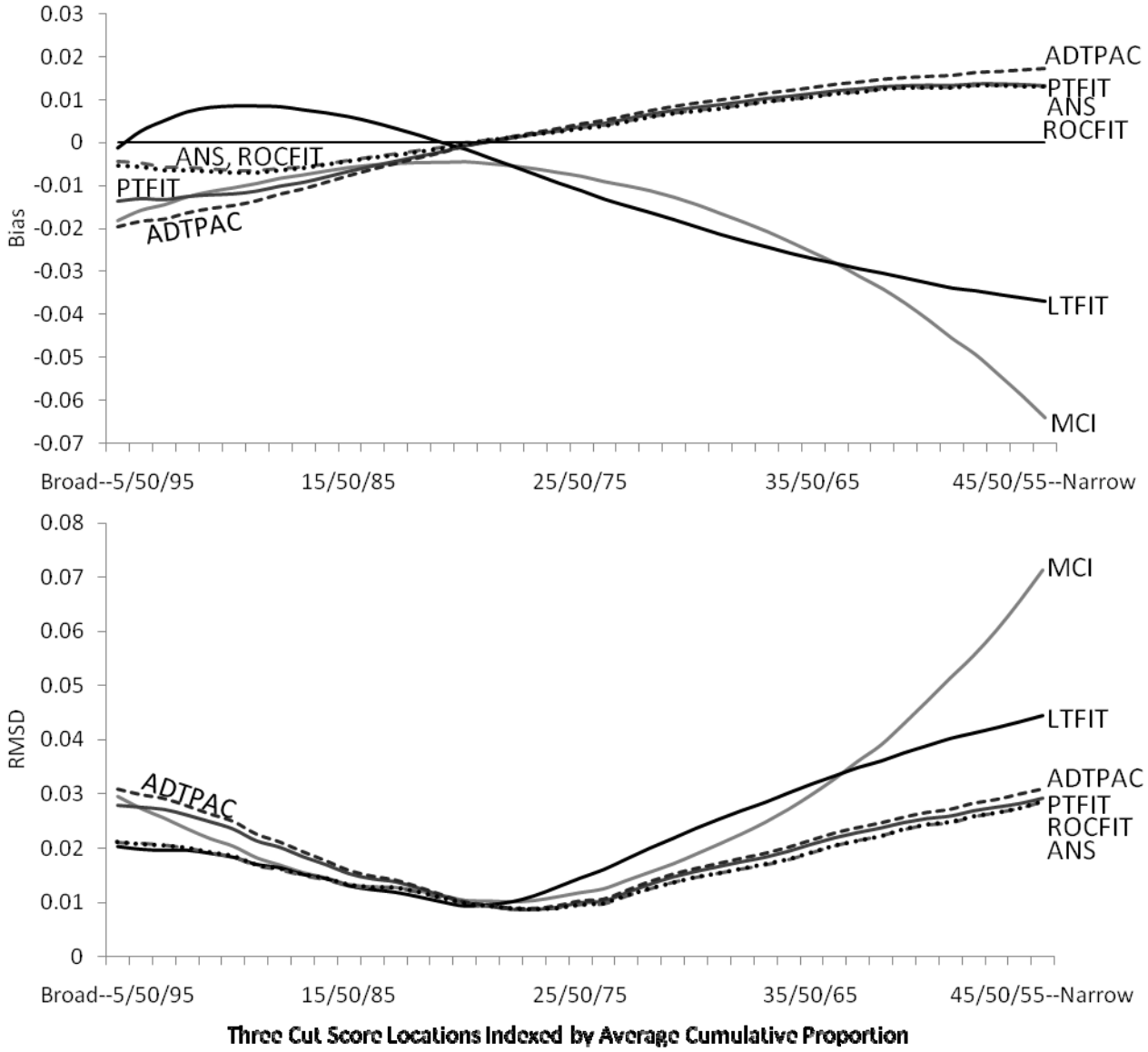


FIGURE 8. *Bias and Root Mean Squared Deviation of gap estimation approaches for White-Black state achievement gaps on the National Assessment of Educational Progress plotted on the stringency of three cut scores.*