Practical Issues in Estimating Achievement Gaps from Coarsened Data

Sean F. Reardon
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

February, 2014

Direct correspondence to sean.reardon@stanford.edu. This research was supported by a grant from the Institute of Education Sciences (R305D110018) to Stanford University (Sean F, Reardon, Principal Investigator). We thank Demetra Kalogrides for excellent research assistance, and Ed Haertel, Erica Greenberg, Rachel Valentino, Ken Shores, and Katherine Castellano for their helpful feedback on earlier drafts of this work. The opinions expressed are ours and do not represent views of the Institute or the U.S. Department of Education. We claim responsibility for any errors.

Practical Issues in Estimating Achievement Gaps from Coarsened Data

Abstract

Ho and Reardon (2012) present methods for estimating achievement gaps when test scores are coarsened into a small number of ordered categories, preventing fine-grained distinctions between individual scores. They demonstrate that gaps can nonetheless be estimated with minimal bias across a broad range of simulated and real coarsened data scenarios. In this paper, we extend this previous work to obtain practical estimates of the imprecision imparted by the coarsening process and of the bias imparted by measurement error. In the first part of the paper, we derive standard error estimates and demonstrate that coarsening leads to only very modest increases in standard errors under a wide range of conditions. In the second part of the paper, we describe and evaluate a practical method for disattenuating gap estimates that are biased toward zero due to measurement error.

Introduction

Ho and Reardon (2012) consider the challenge of measuring the "achievement gap" between two population groups when achievement is reported in a small number of ordinal categories, rather than on a many-valued, more continuous scale. Examples of such "coarsened" data include, for example, when we do not know students' exact test scores, but only in which of, say, four or five ordered proficiency levels their scores fall. These levels may have descriptors similar to the National Assessment of Educational Progress (NAEP) achievement levels: "below basic," "basic," "proficient," and "advanced," or they may simply be numeric, as in the 1 to 5 scale of the Advanced Placement examinations. More generally, the problem Ho and Reardon consider is that of constructing a summary statistic that describes the difference between two distributions measured on a common, continuous scale when only the coarsened data are available. In such cases, familiar statistics for comparing two distributions, such as the standardized difference in means, are not available, because the means and standard deviations of the continuous distributions cannot be readily estimated from the observed data. Coarse data scenarios are common beyond education as well, from political science, where opinions are routinely measured on ordered scales, to health, where Apgar scores, cancer stages, and health questionnaires all represent coarse measurement scales.

To address this problem, Ho and Reardon (2012) propose an approach to constructing a readily interpretable measure of the difference in two distributions given only coarse ordinal data. In their approach, the target statistic for comparing two distributions of some random variable x is the V-statistic, which is defined as a monotone transformation of $P_{a>b}$, the probability that a randomly chosen observation from distribution a has a higher value of x than that of a randomly chosen observation from distribution b:

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

where Φ^{-1} is the probit function, the inverse of the cumulative distribution function for the standard normal distribution. The V-statistic has several desirable features. First, it is invariant under monotone transformations of x, because $P_{a>b}$ depends only on the ordered nature of x. Second, if x is normally distributed in both distributions, with equal or unequal variances, then V is equal to d, the standardized mean difference between the two distributions. That is,

$$V = d \equiv \frac{\mu_a - \mu_b}{\sigma_p}, \tag{2}$$

where μ_a and μ_b are the mean of the two groups' distributions and $\sigma_p \equiv \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2}}$ is the pooled standard deviation. This connection between d and V supports an interpretation of V as a transformation-invariant effect size. As long as some transformation exists that renders x normally distributed in both a and b, an assumption known as "respective normality" (Ho & Haertel, 2006) or "binormality" (e.g., Green & Swets, 1966), V will equal d when d is estimated using those appropriately transformed scores.

A third feature of V that makes it particularly appealing is that it can be estimated quite accurately from highly coarsened data. Ho and Reardon (2012) show that it is possible to recover unbiased estimates of V from coarsened data, even when the observed data include only four ordered categories. They also show that recovery of V is robust under coarsened data scenarios even when the assumption of respective normality is violated. In a different context, Hanley (1988) demonstrated a similar robustness of goodness of fit statistics to violations of the binormal assumption. These features of V is very useful, given the ubiquity of situations in the behavioral sciences in which underlying continuous data are recorded in small numbers of ordered categories.

Ho and Reardon (2012) do not, however, address two important issues in estimating V. First, they do not assess the sampling variance of estimates of V (whether based on continuous or coarsened data), nor do they provide any method for constructing standard errors of such

estimates. Standard errors are necessary for statistically warranted inferences regarding changes in V over time and differences in V between contexts or groups. Second, Ho and Reardon (2012) do not address the potential implications of measurement error for the estimation of V. Although measurement error will tend to attenuate V, classical approaches to correcting gaps for measurement error attenuation may not hold when the underlying metric is not assumed to possess defensible equal-interval properties.

In this paper, we build on the work of Ho and Reardon (2012) and address these two issues in turn. First, we consider the sampling variance of estimates of V. We begin this part of the paper by reviewing the formulae for the sampling variance of estimates of d when the distributions are normal but when their variances must be estimated. This provides a benchmark for assessing the relative magnitude of the sampling variance of $\hat{V}$. We then describe, using simulations, the sampling variance of several estimators of V, including both those that require continuously-measured test score data and those that require only coarsened data. Building on the results of these simulations, we describe methods of computing standard errors of $\hat{V}$, and we describe the amount of precision that is lost in common coarsened data scenarios.

In the second part of the paper, we evaluate methods for correcting gap estimates to take into account the attenuating influence of measurement error. We review conventional methods for disattenuating standardized mean differences like d, then we extend this review to methods for disattenuating V in full and coarsened data scenarios. We test the robustness of a simple disattenuation approach that requires only the reported reliabilities of the test, rather than raw or item-level data.

In both parts of the paper, we relate the methods to features of tests and tested samples observed in practice. In the first part, we review cut score locations in state test score distributions that we observe in practice. This allows us to assess the increase in sampling variability for realistic data coarsening scenarios. We find that the added imprecision is often small relative to the

sampling variability of estimators based on uncoarsened data. In the second part, we review reliability coefficients from operational state testing programs to predict the amount of attenuation bias present in gap estimates and cross-test gap comparisons. We intend for this balance of theoretical results with practical context to support not only the theoretical development of these procedures but also an understanding of the practical difference that they may make in practice.

Notation and Definitions

We have two populations, denoted a and b. Let $F_a(x)$ and $F_b(x)$ denote the cumulative density functions of some random variable x in groups a and b, respectively. In some of the discussion below we will consider K "threshold" values of x, denoted $x_1 < x_2 < \cdots < x_K$. We denote the proportion of cases in group a with values of $x < x_k$ as $p_a^k = F_a(x_k)$. We define $p^k = \frac{1}{2}\left(p_a^k + p_b^k\right)$. We are interested in the "gap" in x between groups a and b. By "gap" we mean a measure of the difference in central tendencies of the two distributions, expressed on a metric that is "standardized" in the sense that differences in central tendencies are expressed with respect to the spread of the distributions.

First, we consider the case where x is normally distributed within both a and b (albeit with different means and standard deviations):

$$x|a \sim N(\mu_a, \sigma_a^2)$$
$$x|b \sim N\left(\mu_b, \sigma_b^2\right). \tag{3}$$

In this case, $F_a(x) = \Phi\left(\frac{x-\mu_a}{\sigma_a}\right)$ and $F_b(x) = \Phi\left(\frac{x-\mu_b}{\sigma_b}\right)$, where Φ is the cumulative normal density function.

We define the pooled standard deviation of x as

$$\sigma_p = \sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}.$$

(4)

Second, we consider the case where x is *respectively normal*, meaning that there is some increasing monotonic function f such that $x^* = f(x)$ is normally distributed within both a and b:

$$x^*|a \sim N(\mu_a^*, \sigma_a^{*2})$$

$$x^*|b \sim N(\mu_b^*, \sigma_b^{*2})$$

(5)

In this case, the pooled standard deviation in the metric defined by f will be

$$\sigma_p^* = \sqrt{\frac{\sigma_a^{*2} + \sigma_b^{*2}}{2}}.$$

(6)

This unweighted, pooled, within-group standard deviation differs from two alternative specifications used in the literature. The first is the standard deviation of the mixture, or total distribution, and the second is the weighted, pooled, within-group standard deviation. We choose the unweighted option both because large-scale educational test score distributions commonly allow for decisive rejection of null hypotheses that variances are equal across groups (implying that one should not use the weighted, pooled, within-group standard deviation) and because it is insensitive to sample size (unlike the mixture distribution). If two groups have different population standard deviations, the gap between them should not logically vary as a function of their relative sample sizes.

Gap Measures

A useful measure of the difference in central tendencies of two distributions relative to the spread of the distributions is Cohen's d (Cohen 1988; Hedges and Olkin 1985), the standardized

difference in means between groups a and b:

$$d = \frac{\mu_a - \mu_b}{\sigma_p}. \tag{7}$$

An alternate measure of the difference between two distributions is the V statistic (Ho and Haertel, 2006). V is defined as

$$V = \sqrt{2}\Phi^{-1}(P_{a>b}), \tag{8}$$

where $P_{a>b} = \int_0^1 F_b\left(F_a^{-1}(q)\right) dq$ is the probability that a randomly chosen observation from group a has a value of x higher than that of a randomly chosen member of group b. An important property of V is that, if x is normally distributed in both groups a and b, then $V = d$ (Ho and Haertel 2006; Ho and Reardon 2012). However, a non-linear monotonic transformation of x will, in general, alter d but leave V unchanged. This is because V depends only on the ordered ranking of x; d depends on the interval nature of x. The metric-free nature of V renders it a more robust measure of distributional differences than d. If x does not have a natural interval-scaled metric (or if it has one, but it is not expressed in that metric), d will be dependent on the arbitrary metric in which x is measured, but V will not.

While both V and d can be easily estimated from micro-data, both can also be readily estimated from certain types of aggregate data. If we have estimates of the group-specific means (μ_a and μ_b) and standard deviations (σ_a and σ_b), we can estimate d by substituting these estimates into Equations (2) and (5) above. If we have estimates of the proportions of each group that fall below a set of one or more threshold values of x ($p_a^1, .. p_a^K, p_b^1, \dots, p_b^K$), we can estimate V using the methods described by Ho and Reardon (2012).

Part 1: Sampling Variance of Gap Measure Estimates

Our aims in this section of this paper are 1) to describe the sampling variance of a set of estimators of d and V; and 2) to describe methods for computing standard errors for estimates of these. We first consider the sampling variance of estimators of d.

Suppose we have a sample of size n, with n_a cases drawn from group a and n_b cases drawn from group b, so that $n = n_a + n_b$. Let $p = n_a/n$ denote the proportion of cases from group a; let $r = \sigma_a^2/\sigma_b^2$ denote the ratio of the population variances of x in groups a and b. Let $\hat{\mu}_a$ and $\hat{\mu}_b$ be the sample means of x in groups a and b, and let $\hat{\sigma}_a$ and $\hat{\sigma}_b$ be the estimated standard deviations of x in groups a and b, respectively.

Sampling Variance of Estimators of d

If the pooled standard deviation σ_p is known (rather than estimated from the sample), then we estimate d with

$$\hat{d} = \frac{\hat{\mu}_a - \hat{\mu}_b}{\sigma_p}. \tag{7}$$

The sampling variance of this estimator will be

$$Var(\hat{d}) = \frac{1}{\sigma_p^2}\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right) = \frac{2(r + p - pr)}{np(1-p)(1+r)}. \tag{8}$$

Note that if $r = 1$ or $p = \frac{1}{2}$, $Var(\hat{d}) = \frac{1}{np(1-p)}$. Note also that the sampling variance in this case does not depend on the magnitude of d.

If the pooled standard deviation σ_p is not known, however, it must be estimated from the sample, which will add some additional sampling variance to the estimator. Specifically, we show in the appendix that if we estimate d as the difference in estimated means divided by the estimated pooled standard deviation,

$$\hat{d}^* = \frac{\hat{\mu}_a - \hat{\mu}_b}{\hat{\sigma}_p},$$

(9)

then the sampling variance of this estimator will be approximately

$$Var(\hat{d}^*) \approx \lambda \cdot Var(\hat{d}),$$

(10)

where

$$\lambda = 1 + \frac{d^2[p + (1-p)r^2]}{4(1+r)[p + (1-p)r]} + \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2}.$$

(11)

Because $r \geq 0$ and $0 < p < 1$, it follows that $\lambda > 1$. Note that, as n get large, the third term in Equation (11) goes to zero, but the second term does not. Note also that the second term depends on the size of the true gap, d. When the true gap is large, the sampling variance inflation factor grows. If $\sigma_a^2 = \sigma_b^2$ (so that $r = 1$), then the sampling variance is simply

$$Var(\hat{d}^*) \approx \left(1 + \frac{d^2}{8} + \frac{1}{8np(1-p)}\right)\left(\frac{1}{np(1-p)}\right).$$

(12)

So, when $r = 1$ and n is moderately large, estimating σ_p^2 increases the sampling variance by a factor of $\left(1 + \frac{d^2}{8}\right)$ (and so should increase the standard error by a factor of $\sqrt{1 + \frac{d^2}{8}}$, where d is the true value of Cohen's d). For $d = 1$, for example, the standard error would increase by about 6% due to the estimation of the variances (because $\sqrt{1.125} = 1.061$).

Estimators of V Using Complete (Non-coarsened) Data

We first consider the case where we have micro-data, that is, individual observations of x for each of the n cases in the sample. With such data, we can estimate V in several ways.

First, we can estimate V non-parametrically, by estimating $P_{a>b}$ from the micro data. This is done by constructing all possible pairs of observations $\{x_a, x_b\}$ and computing the proportion of pairs in which $x_a > x_b$. With noncontinuous variables, we can break ties by adding half the cases in which $x_a = x_b$:

$$\hat{P}_{a>b} = \frac{1}{n_a n_b} \sum_{x_a} \sum_{x_b} \left[I[x_a > x_b] + \frac{1}{2} I[x_a = x_b] \right] \tag{13}$$

We then construct a non-parametric estimate of V from the full data as

$$\hat{V}_{np}^{full} = \sqrt{2}\Phi^{-1}\left(\hat{P}_{a>b}\right). \tag{14}$$

This is a non-parametric estimator of V because it requires no distributional assumptions. This estimator does not provide a standard error, however.

Second, under the assumption that the two distributions are respectively normal (as defined in Equation 5 above), we can use methods described by Ho and Reardon (2012) to estimate V from the complete data. We focus here on two methods they describe, the "PTFIT" ("probit-transform-fit-inverse-transform") method and the maximum likelihood "ROCFIT" method (which we refer to as the maximum likelihood (ML) method hereafter). We also describe a third method, available under the assumption that the two distributions are respectively normal and have equal variances in the metric in which they are both normal, and which Ho and Reardon refer to as ADTPAC ("average difference in transformed percentages-above-cutscore"). Although Ho and Reardon describe these methods as applied to coarsened data, they can readily be applied to complete data (instead of having only a small number of ordered categories, we now have n distinct categories, one for each observed test score value).

Ho and Reardon (2012) describe the ML, PTFIT, and ADTPAC methods in detail; we provide only a quick summary here. Relying on the assumption of respective normality, the ML method

uses maximum likelihood to estimate the relative means and standard deviations of the normal distributions of x in groups a and b most likely to have given rise to the observed numbers of each group scoring below each of K observed values. Cohen's d is then computed from these estimated means and standard deviations; because $V = d$ under respective normality, this provides an estimate of V. The PTFIT method also relies on the respective normality assumption, and makes use of the fact that a plot of $\Phi^{-1}\big(F_b(x)\big)$ against $\Phi^{-1}\big(F_a(x)\big)$ will be linear if F_a and F_b describe respectively normal distributions. Moreover, the slope and intercept of this line are determined by the relative means and standard deviations of the two distributions in the metric in which both are normal, and so are sufficient information to compute V. The PTFIT method fits this line and computes V from its estimated slope and intercept. The ADTPAC method is similar to the PTFIT method, but relies on a stronger assumption—that the two distributions are equivariant respectively normal (meaning they would have the same variance in the metric in which both were normal). This constraint implies that the slope of the line fitted through the probit-transformed percentiles must have a slope of 1. Thus, the ADTPAC method estimates V using the same approach as the PTFIT method, but does so under the constraint that the fitted line has a slope of 1.

Ho and Reardon (2012) showed that, under conditions of respective normality, both the ML and PTFIT methods produce unbiased estimates of V. When the two distributions have equal variance in their normal metric, the ADTPAC method is also unbiased. The PTFIT method, however, is computationally simpler (and considerably faster) to implement than the ML method, particularly when the scores are not highly coarsened. The maximum likelihood method, however, can readily provide standard errors, while the PTFIT method does not. We refer to the PTFIT estimator that uses complete data as $\hat{V}_{ptfit}^{full}$; we refer to the corresponding ML estimator that uses complete data as $\hat{V}_{ml}^{full}$;[1] and we refer to the ADTPAC estimator that uses complete data as $\hat{V}_{adtpac}^{full}$.

[1] Note that in our simulations below we do not estimate $\hat{V}_{ml}^{full}$ using the complete data, because it is computationally very intensive. Rather, we coarsen the observed data into 20 ordered categories of equal

Third, Ho and Reardon (2012) describe an additional set of methods for estimating V that do not require the assumption of respective normality. These methods typically rely on some other parametric assumption (such as an assumption that the relationship between F_b and F_a can be described by a piecewise cubic function). In general, Ho and Reardon found that these methods do not perform as well as the ML, PTFIT, and ADTPAC methods, even when the distributions are not respectively normal. Moreover, examination of many real world test score distributions suggest that many distributions are sufficiently close to respectively normal that the ML, PTFIT, and ADTPAC methods are nearly unbiased. As a result, we do not consider these other methods further in this paper.

Sampling Variance of Estimators of V Using Complete (Non-coarsened) Data

Although we do not have analytic methods of computing the sampling variances of the $\hat{V}^{full}$ estimators, we can examine their sampling variances using simulations. Below we report the results of a set of such simulations. In each simulation, we select values of V, r, and p, and then draw a sample of size $n = 2000$ from a population in which $n_a = pn$ cases are drawn from a normal distribution with mean 0 and variance 1, and in which $n_b = (1-p)n$ cases are drawn from a normal distribution with mean $V[(1+r)/2r]^{1/2}$ and variance $1/r$.[2] We conduct simulations in which we set $V = 0, 0.5, 1.0, 1.5,$ or 2.0; we set $p = 0.90, 0.75, 0.67,$ or 0.50; and we set $r = 0.67, 0.80, 1.0, 1.25$ or 1.5.[3] For each combination of V, p, and r, we draw 1000 samples, and then compute $\hat{d}^*$, $\hat{V}_{np}^{full}$, $\hat{V}_{ml}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{adtpac}^{full}$ (the last we compute only when $r = 1$). We then

size, and then estimate $\hat{V}_{ml}^{coarse}$ (described below) from the sample counts in these 20 categories. There is virtually no gain in precision by using more categories, but great loss in computational time, because the ML estimator must estimate $K+1$ parameters, where K is the number of categories ($K-1$ threshold scores, plus a parameter describing the difference in means and a parameter describing the ratio of the variances in the two distributions).

[2] These parameters ensure that the true population gap is $d \equiv [V[(1+r)/2r]^{1/2} - 0]/[(1+1/r)/2]^{1/2} = V$.

[3] For reference, note that NAEP data yield estimated values of black-white and Hispanic-white V gaps ranging from roughly 0.25 to 1.0; the proportion black across states ranges from 2 to 68%; and the black/white variance ratio r ranges from roughly 0.6 to 1.2.

examine the standard deviations of each of these estimators over the 1000 samples.

Figure 1 about here

Figure 1 summarizes the results of these simulations. Several things are striking about this figure. First, the sampling variances of the five estimators are virtually identical to each other within each scenario. This is an important finding because it means that we can use Equation (10) above, which describes the sampling variance of $\hat{d}^*$ as a function of d, p, r, and n, to approximate the sampling variances of $\hat{V}_{np}^{full}$, $\hat{V}_{ml}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{adtpac}^{full}$ as

$$Var\left(\hat{V}^{full}\right) = \hat{\lambda} \cdot \frac{2(\hat{r} + p - p\hat{r})}{np(1-p)(1+\hat{r})}$$

$$= \left[1 + \frac{\hat{V}^{full^2}[p + (1-p)\hat{r}^2]}{4(1+\hat{r})[p + (1-p)\hat{r}]} + \frac{p + (1-p)\hat{r}^2}{2np(1-p)(1+\hat{r})^2}\right] \frac{2(\hat{r} + p - p\hat{r})}{np(1-p)(1+\hat{r})}. \quad (15)$$

Using Equation (15) to compute a standard error of $\hat{V}^{full}$ requires an estimate of r (p and n are observed so need not be estimated). The ML and PTFIT methods provide estimates of r; the ADTPAC method assumes $r = 1$, so Equation (15) provides a method of obtaining standard errors of $\hat{V}_{ml}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{adtpac}^{full}$ (note, however, that the ML estimator provides a standard error directly, so we do not need this approach). The $\hat{V}_{np}^{full}$ estimator, however, does not provide a method of estimating r. However, Equation (15) is only valid, and r is only defined, when the distributions are respectively normal (r is the ratio of variances of the distributions when the variances are computed in the metric in which both distributions are normal). If we are able to assume respective normality, then we can use the PTFIT or ML estimator to obtain an estimate of r to using Equation (15).

The second thing to note about Figure 1 is that the sampling variance of $\hat{d}^*$ and the $\hat{V}^{full}$ estimators is smallest when $p = 0.5$, $r = 1$, and $V = 0$. The sampling variance is most sensitive to departures from $p = 0.5$; as the group sample sizes grow more uneven, the sampling variance

increases substantially.

Sampling Variance of Estimators of V Using Coarsened Data

When we have coarsened data, we can use the ML, PTFIT, and ADTPAC methods to estimate gaps, under the assumption that the distributions are respectively normal. We refer to these estimators as $\hat{V}_{ml}^{coarse}$, $\hat{V}_{ptfit}^{coarse}$, and $\hat{V}_{adtpac}^{coarse}$. Although these estimators will be unbiased under the assumption of respective normality, they may be substantially less efficient than the $\hat{V}^{full}$ estimators, if the coarsening results in a significant loss of information. As with the $\hat{V}^{full}$ estimators, we can investigate the sampling variance of these estimators using simulations.

We conduct simulations as above, but after drawing each sample we coarsen the data by computing the proportions of cases in each group that fall within each of four ordered categories. We define the categories such that, in the full population, the threshold values of x defining these categories fall at three pre-specified percentiles. We run simulations using 14 different sets of percentiles, ranging from widely-spaced (5th, 50th, and 95th percentiles) to narrowly spaced (45th, 50th, and 55th percentiles), and including some sets that are centered on the 50th percentile and others that are all to one side of the median (e.g., 20th, 30th, and 40th percentiles).

The results of these simulations are shown in Table 1. Each cell in Table 1 reports the ratio of the sampling standard deviation of $\hat{V}_{ml}^{coarse}$ to that of $\hat{V}_{ml}^{full}$ under the same conditions (Table 1 shows only sampling variation for $\hat{V}_{ml}^{coarse}$; the results are nearly identical for $\hat{V}_{ptfit}^{coarse}$ and $\hat{V}_{adtpac}^{coarse}$). Values close to 1 indicate that the sampling variance of the estimates based on the coarsened data are not substantially different than that of the estimates based on the complete data: little information is lost from the coarsening in these cases. Values much greater than 1 indicate that a great deal of precision is lost in the coarsening. The rows are sorted, from lowest to highest, by the average value of the sampling variance ratio across the nine displayed scenarios.

Table 1 about here

Table 1 shows that the sampling variance of $\hat{V}_{ml}^{coarse}$ is minimized, relative to that of $\hat{V}_{ml}^{full}$ when the cutscores are placed near the 20th, 50th, and 80th percentiles of the unweighted combination of the two distributions. The standard errors of $\hat{V}_{ml}^{coarse}$ in this case are generally only 3-7% larger than the standard errors of $\hat{V}_{ml}^{full}$, implying there is very little loss of information when the scores are placed widely (but not too widely) and symmetrically. We estimate gaps almost as precisely with coarsened data in such cases as we can with complete data. Indeed, so long as the cutscores are placed relatively symmetrically and not extremely narrowly or extremely widely, the standard errors of $\hat{V}_{ml}^{coarse}$ are generally less than 10% larger than those of $\hat{V}_{ml}^{full}$. Cutscores at the 10/50/90th percentiles, the 30/50/70th percentiles, the 10/40/70th percentiles, and even the 5/40/75th percentiles all provide little loss of information. However, when the cutscores become too close together and/or too far from symmetrical, the standard errors of $\hat{V}_{ml}^{coarse}$ are significantly larger than those of $\hat{V}_{ml}^{full}$.

Figure 2 shows the location of the percentiles of the high and low cutscores for state accountability tests over the years 1997-2011. In the figure, the diamonds indicate the points that correspond to the simulations shown in Table 1. Each small dot represents a state-grade-year-subject combination. Every such combination for which we were able to obtain data is represented here (though we only show observations in which there were three cutscores; the patterns are similar when there are 2 or 4 cutscores). As is evident in Figure 3, there are few cases in which the high and low percentiles are separated by fewer than 50 percentile points. There are, however, many cases in which one or the other of the cutscores is near the edge of the figure, meaning that there is at least one cutscore that provides little information (a category that has almost no one in it provides almost no information on the relative distributions). Nonetheless, there are many cases in which the cutscores fall in ranges likely to lead to no more than a 10-20% increase in standard errors, relative to what we would find if we had access to full data.

Figure 2 about here

Tables 2 and 3 present the results of several additional simulations. Table 2 reports the sampling variance of $\hat{V}_{ml}^{coarse}$ under different sample sizes. Table 2 illustrates that the sampling standard deviation of $\hat{V}_{ml}^{coarse}$ is roughly proportional to $1/\sqrt{n}$, just as is that of Cohen's d. Moreover, the relative loss of precision from the coarsening of the data is roughly the same, regardless of sample size.

Table 3 reports the ratio of the sampling standard deviation of $\hat{V}_{ml}^{coarse}$ to that of $\hat{V}_{ml}^{full}$ when the coarsening uses different numbers of categories. With six categories (5 cutscores), the coarsening of the data increases standard errors by only 2 percent, on average. With three categories (2 cutscores), the coarsening inflates the standard errors by roughly 10 percent. Although the precision of $\hat{V}_{ml}^{coarse}$ is greater when there are more categories, the estimator is surprisingly precise even when there are very few categories. Note that in these simulations we use evenly spaced cutscores. As Table 1 shows, the precision will generally be less with less evenly spaced cutscores.

Summary of Sampling Variance Analysis

In sum, we have shown that, under conditions of respective normality, the sampling variance of the estimators of V based on full information is essentially identical to the sampling variance of more conventional estimators, such as Cohen's d. This sampling variance can be computed easily from the observed data, allowing us to compute standard errors for the V estimators under full information. This is useful, because the ML estimator is the only one of the estimators of V that produces a standard error estimate, but the ML estimator is too computationally intensive to use with uncoarsened data. Equation (15) provides a method of estimating standard errors of V^{full} based on the other estimators.

Second, we have shown that the sampling variance of estimators of V based on coarsened data will often be only slightly larger than that of the estimators using full information. This

suggests that V statistics estimated from readily-available coarsened data can be nearly as precise as conventional Cohen's d estimators (based on means and standard deviations) and as V statistics based on full information. This means it is possible to recover unbiased and reasonably precise estimates of distributional differences even when the data are highly coarsened.

These findings lead us to several recommendations. First, when using uncoarsened data, one can plot $\Phi^{-1}\big(F_b(x)\big)$ against $\Phi^{-1}\big(F_a(x)\big)$; if this plot shows a roughly linear pattern, then the assumption of respective normality is met. One can then estimate the gap from the complete data using either the PTFIT estimator (if variances are not assumed equal) or the ADTPAC estimator (if variances are assumed equal). Both will produce unbiased estimates, and their standard errors can be computed from Equation (15). Although the ML estimator can also be used in this case—it is unbiased and produces standard errors—it can be very computationally intensive, and appears to be no more efficient than the PRFIT or ADTPAC estimators.

If the plot of $\Phi^{-1}\big(F_b(x)\big)$ against $\Phi^{-1}\big(F_a(x)\big)$ is far from linear, indicating that the respective normality assumption is not met, one case still use the PTFIT or ML estimators; Ho and Reardon (2012) show that these estimators are only slightly biased even when the distributions are far from respectively normal. Alternately, one can use the non-parametric estimator, which makes no distributional assumptions. The disadvantage to this approach is that Equation (15) no longer applies, so computing a standard error would require bootstrapping. Alternately, one could compute $\hat{V}_{np}^{full}$ and $\hat{V}_{ptfit}^{full}$ (or $\hat{V}_{ml}^{full}$); if the non-parametric and PTFIT or ML estimates of V are close, this suggests the bias due to the lack of respective normality is not large, so that the PTFIT or ML estimator can be used.

Second, when using coarsened data, one can be confident that the precision of the estimates of V will not be significantly weakened due to coarsening, so long as the cutscores are not close together and/or highly asymmetric. When using coarsened data, the ML estimator is the most preferable, because it produces standard error estimates. Although our other methods have

sampling variance similar to the ML method under coarsened data, we have no straightforward way to compute their standard errors.

Part 2: Disattenuating V for Measurement Error

Gap statistics like d and V are expressed in terms of units of variation. Observed variation is increased by measurement error, thus d and V will be biased toward zero when measurement error is present. In Part 2 of this paper, we review corrections for measurement error for d and V when full, uncoarsened data are available. Then, we evaluate a practical method for correction when full data are not available that relies only on reported reliabilities or subgroup reliabilities that are likely to be readily available.

Disattenuation of d

We begin with some additional notation. In this section of the paper, we consider x to be an error-prone measure of some true score t (i.e., $x = t + \epsilon$, where $\epsilon \perp t$ and $E[\epsilon] = 0$). The reliability of x, which we denote ρ, is the ratio of true score variance, σ_t^2, to observed score variance, σ_x^2. The standard deviation of true scores in group a is therefore $\sigma_{t_a} = \sqrt{\rho_a}\sigma_a$. It is well understood that measurement error increases observed score standard deviations and biases effect size estimates like d toward zero (e.g., Hedges & Olkin, 1985). Because t and x have the same subgroup means, but the within-subgroup standard deviations of x are larger than those of t, the value of d for the true score t will be larger (in absolute value) than the value of d for the error-prone measure x. We will denote these two values of d as d_t and d_x, respectively, If the subgroup reliabilities of x in both group a and b are known, then d_x can be disattenuated to express the gap in terms of pooled standard deviation units of the true score t:

$$d_t = \frac{\mu_a - \mu_b}{\sqrt{\frac{\rho_a \sigma_a^2 + \rho_b \sigma_b^2}{2}}} = \frac{d_x}{\sqrt{\frac{\sigma_a^2 \rho_a + \sigma_b^2 \rho_b}{\sigma_a^2 + \sigma_b^2}}} = \frac{d_x}{\sqrt{\tilde{\rho}}}. \quad (16)$$

Here, $\tilde{\rho}$ is a weighted average of the reliabilities in groups a and b, where the weights are proportional to the variances of x in the two groups.

Simplified Expressions of $\tilde{\rho}$

Estimating $\tilde{\rho}$ and d_t requires subgroup reliabilities and subgroup variances that are available in full data scenarios. However, given the data-poor context that motivates the methods in this paper, we use real data scenarios to demonstrate here that simplified expressions for $\tilde{\rho}$ often suffice. First, we show that $\tilde{\rho}$ is well approximated by $\bar{\rho} = \frac{\rho_a + \rho_b}{2}$ when either $r = \frac{\sigma_a^2}{\sigma_b^2} \approx 1$ or $\rho_a \approx \rho_b$. Then we show that reliability estimates are generally similar among different subgroups and between subgroups and all students, $\bar{\rho} \approx \rho_a \approx \rho_b \approx \rho_{all}$.

When subgroup score variances are equal across groups ($r = 1$), then $\tilde{\rho} = (\rho_a + \rho_b)/2 = \bar{\rho}$. As a real-data reference, we note that state-level NAEP racial variance ratios (the ratio of the variance of black or Hispanic to white test score distributions) since the 2003 administration have averaged 1.10 and 1.14 for Black-White and Hispanic-White comparisons, respectively. Only 2.5 percent (51 of 2,040 computed variance ratio) are greater than 1.5 or less than 1/1.5. These average and extreme ratios may seem substantial, however, their practical impact on d_t depends on the difference between ρ_a and ρ_b. Note that

$$\tilde{\rho} - \bar{\rho} = \frac{(r-1)}{2(r+1)}(\rho_a - \rho_b). \quad (17)$$

Thus, $\tilde{\rho} \approx \bar{\rho}$ when either $r \approx 1$ or $\rho_a \approx \rho_b$. Figure 3 addresses these subgroup reliability differences in practice. We gathered publicly available subgroup reliability statistics from technical

manuals for state testing programs, from 38 states, grades 3-8, mathematics and reading/English language arts, for all students, White students, Black students, and Hispanic students, from 2009 to 2012. Figure 3 shows the distribution of 1,438 available reliability coefficients from the 38 states, 6 grades, 2 subjects, 4 groups, and 4 years. Almost all states reported classical, internal consistency-type reliability statistics like Cronbach's alpha and, rarely, stratified alpha. A few states reported marginal reliabilities estimated from Item Response Theory models. We do not review the details of this data collection process for space considerations and because the relevant results have straightforward implications.

Figure 3 shows that average subgroup reliabilities are not substantially different from each other, with average White, Hispanic, and Black subgroup reliabilities of .895, .897, and .899. The reliability for all students was higher on average at .903. Although we can reject the null hypothesis that the average reliabilities are equal across subgroups, the practical differences between averages are very small. Addressing the skewed distribution of reliability coefficients by applying a Fisher's z transformation makes no appreciable difference to these findings.

The embedded table in Figure 3 shows correlations between subgroup reliabilities below the diagonal and pairwise root mean square deviations above the diagonal. The RMSDs between subgroups are never greater than .02. In short, although most reliabilities range from .80 to .95 across state-subject-grade-year combinations, pairwise differences between subgroup reliabilities are small in magnitude.

Returning to Equation 17, even granting a variance ratio $r = 1.5$, and even if subgroup reliabilities differ by as much as $\rho_a - \rho_b = 0.10$, the ratio $\sqrt{\tilde{\rho}/\bar{\rho}} < 1.01$ unless $\bar{\rho} < 0.5$, which does not occur in our dataset. Thus, although $\tilde{\rho}$ is ideal, $\bar{\rho}$ will be a very close approximation for the purpose of disattenuating d. Further, Figure 3 shows that, when subgroup reliabilities are not available, using the reliability for all students will also be a reasonable approximation. Average subgroup reliability would have to differ by more than .02 from all-student reliability (an extremely

rare occurrence in practice) to change d_t estimates by much more than 1% (e.g., $\bar{\rho} - \rho_{all} = .02, \sqrt{\bar{\rho}/\rho_{all}} < 1.013$ for $\rho_{all} > 0.7$). Clearly, these data are from the particular context of large-scale achievement testing, and subgroup variances may differ, reliabilities may be lower, or subgroup/all-student reliabilities may show greater differences in other contexts. However, for situations with variance ratios and reliabilities in the ranges we show here, d_t will not differ greatly whether we use the preferable $\tilde{\rho}$, the more approximate $\bar{\rho}$, or the less approximate ρ_{all}.

Reliability in an Ordinal Framework

Conventional expressions of reliability are inherently scale-dependent. Both variances and correlations differ under nonlinear monotone transformations, thus so will reliability coefficients, whether expressed in terms of variances (σ_t^2/σ_x^2) or equivalently as a correlation between replications. When nonlinear monotone transformations are deemed permissible, ordinal frameworks for reliability have been proposed that we review briefly here.

A first approach involves procedures that view the full test score scale as ordinal. Lord and Novick (1968) note that statistics that estimate reliability should be invariant under "a certain class of transformations on the observed scores that do not change the meaning and the use of the scores in a certain testing situation" (p. 215). Following this logic, when transformations of observed scores are permissible, Lord and Novick propose Spearman rank correlations (correlations of ranks) as a framework for expressing reliability, given the invariance of rank correlations under monotone transformations. Reliability estimation can then proceed through split-half reliability estimation techniques (e.g., Haertel, 2006), substituting ranks instead of scores for the split halves. This approach is theoretically useful but impractical in our motivating context, as item level data necessary for split-half analyses are even more difficult to obtain than full score distributions.

A second approach involves procedures that view the scales of individual items as ordinal. Zimmerman, Zumbo, and Lalonde (1993) motivate this work by describing poor performance of

Cronbach's alpha when item scales are skewed. Zumbo, Gadermann, and Zeisser (2007) propose an ordinal version of coefficient alpha for Likert rating scales, where inter-item correlations are calculated as rank or polychoric correlations. Wilcox (1992) proposes a different approach for rank-based item scales that use the "biweight midvariance," a variance statistic that is more robust to scale transformations. This work is thoughtful and has particular utility for Likert-scaled items, but our methods again assume a practical scenario where these item-level data are often dichotomous and, more relevantly, unavailable.

Disattenuation of V With Full Score Distributions

Absent ordinal approaches to reliability, the use of reported reliability statistics can be used to adjust ordinal statistics like V. When full score distributions are available, one approach to addressing measurement error is to shrink individual scores toward their respective group means in accordance with their unreliability:

$$x'_g = \sqrt{\rho_g}(x_g - \mu_g) + \mu_g = (\sqrt{\rho_g})x_g + (1 - \sqrt{\rho_g})\mu_g. \quad (18)$$

The variance of x'_g will equal the variance of the true scores. Following this shrinkage, calculation of ordinal statistics like V can proceed. Reardon and Galindo (2009) take this approach in their measurement of Hispanic-White test score gaps. Note that x'_g is not the same as the least-squares estimate of individual true scores (e.g., Kelley 1947). The latter uses ρ_g in Equation (18) instead of its square root, and, as a result, its variance will underestimate true score variance by a factor of ρ_g.

Disattenuation of V With Coarsened Score Distributions

When full data distribution are available, the two-step procedure described above is possible. The procedure neatly separates the scale-dependent procedures, including the scores, the

reliability estimation, and the adjustment in Equation 18, from the ordinal estimation of V. When only the coarsened score distributions are available, however, the data for which the reliability estimate is relevant is not available. We review two approaches that can be used in this case. First, we can disattenuate the estimated V statistic directly, akin to Equation 16. Second, we can adjust the cumulative proportions that represent the raw data supporting V estimation, akin to Equation 18. Both of these approaches share the assumption that the reliability parameter is relevant to the scale in which score distributions are normal. We evaluate the practical impact of violating this assumption in the subsequent section.

When subgroup distributions are normal, $V = d$, and the ordinal corollary of Equation 16 follows:

$$V_t = \frac{V_x}{\sqrt{\tilde{\rho}}}. \tag{19}$$

Both the ML and PTFIT estimators of V provide an estimate of $r = \sigma_a^2/\sigma_b^2$, the ratio of the group variances of x, in a metric in which the group a and b distributions are normal. Assuming that ρ_a and ρ_b are applicable to the same metric, this is sufficient to compute the average reliability weighted by variance as $\tilde{\rho} = \frac{r\rho_a + \rho_b}{1+r}$. We have already shown in Figure 3 that $\rho_a \approx \rho_b \approx \rho_{all}$. The same approximations for d_t hold for V_t, where $V_x/\sqrt{\tilde{\rho}}$ is preferable, $V_x/\sqrt{\tilde{\rho}}$ is unlikely to deviate by more than 1%, and $V_x/\sqrt{\rho_{all}}$ is unlikely to deviate by more than 1% yet again. The only additional assumption, tested in the next section, is whether these estimates of ρ are applicable to the scale in which subgroup distributions are normal.

An alternative approach involves the disattenuation of cumulative proportions by assuming a distributional form. If we assume score distributions are normal, as assumption consistent with the interpretation of V, then the corrected cumulative proportion below a cut score k for group g is,

$$p_g^{k\prime} = \Phi\left(\Phi^{-1}\left(p_g^k\right)/\sqrt{\rho_g}\right).$$

(20)

Following this correction, we can estimate V_t using these corrected cumulative proportions. Unlike Equation 18, where reliability-based shrinkage occurs on an observed score scale, the score scale in Equation 18 is assumed to be one where score distributions are normal. This scale may be distinct from the scale for which the reliability parameter is actually relevant.

The approaches illustrated by Equations 19 and 20 are equal in expectation if assumptions hold, and they differ neglibly from each other in practice. As a result, we focus on the former approach, given that the analysis of secondary data is more likely to enable disattenuation of a reported V statistic by Equation 19 than correction of all cumulative proportions by Equation 20.

The Scale-Dependence of Reliability

Equations 19 and 20 are combinations of V, an expression that does not vary across transformations of score scales, and ρ_a and ρ_b, reliability parameters that describe the ratios of true score variances to observed score variances—but only for a particular score scale and linear transformations thereof. These expressions implicitly assume that the reliability coefficients are applicable to the respective normal score distributions that underlie the interpretation of V. To evaluate the potential bias imparted by violations of this assumption, we provide an illustrative example that establishes practical limits on the transformation-dependence of a reliability parameter ρ.

Following Equation 5, we assume that the variable x is not necessarily normally distributed in both groups, but a transformation $x^* = f(x)$ renders both distributions normal in x^*. In this illustrative example, we first define a reliability parameter, ρ, that expresses the ratio of true score variance to observed score variance in the non-normal metric, x. Second, we define a reliability parameter, ρ^*, that expresses the ratio of true score variance to observed score variance in the respectively normal metric, x^*, that supports interpretations of V. This parameter is not estimable

in practice, because, without item- and student-level data, we cannot be sure that any reported reliability coefficients are applicable to this normal metric. Third, for convenience in upcoming expressions, we define $g(x^*) = f^{-1}(x^*)$, such that f is the normalizing transformation, and g is the transformation that maps normally distributed scores back to the metric of reported scores.

$$f(x) = g^{-1}(x) \sim N(0,1) \tag{21}$$

In this example, we operationalize the reliability parameter as a correlation, following the definition of reliability as a correlation between scores from two replications of a measurement procedure. The same transformation, g, is applied to both marginal distributions of a bivariate normal distribution that is normal in x^* by construction. The bivariate normal correlation is ρ^*, and the correlation following the transformation is ρ. The target ratio is ρ^*/ρ, which can be interpreted as an adjustment factor for reliability under normalized conditions. Following Equations (19) and (20), and with the additional assumption that subgroups a and b share the same metric for simplicity, the correction factor for any V estimated with reliability ρ is $\sqrt{\rho^*/\rho}$. This factor answers the question, how much of a correction is required when disattenuating V with a parametric reliability parameter? Answering this question requires a definition for the transformation g.

We establish criteria for a convenient, plausible transformation g (equivalently, a plausible distribution of variables x) as a monotonic transformation that, when applied to a normal metric x^*, 1) yields a distribution with a level of skewness that exists in practical data contexts, 2) possesses an intuitively plausible ratio of slopes at the 95th and 5th percentiles, $g'(P_{95})/g'(P_5)$, and 3) is mathematically tractable. Following the third criterion, we select an exponential transformation of the form, $g(x^*) = a + be^{cx^*}$. Under the constraints that the transformation preserves mean (0) and variance (1), the transformation is controlled by the parameter c, and can be written as:

$$g(x^*) = -\frac{sgn(c)}{\sqrt{e^{c^2} - 1}}\left(1 - e^{cx^* - \frac{c^2}{2}}\right).$$

(22)

To challenge the procedure while also meeting the first two criteria, we select a c of 0.4892. The ratio of slopes at the 95th and 5th percentiles is 5, that is, an incremental difference at the 95th percentile is, after transformation, five times the same incremental difference at the 5th percentile. This results in a distribution of x with a skewness of 1.73. This considerable level of skewness is occasionally seen in some observed test score distributions, for example, in New York and Colorado (CTB/McGraw-Hill 2011a; CTB/McGraw-Hill 2011b).

Figure 4 shows the results of the transformation (using $c = .4892$) applied to bivariate standard normal data with correlation 0.8. The marginal densities of the transformed variables are shown along the axes.

Conveniently, ρ, the correlation of the transformed variables, has a closed-form approximation following application of the multivariate delta method (Taylor approximation, e.g., Casella and Berger 2002):

$$\rho \approx \frac{e^{c^2\rho^*} - 1}{e^{c^2} - 1}$$

Table 4 shows reliability parameters under normalized and reported metrics when $c = .4892$. Reliabilities are always attenuated by this transformation. As Figure 3 shows, reliabilities from large scale testing programs generally fall between .80 and .95. If all of these reliabilities arose from skewed distributions with densities following Equation (23), Table 4 shows that the reliabilities after normalization (ρ^*) would be around 2.5% to 0.62% larger. If V were calculated using reliabilities based on the reported metric, the maximal correction necessary under this family of transformations would be 1.25% to .31%. Given the considerable skewness necessary to warrant this correction, we find this level of potential distortion tolerable. On the basis of this illustrative example, we consider the interpretation of disattenuated V statistics from Equations 19 and 20 to be unlikely to be affected by any mismatch between the reported scale for ρ and the

respectively normal scale for V.

Summary of Measurement Error Analysis

Correct disattenuation using $V_x/\sqrt{\tilde{\rho}}$ requires 1) known subgroup reliabilities, 2) known subgroup variances or their ratio, and 3) applicability of these reliabilities and variances to the metric in which subgroup distributions are normal. Figure 3 showed that $\tilde{\rho}$ is well approximated by $\bar{\rho}$ and even ρ_{all}, in cases where requirements (2) and (1) are not met, respectively. Table 4 showed that V_t is robust when requirement (3) is not met.

Of the three requirements, the third is least likely to be satisfied or verifiable in the data-poor environments that motivate this paper. The second requirement is the easiest to satisfy in practice; the ratio of subgroup variances is estimable from coarsened data even if it is not publicly available. The first requirement is generally satisfiable in the context of state-level educational tests. Seven of 38 states reported reliability for all students but reported no subgroup reliabilities in 2009. By 2011, this number of states dropped to 4, as reporting subgroup reliabilities is increasingly standard practice. Nonetheless, we have demonstrated that the use of $V_x/\sqrt{\rho_{all}}$ will be a reasonable disattenuation in most cases.

Disattenuation is particularly essential when comparing gaps across tests with different reliabilities. One example of this is gap comparison from state tests to NAEP, where NAEP explicitly incorporates item sampling into its estimates of standard deviations (Mislevy, Johnson, & Muraki, 1992). Without correction, all else equal, NAEP gaps would be expected to be larger due to their correction for measurement error. An additional factor is the nature of the model for reliability estimation. If the reliabilities are estimated by internal consistency measures like Cronbach's alpha, then only measurement error due to item sampling is incorporated. Large-scale testing programs rarely include other sources of measurement error, such as replications over occasions or raters. To the degree that these sources of error are dramatically different across tests, comparisons may be further biased, and without hope of correction in absence of a

generalizability study (e.g., Brennan 2001) that disentangles these sources of error.

Conclusion

As Ho and Reardon (2012) argued, large-scale gap analyses need not be hindered by coarsened data. However, practical use of these methods requires attention to sampling variation and biases due to measurement error. One might imagine that the sampling variance of gap estimates based on coarsened data is so large as to render the estimates largely useless in practical applications. Likewise, if the measurement error-induced attenuation bias in coarsened gap estimates were large and/or unpredictable in magnitude, one might worry about comparing gap estimates across tests with different and unknown measurement error properties.

Our analyses here suggest, however, that these concerns will be largely unfounded in a wide range of practical data applications. With regard to the sampling variance of estimates of V, we show that 1) it is possible to estimate the sampling variance of estimates of V from the observed data; and 2) the sampling variance of estimates of V based on coarsened data is only slightly larger than the variance of estimates of V based on full data or than that of estimates of d. So long as the thresholds used to coarsen the data are not very closely or very asymetrically located, there is very little loss of precision from the coarsening of the data. The somewhat surprising precision of estimates of V results from the fact that estimating V is equivalent to computing the area under a monotonic curve fitted to the points representing the paired cumulative proportions of each group below each threshold (Ho and Reardon 2012). Given the constraint that this curve must go through the origin and the point (1,1), sampling variability in the paired cumulative proportions will result in little sampling variance in the estimated error under the curve so long as the paired proportions are not tightly clustered in one part of the curve. As a result, estimates of V based on coarsened data are surprisingly precise under a wide range of types of coarsening.

Second, we show that easily applied measurement error bias disattenuation corrections

provide accurate disattenuation over a wide range of data generating scenarios. Although correct disattenuation requires the analyst to know 1) the subgroup specific reliabilities; 2) the ratio of the subgroup score variances in the metric in which the distributions are normal; and 3) that the reliabilities apply to the scores in their respectively normal metric, all but the most extreme violations of these conditions have very little practical effect on the disattenuated estimates. As a result, disattenuating V with reliability estimates for all students—rather than subgroups—will provide accurate disattenuation of gaps, even when reliability is estimated in a metric in which the distributions are not normal.

Together, our findings suggest that issues of sampling variance and measurement error pose no more significant barrier to the estimation of V than they do to more conventional gap measurement statistics. This is not to say that there are not cases where estimation of V is problematic, of course. But the conditions under which sampling variance and measurement error become worrisome—when the thresholds defining the coarsening are too close together or when subgroup reliabilities are very low and differ substantially from each other—do not appear with any frequency in the data we examined. Certainly analysts should be cautious in applying these methods, and we have identified the situations that should cause the most concern. However, our results also suggest that sampling variance inflation is low and measurement error corrections are appropriate under a wide range of conditions common in the analysis of educational achievement gaps.

References

Ahn, Sangtae and Jeffrey A. Fessler. 2003. "Standard errors of mean, variance, and standard deviation estimators." Communications and Signal Processing Lab, EECS Department, University of Michigan.

Brennan, Robert L. 2001. *Generalizability Theory*. New York: Springer-Verlag.

Casella, G. and R. L. Berger. 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

CTB/McGraw-Hill. 2011a. "Colorado Student Assessment Program Technical Report 2011." CTB/McGraw-Hill, Monterey, CA.

—. 2011b. "New York State Testing Program 2011: Mathematics, Grades 3–8 " CTB/McGraw-Hill, Monterey, CA.

Hanley, James A. 1988. "The robustness of the" binormal" assumptions used in fitting ROC curves." *Medical Decision Making* 8:197-203.

Hedges, Larry V. and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Ho, Andrew Dean and Edward Henry Haertel. 2006. "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples." Technical Report for the National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Ho, Andrew Dean and Sean F. Reardon. 2012. "Estimating Achievement Gaps From Test Scores Reported in Ordinal “Proficiency” Categories." *Journal of Educational and Behavioral Statistics* 37:489-517.

Kelley, T. L. 1947. *Fundamentals of Statistics*. Cambridge, MA: Harvard University Press.

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

Reardon, Sean F. and Claudia Galindo. 2009. "The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades." *American Educational Research Journal* 46:853-891.

Figure 1

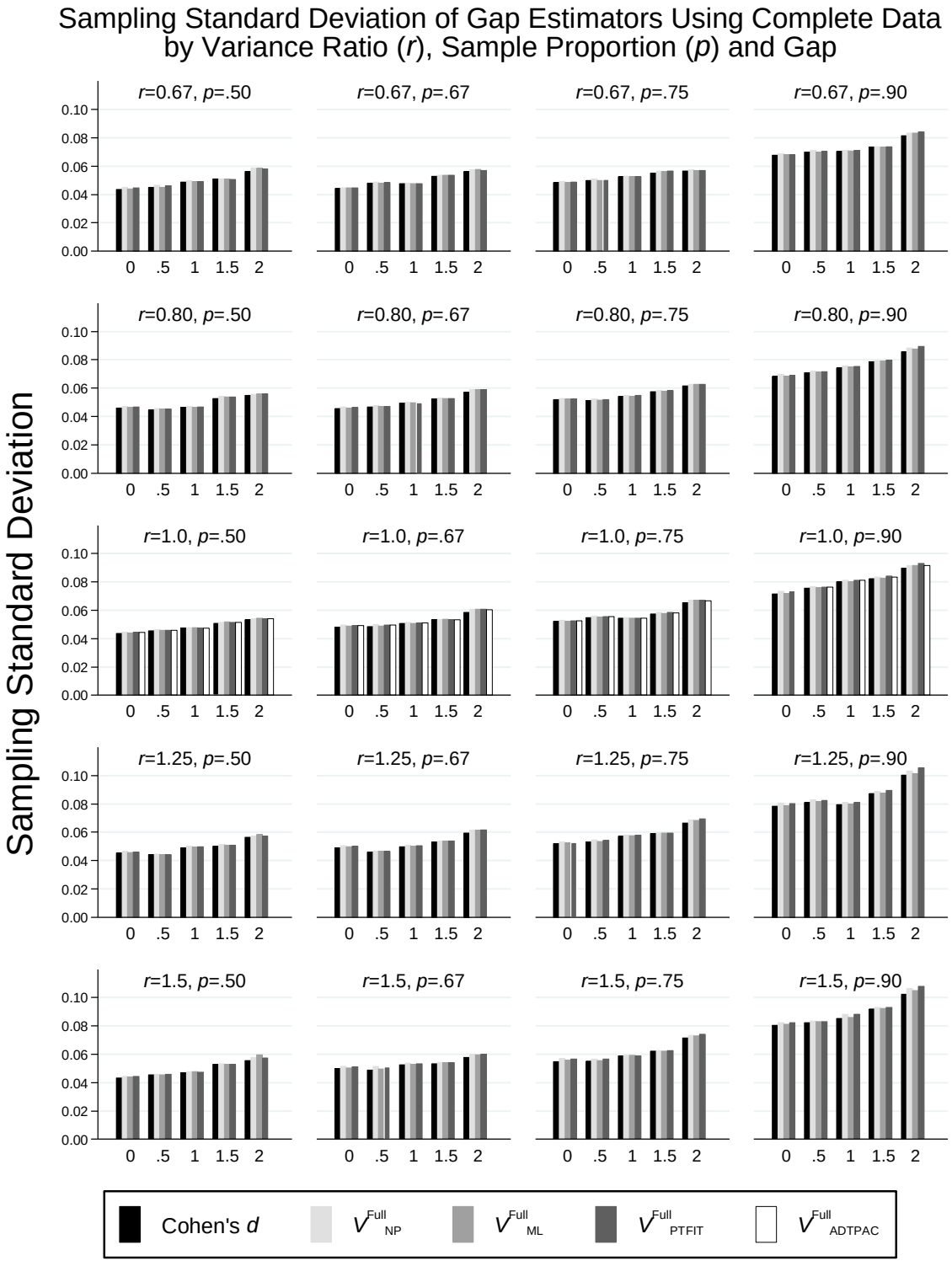

Sampling Standard Deviation of Gap Estimators Using Complete Data
by Variance Ratio (r), Sample Proportion (p) and Gap
Sampling Standard Deviation
r=0.67, p=.50
r=0.67, p=.67
r=0.67, p=.75
r=0.67, p=.90
r=0.80, p=.50
r=0.80, p=.67
r=0.80, p=.75
r=0.80, p=.90
r=1.0, p=.50
r=1.0, p=.67
r=1.0, p=.75
r=1.0, p=.90
r=1.25, p=.50
r=1.25, p=.67
r=1.25, p=.75
r=1.25, p=.90
r=1.5, p=.50
r=1.5, p=.67
r=1.5, p=.75
r=1.5, p=.90
0.00
0.02
0.04
0.06
0.08
0.10
0 .5 1 1.5 2
Cohen's d
V Full NP
V Full ML
V Full PTFIT
V Full ADTPAC

Figure 2

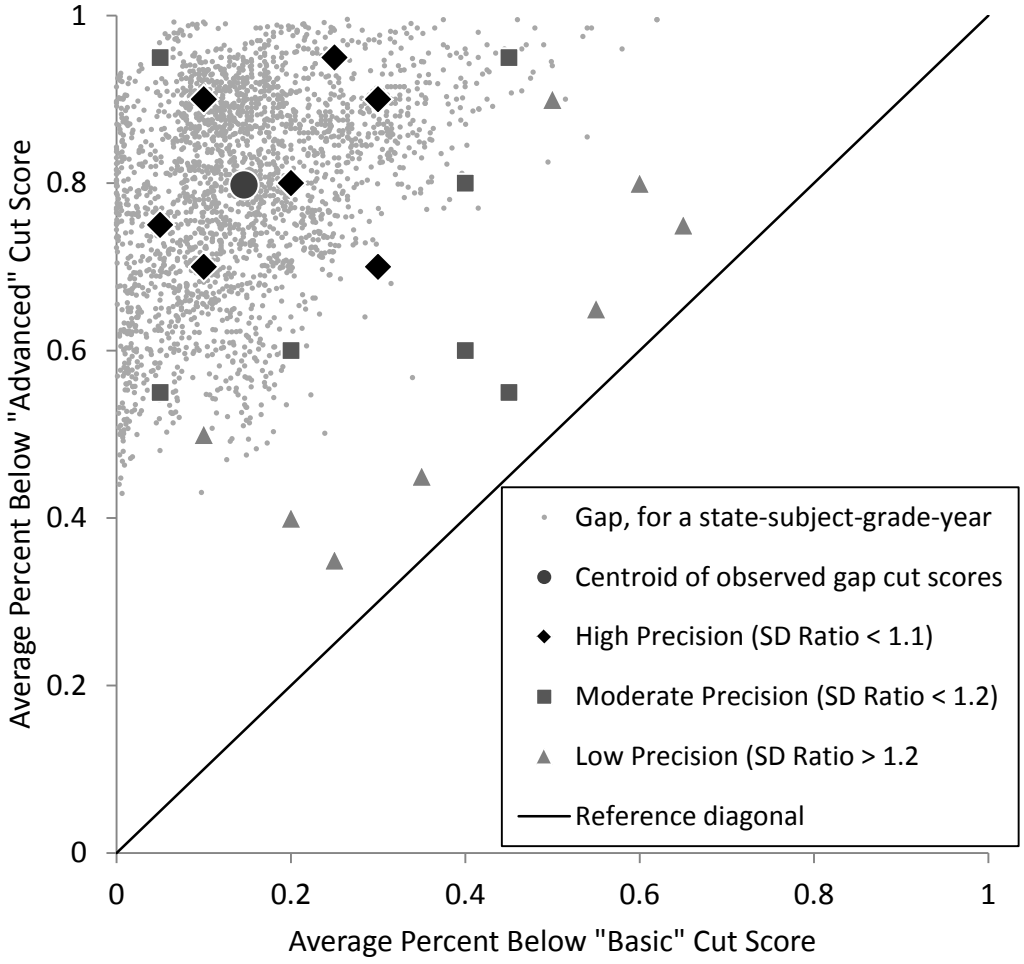

Average Percent Below "Advanced" Cut Score
Average Percent Below "Basic" Cut Score
0
0.2
0.4
0.6
0.8
1
Gap, for a state-subject-grade-year
Centroid of observed gap cut scores
High Precision (SD Ratio < 1.1)
Moderate Precision (SD Ratio < 1.2)
Low Precision (SD Ratio > 1.2
Reference diagonal

Figure 3. Distribution of reported subgroup reliability statistics for White, Black and Hispanic students on state accountability tests, from 38 states, grades 3-8, mathematics and English language arts, 2009-2012 (n=4240). Embedded table shows pairwise correlations below the diagonal and pairwise root mean square deviations above the diagonal.

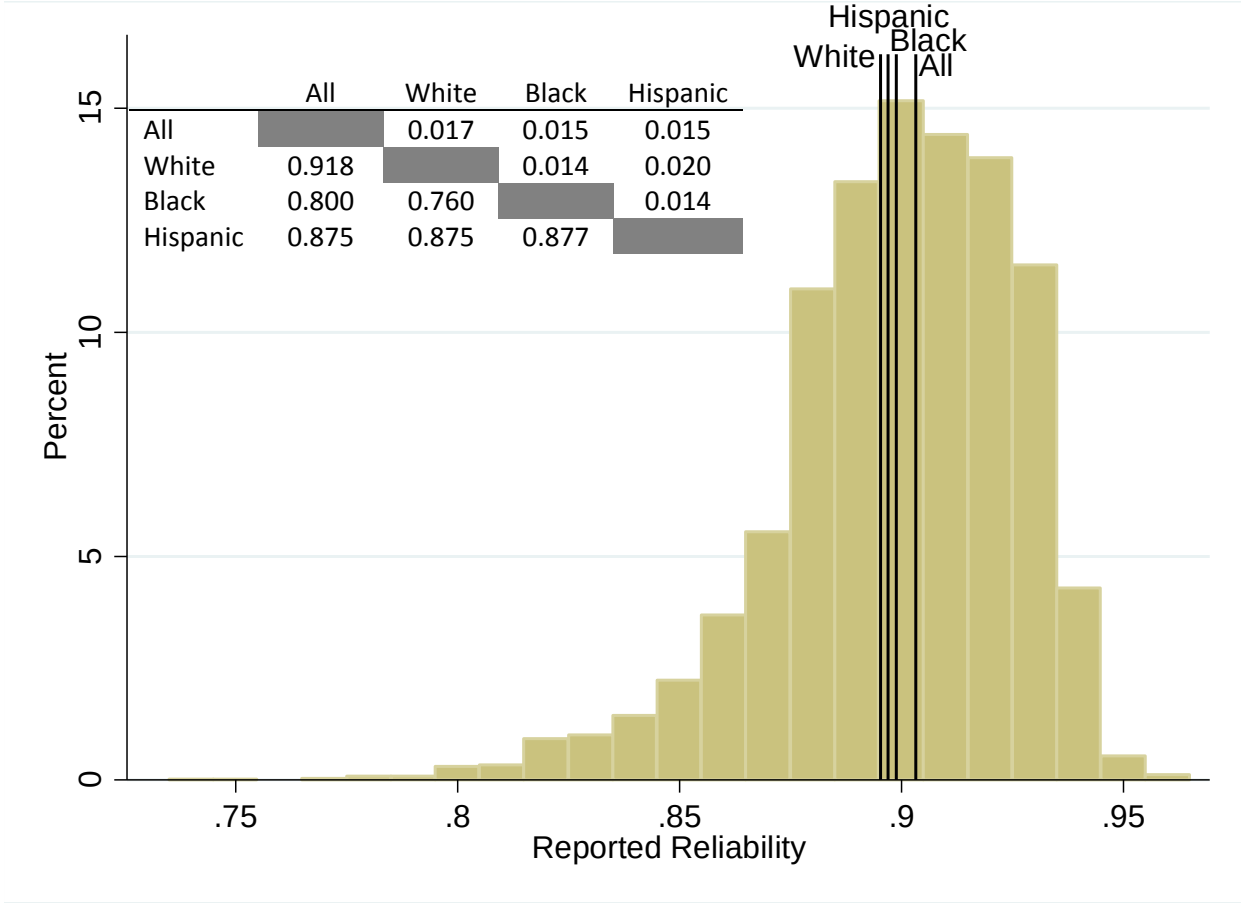
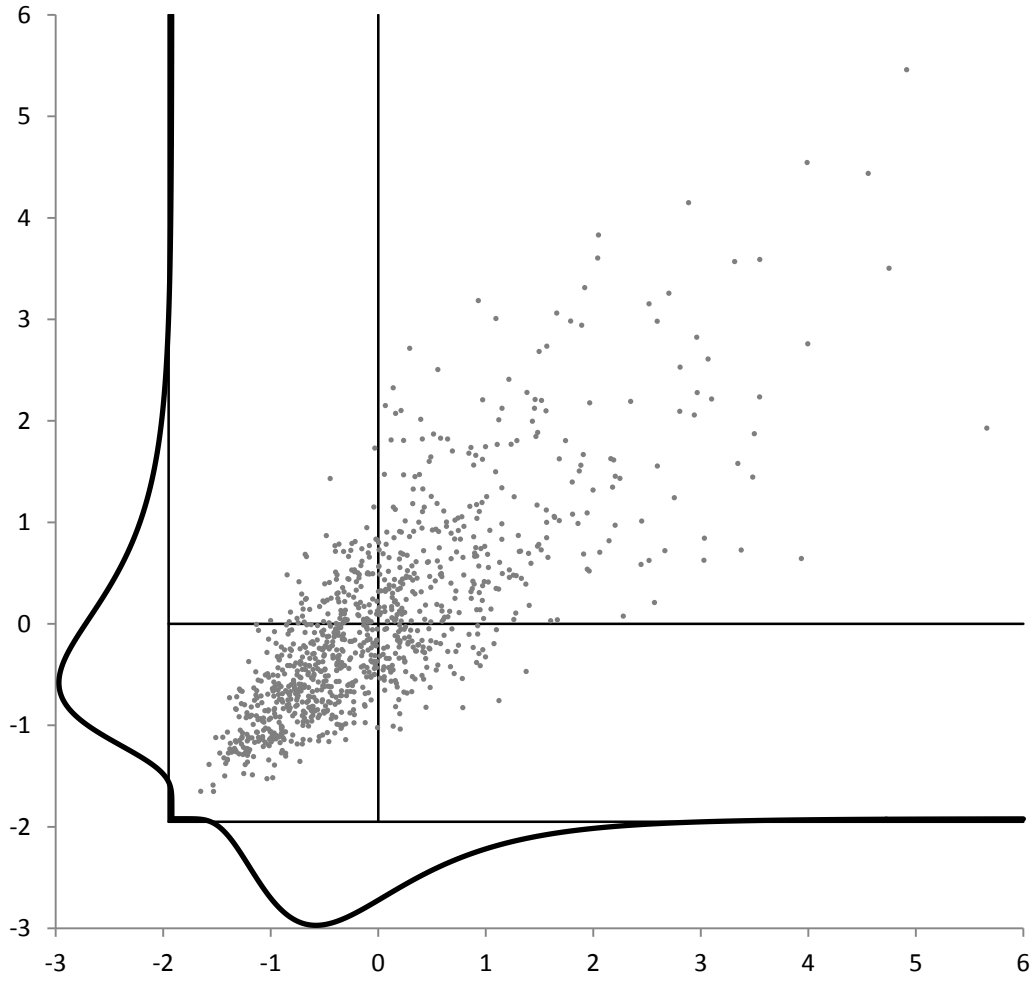


Figure 4. Bivariate normal distribution with correlation 0.8 after application of a mean- and variance-preserving exponential transformation with a skew-inducing factor of $c = 0.4892$ (skewness = 1.73). Marginal densities are shown along the axes.



Notes: Scatterplot shows a random draw of $n = 1000$ from the joint distribution. Population correlation following the transformation is 0.78.

Table 1

Ratio of Sampling Standard Deviation of Maximum Likelihood *V* Estimate (from coarsened data) to Non-Parametric *V* Estimate (from full data), by Variance Ratio (*r*), Location of Cut Points, Proportion of Sample in One Group (*p*), and Size of True Gap

	Sample Ratio and Size of True Gap									
Variance Ratio and	**ratio = 10:90**			**ratio = 25:75**			**ratio=50:50**			Average
Location of Cut Points	gap=0	gap=0.5	gap=1	gap=0	gap=0.5	gap=1	gap=0	gap=0.5	gap=1	SD Ratio
Variance Ratio = 0.80										
(20/50/80)	1.04	1.06	1.07	1.06	1.05	1.05	1.05	1.04	1.05	1.05
(10/40/70)	1.07	1.07	1.06	1.07	1.06	1.07	1.06	1.07	1.05	1.06
(10/50/90)	1.04	1.08	1.09	1.06	1.04	1.08	1.05	1.07	1.08	1.07
(5/40/75)	1.07	1.07	1.09	1.07	1.06	1.08	1.08	1.05	1.07	1.07
(30/50/70)	1.09	1.08	1.06	1.07	1.08	1.07	1.06	1.08	1.07	1.07
(20/40/60)	1.10	1.11	1.09	1.09	1.10	1.11	1.11	1.11	1.10	1.10
(5/50/95)	1.11	1.12	1.15	1.08	1.10	1.14	1.09	1.11	1.16	1.12
(40/50/60)	1.13	1.14	1.12	1.11	1.11	1.12	1.13	1.14	1.14	1.13
(5/30/55)	1.16	1.17	1.15	1.15	1.14	1.13	1.15	1.16	1.13	1.15
(45/50/55)	1.18	1.20	1.17	1.16	1.15	1.16	1.16	1.19	1.19	1.17
(10/30/50)	1.22	1.22	1.23	1.20	1.20	1.20	1.20	1.22	1.22	1.21
(35/40/45)	1.37	1.38	1.39	1.37	1.32	1.32	1.35	1.37	1.39	1.36
(20/30/40)	1.54	1.47	1.48	1.45	1.42	1.45	1.47	1.54	1.51	1.48
(25/30/35)	1.96	1.93	1.86	1.85	1.85	1.87	1.86	1.93	2.02	1.90
Variance Ratio = 1.00										
(20/50/80)	1.03	1.04	1.05	1.04	1.03	1.04	1.07	1.06	1.05	1.05
(30/50/70)	1.08	1.06	1.06	1.07	1.06	1.05	1.06	1.06	1.06	1.06
(5/40/75)	1.06	1.07	1.08	1.06	1.05	1.05	1.06	1.07	1.06	1.06
(10/40/70)	1.08	1.07	1.09	1.06	1.08	1.08	1.06	1.06	1.05	1.07
(10/50/90)	1.07	1.06	1.08	1.06	1.06	1.08	1.07	1.06	1.10	1.07
(20/40/60)	1.10	1.10	1.10	1.09	1.10	1.09	1.09	1.11	1.09	1.10
(5/50/95)	1.11	1.08	1.12	1.10	1.10	1.12	1.09	1.09	1.14	1.11
(40/50/60)	1.14	1.12	1.12	1.12	1.12	1.10	1.11	1.13	1.12	1.12
(5/30/55)	1.16	1.13	1.17	1.17	1.15	1.14	1.14	1.16	1.14	1.15
(45/50/55)	1.19	1.15	1.17	1.17	1.15	1.16	1.17	1.16	1.16	1.16
(10/30/50)	1.22	1.19	1.19	1.21	1.19	1.18	1.23	1.20	1.22	1.20
(35/40/45)	1.43	1.37	1.38	1.38	1.37	1.40	1.40	1.37	1.41	1.39
(20/30/40)	1.55	1.47	1.45	1.49	1.47	1.51	1.51	1.47	1.53	1.49
(25/30/35)	1.89	1.89	1.87	1.88	1.79	1.92	1.90	1.89	1.99	1.89
Variance Ratio = 1.25										
(20/50/80)	1.04	1.03	1.06	1.04	1.04	1.07	1.04	1.05	1.05	1.05
(10/50/90)	1.05	1.04	1.07	1.06	1.06	1.05	1.05	1.08	1.07	1.06
(10/40/70)	1.06	1.06	1.08	1.07	1.07	1.07	1.07	1.07	1.05	1.07
(30/50/70)	1.06	1.06	1.07	1.08	1.08	1.06	1.08	1.06	1.05	1.07
(5/40/75)	1.05	1.06	1.09	1.06	1.08	1.07	1.07	1.07	1.08	1.07
(5/50/95)	1.08	1.08	1.12	1.09	1.09	1.11	1.09	1.10	1.13	1.10
(20/40/60)	1.13	1.11	1.14	1.11	1.12	1.12	1.12	1.11	1.10	1.12
(40/50/60)	1.14	1.12	1.13	1.12	1.14	1.12	1.14	1.12	1.12	1.13
(5/30/55)	1.14	1.17	1.16	1.14	1.16	1.20	1.17	1.13	1.12	1.15
(45/50/55)	1.18	1.17	1.18	1.16	1.19	1.13	1.17	1.16	1.17	1.17
(10/30/50)	1.22	1.20	1.23	1.21	1.23	1.19	1.21	1.21	1.23	1.21
(35/40/45)	1.37	1.37	1.47	1.34	1.38	1.38	1.35	1.39	1.43	1.39
(20/30/40)	1.44	1.45	1.51	1.46	1.51	1.46	1.48	1.51	1.51	1.48
(25/30/35)	1.78	1.82	1.98	1.84	1.93	1.87	1.84	1.94	1.98	1.89

Notes: Sampling standard deviations are computed based on 1000 replications. Each replication includes 2000 observations. The sample ratio is the ratio of the number of observations of group B to those of group A, where A is the higher scoring group. The variance ratio is the ratio of the variance of the test score distribution of group B to that of group A. The cut scores are located at the percentiles of combined test score distribution of groups A and B, in a population in which A and B are equal size.

Table 2

Ratio of Sampling Standard Deviation of Maximum Likelihood *V* Estimate (from coarsened data) to Non-Parametric *V* Estimate (from full data), by Variance Ratio (*r*), Location of Cut Points, Proportion of Sample in One Group (*p*), and Sample Size

	Sample Ratio and Sample Size (True Gap = 1, Cut Scores = 20/50/80)								
	ratio = 10:90			ratio = 25:75			ratio=50:50		
Variance Ratio	n=100	n=500	n=2000	n=100	n=500	n=2000	n=100	n=500	n=2000
Variance Ratio = 0.80									
Coarsened V Sampling Standard Deviation	0.35	0.17	0.08	0.25	0.12	0.06	0.23	0.10	0.05
Ratio: Coarsened V SD to V Full SD	1.00	1.06	1.05	1.04	1.06	1.06	1.02	1.04	1.04
Variance Ratio = 1.00									
Coarsened V Sampling Standard Deviation	0.38	0.17	0.09	0.27	0.12	0.06	0.22	0.10	0.05
Ratio: Coarsened V SD to V Full SD	1.00	1.05	1.04	1.05	1.05	1.05	1.05	1.05	1.03
Variance Ratio = 1.25									
Coarsened V Sampling Standard Deviation	0.41	0.18	0.09	0.28	0.12	0.06	0.24	0.10	0.05
Ratio: Coarsened V SD to V Full SD	1.03	1.06	1.06	1.07	1.04	1.03	1.04	1.05	1.05

Notes: Sampling standard deviations are computed based on 1000 replications. Each replication includes 2000 observations. The sample ratio is the ratio of the number of observations of group B to those of group A, where A is the higher scoring group. The variance ratio is the ratio of the variance of the test score distribution of group B to that of group A. The cut scores are located at the percentiles of combined test score distribution of groups A and B, in a population in which A and B are equal size.

Table 3

Ratio of Sampling Standard Deviation of Maximum Likelihood *V* Estimate (from coarsened data) to Non-Parametric *V* Estimate (from full data), by Variance Ratio (*r*), Number of Cut Points, Proportion of Sample in One Group (*p*), and Size of True Gap

	Sample Ratio and Size of True Gap									
Variance Ratio and	**ratio = 10:90**			**ratio = 25:75**			**ratio=50:50**			Average
Location of Cut Points	gap=0	gap=0.5	gap=1	gap=0	gap=0.5	gap=1	gap=0	gap=0.5	gap=1	SD Ratio
Variance Ratio = 0.80										
(33/67)	1.11	1.11	1.08	1.10	1.08	1.07	1.10	1.08	1.07	1.09
(25/50/75)	1.05	1.05	1.04	1.06	1.06	1.03	1.05	1.04	1.06	1.05
(20/40/60/80)	1.04	1.02	1.02	1.04	1.03	1.04	1.03	1.03	1.03	1.03
(16/33/50/67/84)	1.03	1.01	1.03	1.03	1.02	1.01	1.02	1.02	1.02	1.02
Variance Ratio = 1.00										
(33/67)	1.11	1.11	1.09	1.10	1.09	1.11	1.10	1.08	1.09	1.10
(25/50/75)	1.04	1.05	1.04	1.06	1.06	1.05	1.05	1.06	1.06	1.05
(20/40/60/80)	1.03	1.04	1.04	1.04	1.04	1.02	1.02	1.05	1.03	1.03
(16/33/50/67/84)	1.03	1.02	1.02	1.02	1.00	1.03	1.02	1.00	1.03	1.02
Variance Ratio = 1.25										
(33/67)	1.07	1.08	1.09	1.11	1.10	1.11	1.11	1.09	1.09	1.09
(25/50/75)	1.05	1.05	1.05	1.05	1.06	1.03	1.08	1.04	1.05	1.05
(20/40/60/80)	1.04	1.03	1.03	1.01	1.02	1.02	1.03	1.04	1.03	1.03
(16/33/50/67/84)	1.02	1.00	1.02	1.02	1.02	1.03	1.03	1.03	1.01	1.02

Notes: Sampling standard deviations are computed based on 1000 replications. Each replication includes 2000 observations. The sample ratio is the ratio of the number of observations of group B to those of group A, where A is the higher scoring group. The variance ratio is the ratio of the variance of the test score distribution of group B to that of group A. The cut scores are located at the percentiles of combined test score distribution of groups A and B, in a population in which A and B are equal size.

Table 4. The difference between reliabilities (operationalized as correlations) under normal and reported (exponentially transformed with skew-inducing factor of $c = 0.4872$) metrics, along with a percentage adjustment factor for normalization and a correction factor for V.

Reliability in Normal Metric (ρ^*)	Reliability in Reported Metric (ρ)	% Adjustment $\left(\frac{\rho^*}{\rho} - 1\right)$	% Correction for V: $\left(\sqrt{\frac{\rho^*}{\rho}} - 1\right)$
0.500	0.470	6.36%	3.13%
0.550	0.520	5.71%	2.81%
0.600	0.571	5.06%	2.50%
0.650	0.622	4.42%	2.19%
0.700	0.674	3.78%	1.87%
0.750	0.727	3.14%	1.56%
0.800	0.780	2.51%	1.25%
0.850	0.834	1.88%	0.93%
0.900	0.889	1.25%	0.62%
0.950	0.944	0.62%	0.31%

Appendix: the sampling variance of $\hat{d}^* = \frac{\hat{\mu}_a - \hat{\mu}_b}{\hat{\sigma}_p}$.

First, define $e_a = \hat{\mu}_a - \mu_a$ as the error with which the mean μ in group a is estimated. The variance of e_a, given a sample of size n_a, will be $\frac{\sigma_a^2}{n_a}$. As above, we define $p = \frac{n_a}{n}$ and $r = \sigma_a^2/\sigma_b^2$. The sampling variance of $\hat{d}$, when σ_p is known, is

$$
\begin{aligned}
Var(\hat{d}) &= Var\left(\frac{\hat{\mu}_a - \hat{\mu}_b}{\sigma_p}\right) \\
&= Var\left(\frac{\mu_a - \mu_b + e_a - e_b}{\sigma_p}\right) \\
&= \left(\frac{1}{\sigma_p^2}\right) \cdot Var(e_a - e_b) \\
&= \left(\frac{1}{\sigma_p^2}\right) \cdot [Var(e_a) + Var(e_b)] \\
&= \left(\frac{1}{\sigma_p^2}\right) \cdot \left[\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right] \\
&= \left(\frac{\sigma_b^2}{\sigma_p^2}\right) \cdot \left[\frac{n_b r + n_a}{n_a n_b}\right] \\
&= \frac{2(p + (1-p)r)}{np(1-p)(1+r)}
\end{aligned}
$$

(A.1)

Now, the sampling variance of $\hat{d}^*$, when σ_p is estimated, is

$$
\begin{aligned}
Var(\hat{d}^*) &= Var\left(\frac{\hat{\mu}_a - \hat{\mu}_b}{\hat{\sigma}_p}\right) \\
&= Var\left(\hat{d} \cdot \frac{\sigma_p}{\hat{\sigma}_p}\right) \\
&= Var(\hat{d}) \cdot E\left[\frac{\sigma_p}{\hat{\sigma}_p}\right]^2 + Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right) \cdot E[\hat{d}]^2 + Var(\hat{d}) \cdot Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right)
\end{aligned}
$$

$$\approx Var(\hat{d}) + \left[d^2 + Var(\hat{d})\right] \cdot Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right)$$

$$= Var(\hat{d})\left[1 + Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right)\right] + d^2 \cdot Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right)$$

(A.2)

Next we derive an expression for $Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right)$. For this we use both the delta method and the approximation that $Var(\hat{\sigma}_a) \approx \frac{\sigma_a^2}{2(n-1)}$ (Ahn and Fessler 2003).

$$Var\left(\frac{\sigma_p}{\hat{\sigma}_p}\right) \approx Var\left(\frac{\hat{\sigma}_p}{\sigma_p}\right) \quad \text{(Delta Method)}$$

$$= Var\left(\left(\frac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{\sigma_a^2 + \sigma_b^2}\right)^{\frac{1}{2}}\right)$$

$$\approx \frac{1}{4} Var\left(\frac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{\sigma_a^2 + \sigma_b^2}\right) \quad \text{(Delta Method)}$$

$$= \frac{1}{16\sigma_p^4} Var\left(\hat{\sigma}_a^2 + \hat{\sigma}_b^2\right)$$

$$= \frac{1}{16\sigma_p^4}\left[Var(\hat{\sigma}_a^2) + Var\left(\hat{\sigma}_b^2\right)\right]$$

$$\approx \frac{1}{16\sigma_p^4}\left[4\sigma_a^2 Var(\hat{\sigma}_a) + 4\sigma_b^2 Var(\hat{\sigma}_b)\right] \quad \text{(Delta Method)}$$

$$\approx \frac{1}{4\sigma_p^4}\left[\frac{\sigma_a^4}{2(n_a - 1)} + \frac{\sigma_b^4}{2(n_b - 1)}\right] \quad \text{(Ahn and Fessler 2003)}$$

$$\approx \frac{1}{4\sigma_p^4}\left[\frac{n_b\sigma_a^4 + n_a\sigma_b^4}{2n_a n_b}\right] \quad \text{(if } n_a \text{ and } n_b \text{ are large)}$$

$$= \frac{1}{4\sigma_p^4}\left[\frac{n\left(\sigma_a^4 + 2\sigma_a^2\sigma_b^2 + \sigma_b^4\right) - \left(n_a\sigma_a^4 + 2n\sigma_a^2\sigma_b^2 + n_b\sigma_b^4\right)}{2n_a n_b}\right]$$

$$= \frac{1}{4\sigma_p^4}\left[\frac{4n\left(\sigma_p^4\right) - \left(n_a\sigma_a^4 + 2n\sigma_a^2\sigma_b^2 + n_b\sigma_b^4\right)}{2n_a n_b}\right]$$

$$= \frac{n}{2n_a n_b} - \frac{n_a\sigma_a^4 + 2n\sigma_a^2\sigma_b^2 + n_b\sigma_b^4}{8\sigma_p^4 n_a n_b}$$

$$= \frac{n}{2n_a n_b} - \frac{\sigma_b^4}{\sigma_p^4} \cdot \frac{n_a r^2 + 2nr + n_b}{8n_a n_b}$$

$$= \frac{n}{2n_a n_b} - \frac{n_a r^2 + 2nr + n_b}{2(1+r)^2 n_a n_b}$$

$$= \frac{1}{2n_a n_b}\left[n - \frac{n_a r^2 + 2nr + n_b}{(1+r)^2}\right]$$

$$= \frac{1}{2np(1-p)}\left[\frac{(1+r)^2 - pr^2 - 2r - (1-p)}{(1+r)^2}\right]$$

$$= \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2}$$

(A.3)

Substituting (A.1) and (A.3) into (A.2) yields

$$Var(\hat{d}^*) = \frac{2(r+p-pr)}{np(1-p)(1+r)}\left[1 + \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2}\right] + \frac{d^2(p + (1-p)r^2)}{2np(1-p)(1+r)^2}$$

$$= \frac{2(r+p-pr)}{np(1-p)(1+r)}\left[1 + \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2} + \frac{d^2(p + (1-p)r^2)}{4(1+r)(r+p-pr)}\right]$$

$$= Var(\hat{d}) \cdot \left[1 + \frac{d^2(p + (1-p)r^2)}{4(1+r)(r+p-pr)} + \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2}\right]$$

$$= \lambda \cdot Var(\hat{d})$$

(A.4)