

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

Andrew D. Ho

Harvard Graduate School of Education

Sean F. Reardon

Stanford University

Final pre-publication version; published as

Ho, Andrew D. and Sean F. Reardon. 2012. "Estimating Achievement Gaps from Test Scores Reported in Ordinal 'Proficiency' Categories." *Journal of Educational and Behavioral Statistics* 37(4):489-517.

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. Examples of such reporting 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 (U.S. Department of Education, 2010). Gaps and gap trends are 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 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 overcome and circumvent 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 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 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 unfamiliar 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  $s_b$ . Although a weighted average of variances or a single standard deviation could also be used in the denominator, we choose an expression that does not depend on relative sample size and incorporates both variances:

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

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” follows:

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

This method implicitly assumes that the test scores in groups  $a$  and  $b$  are both normally distributed with equal variance, or that a common transformation exists that can render them normal with equal variance. The resulting gap can be interpreted in terms of standard deviation units. If the distributions meet this 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  $d$ . Formal demonstrations of the logic of this transformation are widespread (e.g., Hedges and Olkin, 1985; Ho, 2009).

Each of these four metrics— $d^{avg}$ ,  $d^{coh}$ ,  $d^{pac}$ , and  $d^{tpac}$ —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). Without them, 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 PAC-based gaps are maximized when the cut score is at the midpoint between the modes of the two normal distributions and diminishes towards zero for extreme cut scores. This cut-score dependence would be acceptable if it supported defensible contrasts between gap sizes at different cut scores. However, a maximized gap at a central cut score is more appropriately interpreted as an interaction—between a gap and the non-uniform density of the distributions—that happens to contrast the two groups best when the modes are on either side. A less confounded approach to comparing gaps at different levels of the distribution would be to compare gaps in higher or lower percentiles. When distributions are normal with equal variance, PAC-based gaps will vary whereas percentile-based gaps will not (Holland, 2002).

The  $d^{tpac}$  approach helps to address the confounding of PAC-based gaps and the locations of cut scores, but it rests on the assumption that the score distributions are normal with equal variance (or share a transformation that renders them so, Ho, 2008, 2009). As Ho (2009) shows, NAEP gap trends calculated on the  $d^{tpac}$  metric vary wildly from the Basic to the Proficient cut score, and neither aligns with  $d^{coh}$  with any regularity. We replicated this analysis with 2009 NAEP data, and the degree of cut-score dependence remains substantial.

Taken together, these shortcomings raise serious concerns about the four traditional gap and gap trend metrics above. The first two,  $d^{avg}$  and  $d^{coh}$ , assume not only equal-interval scale properties but also, for gap trends, the maintenance of equal-interval properties over time. The second two,  $d^{pac}$  and  $d^{tpac}$ , confound the comparison of score distributions with the density of

students adjacent to the cut score. 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  $F_b(x)$ , that 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 two normal CDFs representing White and Black test score distributions on the NAEP Reading test as an example. These are labeled generically as a higher-scoring reference distribution,  $F_a$  (solid line), and a lower-scoring focal distribution,  $F_b$  (dashed line). The vertical axis expresses the proportion of students 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.

The right panel of Figure 1 is the corresponding PP plot that shows the proportion of Group  $b$  below given percentiles of Group  $a$ :

$$G(p_a) = F_b(F_a^{-1}(p_a)). \quad (5)$$

The paired cumulative proportions, (0.33, 0.69), are derived from the NAEP Basic cut score and shown in the right-hand panel. The PP plot is generated by obtaining all paired cumulative proportions across the score scale underlying the CDFs. Due to this construction, which uses only paired cumulative proportions and no scale information, all statistics generated from a PP plot are transformation-invariant.



One useful statistic is the area under the PP curve. This is equal to  $\Pr(X_a > X_b)$  and denoted  $P_{a>b}$  for short. That is,  $P_{a>b}$  is the probability that a randomly drawn student from group  $a$  has a greater score than a randomly drawn student from group  $b$ . 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 (Green & Swets, 1966; 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 is an alternative that avoids the proportion metric, allows for interpretation in terms of standard deviation units, and has better properties for averaging over multiple gaps. For this purpose, Ho and Haertel (2006) and Ho (2009) propose the  $V$  statistic, a nonlinear monotonic transformation of  $P_{a>b}$ :

$$V = \sqrt{2}\Phi^{-1}(P_{a>b}). \quad (6)$$

The  $V$  statistic has several useful properties. The  $V$  statistic is equal to the Cohen effect size when the two test score distributions are normal, even if they have unequal variances.<sup>1</sup>

---

<sup>1</sup> 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  $\mu_a$ ,  $\mu_b$ ,  $\sigma_a^2$ , and  $\sigma_b^2$ , the relationship follows (Downton, 1973):

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

However, even in these circumstances, Cohen's  $d$  will vary under scale transformations whereas  $V$  will not. The implicit condition under which  $V = d^{coh}$  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 that of distributions that are normal on their extant common scale.

It is, in fact, departures from *respective* equal-variance normality, not just equal-variance 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 allow for respective distributions that are not normal and that may be more than two in number.

The  $V$  statistic can be understood as the difference in mean test scores between two groups, both with standard 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 6,  $V$  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

---

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).

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 discussed previously, 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. Estimated curves allow for nonparametric gap estimates from ordered categorical data alone.

### Estimating Ordinal Gaps from Censored Data

To estimate the gap measure  $V$  using censored data, we apply the PP framework described in Figure 1. Extending previous notation, assume  $K$  cut scores,  $x_1 < x_2 < \dots < x_K$ , that divide students into  $K + 1$  ordinal achievement categories. 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  $K$  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, including using Equation 5 to plot the full PP curve in Figure 1. A 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,  $(p_b^1, p_a^1), (p_b^2, p_a^2), \dots, (p_b^K, p_a^K)$ , 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. The right-hand panel of Figure 1 shows these  $K + 2$  points for the previously used example where  $K = 3$ . 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 empirical points are defined by cumulative proportions for the Proficient and Advanced cut scores respectively, and the theoretical points at the origin and the point (1,1) are also shown.

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  $\hat{V}_{cen} = \sqrt{2}\Phi^{-1}(\hat{P}_{a>b})$ . The contrasting target statistic, computed from the full distributions, is  $V_{full} = \sqrt{2}\Phi^{-1}(P_{a>b})$ . In the next section, we describe six candidate methods that attempt to minimize the distance between  $\hat{V}_{cen}$  and  $V_{full}$  to obtain a usable gap statistic 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 6, a  $V$  statistic calculated using  $P_{a>b}$  will have the opposite sign but the same absolute value as 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, that is, the average difference  $\hat{V}_{cen} - V_{full}$  should be zero. Finally, the magnitude of the average squared 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  $\hat{V}_{cen}$  and  $V_{full}$ . The six candidate methods are ordered loosely from those that make fewer parametric assumptions to those that make more parametric assumptions.

### 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:

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

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 tied values, or  $P(a = b) > 0$ . 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:  $P_{a>b}^{midrank} = P(a > b) + P(a = b)/2$ . 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 describe the underlying distributional shape accurately. The integral will be biased toward 0.5 if the true function  $G$  has an entirely positive or negative second derivative, because the linear spline will truncate portions of the area between  $G$  and the 45-degree line. These situations are common and include all cases where distributions are respectively normal with equal variance, and the result in these situations would be an underreporting of the gap.

### Monotone Cubic Interpolation (MCI)

A natural extension of the PLI approach would be to fit a polynomial curve to the PP points. However, polynomial fits on the unit interval may not be monotonic and may extend outside the unit square. To avoid this, a piecewise cubic spline can be fit 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. Second, the function is monotonic, resolving the problem of negative slopes and unbounded PP curves that can arise under the polynomial approaches. Third, the curve is smooth everywhere on the unit interval, potentially resolving the bias that may arise with PLI. Given our unit interval on the horizontal axis,  $x$ , and the  $K + 1$  sets of cubic polynomial coefficients (the  $\hat{\alpha}_{kp}$ 's, where  $\hat{\alpha}_{kp}$  is the estimated coefficient on the  $p^{th}$ -order term of the fitted cubic function in the  $k^{th}$  interval) returned by the Fritsch-Carlson algorithm, we can compute:

$$\hat{P}_{a>b}^{MCI} = \sum_{k=1}^{K+1} \left[ \int_{p_b^{k-1}}^{p_b^k} (\hat{\alpha}_{k0} + \hat{\alpha}_{k1}x + \hat{\alpha}_{k2}x^2 + \hat{\alpha}_{k3}x^3) dx \right]. \quad (8)$$

A drawback of the MCI approach is asymmetry: MCI will return asymmetrical gaps when groups  $a$  and  $b$  are switched on the axes. We resolve undesirable asymmetry through a straightforward averaging approach on the  $V$  scale. Following Equation 6,

$$\hat{V}_{cen}^{MCI} = \frac{\Phi^{-1}(\hat{P}_{a>b}^{MCI}) - \Phi^{-1}(\hat{P}_{b>a}^{MCI})}{\sqrt{2}}. \quad (9)$$

### Probit Transform-Fit-Inverse Transform (PTFIT).

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 investigate the probit function for this purpose and designate the approach PTFIT, for Probit-Transform-Fit-Inverse-Transform. The probit function

is a monotonic mapping 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 fit a  $J$ th-order polynomial to the  $K$  transformed data points:

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

where  $J < K$ . Moreover,  $J$  should be odd such that the fitted curve goes toward  $(-\infty, -\infty)$  and  $(\infty, \infty)$ . 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:

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

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:

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

Here,  $\phi$  is the normal density function,  $\Phi^{-1}$  is the probit function, and  $1/\phi(\Phi^{-1}(\hat{p}_b))$  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 calculated from point-slope equations.

Although this can be transformed back into PP space and integrated, the linear case allows for a convenient estimate of  $V_{full}$ . It is 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  $\sigma_b$ , 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.

### 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. Therefore, the authors established cut scores by assuming the reference distribution,  $F_a$ , was standard normal, leading to  $K$  cut scores defined by  $\Phi^{-1}(p_a^k)$  for  $k = 1 \dots K$ . These cut scores anchor the cumulative proportions for the focal group,  $p_b^k$ , and are used to estimate the mean and variance,



$\mu_b$  and  $\sigma_b^2$ , via the Wolynetz algorithm. Given the assumed standard normal parameters of the reference distribution, the appropriate effect size estimate is simply

$$\hat{V}_{cen}^{ANS} = \frac{-\hat{\mu}_b}{\sqrt{(1 + \hat{\sigma}_b^2)/2}} \quad (13)$$

A weakness of the NS model is that, like the MCI approach, gap estimates are not symmetric under the choice of reference group. We resolve this by averaging  $\hat{V}_{cen}^{ANS}$  with the negative of its value when the groups are reversed, and we contrast this approach with the Furgol, Ho, and Zimmerman (2010) 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)**

We previously 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.<sup>2</sup>

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 approach uses the logit transformation instead of the probit (Ogilvie & Creelman, 1968). The distributional assumption here is respectively logistic or bilogistic. We evaluated

---

<sup>2</sup> We use the algorithm as implemented in the Stata command `-rocfit-`; it is also available in the R package “pROC.”

this approach and found poor performance due to a mismatch between the functional form and both simulated and real data. We exclude the results due to space limitations.

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

A previous section described  $d^{tpac}$ , a gap measure that expresses the difference between groups  $a$  and  $b$  by taking the difference of probit-transformed PACs. When the two test score distributions are respectively normal with equal standard deviations, this measure will be the same across cut scores. Assuming that the variation in the  $d_k^{tpac}$  over  $k$  is sampling variation, a simple method for obtaining a gap estimate is to average across the  $K$   $d_k^{tpac}$  estimates.

We use an improved approach that takes advantage of the same weighting principles as the PTFIT approach. Using the variance of the transformed PACs from Equation 11, we can obtain an approximation of the variance of the difference in transformed PACs; that is,

$\hat{\sigma}^2(d_k^{tpac}) = \hat{\sigma}^2(\hat{p}_a^k) + \hat{\sigma}^2(\hat{p}_b^k)$ . The inverse of this variance can be used as a weight,  $w_k$ , to

obtain a weighted average difference of  $K$  transformed PACs as follows:

$$\hat{V}_{cen}^{ADTPAC} = \sum_{k=1}^K \frac{w_k}{W} d_k^{tpac}. \quad (14)$$

Here,  $w_k = 1/\sigma^2(\hat{d}_k^{tpac})$  and  $W = \sum_{k=1}^K w_k$ . The average is thus an estimate of  $V$  obtained without directly estimating the PP curve,  $G$ .

Table 1 summarizes the six proposed methods of estimating  $V_{full}$  from the observed censored data. Note that the methods guarantee monotonicity, and half 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. The respective normal assumption for the PTFIT approach applies only for linear models in the transformed space; that is, when  $J = 1$ . The PTFIT approach is 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 1 describes, there are strong a priori reasons to discount seemingly straightforward approaches, such as the anticipated bias of the PLI. 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**

This section presents three simulation scenarios: an equal-variance 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 1 to these cumulative proportions to obtain  $\hat{V}_{cen}$  values, and (6) repeat this 5000 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 centered 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 CDFs. The mixture is unweighted in spite of the sample size differences to keep the cut scores centered with respect to the two distributions. This results in a more realistic set of cut scores and a simpler presentation of results. It also keeps the amount of cut-score information somewhat 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 2 shows selected generating distributions (curved solid, gray, and dashed lines) mapped into PP space along with the PP points (hollow squares) that would be generated in the population.

These three cut scores generate three pairs of cumulative proportions. Following the example in the previous paragraph, for a “Basic” cut score of -0.45, 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. Clearly, these percentages may change in any given sample due to sampling variability. The sampled PP point

will vary around the point (.074, .326) on a PP plot. This point can be found in the top left panel of Figure 2.

The other two cut scores define two more PP points, and these are the data that are fit 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 2 (as gray lines and dashed lines in the top right and lower left panel respectively) and are described in greater detail in the next subsections.

The criteria for the recovery of  $V_{full}$  are bias, the average of  $\hat{V}_{cen} - V_{full}$  over all replications, and the root mean squared deviation (RMSD), the square root of the average of  $(\hat{V}_{cen} - V_{full})^2$  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 a more appropriate criterion than  $d_{coh}$ , 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, equal-variance 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. The upper left panel of Figure 2 displays the population PP curves that result from the normal,

equal-variance model when the mean difference is 0, 0.5, 1.0, and 1.5 standard deviation units. The figure displays these normal, equal-variance PP curves as black, solid lines above the diagonal, and the hollow squares are the “observed” points that would be generated by the cut score algorithm in the population. 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 six proposed approaches is to approximate the full curve using the five observed points alone.

The top half of Figure 3 shows the bias—the average of  $\hat{V}_{cen} - V_{full}$  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 six 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 convex curve, and we narrow the range of the figures to focus on the contrasts between the better performing methods. The MCI approach shows 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 3 shows the RMSD, where there is a clear distinction between the PLI approach and the others. The more parametric methods, ANS and ROCFIT, appear to overfit the data slightly when the gap is zero. That is, they seem to attribute sampling error around a simple diagonal line to respectively normal distributions more often than their less parametric counterparts. However, they perform better when the gap is large. These differences are very small with respect to the size of gaps in practice, and the range of the figures is set to discourage overinterpretation of substantively trivial differences.

**The normal, unequal-variance scenario.**

In Figure 2, the unequal-variance scenario is mapped into PP space and shown in the top right panel. The generating distribution for the low-scoring focal group has a variance of 1.5, whereas the reference group variance is 1. This difference in variance, equivalent to a increasing the standard deviation by 22.5%, is a fairly high variance difference in practice, but differences in observed variances are not uncommon. For example, the absolute White-Black variance ratio,  $\max(\sigma_a^2, \sigma_b^2) / \min(\sigma_a^2, \sigma_b^2)$ , for 2009 NAEP was 1.15 across 172 state-subject-grade combinations, and 4 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 4 shows the bias plotted on the standardized mean difference as defined by  $V_{full}$  in the population. The results are very similar to Figure 3 in spite of the notable variance differences. The PLI approach remains negatively biased. The ROCFIT, PTFIT, and ANS approaches account for variance differences explicitly and perform without bias. A notable difference from Figure 3 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 4 shows the RMSD for the normal, unequal-variance scenario. It is worth noting that the RMSD for  $V_{full}$  recovery by the five best approaches stays below .025, 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 that 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 as dashed lines in the lower left panel of Figure 2.

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 5 shows the bias plotted on the  $V_{full}$  metric as before. The performance of the ADTPAC approach is notably worse. It is clear at this point that the PLI approach is flawed under even the most typical scenarios; it will not be considered further. Taking Figures 4 and 5 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, ANS, ROCFIT, and PTFIT show very small positive bias in their recovery of  $V_{full}$  with biases around .004 for the largest gaps. In this scenario, the ANS approach outperforms ROCFIT and PTFIT by negligible amounts. The MCI approach has a larger amount of negative bias approaching -.015. This is still less than 1% of the largest gaps.

The bottom half of Figure 5 shows the RMSD, where the decline in ADTPAC performance is quite apparent. The efficiency of recovery of the 4 best approaches continues to



hover at around .015 and increases to just over .025 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 sets of plausible values as described 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 achievement gap reporting. We calculate the nonparametric gap measure,  $V_{full}$ , for these 490 White-Black gaps; these are the targets for recovery 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 3-5, 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% (broadly spaced cut scores) to 45%, 50%, and 55% (narrowly spaced cut scores). We refer to these sets of cut scores as simply broad

and narrow for short. 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 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 6 shows the bias of the five best-performing metrics in their recovery of real-data gaps across broad and narrow cut scores. As noted previously, the PLI approach performs poorly in common scenarios and is not considered further. The top half of Figure 6 shows that the overall bias of these five candidate methods can be very low. The MCI approach does 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 3-5 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 6 shows the RMSD across cut score breadth. As before, the MCI approach performs 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 4 and 5. The performances 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 7 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 approach continues to perform worse than its 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 7 shows the RMSD of the approaches and results in similar conclusions. The MCI approach performs 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 6 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 7.

### **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 basis for this robustness may be partially explained by Figure 2.

Although the curve itself may be fitted poorly to respectively non-normal data, the areas beneath estimated and true curves may not differ substantially.

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, when ceiling or floor effects compress a theoretically distinguishable score range into a single undifferentiated score point, the problem is one of censored data. These are cases where an ANS-, PTFIT-, or ROCFIT-estimated  $V$  statistic may be preferred over 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 a square root of the reliability estimate; this is discussed briefly by Ho (2009).

Finally, it may seem straightforward to extend these analyses from gaps to trends. If 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 two 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. This is less of an issue for within-year gap measures, where linkings are generally not necessary.

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 linking error

as 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 linking. 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 introduces tools from multiple statistical literatures to address the problem. We find three approaches—ROCFIT, ANS, and PTFIT—whose performance justifies recommendation. These estimates are dramatic improvements over gap estimates derived from a single cut score. The approaches 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. These are 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.



- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- 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.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.
- 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.
- U.S. Department of Education. (2010). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act*. Washington, DC: Office of Planning, Evaluation, and Policy Development.
- 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, U.S. Department of Education. Washington, DC.
- 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.

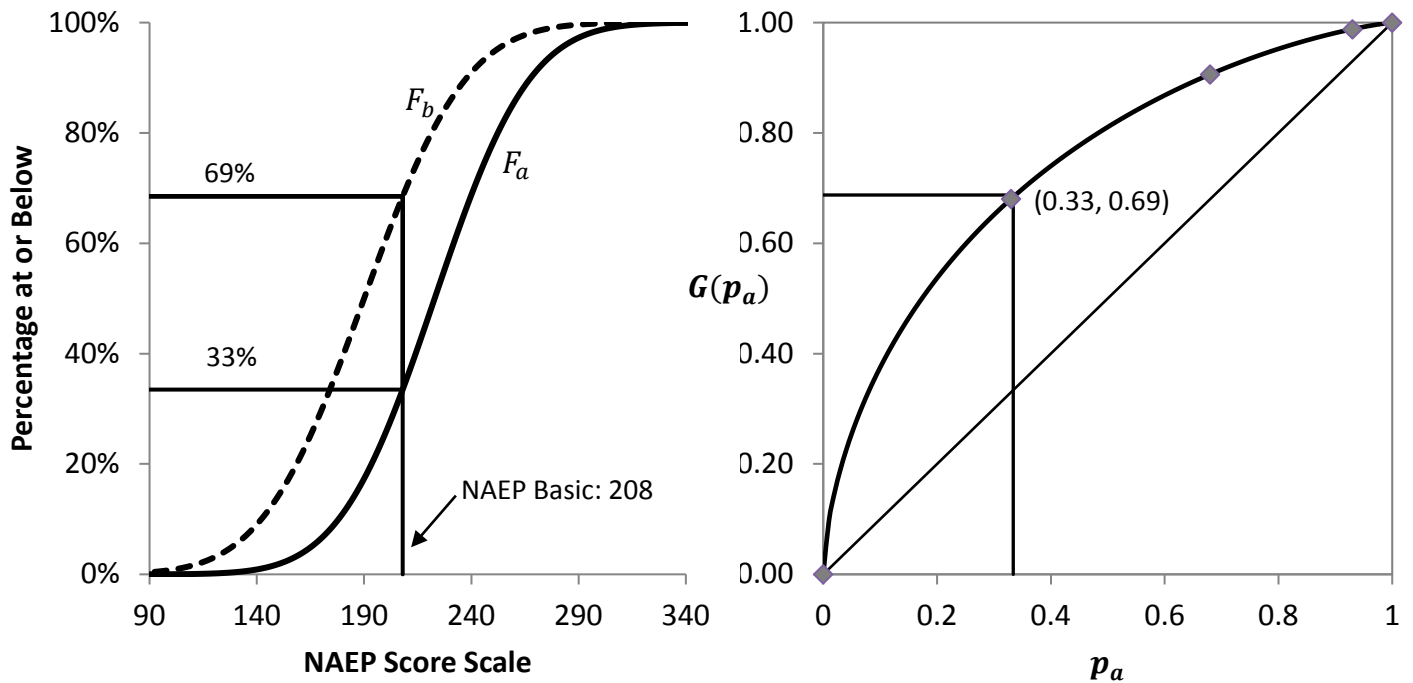
Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data.

*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.

Figure 1. Construction of a Probability-Probability Plot



*Figure 1.* Illustrating the construction of a Probability-Probability (PP) plot from the paired cumulative proportions of distributions. The left-hand panel shows test score distributions for Groups *a* and *b* on a common score scale from the National Assessment of Educational Progress (NAEP). The NAEP Basic cut score is also shown, and the percentages at or below that cut score are labeled. The right-hand panel shows the PP plot that represents the paired cumulative proportions from the two distributions at left. The corresponding PP point from the NAEP Basic cut score is identified along with the PP points for the Proficient and Advanced cut scores.

Table 1

*Characteristics of Proposed Methods of Estimating  $V$  from Censored Data*

Method	Properties		Respective Distributional Assumptions	Notes
	Monotonicity	Symmetry		
PLI	✓	✓		Probable bias toward zero gap.
MCI	✓			Implemented by Matlab's "pchip" spline option.
PTFIT	✓		Normal when $J = 1$	$K < 4$ requires linear constraint.
ANS	✓		Normal	Maximum Likelihood. Not readily available.
ROCFIT	✓	✓	Normal	Maximum Likelihood. Implemented by Stata's <code>-rocfit-</code> command.
ADTPAC	✓	✓	Normal, Equal Variance	Simple to implement.

*Note.* PLI = piecewise linear interpolation; MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit; ADTPAC = average difference in transformed percentages above a cut score.

Figure 2. Generating distributions and the paired cumulative proportions in the population

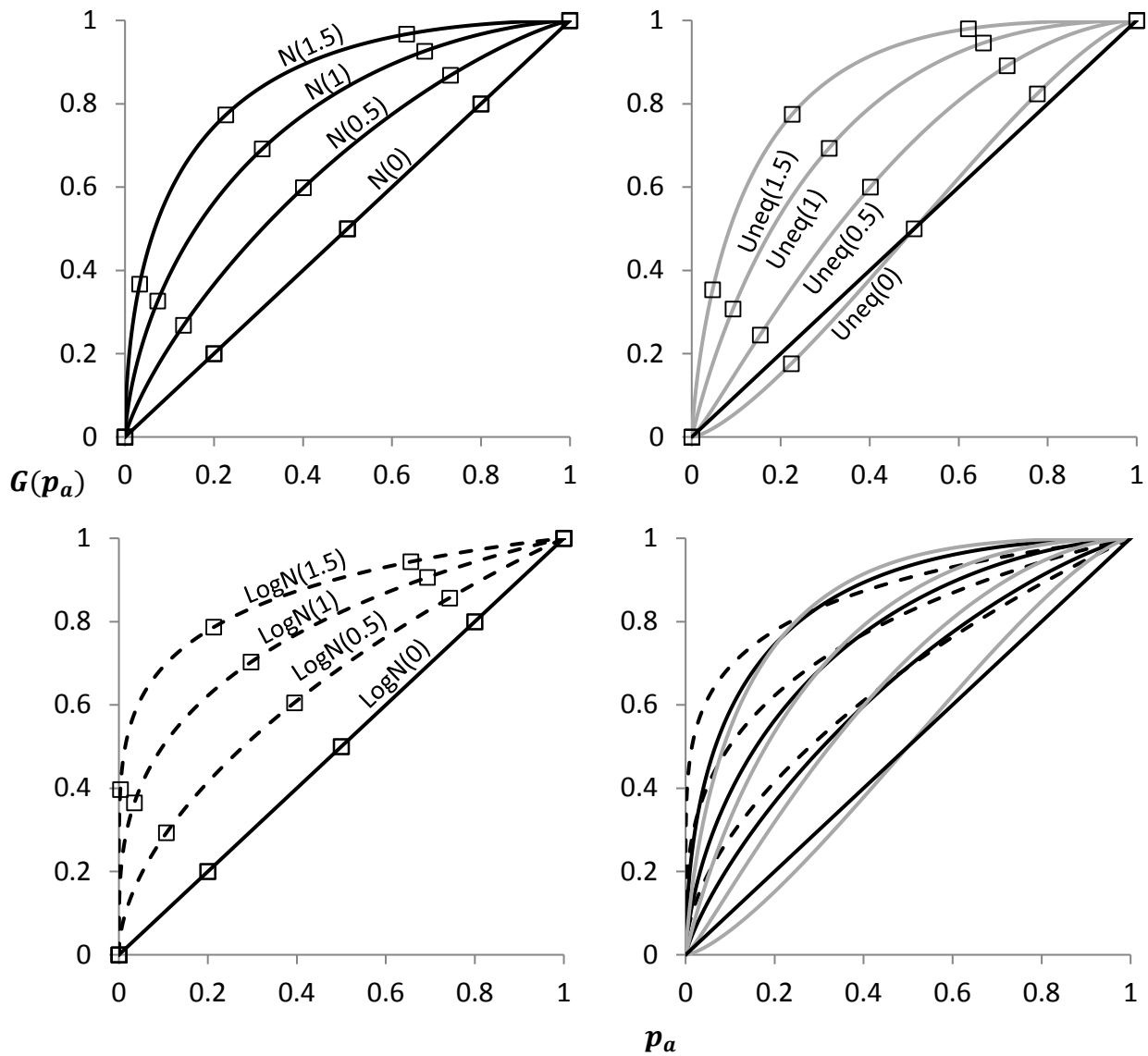


Figure 2. Generating distributions and "observed" proportion-proportion points in the population. The top left panel shows a range of normal, equal-variance distributions with standardized mean differences from 0 to 1.5, abbreviated  $N(0)$ - $N(1.5)$ . The top right panel shows a range of normal, unequal-variance distributions with standardized mean differences from 0 to 1.5, abbreviated  $Uneq(0)$ - $Uneq(1.5)$ . The lower left panel shows a range of lognormal distributions with standardized mean differences from 0 to 1.5, abbreviated  $LogN(0)$ - $LogN(1.5)$ . The observed points in the population are shown in these three panels as hollow squares. The lower right panel overlays these generating distributions to highlight their contrasts.

Figure 3. Recovery of the simulated gap in a normal, equal-variance scenario

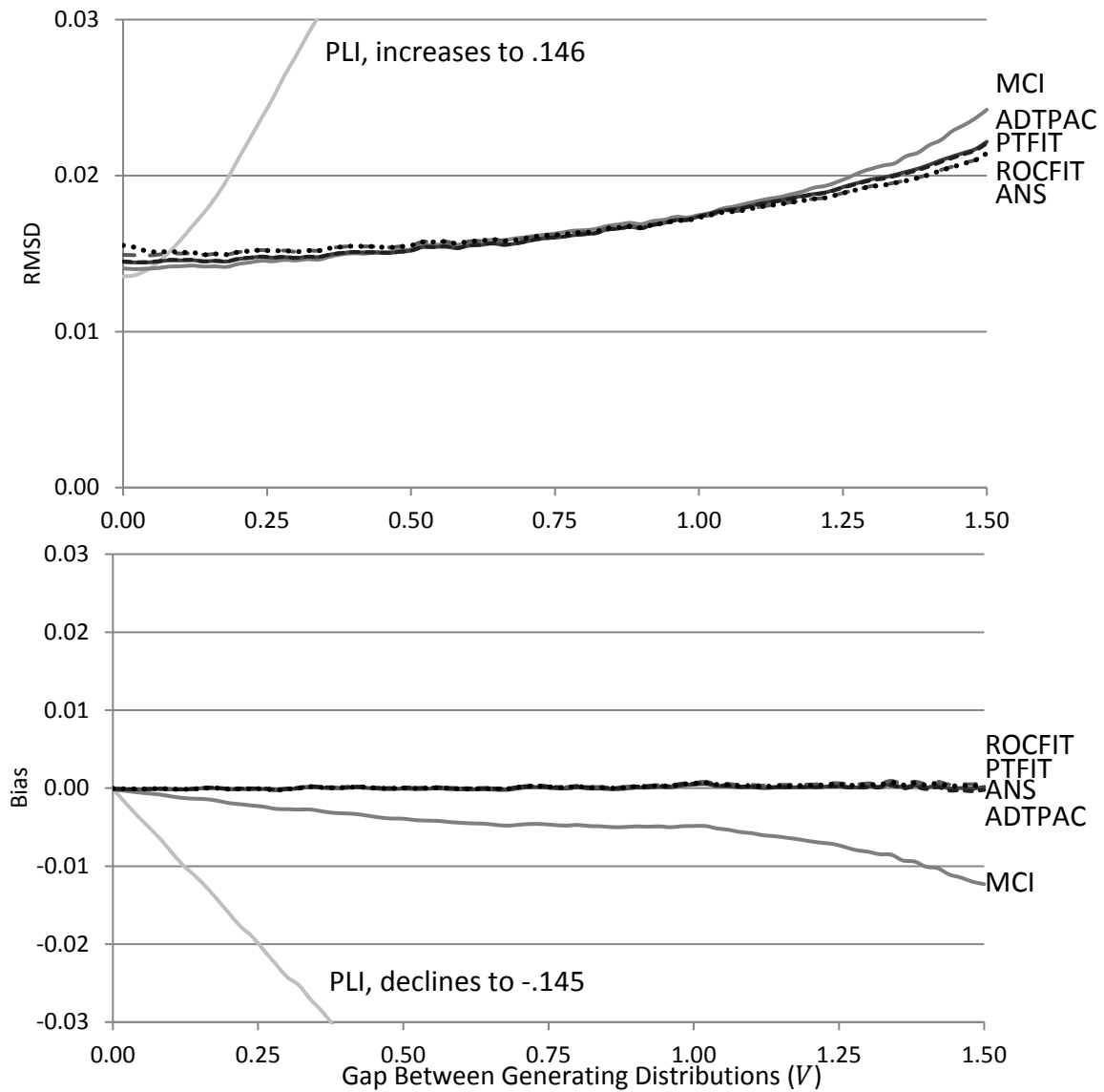
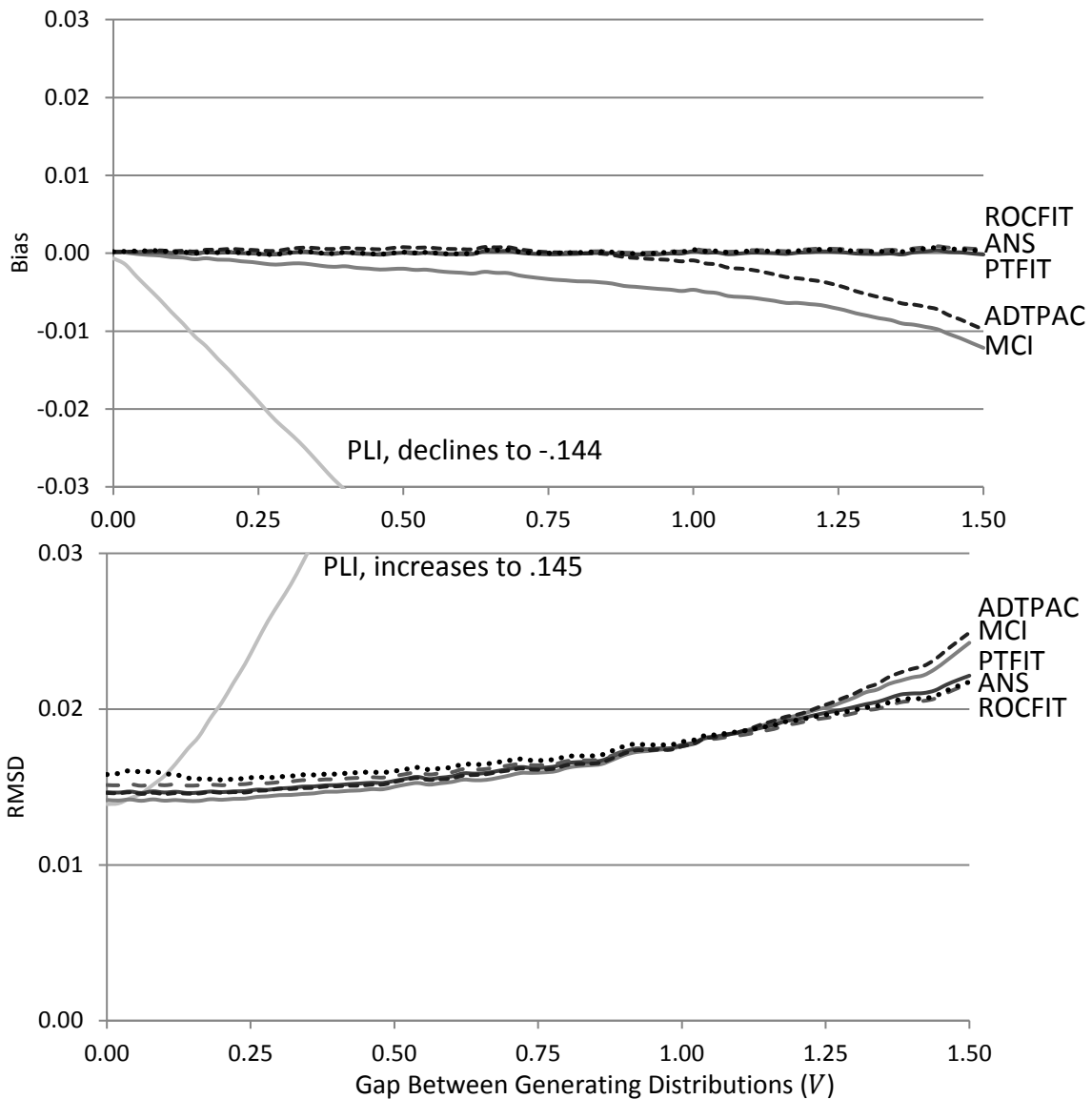


Figure 3. Bias and Root Mean Squared Deviation (RMSD) of six candidate gap estimation approaches using only three paired cumulative proportions from simulated data. Bias and RMSD recovery is plotted on the size of the true, simulated gap in a normal, equal-variance scenario. Curves are smoothed by averaging with nearest neighbors ( $\pm 0.02$ ). PLI = piecewise linear interpolation; MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit; ADTPAC = average difference in transformed percentages above a cut score.

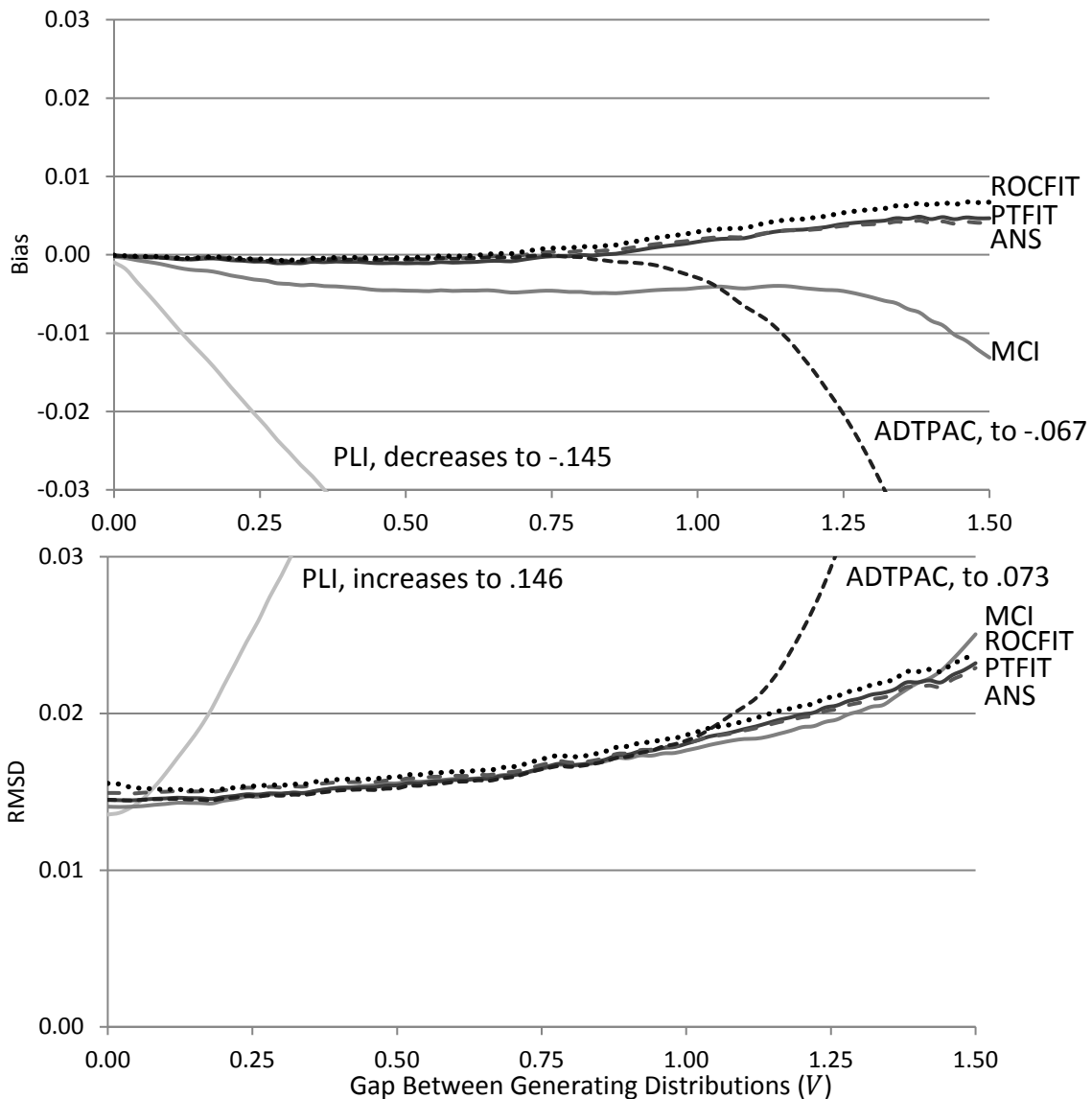


Figure 4. Recovery of the simulated gap in a normal, unequal-variance scenario



*Figure 4.* Bias and Root Mean Squared Deviation (RMSD) of six candidate gap estimation approaches using only three paired cumulative proportions from simulated data. Bias and RMSD recovery is plotted on the size of the true, simulated gap in a normal, unequal-variance scenario. Curves are smoothed by averaging with nearest neighbors ( $\pm 0.02$ ). PLI = piecewise linear interpolation; MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit; ADTPAC = average difference in transformed percentages above a cut score.

Figure 5. Recovery of the simulated gap in a lognormal scenario



*Figure 5.* Bias and Root Mean Squared Deviation (RMSD) of six candidate gap estimation approaches using only three paired cumulative proportions from simulated data. Bias and RMSD recovery is plotted on the size of the true, simulated gap in a lognormal scenario. Curves are smoothed by averaging with nearest neighbors ( $\pm 0.02$ ). PLI = piecewise linear interpolation; MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit; ADTPAC = average difference in transformed percentages above a cut score.

Figure 6. Recovery of the real gap under broadly and narrowly spaced cut-score scenarios.

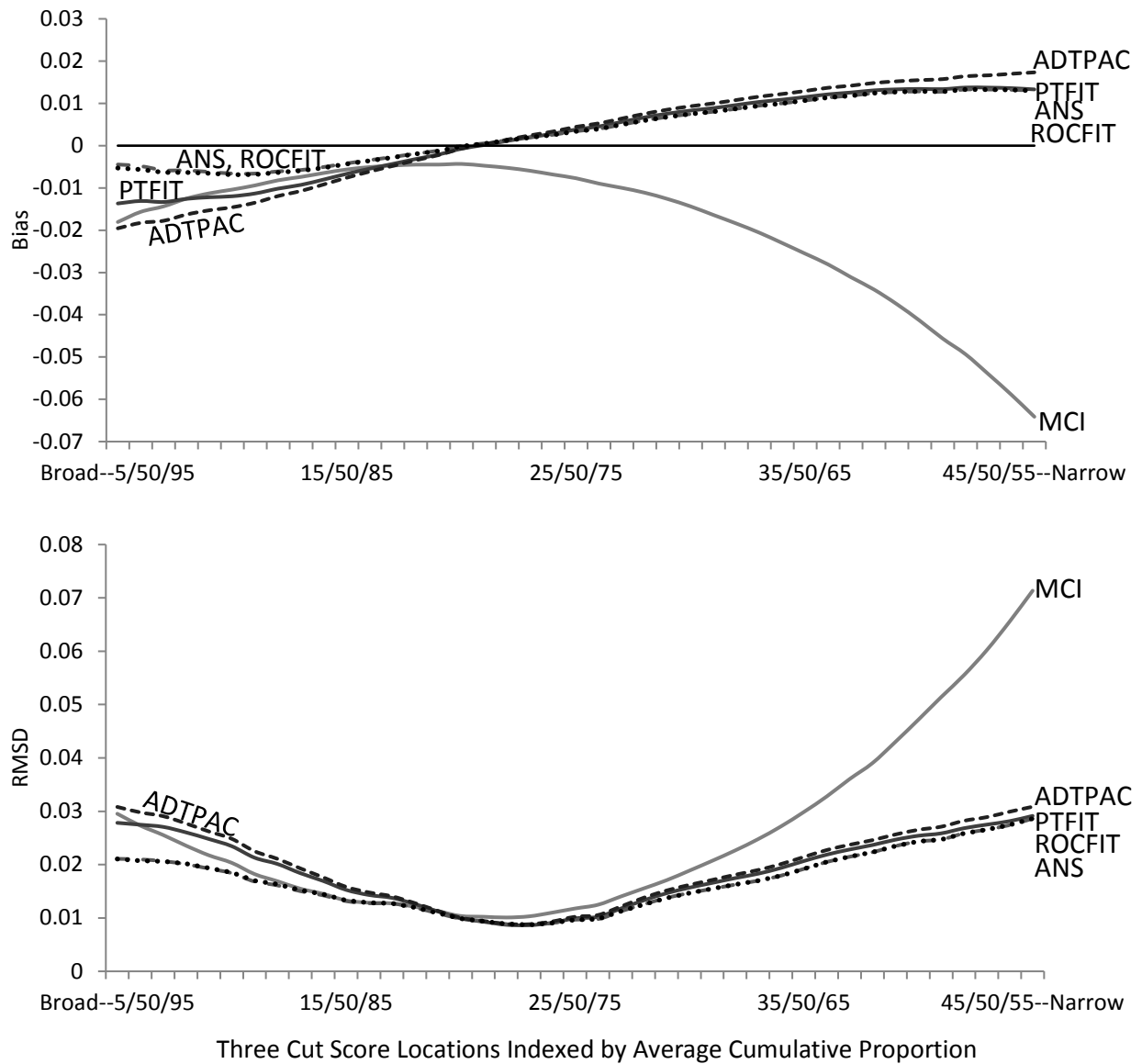


Figure 6. Bias and Root Mean Squared Deviation (RMSD) of four candidate gap estimation approaches using only three paired cumulative proportions from real data from the National Assessment of Educational Progress. Bias and RMSD recovery of the real gap is plotted over broadly-spaced and narrowly-spaced cut score scenarios. MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit.

Figure 7. Recovery of the real gap under less and more stringent cut-score scenarios.

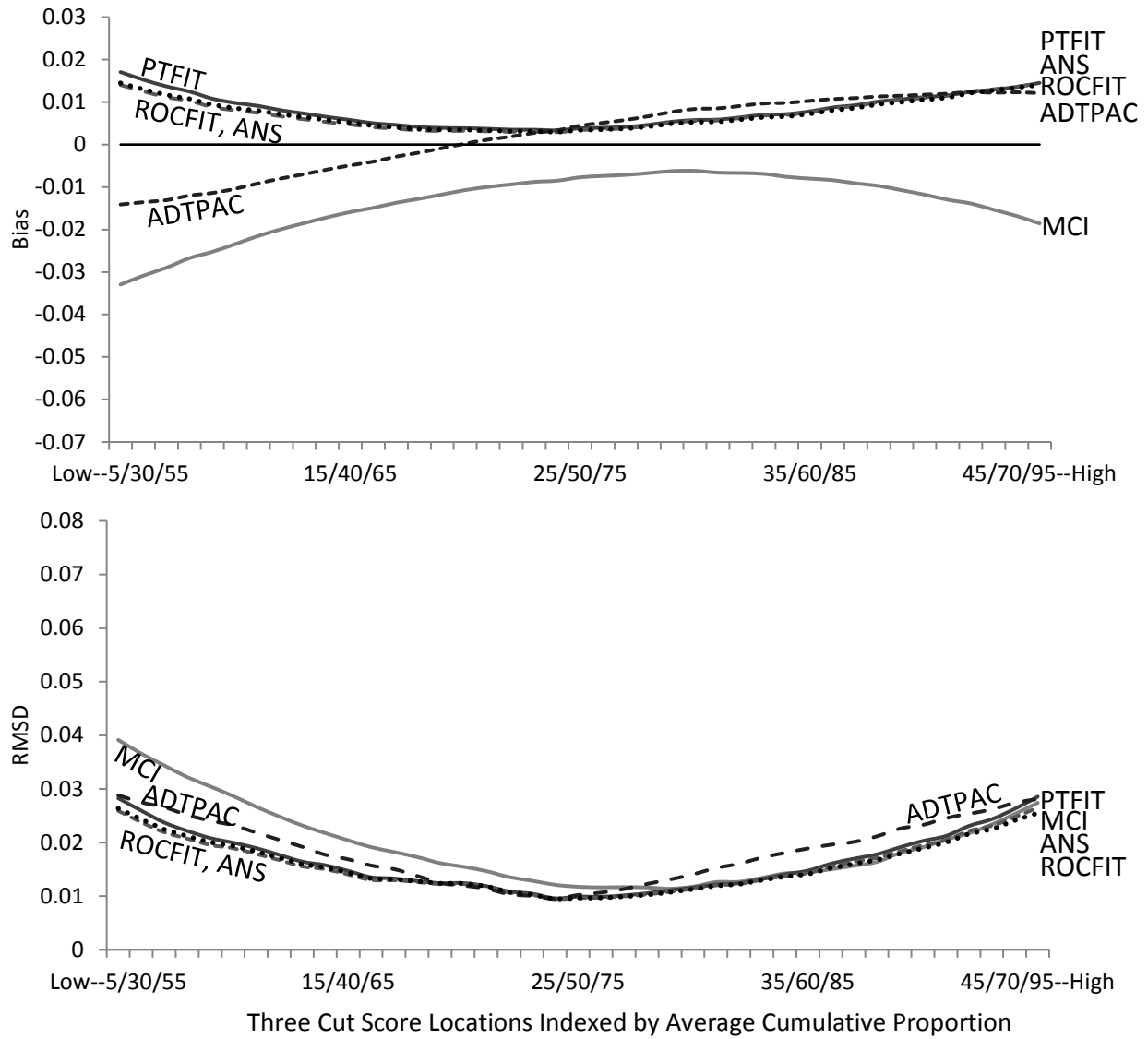


Figure 7. Bias and Root Mean Squared Deviation (RMSD) of four candidate gap estimation approaches using only three paired cumulative proportions from real data from the National Assessment of Educational Progress. Bias and RMSD recovery of the real gap is plotted over less and more stringent cut score scenarios. MCI = monotone cubic interpolation; PTFIT = probit transform, fit, then inverse-transform; ANS = adjusted normal shift; ROCFIT = receiver operating characteristic curve fit.