Practical Issues in Estimating Achievement Gaps from Coarsened Data

Sean F. Reardon
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

August, 2014

Direct correspondence to sean.reardon@stanford.edu. This research was supported by grants from the Institute of Education Sciences (R305D110018 and R305B090016) 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 to account for bias 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, situations when we do not know students' exact test scores, but instead know 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 measure that describes the difference between two distributions measured on a common, continuous scale when only the coarsened data are available. In such cases, familiar measures 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, for example, 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 parameter for comparing two distributions of some random variable x is V, 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 measure V 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 (Ho & Reardon, 2012). 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 are 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 assess 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 difference that they may make in practice.

Notation and Definitions

Let x be a continuous variable indicating a test score (or some other quantity of interest, though our concrete examples here focus on test scores). We have two population groups, denoted a and b. Let $F_a(x)$ and $F_b(x)$ denote the cumulative distribution functions of x in groups a and b, respectively. In some of the discussion below we will consider K ordered "threshold" values of x, denoted $x_1 < x_2 < \cdots < x_K$. We denote the proportion of cases in group $g \in \{a, b\}$ with values of $x \leq x_k$ as $p_g^k = F_g(x_k)$. 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 groups a and b (albeit with different means and standard deviations):

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

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 standard normal distribution function.

We define the pooled within-group standard deviation of x as

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

(4)

Second, we consider the case where the distributions of x in groups a and b are not normal, but rather are *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, $F_a(x) = \Phi\left(\frac{f(x)-\mu_a^*}{\sigma_a^*}\right)$ and $F_b(x) = \Phi\left(\frac{f(x)-\mu_b^*}{\sigma_b^*}\right)$. The pooled within-group standard deviation in the metric defined by f is

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

(6)

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 & Olkin, 1985), the standardized difference in means between groups a and b:

$$d \equiv \frac{\mu_a - \mu_b}{\sigma_p}.$$

(2)

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

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

(1)

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 & Haertel, 2006; Ho & 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 metric 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 continuous data, both can also be readily estimated from certain types of aggregate or coarsened data. If we have estimates of the group-specific means (denoted $\hat{\mu}_a$ and $\hat{\mu}_b$) and standard deviations ($\hat{\sigma}_a$ and $\hat{\sigma}_b$), we can estimate d by substituting these estimates into Equations (2) and (4) above. If we have estimates of the proportions of each group that fall below a set of one or more threshold values of x (denoted $\hat{p}_a^1, .. \hat{p}_a^K, \hat{p}_b^1, \ldots, \hat{p}_b^K$), we can estimate V using the methods described by Ho and Reardon (2012).

Part 1: Sampling Variance of Gap Measure Estimates

In this section of this paper we 1) describe the sampling variance of a set of estimators of d and V; and 2) describe and evaluate methods for computing standard errors for their estimates. 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.

Parametric Estimators of d

If the pooled within-group 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 $p = \frac{1}{2}$, Equation (8) simplifies to $Var(\hat{d}) = \frac{1}{np(1-p)}$. 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 Appendix A that if we estimate d as the difference in estimated means divided by the estimated pooled within-group standard deviation,

$$\hat{d}' = \frac{\hat{\mu}_a - \hat{\mu}_b}{\hat{\sigma}_p}, \tag{9}$$

then the sampling variance of this estimator will be approximately

$$Var(\hat{d}') \approx \lambda \cdot Var(\hat{d}), \tag{10}$$

where

$$\lambda = 1 + \frac{d^2[p + (1-p)r^2]}{4(1+r)[p + (1-p)r]} + \frac{p + (1-p)r^2}{2np(1-p)(1+r)^2}. \tag{11}$$

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

$$Var(\hat{d}') \approx \left(1 + \frac{d^2}{8} + \frac{1}{8np(1-p)}\right)\left(\frac{1}{np(1-p)}\right). \tag{12}$$

So, when $r = 1$ and n is moderately large, estimating σ_p increases the sampling variance by a factor of approximately $\left(1 + \frac{d^2}{8}\right)$ (and so should increase the standard error by a factor of approximately $\sqrt{1 + \frac{d^2}{8}}$, 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 $\sqrt{1.125} = 1.061$).

A Non-Parametric Estimator of V Using Continuous (Non-coarsened) Data

When we observe non-coarsened values of x for each of the n cases in the sample, we can estimate V non-parametrically, by first estimating $P_{a>b}$. This is done by constructing all possible pairs of observed values of x $\{x_a, x_b\}$ (where x_a and x_b here denote values of x drawn from the observed samples from groups a and b, respectively) and computing the proportion of pairs in which $x_a > x_b$. In the case where there are ties (cases where $x_a = x_b$), we add half the proportion of 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 as

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

This is a non-parametric estimator of V because it requires no distributional assumptions. Mee (1990) and Brunner and Munzel (2000) provide formulae for computing confidence intervals of $\hat{P}_{a>b}$,[1] as does Pepe (2003). If the confidence interval of $\hat{P}_{a>b}$ is denoted $[L_{\hat{P}}, U_{\hat{P}}]$, then we can construct confidence intervals of $\hat{V}_{np}^{full}$ as

$$\left[\sqrt{2}\Phi^{-1}(L_{\hat{P}}), \sqrt{2}\Phi^{-1}(U_{\hat{P}})\right]. \tag{15}$$

Parametric Estimators of V Using Continuous or Coarsened Data

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. We focus here on three methods they describe, the "PTFIT" ("probit-transform-fit-inverse-transform") method; the maximum likelihood "ROCFIT" method (which we refer to as the maximum likelihood (ML) method hereafter); and the "ADTPAC" ("average difference in transformed percentages-above-cutscore") method. This third method is only appropriate under the additional assumption that the two distributions have equal variances in the metric in which they are both normal. 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; we provide only a quick summary here. The key to each of these methods is 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 (Green & Swets, 1966; Pepe, 2003).[2] The line will have intercept $n = \frac{\mu_a^* - \mu_b^*}{\sigma_b^*}$ and slope $m = \frac{\sigma_a^*}{\sigma_b^*}$, where μ_a^*, μ_b^*, σ_a^*, and

[1] We thank an anonymous reviewer for pointing out these references.

[2] That is, if we took every possible value of x, and computed $\alpha_x = \Phi^{-1}\big(F_a(x)\big)$ and $\beta_x = \Phi^{-1}\big(F_b(x)\big)$, and then plotted the points (α_x, β_x), the points would fall on a straight line. To see this, note that if the distributions are

σ_b^* are the means and standard deviations of the distributions of x in groups a and b, respectively, in some metric in which the two distributions are both normal (see Equation(5)). Note that the parameters m and n are invariant under any function f that renders both distributions of x^* normal (any function satisfying Equation (5)). Because $V = d$ when both distributions are normal, we have

$$V = d = \frac{\mu_a^* - \mu_b^*}{\sqrt{\frac{\sigma_a^{2*} + \sigma_b^{2*}}{2}}} = \frac{n}{\sqrt{\frac{m^2 + 1}{2}}}. \tag{16}$$

Thus, estimating m and n is sufficient to estimate V. Moreover, we do not need to identify the function f to estimate m and n.

The ML method uses maximum likelihood to estimate the values of m and n that correspond to the distributions most likely to have given rise to the observed numbers of each group scoring below each of K observed values (for detail on the ML estimation methods, see Dorfman & Alf, 1968, 1969; Pepe, 2003). V is then computed from $\hat{m}$ and $\hat{n}$ using Equation (16) (Ho & Reardon, 2012; Equation 12, p. 499). Given the maximum likelihood estimates of $Var(\hat{m})$, $Var(\hat{n})$, and $Cov(\hat{m}, \hat{n})$ (obtained from the observed Fisher information matrix), the sampling variance of $\hat{V}_{ml}$ can be approximated (see Appendix B) as

$$Var(\hat{V}_{ml}) \approx \frac{2}{1 + m^2} Var(\hat{n}) + \frac{2n^2m^2}{(1 + m^2)^3} Var(\hat{m}) - \frac{4mn}{(1 + m^2)^2} Cov(\hat{m}, \hat{n}). \tag{17}$$

The PTFIT method also relies on the respective normality assumption, and uses a weighted least squares regression approach to estimate m and n from the association between $\Phi^{-1}(F_b(x))$ and $\Phi^{-1}(F_a(x))$. The ADTPAC method is similar to the PTFIT method, but relies on a stronger

---

respectively normal, then $\Phi\left(\frac{x^* - \mu_a^*}{\sigma_a^*}\right) = F_a(x)$ and $\Phi\left(\frac{x^* - \mu_b^*}{\sigma_b^*}\right) = F_b(x)$, which implies $\beta_x = \left(\frac{x^* - \mu_b^*}{\sigma_b^*}\right) = \frac{\mu_a^* - \mu_b^*}{\sigma_b^*} + \frac{\sigma_a^*}{\sigma_b^*}\left(\frac{x^* - \mu_a^*}{\sigma_a^*}\right) = \frac{\mu_a^* - \mu_b^*}{\sigma_b^*} + \frac{\sigma_a^*}{\sigma_b^*}\alpha_x$.

assumption—that the two distributions are equivariant respectively normal (meaning they would have the same variance in the metric in which both were normal). Relying on this assumption, 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 $m = 1$. Because the PTFIT and ADTPAC methods do not produce estimates of the sampling covariance matrix of $\hat{m}$ and $\hat{n}$, however, we cannot use Equation (17) to compute the sampling variance of the $\hat{V}_{ptfit}$ and $\hat{V}_{adtpac}$ estimators. We show below, however, that Equation (10) can be used to produce accurate standard errors for both estimators when they are used with continuous, non-coarsened data.

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. 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 $\hat{V}_{ml}^{full}$;[3] and we refer to the ADTPAC estimator that uses complete data as $\hat{V}_{adtpac}^{full}$.

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 $F_b(x)$ and $F_a(x)$ can

[3] 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 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).

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 the other methods further in this paper.

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

Because we do not have analytic methods of computing the sampling variances of all of the $\hat{V}^{full}$ estimators (specifically, we are not aware of formulae for computing the sampling variances of $\hat{V}^{full}_{ptfit}$ and $\hat{V}^{full}_{adtpac}$), we examine their sampling variances using 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 $1/r$.[4] We conduct simulations based on 100 different data generating models (corresponding to each possible combination of V, p , and r, where $V \in \{0, 0.5, 1.0, 1.5, 2.0\}$; $p \in \{0.90, 0.75, 0.67, 0.50\}$; and $r \in \{0.67, 0.80, 1.0, 1.25, 1.5\}$).[5] For each combination of V, p, and r, we draw 1000 samples, and then compute $\hat{d}'$, $\hat{V}^{full}_{np}$, $\hat{V}^{full}_{ml}$, $\hat{V}^{full}_{ptfit}$, and $\hat{V}^{full}_{adtpac}$ (the last we compute only when $r = 1$). We then 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 notable about this figure. First, the sampling standard deviations of the five estimators are virtually identical to each

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

[5] 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 0.67 to 1.5.

other within each data generating scenario. This suggests 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 compute approximate standard errors of $\hat{V}_{ml}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{adtpac}^{full}$ as

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

$$= \sqrt{\left[1 + \frac{\hat{V}^{full^2}[p + (1-p)\hat{r}^2]}{4(1+\hat{r})[p + (1-p)\hat{r}]} + \frac{p + (1-p)\hat{r}^2}{2np(1-p)(1+\hat{r})^2}\right] \frac{2(\hat{r} + p - p\hat{r})}{np(1-p)(1+\hat{r})}}. \quad (18)$$

Using Equation (18) 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 (18) provides a method of obtaining standard errors of $\hat{V}_{ml}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{adtpac}^{full}$. Note that Equation (17) provides an alternate standard error estimate for $\hat{V}_{ml}^{full}$.

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 particularly 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[6] 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

[6] We report ratios of standard deviations because these indicate the extent to which standard errors and confidence intervals will be inflated due to coarsening of the data. The square of this ratio—the ratio of sampling variances—is a measure of relative efficiency, and can be interpreted as the proportion by which the sample size would need to be increased to offset the precision loss from coarsening.

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 $10/50/90^{th}$ percentiles, the $30/50/70^{th}$ percentiles, the $10/40/70^{th}$ percentiles, and even the $5/40/75^{th}$ 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, squares, and triangles 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 2, 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 variability 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.

Confidence Interval Coverage Rates

Above we have suggested a variety of methods of obtaining standard errors from the different estimators of V. To review, when we have (non-coarsened) continuous test score data, we can compute confidence intervals for $\hat{V}_{np}^{full}$ using Equation (15). For $\hat{V}_{adtpac}^{full}$, $\hat{V}_{ptfit}^{full}$, and $\hat{V}_{ml}^{full}$, the similarity of their sampling variances to that of $\hat{d}'$ shown in Figure 1 suggests that we can compute their standard errors using Equation (18) (setting $\hat{r} = 1$ when using $\hat{V}_{adtpac}^{full}$). Finally, for either continuous or coarsened data, standard errors for the maximum likelihood estimators $\hat{V}_{ml}^{full}$ and $\hat{V}_{ml}^{coarse}$ can be computed from Equation (17). We have not identified methods of computing standard errors of the $\hat{V}_{adtpac}^{coarse}$ and $\hat{V}_{ptfit}^{coarse}$ estimators.[7]

In Table 4 we assess the accuracy of these different standard error estimation methods. Specifically, for each of the simulations used in Figure 1 and Tables 1-3, we construct a 95% confidence interval and then report the proportion of cases in which the true value of V is contained

[7] In principle, Equation (17) could be used to compute standard errors for these estimators if the covariance matrix of $\hat{m}$ and $\hat{n}$ were obtained. However, we are not aware of a method to do so. Moreover, given the adequacy of the $\hat{V}_{ml}^{coarse}$ estimator, there is little need to use the $\hat{V}_{adtpac}^{coarse}$ and $\hat{V}_{ptfit}^{coarse}$ estimators.

in that interval. As Table 4 illustrates, each of the methods of constructing standard errors and confidence intervals discussed above yields a coverage rate of almost exactly 95% over the range of simulations. In more detailed analyses not shown here, we find the coverage rate appears constant across different values of the population parameters V, r, p, and n and across different locations of the cutscores.

Table 4 about here

Part 2: Disattenuating V for Measurement Error

Gap measures like d and V are expressed in terms of units of observed variation. Observed variation is increased by measurement error, thus estimates of d and V will be biased toward zero when measurement error is present. In this section, we review bias correction procedures for d and V that allow gap expression in units of variation uninflated by measurement error. These procedures are particularly important for gap comparisons where differences in measurement error may be confounded with differences in gaps. We demonstrate that group-specific reliability coefficients and variances are necessary for exact corrections to d, and that exact corrections to V require the same parameters, except estimated on a scale where distributions are normal. These values are rarely available in practice. We show that a practical approximation using readily available reliability coefficients is generally sufficient for corrections to d and V alike.

Reliability-based 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 as ρ, is defined as the ratio of true score variance, σ_t^2, to observed score variance, σ^2. The standard deviation of true scores for group g is $\sigma_{t_g} = \sqrt{\rho_g}\sigma_g$, where $\rho_g = \sigma_{t_g}^2/\sigma_g^2$ is the reliability of x for group g. Measurement error increases observed score standard deviations and biases effect size estimates like d toward zero (e.g., Hedges & Olkin, 1985). If the group-specific

reliabilities of x in both group a and b are known, then, following Equation (2), d_x can be disattenuated to express the gap in terms of pooled standard deviation units of the true score t:

$$d_t = \frac{\mu_a - \mu_b}{\sqrt{\frac{\sigma_{t_a}^2 + \sigma_{t_b}^2}{2}}} = \frac{\mu_a - \mu_b}{\sqrt{\frac{\rho_a\sigma_a^2 + \rho_b\sigma_b^2}{2}}} = \frac{d_x}{\sqrt{\frac{\sigma_a^2\rho_a + \sigma_b^2\rho_b}{\sigma_a^2 + \sigma_b^2}}} = \frac{d_x}{\sqrt{\frac{r\rho_a + \rho_b}{r+1}}} = \frac{d_x}{\sqrt{\tilde{\rho}}}, \tag{19}$$

where $r = \sigma_a^2/\sigma_b^2$ is the ratio of the group-specific variances as above. This shows that d_t and d_x differ by a constant, $1/\sqrt{\tilde{\rho}}$, where $\tilde{\rho}$ is the weighted average of the reliabilities for groups a and b, and the weights are proportional to the variances of x for the two groups. In short, accurate disattenuation of d_x requires group-specific reliabilities and the ratio of group-specific variances.

Practical approximations for $\tilde{\rho}$

Approximations for $\tilde{\rho}$ are necessary when any of r, ρ_a, or ρ_b are unavailable. Group-specific variances are reported rarely in publicly available documentation (Center on Education Policy, 2007), thus r will generally not be available. In contrast, reliability estimates for large-scale educational tests are generally reported in accordance with professional standards (AERA, APA, & NCME, 1999), for all students (ρ) and sometimes for specific groups (ρ_g). However, as we show below, $\bar{\rho}$, the average of ρ_a and ρ_b, is generally sufficient when r is unavailable, and ρ is generally sufficient when ρ_g are unavailable.

If we disattenuated d_x using $\bar{\rho}$ rather than $\tilde{\rho}$ in Equation (19), d_t will be incorrect by a factor of $\sqrt{\tilde{\rho}/\bar{\rho}}$. This ratio is a function of r and the difference in the group-specific reliabilities, ρ_a and ρ_b. Note that

$$\frac{\tilde{\rho}}{\bar{\rho}} = 1 + \frac{(r-1)}{(r+1)} \cdot \frac{(\rho_a - \rho_b)}{(\rho_a + \rho_b)}. \tag{20}$$

Thus, if either group-specific variances are approximately equal across groups ($r \approx 1$) or group-specific reliabilities are approximately equal across groups ($\rho_a \approx \rho_b$), then the average

reliability will approximately equal the variance-weighted average reliability ($\tilde{\rho}/\bar{\rho} \approx 1$). Both assumptions tend to hold in practice. In the case of group-specific variance ratios, note that state-level NAEP ratios of the variance of Black or Hispanic test scores to White test scores since the 2003 administration have averaged 1.10 and 1.14 for Black-White and Hispanic-White ratios, respectively. Only 2.5 percent (51 of 2,040 computed variance ratios) are greater than 1.5 or less than 0.67.[8] In the case of group-specific reliability differences, Figure 3 shows that they are similar on average. To construct Figure 3, we gathered publicly available group-specific reliability statistics[9] 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. We do not review further details of this data collection process for space considerations and because the relevant results have straightforward implications. Average group-specific reliabilities are very similar, with average White, Hispanic, and Black reliabilities of .895, .897, and .899, respectively. The reliability for all students was slightly higher on average at .903, but the magnitude of average differences is trivial.

The embedded table in Figure 3 shows standard deviations of the pairwise differences in group-specific reliabilities above the diagonal. These standard deviations are never greater than 0.02. At least in the case of tests like these, then, any two group-specific reliabilities that might be used to disattenuate a gap are unlikely to differ by more than 0.04 (two standard deviations).

---

[8] Authors' calculations from Main NAEP data available from the NAEP Data Explorer, available at http://nces.ed.gov/nationsreportcard/naepdata/.

[9] 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.

Returning to Equation (20), even with a variance ratio $r = 1.5$, and even if group 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.

Finally, Figure 3 suggests that, when group-specific reliabilities (ρ_g) are not available, using the reliability for all students (ρ) will also be a reasonable approximation. Given that average reliabilities are about 0.90, and that the standard deviation of differences between group-specific and total reliabilities is less than 0.02 (implying that the standard deviation of differences between $\bar{\rho}$ and ρ will also typically be less than 0.02), the ratio $\sqrt{\bar{\rho}/\rho}$ will typically be between 0.98 and 1.02. That is, using ρ in place of $\bar{\rho}$ in Equation (19) will change d_t by less than 2%. Certainly, group-specific variances may differ, reliabilities may be lower, or differences between group-specific and overall reliabilities may be larger in other contexts. However, for situations with variance ratios and reliabilities in the ranges we show here, d_t is unlikely to differ by more than 3% whether corrected by $\tilde{\rho}$ or by approximations necessitated by limited data ($\bar{\rho}$ when missing r; ρ when missing ρ_g).

Reliability-based disattenuation of V

The interpretation of V as a quasi-effect size relies upon pooled standard deviation units of normal distributions as a basis, even though the distributions themselves need only be respectively normal. Reliability-based correction (attenuation) of these pooled standard deviation units is strictly appropriate only when the reliability coefficient is applicable to the metric in which the distributions are normal. Reported reliability coefficients may apply to the normal metric if group-specific distributions are in fact normal, however, test-score distributions are rarely normal in practice (Ho & Yu, in press; Micceri, 1989). The practical problem for disattenuating V can therefore be summarized as having access to reliability and variances in the observed metric but requiring reliability and variances applicable to the scale in which distributions are normal.

Following our notation above, we denote parameters for this normal metric with an asterisk—the ratio of group-specific variances (r^*), the group-specific reliabilities (ρ_a^* and ρ_b^*), the variance-weighted average reliability ($\tilde{\rho}^*$), and average and overall approximations ($\bar{\rho}^*$ and ρ^*)—all derived from variances applicable to the scale in which distributions are normal. Following Equation (19) and our argument from the previous section:

$$\tilde{\rho}^* = \frac{\sigma_a^{*2}\rho_a^* + \sigma_b^{*2}\rho_b^*}{\sigma_a^{*2} + \sigma_b^{*2}} = \frac{r^*\rho_a^* + \rho_b^*}{r^* + 1} \approx \bar{\rho}^* \approx \rho^* \tag{21}$$

Given these parameters, we can disattenuate V directly as we would d:

$$V_t = \frac{V_x}{\sqrt{\tilde{\rho}^*}} \approx \frac{V_x}{\sqrt{\bar{\rho}^*}} \approx \frac{V_x}{\sqrt{\rho^*}}. \tag{22}$$

It is not standard practice to report variance ratios or reliabilities in the metric in which distributions are normal, however. Practical disattenuation of V hinges upon how well available estimates of reliability parameters like ρ can approximate the desired reliability parameter in the normal metric, ρ^*. We address this generally by taking advantage of the definition of reliability as an expected correlation between strictly parallel tests (Haertel, 2006). Because reliability is a correlation, and correlations vary under transformations of variables, we can use the robustness of correlations to transformations to indicate the robustness of reliability to transformations.

Our strategy is to consider the case where we know the reliability ρ of a variable x when x is expressed in a metric in which it is not normally distributed and where we know the function f that will render the distribution of $x^* = f(x)$ normal. In particular, we consider the case where x has a generalized log-normal distribution such that

$$x = f^{-1}(x^*) = a + be^{cx^*},$$

or, equivalently,

$$x^* = f(x) = \frac{1}{c}\ln\left(\frac{x-a}{b}\right)$$

(23)

The generalized log-normal distribution is useful because it can approximate a wide range of distributions of varying degrees of skewness and is mathematically tractable for our purposes here. Without loss of generality, if $x^* \sim N(0,1)$, we can set

$$a = -\left(e^{c^2} - 1\right)^{-\frac{1}{2}}$$

$$b = sgn(c)\left(e^{2c^2} - e^{c^2}\right)^{-\frac{1}{2}}.$$

(24)

These constraints will result in x having a standardized (mean 0, variance 1) log-normal distribution with skewness, $\gamma = sgn(c)\left(e^{c^2} + 2\right)\sqrt{e^{c^2} - 1}$), that is determined by the parameter c.

We use Figure 4 to illustrate a particularly extreme distribution from this family, using $c = 0.55$, which corresponds to a skewness of $\gamma = 2$. Ho and Yu (in press) find that skewness statistics for state-level test-score distributions rarely exceed ± 0.5 and almost never exceed ± 2.0 (only 2 of their 330 publicly available scale score distributions have skewness exceeding 2.0), thus this represents an extremely skewed distribution by the standards of state tests.

Figure 4 shows an observed correlation of 0.8, corresponding to a fairly low observed reliability parameter ρ. When f is a logarithmic transformation of the type in Equation (23), we can derive the relationship between correlations on the scale of x (ρ) and correlations on the scale of x^* (ρ^*) in closed form:

$$\rho^* = \frac{\ln\left(\rho\left(e^{c^2} - 1\right) + 1\right)}{c^2}.$$

(25)

Table 5 shows the theoretical relationship between reliabilities from distributions with a normalizing transformation f and reliabilities after the scales have been normalized. The

reliabilities on the normalized scale, ρ^*, are always higher, however, the differences are small. Using ρ instead of ρ^* will underestimate the desired reliability, ρ^*, and overestimate the corrected gap, V_t. For the difference to amount to even 1% of the gap, Table 5 shows test score distributions must be extremely skewed, and reliabilities must be low. We conclude that, for practical purposes, disattenuating V using generally available reliability coefficients, ρ, instead of reliability coefficients for normalized scales, ρ^*, is not consequential unless there is evidence that reliabilities are low and test score distributions are heavily skewed. In these rare cases, we may expect that disattenuated V gaps will be slightly overcorrected.

Alternative approaches to disattenuating V

There are three alternative approaches to disattenuating V that may be appropriate in particular scenarios. First, a number of methods exist for estimating reliability in an ordinal framework, where the impact of transformations on reliability may be lessened or negated. Lord and Novick (1968) describe an ordinal reliability estimation procedure using pairwise Spearman rank correlations among split-half test scores. For ordinal item scales, a number of robust procedures have been proposed (Wilcox, 1992; Zimmerman, Zumbo, & Lalonde, 1993; Zumbo, Gadermann, & Zeisser, 2007). These procedures are compelling theoretically but impractical in our coarsened data scenarios, wherein means and standard deviations are generally unavailable, let alone individual scores and item-response data.

Second, when full score distributions are available, we can shrink individual scores toward respective group means in accordance with their unreliability. For example, for group g, the shrunken test scores will be:

$$\hat{x} = \sqrt{\rho_g}(x - \mu_g) + \mu_g = \left(\sqrt{\rho_g}\right)x + \left(1 - \sqrt{\rho_g}\right)\mu_g. \quad (26)$$

Following this shrinkage,[10] calculation of measures like V can proceed. Reardon and Galindo (2009) take this approach in their measurement of Hispanic-White test score gaps. Again, individual scores are generally not available in the coarsened data scenario that motivates this paper, thus we do not consider this further.

Third, one could apply a reliability adjustment directly to individual cumulative proportions, p_g^k, for each group g, assuming that the cumulative proportions arise from a normal distribution. If ρ_g^* is the reliability in the normalized metric, then the corrected cumulative proportion below a cut score k is

$$\hat{p}_g^k = \Phi\left(\Phi^{-1}(p_g^k)/\sqrt{\rho_g^*}\right). \tag{27}$$

After this adjustment, V can be estimated from the $\hat{p}_g^k$'s. A small complication arises when certain estimation procedures require count data (e.g., ML, though not PTFIT), as the adjustment in Equation (27) will ultimately require rounding. We find it simpler to disattenuate V directly, following Equation (22).

Conclusion

As Ho and Reardon (2012) argued, coarsened data may challenge, but does not prevent, the estimation of achievement gaps . They show that is possible to obtain unbiased estimates of gaps under a wide range of conditions; even substantial violations of the respective normality assumption do not lead to large biases in gap estimates. In practice, however, one might imagine that the sampling variance of gap estimates based on coarsened data is so large as to render the

[10] We note that this correction differs from a similar, well known correction proposed by Kelley (1947). The Kelley shrinkage estimate supports prediction of individual true scores, whereas Equation (26) allows the variance of $\hat{x}$ to equal the variance of the true scores.

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 the $\hat{V}$ estimators, three results here are noteworthy. First, under conditions of respective normality, the sampling variance of the estimators of the V^{full} estimators is essentially identical to the sampling variance of the more conventional Cohen's d estimator, despite the fact that V does not rely on the interval properties of the test metric. Second, the sampling variance of estimators of V based on coarsened data will often be only slightly larger than that of the V^{full} estimators. Estimates of V derived from highly coarsened data can be nearly as precise as conventional Cohen's d estimators (which rely on sample means and standard deviations) and as V estimates based on full information (which rely on individual rank information). Third, we provide formulas for computing standard errors and confidence intervals for many of the V estimators, and we show that these formulas provide accurate confidence interval coverages. In wide range of scenarios, then, it is possible to recover unbiased and reasonably precise estimates of gaps and their standard errors, even when the data are highly coarsened.

The relative precision of the $\hat{V}^{coarse}$ estimators may be somewhat surprising. After all, coarsening a continuous test score into four ordered categories represents a potentially substantial loss of information regarding individual scores. Nonetheless, so long as the thresholds used to coarsen the data are not very closely or very asymmetrically located, the coarsening results in very little loss of precision in $\hat{V}^{coarse}$ relative to $\hat{V}^{full}$. This 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 & 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 coarsening conditions.

With regard to the effects of measurement error on estimates of V, we show that an easily applied measurement error bias correction of $1/\sqrt{\rho}$ can provide accurate disattenuation over a wide range of common data scenarios. Although the exact disattenuation factor , $1/\sqrt{\tilde{\rho}^*}$, requires the analyst to know group-specific reliabilities and variance ratios in the metric in which distributions are normal, we demonstrate that 1) average reliabilities often closely approximate variance-weighted reliabilities ($\bar{\rho} \approx \tilde{\rho}$), 2) overall reliabilities often approximate average reliabilities ($\rho \approx \bar{\rho}$), and 3) reliabilities are robust to transformations ($\rho \approx \rho^*$). Unless there is some combination of low reliabilities, differing group-specific reliabilities and variances, and extremely non-normal distributions, the approximate correction factor of $1/\sqrt{\rho}$ will provide quite accurate disattenuation of gaps.

These findings are useful given the importance of disattenuation for 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, we expect NAEP gaps to be larger due to their correction for measurement error. Nonetheless, corrections for measurement error are generally incomplete. Most reliability estimates are internal consistency measures like Cronbach's alpha, where 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, such as those from occasions or raters, are dramatically different across tests, comparisons may be further biased. Models for these

sources of error, such as those offered by generalizability theory (e.g., Brennan, 2001) can help to disattenuate and compare gaps to account for these additional sources of error.

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 measures. 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 group reliabilities are very low and differ substantially from each other—do not appear with any frequency in the standardized test score 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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brunner, E., & Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal, 42*(1), 17-25.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Retrieved from http://www.cep-dc.org/cfcontent_file.cfm?Attachment=CEP%5FReport%5FStudentAchievement%5F053107%2Epdf

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

Dorfman, D. D., & Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory—a direct solution. *Psychometrika, 33*(1), 117-124.

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*(3), 487-496.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association, 55*(292), 708-713.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1): Wiley New York.

Haertel, E. H. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 65-110). Westport, CT: American Council on Education / Praeger Publishers.

Hanley, J. A. (1988). The robustness of the" binormal" assumptions used in fitting ROC curves. *Medical Decision Making, 8*(3), 197-203.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Ho, A. D., & Haertel, E. H. (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, A. D., & Reardon, S. F. (2012). Estimating Achievement Gaps From Test Scores Reported in Ordinal 'Proficiency' Categories. *Journal of Educational and Behavioral Statistics, 37*(4), 489-517.

Ho, A. D., & Yu, C. C. (in press). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*.

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

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association, 85*(411), 793-800.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental designs* (3rd ed.). Homewood, IL: Richard D. Irwin, Inc.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*: Oxford University Press.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades. *American Educational Research Journal, 46*(3), 853-891.

Wilcox, R. R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. *British Journal of Mathematical and Statistical Psychology, 45*(2), 239-254.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*(1), 33-49.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21-29.

Figure 1

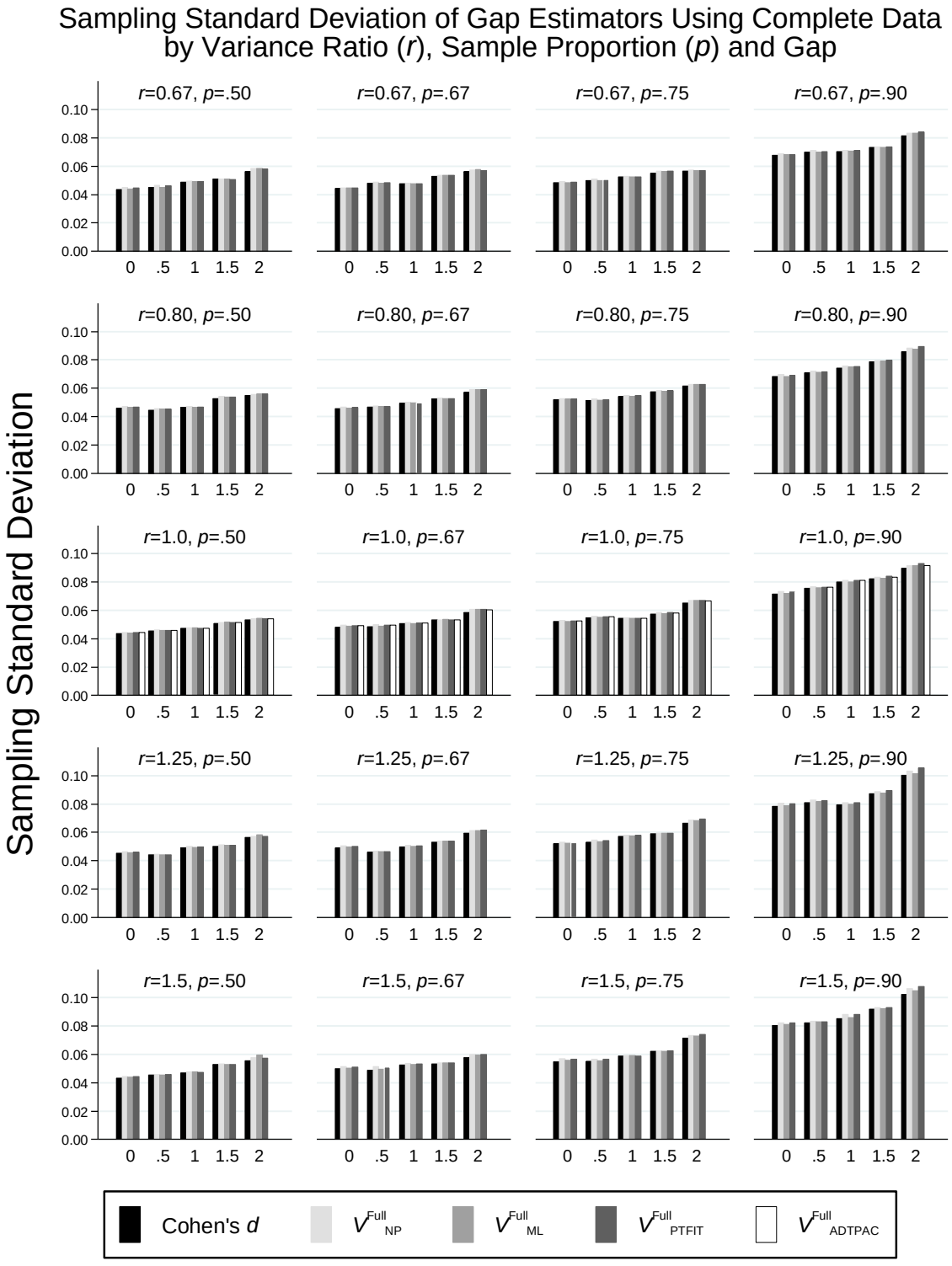

Sampling Standard Deviation of Gap Estimators Using Complete Data
by Variance Ratio (r), Sample Proportion (p) and Gap
Sampling Standard Deviation
r=0.67, p=.50
r=0.67, p=.67
r=0.67, p=.75
r=0.67, p=.90
r=0.80, p=.50
r=0.80, p=.67
r=0.80, p=.75
r=0.80, p=.90
r=1.0, p=.50
r=1.0, p=.67
r=1.0, p=.75
r=1.0, 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. Locations of high and low cut scores from a sample of state accountability tests, 1997-2011.

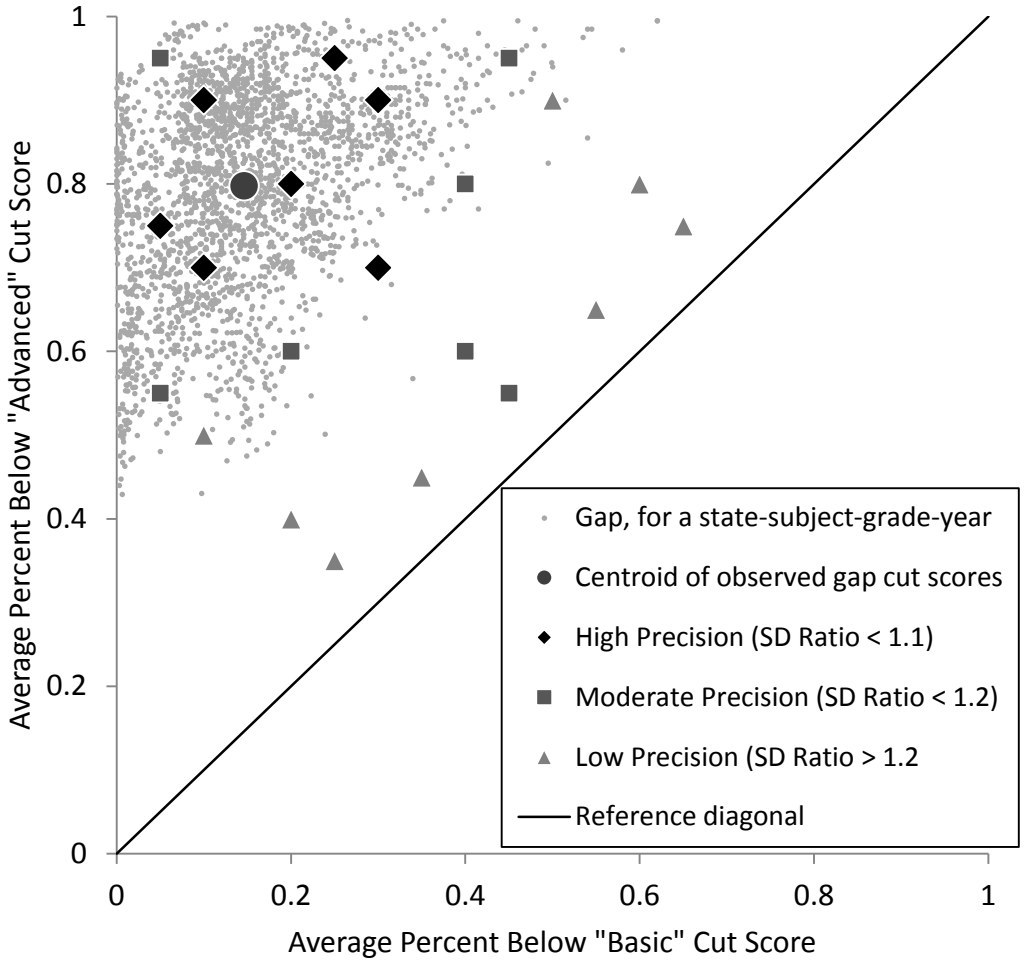


Figure 3. Distribution of reported group-specific reliability statistics for White, Black and Hispanic students on state accountability tests, from 38 states, grades 3-8, mathematics and English language arts, 2009-2012 (n=4240). Embedded table shows pairwise correlations below the diagonal and standard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

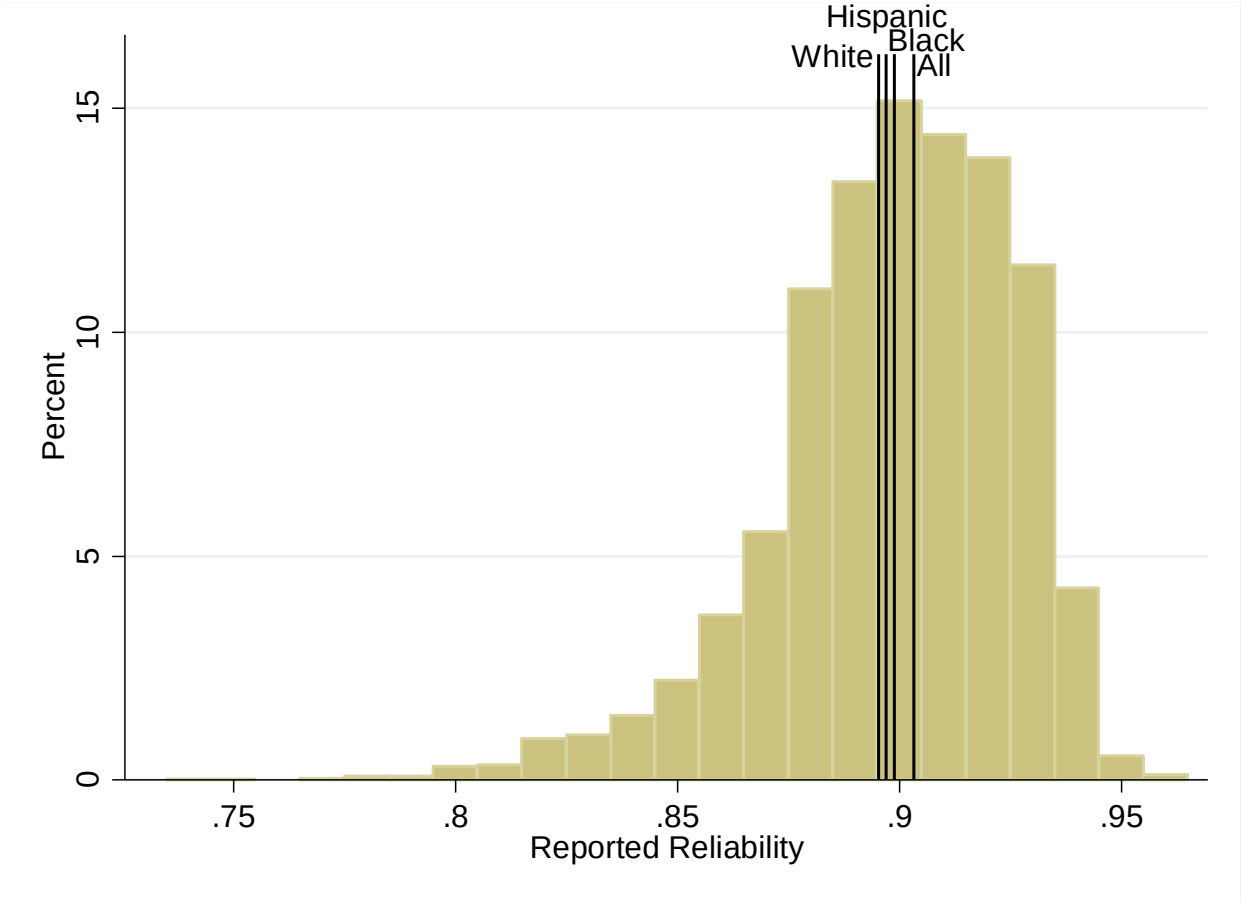
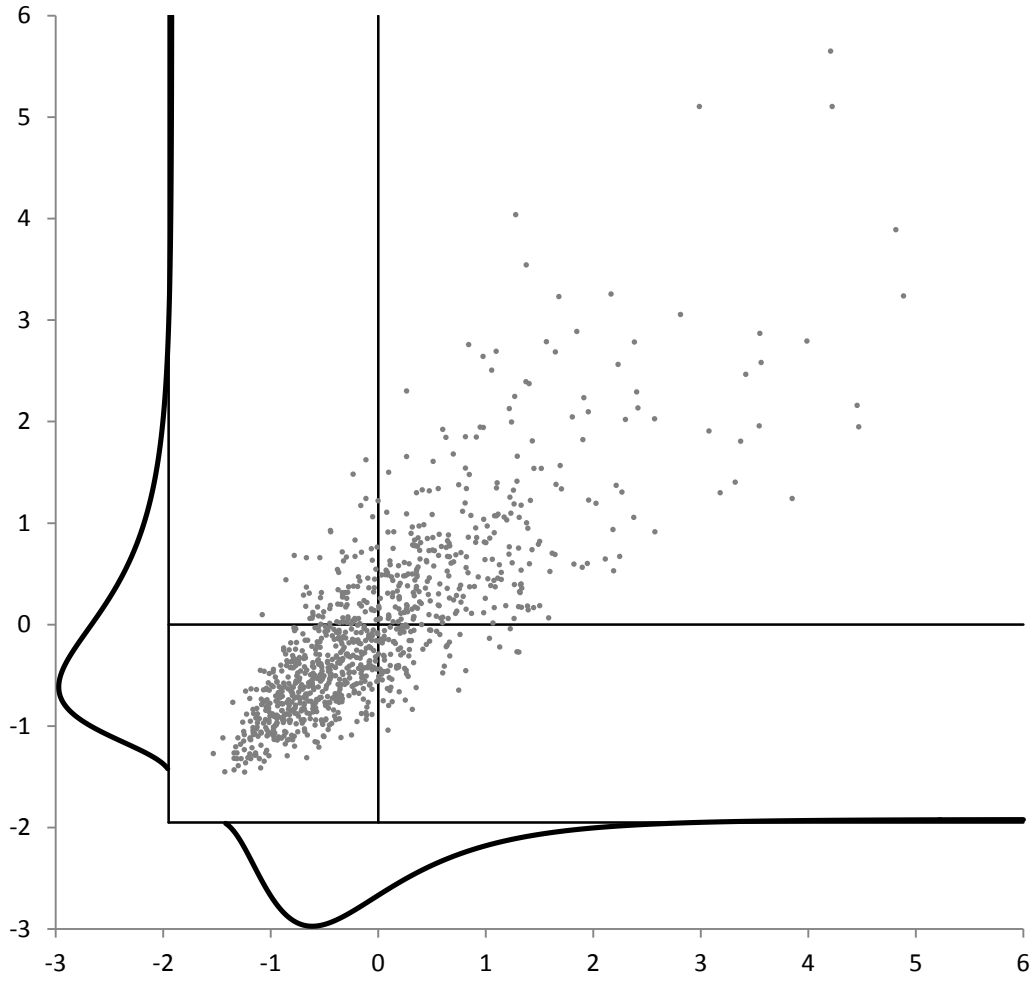


Figure 4. An illustration of reliability of 0.80 for a standardized log normal test score distribution (skewness = 2) that is normalizable by a function f (see Equation 23 in text) with a skew-correcting factor of $c = 0.55$.



Notes: Scatterplot shows a random draw of $n = 1000$ from the joint distribution. Population reliability following normalization is 0.823.

Table 1

Ratio of Sampling Standard Deviation of Maximum Likelihood *V* Estimate From Coarsened Data to Maximum Likelihood *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

Variance Ratio and Location of Cut Points	Sample Ratio and Size of True Gap: ratio = 10:90 gap=0	ratio = 10:90 gap=0.5	ratio = 10:90 gap=1	ratio = 25:75 gap=0	ratio = 25:75 gap=0.5	ratio = 25:75 gap=1	ratio=50:50 gap=0	ratio=50:50 gap=0.5	ratio=50:50 gap=1	Average SD Ratio
Variance Ratio = 0.80										
(20/50/80)	1.06	1.07	1.08	1.07	1.06	1.06	1.06	1.05	1.06	1.06
(10/40/70)	1.09	1.09	1.07	1.08	1.08	1.08	1.08	1.08	1.06	1.08
(10/50/90)	1.06	1.10	1.10	1.07	1.06	1.09	1.06	1.08	1.09	1.08
(30/50/70)	1.08	1.09	1.11	1.09	1.08	1.09	1.10	1.05	1.07	1.08
(5/40/75)	1.11	1.09	1.07	1.08	1.09	1.09	1.07	1.09	1.08	1.09
(20/40/60)	1.12	1.12	1.10	1.10	1.12	1.12	1.12	1.12	1.11	1.11
(5/50/95)	1.13	1.14	1.16	1.09	1.12	1.16	1.10	1.12	1.18	1.13
(40/50/60)	1.15	1.15	1.13	1.12	1.13	1.14	1.15	1.15	1.15	1.14
(5/30/55)	1.17	1.20	1.17	1.16	1.16	1.13	1.17	1.17	1.14	1.16
(45/50/55)	1.20	1.22	1.18	1.17	1.17	1.17	1.18	1.20	1.20	1.19
(10/30/50)	1.24	1.24	1.24	1.21	1.22	1.22	1.21	1.23	1.24	1.23
(35/40/45)	1.40	1.40	1.40	1.38	1.34	1.34	1.37	1.38	1.40	1.38
(20/30/40)	1.57	1.49	1.49	1.47	1.45	1.47	1.48	1.55	1.52	1.50
(25/30/35)	2.00	1.95	1.88	1.87	1.88	1.89	1.89	1.95	2.04	1.93
Variance Ratio = 1.00										
(20/50/80)	1.05	1.06	1.07	1.05	1.04	1.05	1.08	1.07	1.05	1.06
(10/40/70)	1.10	1.07	1.07	1.08	1.08	1.06	1.07	1.07	1.06	1.07
(10/50/90)	1.08	1.08	1.09	1.08	1.07	1.05	1.07	1.08	1.07	1.07
(30/50/70)	1.10	1.08	1.11	1.07	1.09	1.08	1.07	1.07	1.06	1.08
(5/40/75)	1.09	1.08	1.09	1.07	1.07	1.08	1.08	1.07	1.10	1.08
(20/40/60)	1.12	1.12	1.12	1.10	1.12	1.09	1.11	1.13	1.09	1.11
(5/50/95)	1.13	1.09	1.14	1.11	1.12	1.13	1.10	1.10	1.15	1.12
(40/50/60)	1.16	1.13	1.14	1.13	1.13	1.11	1.13	1.14	1.12	1.13
(5/30/55)	1.18	1.15	1.19	1.18	1.16	1.14	1.14	1.17	1.15	1.16
(45/50/55)	1.22	1.17	1.18	1.18	1.17	1.16	1.18	1.17	1.16	1.18
(10/30/50)	1.25	1.20	1.21	1.22	1.20	1.19	1.24	1.21	1.22	1.22
(35/40/45)	1.45	1.39	1.40	1.39	1.39	1.40	1.41	1.39	1.41	1.40
(20/30/40)	1.58	1.49	1.47	1.50	1.49	1.52	1.53	1.49	1.54	1.51
(25/30/35)	1.93	1.91	1.89	1.89	1.81	1.93	1.92	1.91	1.99	1.91
Variance Ratio = 1.25										
(20/50/80)	1.07	1.04	1.08	1.05	1.07	1.07	1.06	1.05	1.07	1.06
(10/40/70)	1.07	1.06	1.08	1.08	1.09	1.06	1.08	1.09	1.09	1.08
(30/50/70)	1.09	1.08	1.08	1.10	1.10	1.06	1.10	1.06	1.07	1.08
(5/40/75)	1.06	1.08	1.10	1.06	1.10	1.09	1.09	1.08	1.09	1.08
(10/50/90)	1.08	1.08	1.10	1.08	1.10	1.08	1.10	1.08	1.07	1.09
(20/40/60)	1.11	1.10	1.14	1.11	1.12	1.11	1.11	1.11	1.15	1.12
(5/50/95)	1.15	1.13	1.16	1.13	1.15	1.12	1.14	1.12	1.12	1.14
(40/50/60)	1.17	1.14	1.15	1.14	1.16	1.12	1.16	1.13	1.13	1.14
(5/30/55)	1.15	1.19	1.17	1.15	1.18	1.22	1.19	1.15	1.12	1.17
(45/50/55)	1.21	1.19	1.20	1.18	1.22	1.14	1.19	1.17	1.19	1.19
(10/30/50)	1.25	1.22	1.25	1.22	1.26	1.20	1.23	1.22	1.25	1.23
(35/40/45)	1.41	1.40	1.49	1.36	1.42	1.39	1.38	1.40	1.45	1.41
(20/30/40)	1.48	1.48	1.53	1.48	1.55	1.47	1.51	1.52	1.53	1.51
(25/30/35)	1.82	1.86	2.01	1.87	1.98	1.88	1.88	1.95	2.01	1.92

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 Maximum Likelihood *V* Estimate From Full Data, by Variance Ratio (*r*), 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	100	500	2000	100	500	2000	100	500	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	1.04	1.07	1.06	1.06	1.07	1.06	1.05	1.04	1.05
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	1.06	1.08	1.05	1.08	1.06	1.05	1.07	1.05	1.04
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	1.10	1.09	1.07	1.10	1.06	1.04	1.06	1.06	1.06

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 Maximum Likelihood *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

Variance Ratio and Location of Cut Points	Sample Ratio and Size of True Gap: ratio = 10:90 gap=0	ratio = 10:90 gap=0.5	ratio = 10:90 gap=1	ratio = 25:75 gap=0	ratio = 25:75 gap=0.5	ratio = 25:75 gap=1	ratio=50:50 gap=0	ratio=50:50 gap=0.5	ratio=50:50 gap=1	Average SD Ratio
Variance Ratio = 0.80										
33/67	1.12	1.12	1.08	1.10	1.09	1.07	1.12	1.09	1.07	1.10
25/50/75	1.06	1.06	1.04	1.07	1.07	1.03	1.06	1.05	1.07	1.06
20/40/60/80	1.06	1.03	1.02	1.04	1.04	1.04	1.04	1.04	1.04	1.04
16/33/50/67/84	1.04	1.02	1.03	1.03	1.03	1.02	1.04	1.02	1.03	1.03
Variance Ratio = 1.00										
33/67	1.12	1.14	1.10	1.11	1.11	1.12	1.10	1.09	1.10	1.11
25/50/75	1.05	1.08	1.05	1.07	1.07	1.05	1.05	1.08	1.06	1.06
20/40/60/80	1.04	1.06	1.05	1.04	1.05	1.02	1.02	1.06	1.04	1.04
16/33/50/67/84	1.04	1.04	1.03	1.03	1.02	1.04	1.02	1.02	1.04	1.03
Variance Ratio = 1.25										
33/67	1.08	1.10	1.09	1.12	1.11	1.13	1.13	1.10	1.11	1.11
25/50/75	1.06	1.06	1.05	1.06	1.07	1.05	1.10	1.05	1.07	1.06
20/40/60/80	1.04	1.04	1.03	1.02	1.04	1.04	1.05	1.04	1.05	1.04
16/33/50/67/84	1.03	1.02	1.03	1.03	1.03	1.04	1.04	1.04	1.03	1.03

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

95% Confidence Interval Coverage Rates, by Estimator and Confidence Interval Construction Method

Estimator	Standard Error/ Confidence Interval Formula	95% Confidence Interval Coverage Rate
$\hat{V}_{np}^{full}$	Equation (15)	94.8%
$\hat{V}_{adtpac}^{full}$	Equation (18)	94.7%
$\hat{V}_{ptfit}^{full}$	Equation (18)	94.7%
$\hat{V}_{ml}^{full}$	Equation (18)	94.7%
$\hat{V}_{ml}^{full}$	Equation (17)	95.0%
$\hat{V}_{ml}^{coarse}$	Equation (17)	94.9%

Note: 95% confidence intervals coverage rates for each of the V(full) estimators are averaged over all simulations used in Figure 1. Coverage rates for the V(coarse) estimator are averaged over all simulations used in Tables 1-3.

Table 5.

Consequences of using reliability parameter ρ as an approximation of the reliability parameter ρ^* when distributions of x are generalized log normal distributions.

	Reliability Parameter (ρ)	Normalized Reliability (ρ^*)	Underestimates ρ^* by $\left(\frac{\rho^*}{\rho}-1\right)$	Overestimates V by $\left(\sqrt{\frac{\rho^*}{\rho}}-1\right)$
Skewness = 2	0.750	0.777	3.61%	1.79%
	0.800	0.823	2.86%	1.42%
	0.850	0.868	2.12%	1.06%
	0.900	0.913	1.40%	0.70%
	0.950	0.957	0.69%	0.35%
Skewness = 0.5	0.750	0.753	0.33%	0.17%
	0.800	0.802	0.27%	0.13%
	0.850	0.852	0.20%	0.10%
	0.900	0.901	0.13%	0.07%
	0.950	0.951	0.07%	0.03%

Appendix A: 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 $\sigma_p \equiv \frac{\sigma_a^2+\sigma_b^2}{2}$ is known, is

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

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

$$
\begin{aligned}
Var(\hat{d}') &= Var\left(\frac{\hat{\mu}_a - \hat{\mu}_b}{\hat{\sigma}_p}\right) \\
&= Var\left(\hat{d} \cdot \frac{\sigma_p}{\hat{\sigma}_p}\right)
\end{aligned}
$$

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

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

(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 fact that $Var(\hat{\sigma}^2) \approx \frac{2\sigma^4}{n-1}$ (and the expression is exact if there is no excess kurtosis) (Casella & Berger, 2002; Neter, Wasserman, & Kutner, 1990).

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

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

$$\approx \frac{1}{16\sigma_p^4} Var(\hat{\sigma}_a^2 + \hat{\sigma}_b^2) \qquad \text{(Delta Method)}$$

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

$$\approx \frac{1}{16\sigma_p^4}\left[\frac{2\sigma_a^4}{(n_a - 1)} + \frac{2\sigma_b^4}{(n_b - 1)}\right]$$

$$\approx \frac{1}{8\sigma_p^4}\left[\frac{n_b\sigma_a^4 + n_a\sigma_b^4}{n_a n_b}\right] \qquad \text{(if } n_a \text{ and } n_b \text{ are large)}$$

$$= \frac{1}{8\sigma_p^4}\left[\frac{n(\sigma_a^4 + 2\sigma_a^2\sigma_b^2 + \sigma_b^4) - (n_a\sigma_a^4 + 2n\sigma_a^2\sigma_b^2 + n_b\sigma_b^4)}{n_a n_b}\right]$$

$$= \frac{1}{8\sigma_p^4}\left[\frac{4n(\sigma_p^4) - (n_a\sigma_a^4 + 2n\sigma_a^2\sigma_b^2 + n_b\sigma_b^4)}{n_a n_b}\right]$$

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

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

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

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

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

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

(A.3)

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

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

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

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

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

(A.4)

Appendix B: The sampling variance of $\widehat{V}_{ml}$

We wish to compute

$$Var(\hat{V}_{ml}) = Var\left(\frac{\sqrt{2}\hat{n}}{\sqrt{1+\hat{m}^2}}\right) = Var(xy),$$

(B.1)

where

$$x = \hat{n};$$

$$y = \sqrt{\frac{2}{1+\hat{m}^2}}.$$

(B.2)

Assuming $Var(\hat{m}) \ll m^2$, we have

$$E[y] \approx \sqrt{\frac{2}{1+m^2}}.$$

(B.3)

The delta method yields

$$\sigma_y^2 \approx \frac{2m^2}{(1+m^2)^3}\sigma_m^2$$

(B.4)

and

$$\sigma_{xy} = Cov\left(\hat{n}, \sqrt{\frac{2}{1+\hat{m}^2}}\right) \approx -.5m\left(\frac{2}{1+m^2}\right)^{1.5} \cdot \sigma_{mn}.$$

(B.5)

The variance of the product of two correlated random variables x and y can be approximated as

$$var(xy) \approx E[y]^2\sigma_x^2 + E[x]^2\sigma_y^2 + 2E[x]E[y]\sigma_{xy}$$

(B.6)

(Goodman, 1960). So we have

$$Var(\hat{V}_{ml}) \approx \frac{2}{1+m^2}\sigma_n^2 + \frac{2n^2m^2}{(1+m^2)^3}\sigma_m^2 - \frac{4mn}{(1+m^2)^2}\sigma_{mn}.$$

(B.7)