Article

Journal of Educational and Behavioral Statistics
2017, Vol. 42, No. 1, pp. 3–45
DOI: 10.3102/1076998616666279

Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions From Coarsened Data

Sean F. Reardon
Benjamin R. Shear
Stanford University

Katherine E. Castellano
Educational Testing Service

Andrew D. Ho
Harvard Graduate School of Education

Test score distributions of schools or demographic groups are often summarized by frequencies of students scoring in a small number of ordered proficiency categories. We show that heteroskedastic ordered probit (HETOP) models can be used to estimate means and standard deviations of multiple groups' test score distributions from such data. Because the scale of HETOP estimates is indeterminate up to a linear transformation, we develop formulas for converting the HETOP parameter estimates and their standard errors to a scale in which the population distribution of scores is standardized. We demonstrate and evaluate this novel application of the HETOP model with a simulation study and using real test score data from two sources. We find that the HETOP model produces unbiased estimates of group means and standard deviations, except when group sample sizes are small. In such cases, we demonstrate that a "partially heteroskedastic" ordered probit (PHOP) model can produce estimates with a smaller root mean squared error than the fully heteroskedastic model.

Keywords: *heteroskedastic ordered probit model; test score distributions; coarsened data*

The widespread availability of aggregate student achievement data provides a potentially valuable resource for researchers and policy makers alike. Often, however, these data are only publicly available in "coarsened" form in which students are classified into one or more ordered "proficiency" categories (e.g., "basic," "proficient," "advanced"). Although proficiency category data are clearly useful when proficiency status itself is of substantive interest, coarsened data pose challenges for the analyst when moments of the underlying test score distributions are of interest. Proficiency rates convey information about a single point in a cumulative test score distribution. This not only limits the information available to the analyst about the underlying test score distribution, but it also complicates inferences about relative changes in achievement levels in different population subgroups, a point illustrated by Ho (2008) and Holland (2002).

For example, suppose one wants to compare the average test scores among multiple schools, but one knows only the proportion of students scoring in each of several ordered proficiency categories. If the underlying test score distributions have unequal variances among schools, then rankings of schools on the basis of the percentages scoring at or above a given proficiency category will depend on which threshold is chosen. Moreover, rankings of schools based on percentages above some threshold will not, in general, match rankings based on mean scores. The same problem holds if one wishes to compare average test scores among multiple student subgroups (such as racial/ethnic groups) or to compare average test scores in a given school over time. In each case, judgments about the relative magnitude of between-group differences and even the ordering of groups' average performance will be dependent on what proficiency category threshold is used. These and other limitations posed by the coarsening of standardized test scores have been described extensively (Ho, 2008; Ho & Reardon, 2012; Holland, 2002; Jacob, Goddard, & Kim, 2013; Jennings, 2011).

With access to only coarsened test score data, therefore, comparisons of average performance among groups of students may be ambiguous. Unfortunately, most publicly available data on student performance on state standardized tests consist of coarsened test scores. Most states, for example, do not report school- or district-level test score means (and very few report standard deviations). The ED*Facts* Assessment Database (U.S. Department of Education, 2015), for example, provides test score data for every public school in the United States but does not include means and standard deviations. Rather, it contains the counts of students (by school, grade, subject, and student subgroup) scoring in each of the two to five state-defined performance levels, as required under the Elementary and Secondary Education Act. While these data are a valuable resource for educators, policy makers, and researchers, their utility is severely hampered by the absence of test score means and standard deviations.

In this article, we describe an approach that allows the analyst to recover more complete information about continuous test score distributions when only coarsened test score data are available. To achieve this, we propose a novel application of the heteroskedastic ordered probit (HETOP) model (e.g., Alvarez & Brehm, 1995; Greene & Hensher, 2010; Keele & Park, 2006; Williams, 2009). As we describe, the HETOP model can be used to recover means and standard deviations of continuous test score distributions of multiple groups from coarsened data. These groups may be schools, districts, or demographic subgroups. Estimates of these group means and standard deviations can be used to estimate intraclass correlations (ICCs), between-group achievement gaps, and other theoretically interesting or policy-relevant statistics, just as if each group's mean and standard deviation were provided directly.

The methods we describe generalize prior work quantifying achievement gaps in an ordinal or nonparametric framework, both with continuous (Ho, 2009) and with coarsened (Ho & Reardon, 2012; Reardon & Ho, 2015) test scores. Although we describe the use of such models to recover moments of test score distributions from aggregate proficiency data, the methods are applicable to other educational testing contexts when only coarsened scores are reported, such as Advanced Placement (AP) or English language proficiency exams. Aggregate data on coarse, ordered scales can also arise in college rankings, health research and practice scales, and income reporting. In all of these cases, our methods enable the estimation of group means and standard deviations from ordered data.

This article is organized into four main sections. In Section 1, we describe the statistical and conceptual framework for our application of the HETOP model. In Section 2, we use Monte Carlo simulations to evaluate recovery of the parameters of interest across a range of scenarios that might be encountered in practice. In Section 3, we use two real test score data sets, one from the National Assessment of Educational Progress (NAEP) and one from a State testing program, to evaluate the extent to which the key assumption of the HETOP model holds for real data. For these case studies, both student-level scale scores and coarsened proficiency counts are available, allowing us to evaluate the agreement between HETOP model estimates of means and standard deviations and estimates of the same parameters based on uncoarsened scale score data. Section 4 summarizes and discusses the results and offers recommendations for applying the methodology in practice.

1. Background and Statistical Framework

1.1 Canonical Application: Data, Assumptions, and Estimands

In our context of interest—the reporting of large-scale educational test scores in proficiency categories—the data consist of frequencies of students scoring within each of K ordered categories (often called performance levels) for G groups. Groups might be defined by racial/ethnic categories, schools, school districts, or other relevant categories. Such data can be summarized in a $G \times K$ matrix. Each cell in the matrix indicates the observed frequency of students from group $g = \{1, \ldots, G\}$ scoring at performance level $k = \{1, \ldots, K\}$ of a test. The performance levels describe ordered degrees of student proficiency in a content area. In standard current state testing practice, a panel of content experts selects one to four cut scores that divide the score scale into performance levels through a standard setting procedure (e.g., Cizek, 2012).

Let y denote a continuous random variable (scaled test scores, in our example), with μ_g and σ_g denoting the mean and standard deviation, respectively, of y in group g. Although we make no specific distributional assumptions about the shape of the distributions of y in each group, we do make the assumption that the distributions are "respectively normal" (Ho, 2009; Ho & Reardon, 2012). This means we assume there exists a continuous monotonic increasing function f defined for all values of y, such that $y^* = f(y)$ has a normal distribution within each group g:

$$y^*|(G=g) \sim N(\mu_g^*,\ \sigma_g^{*2}). \tag{1}$$

This does not require that the test score y be normally distributed within each group, only that the metric of y can be transformed so that this is true for the resulting transformed scores. Without loss of generality, we assume that f is defined so that y^* is standardized in the population, that is, $E[y^*] = 0$ and $\text{Var}(y^*) = 1$. Hence, we assume that there is a continuous scale for "academic achievement" (y^*) for which all within-group distributions are normal. Note that neither y nor y^* is assumed to be normally distributed in the combined population of all groups. We elaborate on the conceptual distinctions between these two metrics in Section 1.5.

We are interested in the case where neither y nor y^* is observed. Instead, we observe a "coarsened" version of y. This coarsened version, denoted $s \in \{1, \ldots, K\}$, is determined by $K - 1$ distinct ordered threshold values, $c_1, \ldots, c_{K-1}$, where $c_{k-1} < c_k$ for all k:

$$s \equiv k \text{ iff } c_{k-1} < y \leq c_k, \tag{2}$$

where we define $c_0 \equiv -\infty$ and $c_K \equiv +\infty$. Because f is a monotonic increasing function, s is also a coarsened version of y^*:

$$s \equiv k \text{ iff } c_{k-1}^* < y^* \leq c_k^*, \tag{3}$$

where $c_k^* = f(c_k)$. Under our assumption of respective normality, the model-implied proportion of observations in category k for group g is therefore:

$$\pi_{gk} = \Phi\left(\frac{\mu_g^* - c_{k-1}^*}{\sigma_g^*}\right) - \Phi\left(\frac{\mu_g^* - c_k^*}{\sigma_g^*}\right) = \Pr(c_{k-1}^* < y_{ig}^* \leq c_k^*) \equiv \Pr(c_{k-1} < y_{ig} \leq c_k), \tag{4}$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function. The aim is to estimate μ_g^* and σ_g^* for each group based on the observed frequencies of members of group g in each of the K ordered proficiency categories.

Equation 4 is an instance of the HETOP model. Here we think of each ordered proficiency category as the result of a draw from an underlying continuous (normal) distribution of test scores within a group. The HETOP model is an extension of the homoskedastic ordered probit (HOMOP) model that allows for heteroskedasticity in the variances of the underlying continuous variable across groups. In the remainder of the article, we refer to the ordered probit model in which all group variances are assumed equal as the HOMOP model. The ordered probit model is sometimes referred to as an ordered choice model (Williams, 2009) or as a location-scale model (Cox, 1995; McCullagh, 1980). Most broadly, it is an instance of a generalized linear model that parameterizes the multinomial distribution of observations in each group as cumulative probabilities from a normal density function (Agresti, 2002). Use of a HETOP model allows us to relax the often unrealistic assumption that test scores are homoskedastic across groups and to obtain direct estimates of the within-group standard deviations. To our knowledge, the HETOP model has not been used for the recovery of means and standard deviations from the coarsened data of multiple groups.

Our proposed application and interpretation of the HETOP model is analogous to the ordered regression model used in the analysis of receiver operating characteristic (ROC) curves, where the model can be interpreted as estimating the mean and standard deviation of unobserved (latent) normal distributions across multiple groups. Tosteson and Begg (1988) demonstrated that the HETOP model generalizes the binormal model for analyzing ROC curves comparing two groups (Dorfman & Alf, 1969) to scenarios with more than two groups. The binormal model has been used previously as a method to estimate the nonparametric V gap statistic between two groups when only coarsened proficiency data are available (Ho & Reardon, 2012; Reardon & Ho, 2015).[1] The HETOP model also generalizes the maximum likelihood (ML)–based estimator of V recommended by Ho and Reardon (2012). It effectively allows for simultaneous estimation of all pairwise V gaps on a common metric for three or more groups.

1.2 HETOP Model Estimation and Identification

Let $\mathbf{N}$ be an observed $G \times K$ matrix with elements n_{gk} containing the counts of observations in group g for which $s = k$; let $\mathbf{P} = [p_1, \ldots, p_G]$ be the $1 \times G$ vector of the groups' proportions in the population; and let $\mathbf{n} = [n_1, \ldots, n_G]$ be the $1 \times G$ vector of the observed sample sizes in each group, with $N = \sum_g n_g$.[2] We would like to estimate the vectors $\mathbf{M}^* = [\mu_1^*, \ldots, \mu_G^*]^t$, $\mathbf{\Sigma}^* = [\sigma_1^*, \ldots, \sigma_G^*]^t$, and $\mathbf{C}^* = [-\infty, c_1^*, \ldots, c_{K-1}^*, +\infty]$. In practice, it is preferable to estimate $\mathbf{\Gamma}^* = [\gamma_1^*, \gamma_2^*, \ldots, \gamma_G^*]^t$, where $\gamma_g^* = \ln(\sigma_g^*)$. This ensures that the estimates of σ_g^*

will all be positive. Following estimation of $\mathbf{\Gamma}^*$, we have $\widehat{\mathbf{\Sigma}}^* = [e^{\hat{\gamma}_1^*}, \ldots, e^{\hat{\gamma}_G^*}]^t$. Given $\mathbf{M}^*$, $\mathbf{\Gamma}^*$, and $\mathbf{C}^*$, and under the assumption of conditional independence of scores within groups, the log likelihood of drawing a sample with observed counts $\mathbf{N}$ is

$$\begin{aligned} L = \ln[P(\mathbf{N}|\mathbf{M}^*, \mathbf{\Gamma}^*, \mathbf{C}^*)] &= \sum_{g=1}^{G} \left\{ \ln(n_g!) + \sum_{k=1}^{K} [n_{gk} \ln(\pi_{gk}) - \ln(n_{gk}!)] \right\} \\ &= A + \sum_{g=1}^{G} \sum_{k=1}^{K} n_{gk} \ln(\pi_{gk}) \\ &= A + \sum_{g=1}^{G} \sum_{k=1}^{K} n_{gk} \ln \left[\Phi\left(\frac{\mu_g^* - c_{k-1}^*}{e^{\gamma_g^*}} \right) - \Phi\left(\frac{\mu_g^* - c_k^*}{e^{\gamma_g^*}} \right) \right], \end{aligned} \tag{5}$$

where $A = \ln\left(\frac{\prod_{g=1}^{G} n_g!}{\prod_{g=1}^{G} \prod_{k=1}^{K} n_{gk}!} \right)$ is a constant based on the observed counts in $\mathbf{N}$.

Without constraints on the parameters, the scale of $\mathbf{M}^*$, $\mathbf{\Gamma}^*$, and $\mathbf{C}^*$ is indeterminate up to a linear transformation. The constraints $\sum_g p_g \hat{\mu}_g^* = 0$ and $\sum_g p_g \hat{\mu}_g^{*2} + \sum_g p_g e^{2\hat{\gamma}_g^*} = 1$ together imply that y^* has mean 0 and variance 1, as desired. However, these nonlinear constraints are not easily implemented in standard software. Instead, it is easier to fit the model subject to two linear constraints on the parameters. As a default, we recommend the constraints

$$\begin{aligned} \mathbf{P}\widehat{\mathbf{M}}' &\equiv \sum_{g=1}^{G} p_g \hat{\mu}_g' = 0, \\ \mathbf{P}\widehat{\mathbf{\Gamma}}' &\equiv \sum_{g=1}^{G} p_g \hat{\gamma}_g' = 0, \end{aligned} \tag{6}$$

where we use a superscript prime symbol to denote the metric defined by the linear constraints.[3]

To estimate an HOMOP model, we impose the additional constraint that $\hat{\gamma}_1' = \hat{\gamma}_2' = \ldots = \hat{\gamma}_G'$ before maximizing Equation 5, so that all groups have identical standard deviations.[4] In some cases, as we discuss below, we may wish to fit a partially heteroskedastic ordered probit (PHOP) model, in which we constrain some subset of the groups to have identical standard deviations, but we allow the others to vary freely. This is achieved by adding to Equation 6 the constraint that the relevant elements of $\widehat{\mathbf{\Gamma}}'$ are equal to one another.

We can then maximize Equation 5 subject to the constraints, resulting in estimates $\widehat{\mathbf{M}}'$, $\widehat{\mathbf{\Gamma}}'$, and $\widehat{\mathbf{C}}'$ from which we obtain $\widehat{\mathbf{\Sigma}}' = [e^{\hat{\gamma}_1'}, e^{\hat{\gamma}_2'}, \ldots, e^{\hat{\gamma}_G'}]^t$. Note that the constraints listed in Equation 6 (or any set of linear constraints) do not, in general, yield estimates that satisfy the requirement that $\sum_g p_g \hat{\mu}_g'^2 + \sum_g p_g e^{2\hat{\gamma}_g'} = 1$. We can, however, standardize the estimates to recover estimates of $\mathbf{M}^*$, $\boldsymbol{\Sigma}^*$, and $\mathbf{C}^*$, using

$$\widehat{\mathbf{M}}^* = \frac{1}{\hat{\sigma}'}\widehat{\mathbf{M}}'$$

$$\widehat{\boldsymbol{\Sigma}}^* = \frac{1}{\hat{\sigma}'}\widehat{\boldsymbol{\Sigma}}'$$

$$\widehat{\mathbf{C}}^* = \frac{1}{\hat{\sigma}'}\widehat{\mathbf{C}}', \tag{7}$$

where $\hat{\sigma}'$ is an estimate of the population standard deviation in the metric defined by the constraints (the "prime" metric). We show in Appendix A that $\hat{\sigma}'$ can be estimated as

$$\hat{\sigma}' = \sqrt{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}, \tag{8}$$

where

$$\hat{\sigma}_W'^2 = \frac{\mathbf{P}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}}, \tag{9}$$

and

$$\hat{\sigma}_B'^2 = \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{[\mathbf{n}^{\circ -1}\circ(\mathbf{P}^{\circ 2} - \mathbf{P})]\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}}. \tag{10}$$

In these equations, $\widehat{\overline{\omega_g^2}}$ is the estimated average sampling variance of the $\hat{\gamma}_g'$ and the "$(\mathbf{A})^{\circ b}$" notation indicates the matrix whose elements are the corresponding elements of matrix $\mathbf{A}$ raised to the power b. Appendix A shows that for the HETOP model, we can use the approximation $\widehat{\overline{\omega_g^2}} \approx (2\tilde{n})^{-1}$, where $\tilde{n}$ is the harmonic mean of $n_g - 1$: $\tilde{n} = \left(\frac{1}{G}\sum_g \frac{1}{n_g - 1}\right)^{-1}$. For the HOMOP and PHOP models, $\widehat{\overline{\omega_g^2}}$ is approximated slightly differently (see Appendix A).

As we noted above, the model can be estimated with different constraints than those in Equation 6, as long as two independent constraints are used. However, if an alternate set of constraints are used, it is necessary to transform the resulting estimates to the metric defined by Equation 6 before standardizing using the procedure in Equation 7; we describe the necessary transformation in Online Appendix A (available in the online version of the journal). Absent problems maximizing the likelihood function, such as those discussed in Section 1.6, these transformation and standardization procedures will yield the same estimates of $\mathbf{M}^*$ and $\boldsymbol{\Sigma}^*$ regardless of the linear constraints imposed to identify the model.

ML estimation of the HETOP, HOMOP, and PHOP models can be conducted in a number of widely available statistical packages; see Greene and Hensher (2010, p. 179) for a fairly recent list. For all simulations and analyses described in this article, we carry out the ML estimation of Equation 5 using a modification of the `-oglm-` (“ordinal generalized linear models”) routine (Williams, 2010) written for *Stata* (StataCorp, 2013).[5]

1.3 Additional Estimands of Interest

Once we have obtained $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$, estimation of summary statistics like between-group gaps and ICCs is straightforward. First, the achievement gap between any two groups g and h can be computed as the standardized mean difference in y^* between the groups:

$$D_{gh} = \frac{\hat{\mu}_g^* - \hat{\mu}_h^*}{\sqrt{\frac{1}{2}\left(\hat{\sigma}_g^{*2} + \hat{\sigma}_h^{*2}\right)}}. \tag{11}$$

Note that, under the assumption of respective normality, D_{gh} is equal to V, a gap statistic invariant to monotonic scale transformations (Ho & Reardon, 2012).

Second, the ICC (the between-group share of total test score variance) is simply one minus the estimated within-group variance of y^*, because the total variance of y^* is 1:

$$\widehat{\text{ICC}} = 1 - \hat{\sigma}_W^{*2} = 1 - \left[\frac{\mathbf{P}(\widehat{\boldsymbol{\Sigma}}^*)^{\circ 2}}{1 + 2\widehat{\omega_g^2}}\right]. \tag{12}$$

1.4 Computation of Standard Errors

Once we have standardized the estimated group means and standard deviations using Equation 7, we can also compute their standard errors. Because the elements of $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ are the products of error-prone estimates of σ' and error-prone elements of $\mathbf{M}'$ and $\boldsymbol{\Sigma}'$, the standard errors of the elements of $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ will depend on the variances and covariances of $\hat{\sigma}'$ and the elements of $\widehat{\mathbf{M}}'$ and $\widehat{\boldsymbol{\Sigma}}'$. In Appendix B, we derive formulas to estimate $\mathbf{V}^*$ and $\mathbf{W}^*$, the sampling variance–covariance matrices of $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$, respectively, when the model is fit with the constraints $\mathbf{P}\widehat{\mathbf{M}}' = 0$ and $\mathbf{P}\widehat{\boldsymbol{\Gamma}}' = 0$. These derivations take into account the sampling error in $\hat{\sigma}'$.

The standard errors of the gaps described in Equation 11 can be computed from $\widehat{\mathbf{M}}^*$, $\widehat{\mathbf{\Sigma}}^*$, $\widehat{\mathbf{V}}^*$, and $\widehat{\mathbf{W}}^*$, as described in Online Appendix B (available in the online version of the journal). There are generalizations of the formulas used in Reardon and Ho (2015).

The standard error of the ICC is relatively straightforward to compute once we have $\widehat{\mathbf{W}}^*$. Given Equation 12, the variance of the ICC estimator will be

$$\mathrm{Var}(\widehat{\mathrm{ICC}}) \approx \left(\frac{1}{1+2\widehat{\omega_g^2}}\right)^2 \mathrm{Var}[\mathbf{P}(\widehat{\mathbf{\Sigma}}^*)^{\circ 2}] \approx 4\left(\frac{1}{1+2\widehat{\omega_g^2}}\right)^2 \mathbf{P}[\mathrm{diag}(\mathbf{\Sigma}^*)]\mathbf{W}^*[\mathrm{diag}(\mathbf{\Sigma}^*)]\mathbf{P}^t. \tag{13}$$

Substituting $\widehat{\mathbf{\Sigma}}^*$ and $\widehat{\mathbf{W}}^*$ and the appropriate approximation of $\widehat{\omega_g^2}$ (see Appendix A) into Equation 13 yields an estimate of the variance of the ICC estimator.

1.5 A Note on Interpreting the Estimated Parameters

There are two different test score scales relevant to the interpretation of estimated parameters. The first is the continuous scale in which test scores are constructed (i.e., the scale score metric of a test as constructed by a test developer). We denote the scores measured in this metric (the original test metric) with the variable y and denote estimates based on these scores as $\hat{\mu}_k$ and $\hat{\sigma}_k$. The second is the scale of the standardized estimates produced by the HETOP model. We denote the scores measured in this metric with the variable y^* and denote estimates in this metric with a superscript "star" (e.g., $\hat{\mu}_k^*$ and $\hat{\sigma}_k^*$). The estimates in this metric are interpreted relative to a population mean and standard deviation of 0 and 1, respectively. Scores in the prime metric described in Section 1.2 are simply a linear transformation of y^* used in the process of model estimation and are not relevant to the final interpretation of y^*.

If the function f that transforms y into y^* is nonlinear, then the group means and standard deviations in the y^* metric will not be linearly related to those in the original y metric. In other words, the target parameters of our application of the HETOP model are not the test score means and standard deviations in the (potentially observed) test score metric of y. Rather, they are the means and standard deviations in the continuous metric of y^*—the metric in which each group's distribution is normal and in which the population distribution has mean 0 and standard deviation 1. The key assumption of the model is that such a metric exists. That is equivalent to saying that the group distributions of y (if y could be observed) are "respectively normal," as defined above.

In some cases, these parameters may be unsatisfying. If, for example, we want to recover means and standard deviations in the reported metric of y (e.g., if we want to recover group-specific mean SAT scores, expressed in the SAT score scale metric), we could do so if two conditions are met. First, we would need to know the threshold scores (in the original metric) used to coarsen the data (i.e., we would need to know $c_1, \ldots, c_{K-1}$, where $K \geq 3$). Second, we would have to assume that the group-specific distributions of scores are normally distributed in the original metric (rather than assuming only that they could be normalized by some common transformation *f*). If these two conditions are met, we could fit the HETOP model using the frequency counts within each ordered category, as above, except that we would constrain the vector **C** to have values equal to the known threshold scores (rather than imposing constraints on the vectors of estimated means and standard deviations). The vectors $\widehat{\mathbf{M}}'$ and $\widehat{\mathbf{\Sigma}}'$ would then be freely estimated and would be interpretable as group means and standard deviations in the original score metric *y*. From a practical standpoint, if the group distributions in the original metric are already normal (or nearly normal), the *y* and y^* estimates will differ only by a linear transformation (or a nearly linear transformation). While the scale of the means and standard deviations will thus differ, auxiliary statistics such as standardized mean differences or the estimated ICC will be unchanged (or nearly unchanged).

When it is reasonable to assume distributions of the original scores *y* are normal within each group and it is desirable to obtain estimates in that metric, then constraining the thresholds to their known values may be preferable. Unlike physical properties like height or weight, however, it is not clear that there is any natural cardinal scale for cognitive or academic skill (Lord, 1980) or that the test design principles necessary to support cardinality have been addressed for many common test score scales (Briggs, 2013). In many cases, then, the fixed intervals between established cut scores defined in the original scale score metric might have little theoretical justification or relevance and may be unnecessarily restrictive. The y^* metric provides a unique metric interpretable in standard deviation units for comparing test score distributions across groups. The y^* metric also preserves the ordinal structure of the observed data, while remaining invariant to plausible monotonic transformations of the original test score scale (i.e., it does not rely on the cardinality of the original test score scale). We therefore use the parameters on the y^* scale as targets for simulation and interpretation.

1.6 Estimation Issues and the Partially Heteroskedastic Ordered Probit Model

The HETOP model will be unidentified if there are groups in which all observations fall in the highest or lowest proficiency category. In this case, the ML estimates of these groups' means will be $\pm\infty$. The model is also unidentified if there are groups in which all observations fall into only two adjacent categories; in such cases the MLE of the log standard deviation does not exist. In such cases, the HETOP model does not have enough information to provide estimates. In other cases, heteroskedastic multinomial or ordered probit models may suffer from fragile identification (Freeman, Keele, Park, Salzman, & Weickert, 2015; Keane, 1992), meaning that although the model is formally identified, the likelihood function may be nearly flat over a range of the parameter space. This may result in a near-singular Hessian, failure of the estimation algorithm to converge, or convergence with very large standard errors. In addition, ML algorithms can sometimes indicate convergence even when the multinomial or ordinal probit model is formally unidentified, due to approximation errors in estimating the likelihood function (Horowitz, Sparmann, & Daganzo, 1982; Keane, 1992).

In our simulations and in applying the HETOP model to real data, we found evidence of such fragile identification in some cases. This occurred when one or more groups had sparse data—for example, when the coarsened data showed all members of a group scoring in the same one or two ordered categories. This condition is unlikely to occur unless more than one of the following conditions hold: the group has a relatively high or low mean, a small standard deviation, a small sample size, and/or the cut scores are narrowly or unevenly spaced. In such cases, the HETOP algorithm sometimes either failed to converge or converged and returned estimates with very large standard errors for particular groups' parameter estimates (often many orders of magnitude larger than those for other, well-identified groups). In some cases, the algorithm would converge using one set of constraints but not another or would converge with two different sets of constraints but result in differing estimates of μ_g^* and σ_g^*, suggesting these parameters were at best tenuously identified and not to be trusted.[6]

In such cases, one can drop sparsely populated groups from the model and fit the HETOP model only with groups with sufficient data to identify their parameters. One disadvantage of this is that the standardization procedure we describe will no longer include the full population of interest. An alternate solution is to fit a PHOP (or HOMOP) model instead of the HETOP model, imposing some constraints on the standard deviations of the groups with sparse data. For example, constraining all groups with small sample sizes, or with similar values of some covariate, to have the same standard deviation allows the model to use information from multiple groups to estimate a common standard deviation for those groups. As long as at least some of the constrained groups have sufficient data to identify the parameters, the fragile identification problem may be avoided. We describe simulation analyses of such a model in Section 2.3, where we find that the PHOP model often yields a smaller root mean squared error (RMSE) than the HETOP model, even when the groups' true standard deviations are not identical.

2. Evaluating the Performance of the HETOP and PHOP Models Using Simulated Data

We conducted a Monte Carlo simulation to evaluate the accuracy of our proposed use of the HETOP model (and our described standardization procedure) when the data generating procedure matches the model. The first simulation study uses a range of conditions selected to represent those likely to be encountered when analyzing coarsened proficiency data in practice. It builds upon prior simulation studies of HETOP models that sought to recover individual-level parameters rather than group parameters (e.g., Keele & Park, 2006). We focus directly on recovery of the means, standard deviations, and ICCs of the continuous y^* variable after applying our proposed standardization procedure, including evaluation of bias, sampling variability, and confidence interval (CI) coverage of the estimated standard errors. We also evaluate the performance of the PHOP model as a potential way to overcome estimation problems caused by small sample sizes.

2.1 HETOP Model Simulation Conditions and Procedure

We simulated data from populations that differ in the degree to which the true means and standard deviations of test scores vary among groups. We characterize the variation in group means using the ICC (the proportion of total variance in test scores that lies between groups) and the variation in group standard deviations using the coefficient of variation (CV) of group variances (defined as $\mathrm{CV} = SD(\sigma^2)/E[\sigma^2]$). We first created four populations, each defined by an ICC (0.05 or 0.20) and a CV (0.0 or 0.3) and containing 100 groups. In each population, each of the 100 groups has 1 of 10 uniformly spaced true means and 1 of 10 uniformly spaced true standard deviations (when $\mathrm{CV} = 0$, all groups have identical standard deviations), with the set of means and standard deviations defined so that the population has the desired ICC and CV, and an overall mean of 0 and total variance of 1. We selected the ICC and CV values to correspond roughly to the high and low ends of values reported in prior literature on test score variation (Hedberg & Hedges, 2014; Hedges & Hedberg, 2007) and the real test score data we analyze later in this article.

In each of the four populations, we conducted four sets of simulations, each defined by groups of a different sample size ($n = 25, 50, 100, 400$). For each of the 16 resulting simulation scenarios defined by the ICC, CV, and group sample size, we generated random samples of size n from each of the 100 groups. Each group's sample was drawn from normal population distributions with means and standard deviations defined by the parameters for each of the 100 groups. We then coarsened the scores four different ways, each time using a different set of cut score locations (set at the 20th/50th/80th, 5th/50th/95th, 5th/30th/55th, and 5th/25th/50th/75th/95th percentiles of the population test score distribution, and described as "mid," "wide," "skewed," and "many" cut scores, respectively). The cut score locations were chosen to be representative of a wide range of conditions found in empirical coarsened test score data (Reardon & Ho, 2015). Finally, we fit both the HETOP and HOMOP model to the coarsened data and followed the procedures described above to obtain $\widehat{\mathbf{M}}^*$, $\widehat{\boldsymbol{\Sigma}}^*$, the estimated ICC, and their standard errors. For each of the 64 scenarios, we repeated this process 1,000 times.

Although our primary goal is to assess the performance of the HETOP model, we also fit the HOMOP model to each simulated data set in order to compare the relative performance of the two models. Fitting both the HOMOP and HETOP models to data generated from a population that is homoskedastic (CV $= 0.0$) allows us to assess whether using the HETOP model when it is not needed leads to bias or inefficient estimation relative to the more appropriate HOMOP model. Likewise, fitting both models to data generated from a population that is heteroskedastic (CV $= 0.3$) allows us to assess whether and how much the use of the HETOP model improves estimation relative to the HOMOP model.

We evaluated the performance of the HETOP and HOMOP models by computing the bias and RMSE of the estimated means, standard deviations, and ICCs. For the means and standard deviations, we focused primarily on the aggregate bias and RMSE (averaged across all $G = 100$ groups) for each estimator $\hat{\theta}$ (where θ could be a mean or a standard deviation) by computing

$$\begin{aligned} \text{Bias}_{\hat{\theta}} &= \frac{1}{RG}\sum_{r=1}^{R}\sum_{g=1}^{G}(\hat{\theta}_{gr} - \theta_g) \\ \text{RMSE}_{\hat{\theta}} &= \sqrt{\frac{1}{RG}\sum_{r=1}^{R}\sum_{g=1}^{G}(\hat{\theta}_{gr} - \theta_g)^2}, \end{aligned} \tag{14}$$

where R is the number of converged replications (usually 1,000), $\hat{\theta}_{gr}$ is the estimate for group g in replication r, and θ_g is the true value. For the ICC estimates, $\widehat{\text{ICC}}$, bias, and RMSE were computed as

$$\begin{aligned} \text{Bias}_{\widehat{\text{ICC}}} &= \frac{1}{R}\sum_{r=1}^{R}(\widehat{\text{ICC}}_r - \text{ICC}) \\ \text{RMSE}_{\widehat{\text{ICC}}} &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\text{ICC}}_r - \text{ICC})^2}. \end{aligned} \tag{15}$$

To evaluate the accuracy of our formulas for the standard errors of group means and standard deviations, we computed the average ratio of the median[7] estimated standard error of a parameter to its empirical standard error (the standard deviation of the sampling distribution of the parameter, denoted $SD(\hat{\theta}_{gr})$) across all $G = 100$ groups in a condition:

$$\text{Standard error ratio}_{\hat{\theta}} = \frac{1}{G}\sum_{g=1}^{G}\frac{\text{Median}(\widehat{SE}_{\hat{\theta}_{gr}})}{SD(\hat{\theta}_{gr})}. \tag{16}$$

To evaluate the accuracy of the standard error formula of the estimated ICCs, we compute the ratio of the median estimated standard error of the ICC to its empirical standard error:

$$\text{Standard error ratio}_{\widehat{\text{ICC}}} = \frac{\text{Median}(\widehat{SE}_{\widehat{\text{ICC}}_r})}{SD(\widehat{\text{ICC}}_r)}, \tag{17}$$

where $SD(\widehat{\text{ICC}}_r)$ is the observed standard deviation of the sampling distribution of the ICC estimates across the R replications for a given condition. If our standard error formulas in Appendix B are accurate, we expect the ratios in Equations 16 and 17 to be close to 1. We also computed the 95% CI coverage rates for each parameter, computed as the proportion of cases for which $|\hat{\theta}_{gr} - \theta_g| < 1.96 \times \widehat{SE}_{\theta_{gr}}$ or $|\widehat{\text{ICC}}_r - \text{ICC}| < 1.96 \times \widehat{SE}_{\text{ICC}_r}$. If the estimates are biased, the CI coverage rates will not equal 95%, however, even if the standard error formulas accurately reflect sampling variability.

Finally, we present results describing the loss of efficiency (in terms of increased sampling variance) when estimating group means and standard deviations from coarsened rather than full data. For each condition, we estimate relative efficiency as the average efficiency ratio across groups. We define this as the average (across groups) of the ratio of the observed sampling variance of the target parameter in the simulations (using coarsened data) to its theoretical sampling variance if it were estimated from continuous data:

$$\text{Average efficiency ratio}_{\hat{\theta}} = \frac{1}{G}\sum_{g=1}^{G} \frac{\widehat{\text{Var}}\left(\hat{\theta}_{gr}\right)}{\tau^2_{\theta_g}}, \tag{18}$$

where θ is either a mean or a standard deviation, $\widehat{\text{Var}}\left(\hat{\theta}_{gr}\right)$ is the observed variance in estimates of the target parameter across the R replications and $\tau^2_{\theta_g}$ is the theoretical sampling variance of the estimator based on continuous data (when θ is the mean, μ^*_g, then $\tau^2_{\theta_g} = \sigma^2_g/n_g$; when θ is the standard deviation, σ^*_g, then $\tau^2_{\theta_g} = \sigma^2_g \times 2[n_g - 1]^{-1}$). An efficiency ratio of 1.0 would indicate that estimates based on coarsened data have the same sampling variance as estimates based on continuous data; efficiency ratios larger than 1.0 indicate there is greater variability in estimates based on coarsened data. The efficiency ratio can be interpreted as the ratio by which the sample size would need to be increased to estimate the parameters from coarsened data with the same precision as if the parameters were estimated from continuous data.

Because the 100 true group means and standard deviations were held constant across the 1,000 replications within a given scenario, we also examine the bias, RMSE, and standard error performance for individual groups within a particular condition when relevant. Online Appendix C (available in the online version of the journal) contains detailed tables of all aggregate bias, RMSE, and standard error results. We used the constraints $\mathbf{P}\widehat{\mathbf{M}}' = 0$ and $\mathbf{P}\widehat{\mathbf{\Gamma}}' = 0$ to identify the model, and the ML algorithm converged in all but 5 of the total 128,000 replications.

2.2 HETOP Model Simulation Results

2.2.1 Recovery of means. The aggregate bias of means estimated with the HETOP model was indistinguishable from 0 for all conditions, and the bias for individual groups was also indistinguishable from 0 in almost all of the scenarios we explored. The one exception was in the small sample ($n = 25$) simulations with a large ICC (0.20), large CV (0.3), and skewed or wide cut scores; in these cases, we detected nonzero bias for some groups, though the bias was very small, nearly always less than 0.05 standard deviation units for any given group. Moreover, not only was this bias small in absolute terms, but it was also very small in relation to the aggregate RMSE of the estimated means (which in this case was on average approximately 10 times larger than the largest bias we observed). We do not show these results for parsimony. The precision of the estimated means varied primarily as a function of sample size, although when sample sizes were small, the sampling variance was modestly affected by the location of the cut scores; sampling variance was consistently lowest for estimates based on the "many" cut score condition, as one would expect given the light degree of coarsening.

2.2.2 Recovery of standard deviations. The top panel of Figure 1 shows the average bias in standard deviation estimates across all groups and all replications for each condition. This figure illustrates that there is some negative bias in the standard deviation estimates from the HETOP model and that the bias is primarily a function of sample size that is exacerbated when cut scores are skewed or wide. Note the average bias for standard deviations is quite small compared to the true standard deviation of scores, typically less than 1% of the size of the true standard deviations, except when sample sizes are less than 50 and the cut scores are skewed or wide (the average standard deviation is approximately 0.89 when ICC = 0.20 and 0.97 when ICC = 0.05, and the largest absolute bias in any condition is approximately 0.045).

When CV = 0 and the HOMOP model is the correct model, the top panel of Figure 1 indicates a very slight negative bias in HOMOP standard deviation estimates that generally approaches 0 with increasing sample size more quickly than the corresponding HETOP estimates, particularly with skewed or wide cut scores. That is, when the group distributions are truly homoskedastic and the coarsening is done suboptimally, the HOMOP model produces less biased estimates of standard deviations than the HETOP model. Note, however, that the HOMOP model produces modest positive bias in the standard deviation estimates in the CV = 0.3 conditions (where the HOMOP model is not the correct

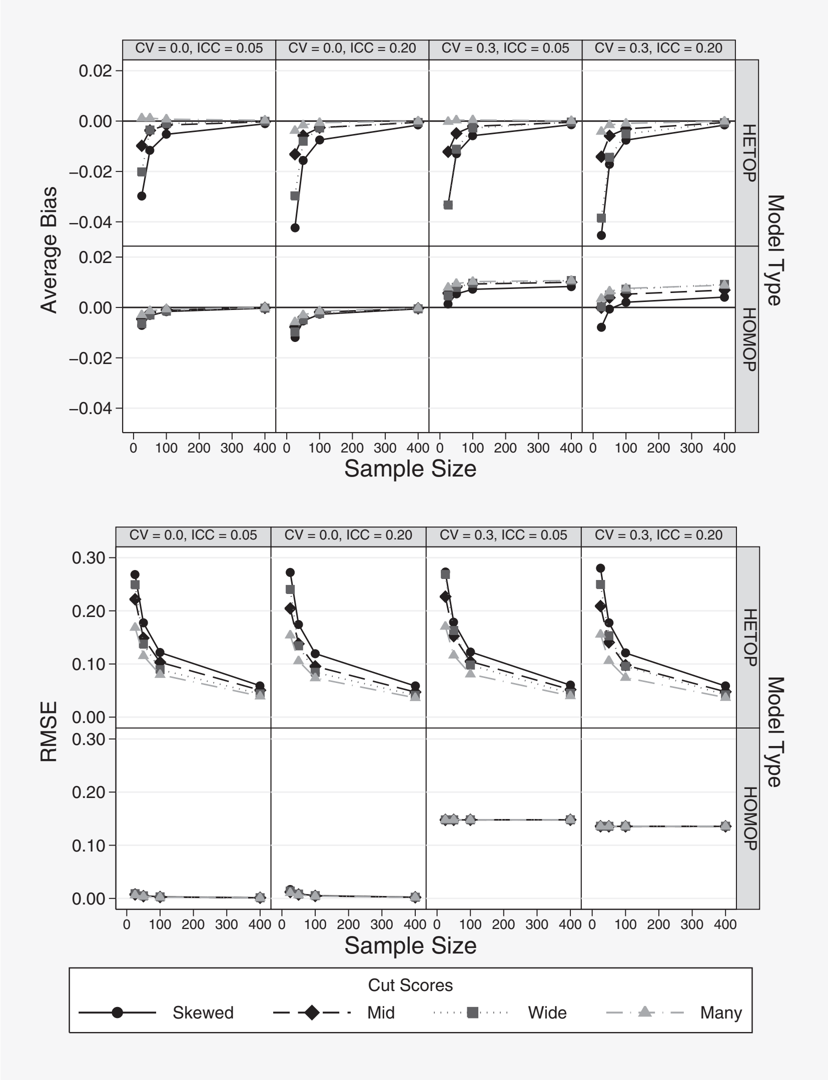

FIGURE 1. *Average bias and aggregate root mean squared error (RMSE) in group standard deviation estimates.*

model), particularly when sample sizes are large. Given the misspecification of the model, such bias is not surprising.

The bottom panel of Figure 1 depicts the RMSE of standard deviation estimates across conditions as defined in Equation 14. The results in Figure 1 together suggest that when the data are homoskedastic, the HOMOP model is generally

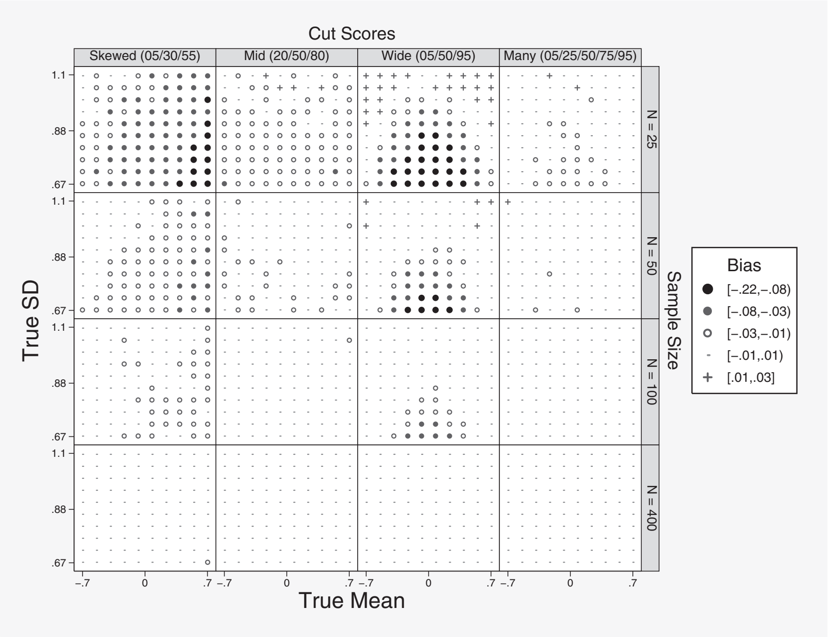

FIGURE 2. *Bias in standard deviation estimates by true group mean and standard deviation (for intraclass correlation = 0.20 and coefficient of variation = 0.3 condition).*

(and unsurprisingly) preferable to the HETOP model. If data are heteroskedastic, however, the HOMOP model will systematically over-/underestimate individual group standard deviations, with bias inversely related to the true standard deviation. Nonetheless, if one wishes to minimize RMSE, it may still be better to use an HOMOP model if sample sizes are small. In the scenarios shown in Figure 1, the HOMOP model generally yields a smaller RMSE than the HETOP model for scenarios with $n < 100$. The sample size at which the HOMOP model is preferable to the HETOP model (in terms of RMSE) will be a function of a number of factors, particularly the CV of group variances and the location of the cut scores. We further investigate this bias/variance trade-off in Section 2.3.

Figure 2 provides more detail on the systematic patterns of bias in group standard deviation estimates from the HETOP model, showing the bias in standard deviation estimates as a function of the true population means and standard deviations for the condition in which ICC = 0.20 and CV = 0.3. Each panel in Figure 2 shows the bias as a function of groups' true mean and standard deviation for a given sample size and cut score condition, with the *x*-axis indicating group means and y-axis indicating group standard deviations. The figure makes clear that the bias in estimated standard deviations varies with a group's true mean and standard deviation, when cut scores are skewed or wide and sample sizes are small. The top

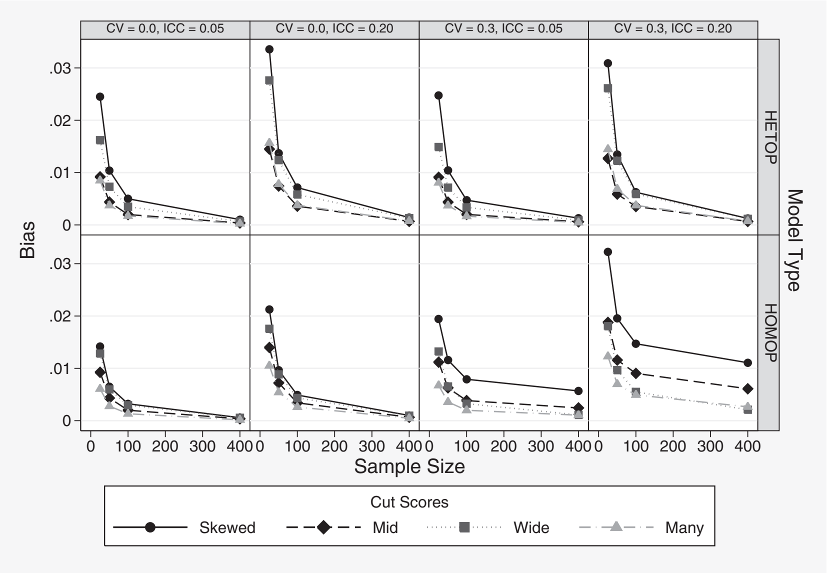

FIGURE 3. *Bias in intraclass correlation estimates.*

left panel, for example, shows that nearly all group standard deviation estimates are negatively biased when $n = 25$ and cut scores are skewed, but the bias is largest for groups with larger true means and smaller true standard deviations. This pattern is a result of the loss of information due to the coarsening of the data. If a group's true standard deviation is small and its true mean is high, then when the cut scores are skewed, the coarsening leads to observed data with little information (most cases will fall in the top category, providing little information about the group's standard deviation) and to underestimation of standard deviations. In larger samples, coarsening leads to much less consequential loss of information, however, as is evident in the bottom ($n = 400$) row of panels, where absolute bias is less than 0.01 for all but one group across all cut score conditions. Although not pictured, the pattern of small negative bias in small groups is similar when CV = 0.0.

2.2.3 Recovery of ICC. Figure 3 shows the bias in ICC estimates across conditions. The ICC estimates are upwardly biased, particularly when sample sizes are small; this is partly a result of the small negative bias in standard deviations in these cases (see Figure 1). The bias in the HETOP ICC estimates does not appear to depend on the true CV but is modestly larger when the true ICC is larger. When CV = 0, the HETOP and HOMOP estimates are similarly biased in most cases, though the bias is slightly lower in the HOMOP model when the cut scores are skewed or wide. When the data are heteroskedastic (CV = 0.3), the HOMOP

TABLE 1.
Ratio of Median Estimated Standard Error to Empirical Standard Error and 95% Confidence Interval Coverage for HETOP Estimates by Parameter, Sample Size, and Cut Scores

		Group Mean		Group Standard Deviation		ICC	
Sample Size	Cut Scores	Ratio	Coverage	Ratio	Coverage	Ratio	Coverage
25	Skewed (05/30/55)	0.923	0.943	0.865	0.892	0.914	0.975
	Mid (20/50/80)	0.944	0.945	0.887	0.906	1.133	0.942
	Wide (05/50/95)	1.608	0.983	1.881	0.995	3.251	0.994
	Many (05/25/50/75/95)	0.965	0.936	0.933	0.927	1.073	0.897
50	Skewed (05/30/55)	0.954	0.946	0.930	0.920	1.084	0.933
	Mid (20/50/80)	0.973	0.948	0.944	0.929	1.084	0.940
	Wide (05/50/95)	1.043	0.957	0.923	0.965	1.476	0.943
	Many (05/25/50/75/95)	0.982	0.943	0.967	0.939	1.044	0.925
100	Skewed (05/30/55)	0.974	0.947	0.964	0.935	1.053	0.924
	Mid (20/50/80)	0.984	0.948	0.972	0.940	1.019	0.933
	Wide (05/50/95)	0.996	0.948	0.957	0.951	1.093	0.929
	Many (05/25/50/75/95)	0.990	0.946	0.983	0.945	1.019	0.932
400	Skewed (05/30/55)	0.995	0.949	0.992	0.946	0.995	0.935
	Mid (20/50/80)	0.999	0.950	0.993	0.948	0.995	0.944
	Wide (05/50/95)	0.998	0.949	0.993	0.949	1.006	0.946
	Many (05/25/50/75/95)	1.000	0.949	0.996	0.948	0.997	0.944

Note. ICC = intraclass correlation coefficient; Ratio = ratio of median estimated standard error to empirical standard error; Coverage = confidence interval coverage rate of an estimated 95% confidence interval; HETOP = heteroskedastic ordered probit model.

ICC estimates are biased even when $n = 400$, due to the misspecification of the model. In all cases, the bias is largest when cut scores are skewed or wide. Overall, however, the bias in ICCs is relatively small, generally less than 0.01 unless $n = 25$ or the cut scores are skewed.

2.2.4 Accuracy of standard errors. The accuracy of the standard errors in the simulations was similar across ICC and CV conditions. For parsimony and to limit sampling variability, Table 1 shows the standard error ratios and CI coverage rates averaged across the four combinations of ICC (0.05 and 0.20) and CV (0.0 and 0.3) conditions. Table 1 indicates that estimated standard errors and CIs for all three parameters were accurate with moderate and large sample sizes ($n = 100$ or more), but less accurate with smaller sample sizes. Standard errors and CIs were least accurate with small sample sizes when cut scores were widely spaced. In such cases, the approximations used to derive the standard error formulas (Appendix B) appear to break down.

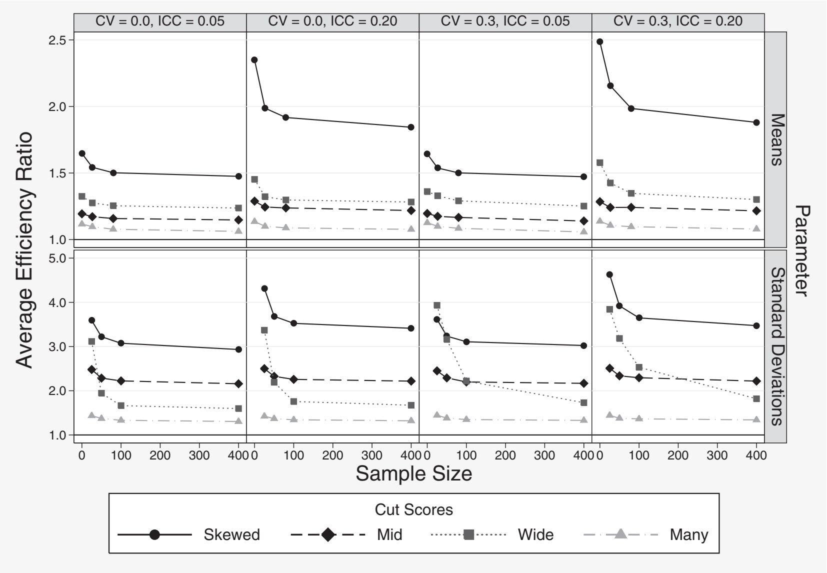

FIGURE 4. *Average efficiency ratios of estimated means and standard deviations using the heteroskedastic ordered probit model.*

2.2.5 Efficiency of estimators. Figure 4 presents the average efficiency ratio across all 100 groups for each condition when using the HETOP model. The top panel shows average efficiency ratios for estimated means while the bottom panel shows the average efficiency ratios for the standard deviations; each panel represents a different ICC and CV condition while each line represents a different cut score condition. For the means, the loss of efficiency is moderate and depends primarily on the cut score locations, with the greatest loss of efficiency when the cut scores are skewed. Within any combination of cut scores, CV, and ICC, the relative loss of precision is largest when samples are small. The average efficiency ratio for estimated means across all groups and conditions is 1.36, ranging from a minimum of 1.06 (in the case where there are many cut scores, a CV of 0.3, an ICC of 0.05, and $n = 400$) to a maximum of 2.49 (in the case where the cut scores are skewed, the CV is 0.3, the ICC is 0.20, and $n = 25$). This indicates that in some conditions, the coarsening of the data results in very little loss of precision, while in others (very small group sizes and skewed cut scores) the loss of precision is more substantial.

The efficiency loss with respect to estimating group standard deviations is larger than when estimating means, but again, the efficiency ratio varies considerably depending on cut score locations and sample size. The skewed cut score condition is consistently the least efficient, and the many cut scores condition is consistently the most efficient, with average efficiency ratios of 3.53 and 1.37, respectively, averaging across all CVs, ICCs, and group sizes. The efficiency ratio in the wide cut score condition appears to be most dependent upon group sample sizes: When group sample sizes are 25, the average efficiency ratios range from 3.12 to 3.93 across the ICC and CV conditions; they are half as large (1.60–1.82) when sample sizes are 400.

2.3 PHOP Model Simulation Conditions and Procedure

When the data are truly homoskedastic, the simulation results in Section 2.2 show that an HOMOP model with all group standard deviations constrained to equality performs better than a fully heteroskedastic model. However, the results also suggest that in some truly heteroskedastic cases with small sample sizes, the HOMOP model may be preferable, as reductions in RMSE could outweigh increases in bias for group standard deviation estimates. In the simulations above, however, all groups in a given simulation scenario had the same sample sizes, a condition that may not often hold in practice.

Anticipating contexts in which sample sizes across groups differ, we evaluate the performance of a PHOP model in which standard deviation estimates for small groups are constrained to equality while those for large groups are freely estimated. Because the efficiency/bias trade-off implicit in constraining group standard deviations will depend on how much true variation in standard deviations there is, we conduct these simulations in populations with different degrees of heteroskedasticity (CV).

We follow the same general simulation methodology as outlined in Section 2.1, but with the following modifications. We generate data from five populations, each with ICC $= 0.20$, and one of five different CVs of group variances (0.0, 0.1, 0.2, 0.3, or 0.4). Each population contains 36 group types whose means and standard deviations are bivariate uniformly distributed with values set to produce the defined ICC and CV. For each group type, we draw seven random samples of sizes $n = 25, 50, 75, 100, 150, 200$, and 400 from normal distributions defined by each of the 36 group mean/standard deviation values. This yields $7 \times 36 = 252$ groups, one of each combination of mean, standard deviation, and sample size. We then coarsen these scores four separate times, using the same cut scores described above (i.e., the “mid,” “many,” “skewed,” and “wide” conditions). For each coarsened sample, we then fit eight different models: the HETOP and HOMOP models as well as PHOP models where groups with sample sizes less than or equal to 25, 50, 75, 100, 150, or 200 were all constrained to be equal. We repeated this process for 1,000 replications. We used the constraints $\mathbf{P}\widehat{\mathbf{M}}' = 0$ and $\mathbf{P}\widehat{\boldsymbol{\Gamma}}' = 0$ to identify the model. The ML algorithm converged in all 40,000 replications using the “many” cuts cores and failed to converge in 9, 13,

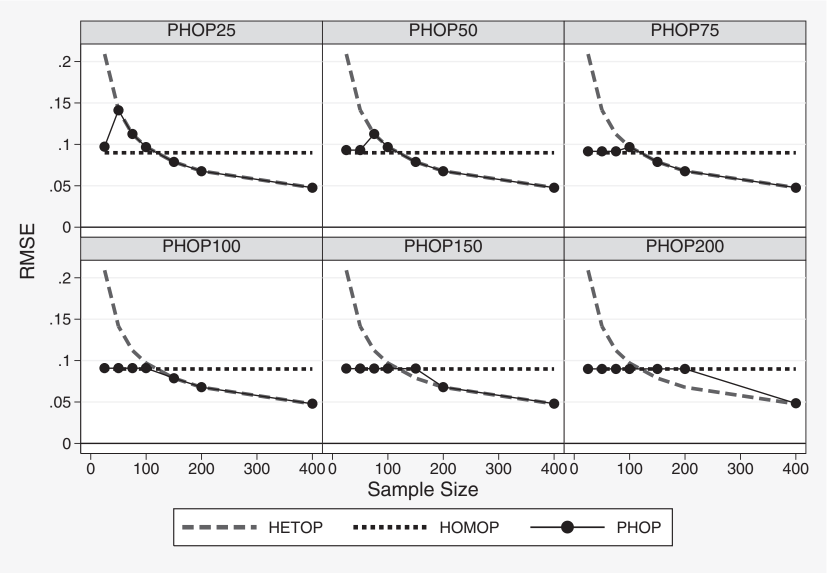

FIGURE 5. *Aggregate root mean squared error (RMSE) of group standard deviation estimates, by group sample size and PHOP model type with CV = 0.2. HETOP = heteroskedastic ordered probit model; HOMOP = homoskedastic ordered probit model; PHOP = partially heteroskedastic ordered probit model; CV = coefficient of variation. The HETOP and HOMOP lines are included for reference and are constant across all six panels.*

and 271 of the 40,000 replications using the “mid,” “skewed,” and “wide” cut scores, respectively.

2.4 PHOP Model Simulation Results

In general, results for the bias, RMSE, and standard errors were similar for the overlapping HETOP and HOMOP conditions here and in Section 2.2, suggesting that the conclusions above remain largely unchanged for conditions with groups of varying sample sizes. However, in some cases, there was less bias in HETOP standard deviation estimates for groups with small sample sizes in the simulations with a range of group sizes. This may result from the fact that the overall standard deviation ($\hat{\sigma}'$) is more accurately estimated when there are some groups with large sample sizes, so that less bias is introduced when we divide by this estimated standard deviation to obtain the $\hat{\sigma}^*$ estimates. For the PHOP models, the accuracy of the estimated standard errors was very good: The ratio of median estimated to empirical standard errors was close to 1 in all cases.

Our primary motivation for testing the PHOP models was to assess whether they reduce the aggregate RMSE of group standard deviation estimates.

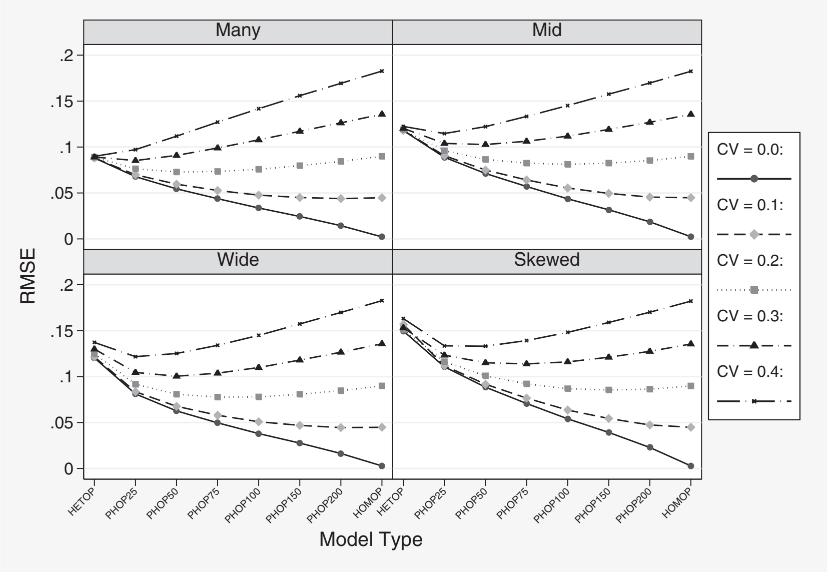

FIGURE 6. *Aggregate root mean squared error (RMSE) of group standard deviation estimates, by cut score location, coefficient of variation (CV), and model type (HETOP = heteroskedastic ordered probit; PHOP = partially heteroskedastic ordered probit, with standard deviations constrained to equality when sample sizes are equal to or below the number noted; HOMOP = homoskedastic ordered probit).*

Estimating a single pooled standard deviation estimate across small groups will yield more precise (but potentially biased) estimates of small groups' group standard deviations; if the increase in bias is outweighed by the reduction in error, the PHOP model may be preferred. Hence, our discussion of the results in this section focuses on the RMSE of the standard deviations of groups of various sizes.

Figure 5 displays the RMSE of group standard deviation estimates for different PHOP models in the condition in which CV = 0.2 and data were coarsened with the "mid" cut scores. Each panel of the figure displays the aggregate RMSE of standard deviation estimates for a different PHOP model (e.g., PHOP25 is a model in which group standard deviations are constrained to equality for groups with $n \leq 25$); each panel also includes results for the HOMOP (dotted line) and HETOP (dashed line) models, which are the same across panels, as they are not affected by the sample size threshold used in the PHOP model. We show RMSE disaggregated by group size here (unlike in Figure 1) because the PHOP model treats groups of different sizes differently by design. Figure 5 shows that the RMSEs of constrained group standard deviation estimates in the PHOP model are nearly identical to HOMOP model RMSEs while the unconstrained group standard deviation estimates are nearly identical to the HETOP RMSEs. The pattern of constrained and unconstrained group RMSEs tracking the HOMOP and HETOP model results was consistent across CV values (not shown).

Figure 5 suggests there will be an optimal sample size threshold at which to constrain standard deviation estimates to minimize the overall RMSE. Figure 6 displays information useful for determining such a threshold for each CV-by-cut scores condition. Each panel of Figure 6 shows RMSE of group standard deviations (aggregated across all sample sizes) for each model type and each CV condition. The upper left panel, for example, shows the results for the "many" cut score condition and includes a line for each CV condition. For a given CV and cut score condition, an optimal threshold can be identified by finding the model that minimizes RMSE for the corresponding line. When the true CV is 0, the HOMOP model minimizes RMSE in all conditions. When the true CV is 0.2, the optimal models (among those we tried) would be PHOP50, PHOP100, PHOP75, and PHOP150 for the many, mid, wide, and skewed cut score conditions, respectively. Although these results do not cover all possible combinations of ICC, CV, and cut score locations, they are suggestive about the conditions under which a PHOP model would minimize the RMSE of group standard deviation estimates. In analysis of real data, analysts will know the location of the cut scores, the number of groups, and the group sizes; they may also have information about the range of plausible values of the ICC and CV. These could be used to conduct customized simulations of the type we show here to make an informed decision about the optimal HETOP/PHOP/HOMOP model to select to minimize RMSE, if that is their goal.

2.5 Summary of Simulation Analyses

These simulations demonstrate that the HETOP model works well when the model matches the data generating process. Unbiased and precise recovery of standard deviations generally requires group sizes of 100 or more. Figure 1 suggests that in some cases where sample sizes are small, a homoskedastic model may produce more efficient (although biased) standard deviation estimates even if the data are truly heteroskedastic. The results in Section 2.4 suggest that using a PHOP model, which constrains small groups to have equal standard deviation estimates, improved the efficiency of standard deviation estimates with only a relatively small increase in average bias, thus reducing the RMSE. Although the optimal group size at which to constrain the group standard deviations to be equal is not a priori clear in any given scenario, the results above suggest that the analyst may be able to make an informed choice to achieve a roughly optimal model.

3. Application of the HETOP Model to Real Data

The simulations in Section 2 indicate that the HETOP model accurately recovers means, standard deviations, and ICCs from coarsened data across a range of scenarios when sample sizes are moderately large and the group distributions are normal. This section analyzes 18 sets of real test score data to investigate whether means and standard deviations can be recovered from real coarsened test score distributions. To carry out these analyses, we selected data sets for which we had access to both the coarsened proficiency data and the scale scores (the uncoarsened, continuous data) for each student: 10 data sets from a midsize state's testing program and eight data sets from the state NAEP administrations in 2009 and 2011. In effect, these analyses assess whether the actual test score distributions in these 18 cases satisfy the respective normality assumption of the HETOP model.

3.1 Data

The first eight data sets contained student-level records for the 2009 and 2011 Grades 4 and 8 Main NAEP mathematics and reading administrations, with each data set containing scores for a single year-by-grade-by-subject combination (e.g., 2009 Grade 4 math scores constitute one data set). The groups in these data sets were states, and the aim was to estimate the means and standard deviations of state test score distributions with the HETOP model. Hence, there were 50 groups in each of the NAEP data sets, with a median group (state) sample size of 3,050 across all eight data sets.[8]

The other 10 data sets consist of mathematics and reading test scores from a medium-sized state for a cross section of approximately 90,000 students in Grades 4 through 8 during the 2005–2006 school year. Each data set contained student-level scores from a single grade-by-subject combination, with scores grouped at the school level, so that the target estimates of interest were the school means and standard deviations of test scores for given grade levels. Across the 10 state data sets, the number of groups (schools) ranged from 428 to 1,244 and the median group (grade within school) sample size ranged from 70 to 194. Both the NAEP and State testing programs use three unique cut scores in each grade and subject level to classify students into one of the four ordered proficiency categories. Online Appendix Table D1 (available in the online version of the journal) provides detailed descriptive information about the 18 data sets.

3.2 Comparison of HETOP and Uncoarsened Estimates

If the test scores in these 18 data sets are respectively normal, and if the group sample sizes are large and the cut scores are well placed, the HETOP model should return precise, unbiased estimates of the group means and standard deviations in the continuous metric of y^*, as our simulation suggests. If, additionally, the function relating the reported scale scores to the metric in which the distributions are normal is linear, then group means and standard deviations based on the student-level scale scores should be perfectly correlated (within the limits of sampling variability), with the group means and standard deviations based on fitting the HETOP model to the coarsened proficiency data. This suggests we could examine the correlation between HETOP estimates and estimates based on observed scale scores to assess the extent to which the empirical test score distributions satisfy the respective normality assumption of the model.

An imperfect correlation, however, not only might result from a failure of respective normality but also might arise if (a) the function f is not linear or (b) the estimates are imprecise because sample sizes are not large or the cut scores are not sufficiently informative. The first condition will lead to a nonlinear association between the two sets of estimates. The second will produce a noisy association. To assess the respective normality assumption in real test score data, then, we must determine whether the less-than-perfect correlation between the HETOP estimates and the estimates based on the observed scale scores can be explained by the error that comes from coarsening and/or the nonlinearity of f. We describe our approach to doing this below. To the extent that these factors do not explain an observed correlation less than one, the test score distributions are not respectively normal.

First, we estimated group means and standard deviations based on the original student-level scale scores, using traditional estimators of means and standard deviations. We refer to these as the "original" scale score estimates. Second, to model the data that researchers may be limited to in practice, we coarsened the scale scores according to the operational NAEP and State cut scores. We then used the HETOP model to estimate means and standard deviations based on the coarsened frequency counts. We refer to these as the "H4" estimates because they are based on four proficiency categories. The correlation between these two sets of estimates will be degraded by imprecision due to the coarsening and by any nonlinearity in f.

Third, to generate HETOP estimates less affected by loss of information due to coarsening, we coarsened each data set a second time, using 19 equal-interval cut scores that classified students into 20 "proficiency" categories (instead of 4). We then estimated means and standard deviations for each group with a HETOP model using the 20 observed frequencies for each group. We refer to these as "H20" estimates, because they are based on 20 proficiency categories.[9]

Finally, we estimate a function f^* that, when applied to the observed student-level scale scores simultaneously, renders all of the within-group distributions as nearly normal as possible. We estimate f^* from the mapping between the 19 cut scores estimated from the H20 model (i.e., $\hat{c}_1^*, \hat{c}_2^*, \ldots, \hat{c}_{19}^*$) and their corresponding values on the reported score scale ($c_1, \ldots, c_{19}$). We estimate a monotonic function that goes through these 19 points (so that $\hat{f}^*(c) = \hat{c}^*$); this function will closely approximate a function that renders the within-group distributions as nearly normal as possible. We then apply this transformation to the observed student-level scale scores, resulting in transformed scores, $\hat{y}^*$, for each student. If test scores are respectively normal, this transformation should render the group

TABLE 2.
Average, Minimum, and Maximum Correlations Between HETOP Estimates and Uncoarsened Score Estimates

		Means				Standard Deviations			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Estimate 1:	H4	H20	Orig.	H4	H4	H20	Orig.	H4
	Estimate 2:	Orig.	Trans.	Trans.	Trans.	Orig.	Trans.	Trans.	Trans.
NAEP	Average	.995	1.000	.999	.996	.851	.995	.955	.910
	Minimum	.988	1.000	.998	.992	.738	.992	.929	.831
	Maximum	.998	1.000	1.000	.998	.923	.998	.978	.967
State	Average	.973	1.000	.999	.973	.779	.987	.932	.759
	Minimum	.941	.999	.998	.941	.667	.979	.835	.626
	Maximum	.991	1.000	1.000	.992	.866	.995	.990	.864

Note. H4 = heteroskedastic ordered probit model with 4 proficiency categories as defined by testing program; H20 = heteroskedastic ordered probit model with 20 categories defined by 19 equally spaced cut scores; Orig. = original score scale metric; Trans. = transformed score scale metric; HETOP = heteroskedastic ordered probit model.

score distributions normal; the group means and standard deviations of $\hat{y}^*$ will be linearly related, within sampling variability, to those estimated from the HETOP model applied to coarsened data. We refer to group means and standard deviations based on these normalized $\hat{y}^*$ scores as "transformed" estimates. The procedure used to estimate f^* is described in Online Appendix E (available in the online version of the journal).

We calculated Pearson correlations between these four sets of estimates for each of the 18 data sets. These correlations are summarized in Table 2, which presents the average, minimum, and maximum correlation among the estimates for the NAEP and State data sets separately (correlations for each of the 18 data sets are in Online Appendix Table D2 [available in the online version of the journal]). Column 1, for example, summarizes correlations between means estimated based on the H4 and uncoarsened original scale scores, while column 5 summarizes the corresponding correlations between the standard deviation estimates.

As mentioned above, these correlations may be less than 1.0 even if score distributions are respectively normal. If the test score data are respectively normal, however, then we expect the correlations in columns 2 and 6 to be near 1.0, because these correlations are based on estimates that adjust for a lack of normality of the reported scale score metric (transformed) and the error due to coarsening into only four categories. Indeed, the average correlations between these estimates are uniformly near 1 for all data sets (the lowest correlation across both columns 2 and 6 is 0.979), suggesting both that the data are respectively normal and that the H20 model accurately recovers the group means and standard deviations in the y^* metric.

To evaluate whether the original scale scores are reported in the metric in which they are normal in each group, we examine two sets of results. First, we inspect the correlations between the original and transformed estimate in columns 3 and 7. If the original test score data were reported in the normal metric, we would expect these correlations to be close to 1, because $\hat{f}^*$ would be linear. Second, we plot the function $\hat{f}^*$ in each case to examine its linearity directly. The correlations are near 1 for the means, but lower (as low as 0.83 in one case) for the standard deviations. The plots of $\hat{f}^*$ (in Online Appendix Figure E1 [available in the online version of the journal]) show very slight nonlinearity in most cases. Both of these patterns indicate that while the original test score scales are generally not one in which the distributions are as near to normal as possible, the original scales are not very different from such a scale. The modest departure from normality appears to cause more discrepancy in the estimated standard deviations (average correlations of .955 and .932 for NAEP and State) than in the estimated means (average correlations of .999 for both the NAEP and State scales).

Finally, it is useful to compare columns 2 and 4 and columns 6 and 8; this comparison indicates the extent to which coarsening into 4 rather than 20 categories reduces the precision of the estimated means and standard deviations. The correlations between the H20 and transformed estimates of means (column 2) are generally only modestly larger than those between the H4 and the transformed estimates (column 4). In the case of the estimated standard deviations, however, coarsening substantially degrades precision: The correlations in column 8 are much lower than in column 6. This is consistent with our simulation results, showing that the HETOP model more reliably estimates group means than standard deviations, particularly when group sizes are small and cut scores are not optimally located, as is the case in the State data sets.

These analyses suggest the assumption of respective normality of test score distributions is reasonable in the data sets we examined, which include both school-level and state-level groups. Moreover, the reported scale scores in these data sets appear to be in a metric that is very close to the latent metric in which the means and standard deviations are estimated by the HETOP model. This may not be true for all empirical test score distributions, of course; it would be useful to test in other cases where continuous scores are available.

4. Discussion

This article introduces a method for estimating the means and standard deviations of continuous test score distributions in multiple groups using only coarsened proficiency data. Through simulations and real data analyses, we demonstrate that accurate estimation of means and standard deviations of test score distributions for multiple groups (states, districts, schools, etc.) is possible under a wide range of scenarios, with modest loss of efficiency, particularly when sample sizes are larger than 50 and when the cut scores are not highly skewed. The analyses also showed that estimates of secondary statistics such as the ICC can be recovered accurately, with slight positive bias when group sizes are small. While estimates of group standard deviations were accurate across all conditions with larger sample sizes, there was evidence of small negative bias in some conditions with smaller sample sizes, particularly when the location of cut scores used to coarsen the data are unequally and/or unevenly spaced, thus providing relatively little information about the original distributions. The bias was very small when group sample sizes were 100, and modestly larger with small samples of size 25, though the average bias was never sizable compared to the true standard deviations or the sampling variance of the estimates. Our analyses of real test score data sets suggest the primary assumption of respective normality is reasonable for these particular test scores and likely those developed under similar conditions. Further simulation studies to evaluate the methodology across a wider range of conditions, including those where data are not respectively normal, would be a useful extension to this work.

The simulation results and real data analyses suggest a few common considerations for researchers to attend to when applying the HETOP model in practice. First, because the quality and reliability of HETOP estimates (particularly for group standard deviations) depend primarily on group sample sizes and cut score locations, an inspection of the overall proportion of students within each proficiency category and the proportion of groups with zero observations in one or more categories can be useful indicators of potential problems. Other indicators include models that will not converge, are slow to converge, or converge but produce abnormally large standard errors. In these cases, our simulations and other work with real test score data suggest a PHOP model is a good way to overcome some data limitations and is generally preferable to a HOMOP model unless the assumption of homoskedasticity is defensible.

In fitting the PHOP model, the analyst must determine a sample size threshold below which to impose the homoskedasticity constraint. This choice can be guided by knowledge of the cut score locations, the number and size of groups, and prior research that provides information about plausible values of the ICC and CV. When the CV of group variances is approximately 0.2 (roughly the average value observed in the data we analyzed), constraining the standard deviations of groups smaller than 100 is generally near optimal in our simulations. Of course, RMSE need not be the only criterion used to determine the best model. For analysts who are less willing to tolerate bias than error variance, a smaller constraint threshold would be preferable and vice versa. In addition, if group means are of primary interest, the choice of a

PHOP model will matter little; if standard deviations are of interest, the bias-precision trade-off is more salient. Further development of practical model fit statistics and diagnostics that can inform PHOP model selection is an important direction for future research.

One benefit of the PHOP model is that it improves estimation for small groups, particularly when cut score locations are suboptimal. The challenges for estimation posed by small sample sizes or extreme cut scores could also be addressed with alternative estimation strategies or frameworks, such as Bayesian or random-effects models. It is possible to estimate a mixed-effects HETOP model (see, e.g., Gu, Fiebig, Cripps, & Kohn, 2009; Hedeker, Demirtas, & Mermelstein, 2009) from which one could obtain shrunken estimates of group means and standard deviations. These Bayesian estimates would have smaller RMSE than our ML estimates but would also contain more bias. The decision of whether to prefer more-biased, lower RMSE shrunken estimates over less-biased, higher RMSE ML estimates depends on how one wants to use the resulting estimates. If the estimates will be used as outcome variables in subsequent models or as descriptive statistics, the (less biased) ML estimates may be preferable to the (more biased) shrunken estimates. If the estimates will be used as predictor variables in subsequent models, however, the shrunken estimates may be preferable (although in this case they should, in principle, be shrunken to their mean conditional on the other covariates to be used in the model). Shear, Castellano, and Lockwood (2016) present some preliminary comparisons of these two approaches in the context of coarsened test score data, but additional work exploring the potential benefits of Bayesian HETOP models would be very useful.

In our discussion here, we have ignored the potential effects of measurement error. If we think of the continuous scores in the y metric as containing measurement error, then the key assumption of the HETOP model is that the observed, error-prone test score distributions are respectively normal. Given this assumption, estimation proceeds as we describe it above, and the resulting estimates are understood as means and standard deviations of the error-prone scores in the y^* metric. To recover means and standard deviations of true scores in the y^* metric, one would need information about the reliability of the test scores in that metric. Although this is not identical to the reliability of scores in the metric y (the metric reported by test score developers) unless the function f is linear, Reardon and Ho (2015) show that using published reliabilities to adjust group means and standard deviations on a transformed scale generally produces only trivial bias, given that widely used standardized tests typically have high reliability. When reliability is high, distortions of measurement error due to the transformation function f are trivial unless f is extremely nonlinear. As a result, standard measurement error adjustments, based on published reliabilities of scores in the y metric, can be made to yield estimates of groups' true test score means and standard deviations in the y^* metric.

Finally, as mentioned above, these methods are applicable whenever data can be conceptualized as coarsened: the result of some form of polychotomization, censoring, binning, or rounding. In the case of aggregate proficiency data, such as that contained in the ED*Facts* database, such a model is clearly applicable, and our results show that the HETOP model can provide estimates of means and standard deviations that can overcome some of the limitations with such data as described by Ho (2008) and others. In the case of AP exams, where scores are only reported on a 1 to 5 ordinal scale, one might still presume the existence of a continuous underlying variable of which the observed scores are a coarsened version. In such cases, our methods provide a way to estimate the distributions of this underlying continuous variable in multiple groups. Ordinal data of many kinds—from Likert-type scale survey data to Apgar scores and from discrete levels of educational attainment to demographic age or income bins—can be thought of as representing coarsened versions of latent continuous variables. In many of these cases, the methods described here could be usefully applied to estimate moments of group distributions.

Appendix A

Estimating the Total Between- and Within-Group Variances

Given $\widehat{\mathbf{M}}'$ and $\widehat{\mathbf{\Gamma}}'$, we wish to estimate the within- and between-group variance of y. As noted in the text, we assume throughout this article that the population consists of a finite number of groups ($g = 1, \ldots, G$), all of which are observed. As above, $\mathbf{P}$ is the $1 \times G$ vector of group population proportions (the p_g's). We observe a sample of size n_g from each group, where n_g may or may not be proportional to p_g. Without loss of generality, we assume the model is fit subject to the constraints that $\mathbf{PM}' = 0$ and $\mathbf{P\Gamma}' = 0$. If it is not, we transform the estimates to obtain $\widehat{\mathbf{M}}'$ and $\widehat{\mathbf{\Sigma}}'$ in this metric, as described in Online Appendix A (available in the online version of the journal).

The between-group and within-group variances are defined as:

$$\begin{aligned} \sigma_B'^2 &= \mathbf{PM}'^2 \\ \sigma_W'^2 &= \mathbf{P\Sigma}'^2. \end{aligned} \tag{A1}$$

We can compute (biased) estimates of these using their sample analogs, $\mathbf{P}\widehat{\mathbf{M}}'^2$ and $\mathbf{P}\widehat{\mathbf{\Sigma}}'^2$. Below we derive the expected values of these estimators to assess their bias. We use the results of these derivations to obtain approximately unbiased estimators.

Estimating $\sigma_W'^2$

Let w_g be the error in $\hat{\gamma}_g : \hat{\gamma}_g = \gamma_g + \hat{w}_g$. Let $\boldsymbol{\Omega}'$ be the sampling variance–covariance matrix of the γ_g''s. The diagonal elements of this are the squared sampling variances (the ω_g^2's). Then:

$$\begin{aligned}
E[\mathbf{P}\widehat{\boldsymbol{\Sigma}}^{\circ 2}] &= E\left[\sum\nolimits_g p_g \hat{\sigma}_g^2\right] \\
&= \sum\nolimits_g E[p_g e^{2\hat{\gamma}_g}] \\
&= \sum\nolimits_g E[p_g e^{2(\gamma_g + \hat{w}_g)}] \\
&= \sum\nolimits_g E[p_g e^{2\gamma_g} e^{2\hat{w}_g}] \\
&= \sum\nolimits_g (p_g \sigma_g^2) E[e^{2\hat{w}_g}] \\
&\approx \sum\nolimits_g (p_g \sigma_g^2) E[1 + 2\hat{w}_g + 2\hat{w}_g^2] \\
&= \sum\nolimits_g (p_g \sigma_g^2)(1 + 2\omega_g^2) \\
&= \mathbf{P}\boldsymbol{\Sigma}^{\circ 2} \cdot (1 + 2\overline{\omega_g^2}) + 2G\text{Cov}(p_g\sigma_g^2, \omega_g^2),
\end{aligned} \tag{A2}$$

where $\overline{\omega_g^2} = \frac{1}{G}\mathbf{1} \cdot \text{vecdiag}(\boldsymbol{\Omega}')$ is the average sampling variance of the $\hat{\gamma}'_g$'s. Under the assumption that $\text{Cov}(p_g\sigma_g^2, \omega_g^2) \approx 0$, we have

$$E[\mathbf{P}\widehat{\boldsymbol{\Sigma}}^{\circ 2}] \approx \mathbf{P}\boldsymbol{\Sigma}^{\circ 2} \cdot (1 + 2\overline{\omega_g^2}). \tag{A3}$$

Therefore, we can compute an approximately unbiased estimate of $\hat{\sigma}_W'^2$ as

$$\hat{\sigma}_W'^2 = \frac{\mathbf{P}\widehat{\boldsymbol{\Sigma}}^{\circ 2}}{1 + 2\overline{\omega_g^2}}. \tag{A4}$$

Equation A4 requires an estimate of $\overline{\omega_g^2}$, the average sampling variance of the $\hat{\gamma}_g$'s, which can be obtained from the estimated sampling covariance matrix of the $\hat{\gamma}_g$'s

$$\widehat{\overline{\omega_g^2}} = \frac{1}{G}\mathbf{1} \cdot \text{vecdiag}(\widehat{\boldsymbol{\Omega}}'). \tag{A5}$$

However, $\widehat{\boldsymbol{\Omega}}'$ is prone to sampling variance (i.e., the estimated sampling variances of the γ_g's themselves have sampling variances). Our simulations show that when n is small, the sampling variance of the elements of $\widehat{\boldsymbol{\Omega}}'$ can be very large, because the sparse coarsened data provide little information from which to estimate the sampling variances. As a result $E\left[\frac{1}{G}\mathbf{1} \cdot \mathit{vecdiag}(\widehat{\boldsymbol{\Omega}}')\right] \gg \frac{1}{G}\mathbf{1} \cdot \mathit{vecdiag}(\boldsymbol{\Omega}')$ in such cases.

An alternate method of estimating $\overline{\omega_g^2}$ is to derive an approximate formula based on group sample sizes. To do so, let $\hat{u}_g$ be the error in $\hat{\sigma}_g^2 : \hat{\sigma}_g^2 = \sigma_g^2 + \hat{u}_g$. If a population variance σ^2 is estimated from a sample of size n (using data that have not been coarsened), the sampling variance of $\hat{\sigma}^2$ is approximately $\frac{2\sigma^4}{n-1}$ (Casella & Berger, 2002; Neter, Wasserman, & Kutner, 1990). Note that, for a normally distributed variable X with mean 0 and standard deviation s, $\mathrm{Var}(X + X^2) \approx s^2 + 2s^4$. We then have

$$\begin{aligned}
\hat{\sigma}_g^2 &= \sigma_g^2 + \hat{u}_g \\
e^{2\hat{\gamma}_g} &= e^{2\gamma_g} + \hat{u}_g \\
e^{2\gamma_g} e^{2\hat{w}_g} &= e^{2\gamma_g} + \hat{u}_g \\
e^{2\gamma_g}\left(e^{2\hat{w}_g} - 1\right) &= \hat{u}_g \\
e^{2\gamma_g}\left(1 + 2\hat{w}_g + 2\hat{w}_g^2 - 1\right) &\approx \hat{u}_g \\
2e^{2\gamma_g}\left(\hat{w}_g + \hat{w}_g^2\right) &= \hat{u}_g \\
\mathrm{Var}\Big(2e^{2\gamma_g}\left(\hat{w}_g + \hat{w}_g^2\right)\Big) &= \mathrm{Var}(\hat{u}_g) \\
4e^{4\gamma_g}\mathrm{var}\left(\hat{w}_g + \hat{w}_g^2\right) &= \mathrm{Var}(\hat{u}_g) \\
4\sigma_g^4[\omega_g^2 + 2\omega_g^4] &\approx \frac{2\sigma_g^4}{n_g - 1} \\
\omega_g^2 + 2\omega_g^4 &= \frac{1}{2(n_g - 1)}.
\end{aligned} \tag{A6}$$

Applying the quadratic formula to solve for ω_g^2 yields one positive root:

$$\begin{aligned}
\omega_g^2 &= -\frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{4}{n_g - 1}} \\
&\approx -\frac{1}{4} + \frac{1}{4}\left(1 + \frac{2}{n_g - 1}\right) \\
&= \frac{1}{2(n_g - 1)},
\end{aligned} \tag{A7}$$

where the approximation holds if n_g is even moderately large.

Given Equation A7, we have

$$\overline{\omega_g^2} \approx \frac{1}{G}\sum_g \frac{1}{2(n_g - 1)} = \frac{1}{2\tilde{n}} \tag{A8}$$

where $\tilde{n}$ is the harmonic mean of $n_g - 1 : \tilde{n} = \left(\frac{1}{G}\sum_g \frac{1}{n_g - 1}\right)^{-1}$.

Note that Equation A8 is based on a formula for the sampling variance of a population variance based on uncoarsened data. When the data are coarsened, the sampling variability of $\hat{\sigma}_g'^2$ will certainly be larger than that given by the formula used above $\left(\frac{2\sigma_g'^4}{n_g-1}\right)$, but the difference may not be large. For example, suppose the true sampling variance of $\sigma_g'^2$ were $\frac{2c_s\sigma_g'^4}{n_g-1}$, where $c_s \geq 1$, then using the approximation in Equation A8 in Equation A4 will inflate our estimate of $\hat{\sigma}_W'^2$ by a factor of $\frac{\tilde{n}+c_s}{\tilde{n}+1}$. Unless c_s is large in relation to $\tilde{n}$, the difference will be trivial.

The approximation in Equation A8 needs to be modified when using either the HOMOP or PHOP (rather than the HETOP) model. When we fit the HOMOP model, the sampling variance in the estimate of the $\hat{\gamma}_g$'s will be smaller, because the estimate is based on the pooled sample of all groups. In this case, $\overline{\omega_g^2}$ might be well estimated by Equation A5. Alternately, because the effective sample size for estimating $\overline{\omega_g^2}$ is N, and we lose a degree of freedom in estimating each group's mean, Equation A8 can be replaced by

$$\overline{\omega_g^2} \approx \frac{1}{2(N-G)}. \tag{A9}$$

In the PHOP model, the average sampling variance of the γ_g's can be approximated as

$$\overline{\omega_g^2} \approx \frac{1}{G}\sum_g \frac{1}{2(\breve{n}_g - 1)}, \tag{A10}$$

where $\breve{n}_g = n_g$ if group g's standard deviation is not constrained, and $\breve{n}_g = \sum_{g \in C}(n_g - 1)$ if group $g \in C$, where C is the set of constrained groups.

When estimating $\hat{\sigma}_W'^2$, we substitute Equations A8, A9, or A10 into Equation A4 depending upon which model was fit.

Estimating $\sigma_B'^2$

To compute the expected value of $\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2}$, first note that estimating the variance of the group means involves error in the overall mean and the individual group means. The estimate of each group's mean has two sources of error in it: $\hat{\mu}_g' = \mu_g' - \hat{u}' + \hat{e}_g'$, where $\hat{u}' = \sum p_g \hat{e}_g'$ and $\hat{e}_g' = \hat{\mu}_g' - \mu_g'$. Then

$$\begin{aligned} E[\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2}] &= E[\mathbf{P}(\mathbf{M}' - \widehat{\mathbf{u}}' + \widehat{\mathbf{e}}')^{\circ 2}] \\ &= E[\mathbf{P}\mathbf{M}'^{\circ 2} - 2\mathbf{P}(\widehat{\mathbf{u}}' \circ \widehat{\mathbf{e}}') + \mathbf{P}\widehat{\mathbf{u}}'^{\circ 2} + \mathbf{P}\widehat{\mathbf{e}}'^{\circ 2}] \\ &= \mathbf{P}\mathbf{M}'^{\circ 2} - 2\mathbf{P}E[\widehat{\mathbf{u}}' \circ \widehat{\mathbf{e}}'] + \mathbf{P}E[\widehat{\mathbf{u}}'^{\circ 2}] + \mathbf{P}E[\widehat{\mathbf{e}}'^{\circ 2}] \\ &= \mathbf{P}\mathbf{M}'^{\circ 2} - \mathbf{P}\mathbf{V}'\mathbf{P}^t + \mathbf{P} \cdot \mathrm{vecdiag}(\mathbf{V}'), \end{aligned} \tag{A11}$$

where vecdiag$(\mathbf{V}')$ is the $G \times 1$ matrix of sampling variances of the means (the diagonal of $\mathbf{V}'$). So we can compute an unbiased estimate of $\hat{\sigma}_B'^2$ as

$$\hat{\sigma}_B'^2 = \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{P}\mathbf{V}'\mathbf{P}^t - \mathbf{P} \cdot \text{vecdiag}(\mathbf{V}'). \tag{A12}$$

Equation A12 requires an estimate of $\mathbf{V}'$, the variance–covariance matrix of the vector of estimated group means, $\widehat{\mathbf{M}}'$. One estimate of this is the estimated matrix $\widehat{\mathbf{V}}'$. However, like $\widehat{\boldsymbol{\Omega}}'$ above, $\widehat{\mathbf{V}}'$ is prone to sampling variance (i.e., the estimated sampling variances of the $\hat{\mu}_g$'s themselves have sampling variances). Our simulations show that when n is small, the sampling variance of the elements of $\widehat{\mathbf{V}}'$ can be very large, because the sparse coarsened data provide little information from which to estimate the sampling variances. As a result $E[\mathbf{P} \cdot \text{vecdiag}(\widehat{\mathbf{V}}')] \gg \mathbf{P} \cdot \text{vecdiag}(\mathbf{V}')$ in such cases.

An alternate method of estimating $\widehat{\mathbf{V}}'$ is to derive an approximate formula for its diagonal elements based on group sample sizes. We begin by assuming that the off-diagonal elements of $\mathbf{V}'$ are approximately 0 (they will not be exactly 0, because the estimated means are dependent on one another, since all are estimated simultaneously and constrained to satisfy $\mathbf{P}\widehat{\mathbf{M}}' = 0$, but they will be close to zero when G and n are moderately large). We then assume the sampling variance of μ is given by the standard formula (based on uncoarsened data) for the sampling variance of a mean: $\text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$. Then the diagonal elements of $\mathbf{V}'$ will be $v_{gg} = \frac{\sigma_g'^2}{n_g}$. Substituting this matrix into Equation A12, we get

$$\begin{aligned}
\hat{\sigma}_B'^2 &= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{P}\mathbf{V}'\mathbf{P}^t - \mathbf{P} \cdot \text{vecdiag}(\mathbf{V}') \\
&= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + (\mathbf{P}^{\circ 2} - \mathbf{P}) \cdot \text{vecdiag}(\mathbf{V}') \\
&= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \sum_g (p_g^2 - p_g) \frac{\sigma_g'^2}{n_g} \\
&= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \left(\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\right)\boldsymbol{\Sigma}'^{\circ 2} \\
&= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\right)\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\omega_g^2}} \\
&= \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\right)\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}},
\end{aligned} \tag{A13}$$

where we substitute in the approximations of $\overline{\omega_g^2}$ from Equations A8, A9, or A10 as appropriate.

Again, if the sampling variance of the $\hat{\mu}_g'$ estimates is greater than they would be if the data were not coarsened, then it may be more appropriate to substitute

$v_{gg} = c_m \frac{\sigma_g'^2}{n_g}$ into Equation A12 above, where $c_m \geq 1$ is a constant. Then if we also use c_s as above in the formula estimating $\boldsymbol{\Sigma}'^{\circ 2}$, Equation A13 becomes

$$\hat{\sigma}_B'^2 = \mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{c_m}{1 + c_s 2\widehat{\overline{\omega_g^2}}}\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\right)\widehat{\boldsymbol{\Sigma}}'^{\circ 2}. \tag{A14}$$

Given that $\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\right)\widehat{\boldsymbol{\Sigma}}'^{\circ 2}$ will be small when the elements of $\mathbf{n}$ are modestly large, however, setting $c_m = 1$ has very little effect of the estimate of $\hat{\sigma}_B'^2$.

Estimating the Population Standard Deviation, σ'

Given estimates of $\sigma_W'^2$ and $\sigma_B'^2$ from Equations A4 and A13, we compute

$$\begin{aligned}
\hat{\sigma}' &= (\sigma_W'^2 + \sigma_B'^2)^{\frac{1}{2}} \\
&= \left(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{\mathbf{n}^{\circ -1} \circ (\mathbf{P}^{\circ 2} - \mathbf{P})\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}} + \frac{\mathbf{P}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}}\right)^{\frac{1}{2}} \\
&= \left(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \frac{\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P} + \mathbf{n} - \mathbf{1}) \circ \mathbf{P}\right)\widehat{\boldsymbol{\Sigma}}'^{\circ 2}}{1 + 2\widehat{\overline{\omega_g^2}}}\right)^{\frac{1}{2}} \\
&= \left(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{Q}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}\right)^{\frac{1}{2}},
\end{aligned} \tag{A15}$$

where

$$\mathbf{Q} = \frac{\left(\mathbf{n}^{\circ -1} \circ (\mathbf{P} + \mathbf{n} - \mathbf{1}) \circ \mathbf{P}\right)}{1 + 2\widehat{\overline{\omega_g^2}},} \tag{A16}$$

and we again substitute one of the approximations from Equations A8, A9, or A10 for $\overline{\omega_g^2}$ depending upon whether a HETOP, HOMOP, or PHOP model is fit.

Appendix B

Computation of Standard Errors of $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$

Once we have constructed $\widehat{\mathbf{M}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ via Equation 7, we must estimate the covariance matrices $\mathbf{V}^* = \text{Cov}(\widehat{\mathbf{M}}^*, \widehat{\mathbf{M}}^*)$, $\mathbf{Z}^* = \text{Cov}(\widehat{\mathbf{M}}^*, \widehat{\boldsymbol{\Sigma}}^*)$, and $\mathbf{W}^* = \text{Cov}(\widehat{\boldsymbol{\Sigma}}^*, \widehat{\boldsymbol{\Sigma}}^*)$, from which we can obtain standard errors for the parameters of interest in the model.

Assuming that $E[\widehat{\mathbf{M}}'] = \mathbf{M}'$ and $E[\hat{\sigma}'] = \sigma'$,[10] the g, h element of $\mathbf{V}^*$ is

$$\begin{aligned} v^*_{gh} &= \mathrm{Cov}(\hat{\mu}^*_g, \hat{\mu}^*_h) \\ &= \mathrm{Cov}\left(\frac{\hat{\mu}'_g}{\hat{\sigma}'}, \frac{\hat{\mu}'_h}{\hat{\sigma}'}\right) \\ &\approx \frac{1}{\sigma'^2} v'_{gh} - \frac{\mu'_g}{\sigma'^3}\mathrm{Cov}(\hat{\sigma}', \hat{\mu}'_h) - \frac{\mu'_h}{\sigma'^3}\mathrm{Cov}(\hat{\mu}'_g, \hat{\sigma}') + \hat{\mu}'_g\hat{\mu}'_h \mathrm{Var}\left(\frac{1}{\sigma'}\right) \\ &\approx \frac{1}{\sigma'^2}[v'_{gh} - \mu^*_g \mathrm{Cov}(\hat{\sigma}', \hat{\mu}'_h) - \mu^*_h \mathrm{Cov}(\hat{\mu}'_g, \hat{\sigma}') + \mu^*_g\mu^*_h \mathrm{Var}(\hat{\sigma}')]. \end{aligned} \tag{B1}$$

Now let $\mathbf{I}_h$ denote the hth column of the $G \times G$ identity matrix. Then define[11]

$$\begin{aligned} r_h &= \mathrm{Cov}(\hat{\sigma}', \hat{\mu}'_h) \\ &= \frac{1}{2\sigma'}\mathrm{Cov}(\hat{\sigma}'^2, \hat{\mu}'_h) \\ &= \frac{1}{2\sigma'}\mathrm{Cov}(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{Q}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}, \hat{\mu}'_h) \\ &= \frac{1}{\sigma'}\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{V}'\mathbf{I}_h + \frac{1}{\sigma'}\mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{Z}'^{t}\mathbf{I}_h. \end{aligned} \tag{B2}$$

Then define the $1 \times G$ vector $\mathbf{R}$, with elements r_h, as

$$\begin{aligned} \mathbf{R} &= \frac{1}{\sigma'}\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{V}' + \frac{1}{\sigma'}\mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{Z}'^{t} \\ &= \mathbf{P}[\mathrm{diag}(\mathbf{M}^*)]\mathbf{V}' + \mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}^*)]\mathbf{Z}'^{t}. \end{aligned} \tag{B3}$$

Now we have

$$\mathbf{V}^* \approx \frac{1}{\sigma'^2}[\mathbf{V}' - (\mathbf{M}^*\mathbf{R} + \mathbf{R}^t\mathbf{M}^{*t}) + \mathbf{M}^*\mathbf{M}^{*t}\mathrm{Var}(\hat{\sigma}')]. \tag{B4}$$

Similarly, assuming that $E[\widehat{\boldsymbol{\Sigma}}'] = \boldsymbol{\Sigma}'$ and $E[\hat{\sigma}'] = \sigma'$, the g, h element of the covariance matrix $\mathbf{W}^*$ of the $\hat{\sigma}^*_g$'s is

$$\begin{aligned} w^*_{gh} &= \mathrm{Cov}(\hat{\sigma}^*_g, \hat{\sigma}^*_h) \\ &= \mathrm{Cov}\left(\frac{\hat{\sigma}'_g}{\hat{\sigma}'}, \frac{\hat{\sigma}'_h}{\hat{\sigma}'}\right) \\ &\approx \frac{1}{\sigma'^2} w'_{gh} - \frac{\sigma'_g}{\sigma'^3}\mathrm{Cov}(\hat{\sigma}', \hat{\sigma}'_h) - \frac{\sigma'_h}{\sigma'^3}\mathrm{Cov}(\hat{\sigma}'_g, \hat{\sigma}') + \hat{\sigma}'_g\hat{\sigma}'_h \mathrm{Var}\left(\frac{1}{\sigma'}\right) \\ &\approx \frac{1}{\sigma'^2}[w'_{gh} - \sigma^*_g \mathrm{Cov}(\hat{\sigma}', \hat{\sigma}'_h) - \sigma^*_h \mathrm{Cov}(\hat{\sigma}'_g, \hat{\sigma}') + \sigma^*_g\sigma^*_h \mathrm{Var}(\hat{\sigma}')]. \end{aligned} \tag{B5}$$

Define

$$\begin{aligned} t_h &= \mathrm{Cov}(\hat{\sigma}', \hat{\sigma}'_h) \\ &= \frac{1}{2\sigma'}\mathrm{Cov}(\hat{\sigma}'^2, \hat{\sigma}'_h) \\ &= \frac{1}{2\sigma'}\mathrm{Cov}(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{Q}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}, \hat{\sigma}'_h) \\ &= \frac{1}{\sigma'}\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{Z}'\mathbf{I}_h + \frac{1}{\sigma'}\mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{W}'\mathbf{I}_h. \end{aligned} \tag{B6}$$

Then define the 1 × G vector $\mathbf{T}$, with elements t_h, as

$$\begin{aligned} \mathbf{T} &= \frac{1}{\sigma'}\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{Z}' + \frac{1}{\sigma'}\mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{W}' \\ &= \mathbf{P}[\mathrm{diag}(\mathbf{M}^*)]\mathbf{Z}' + \mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}^*)]\mathbf{W}'. \end{aligned} \tag{B7}$$

We then have

$$\mathbf{W}^* \approx \tfrac{1}{\sigma'^2}[\mathbf{W}' - (\boldsymbol{\Sigma}^*\mathbf{T} + \mathbf{T}^t\boldsymbol{\Sigma}^{*t}) + \boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{*t}\mathrm{Var}(\hat{\sigma}')]. \tag{B8}$$

Finally, the element z^*_{gh} of the matrix $\mathbf{Z}^*$ is

$$\begin{aligned} z^*_{gh} &= \mathrm{Cov}(\hat{\mu}^*_g, \hat{\sigma}^*_h) \\ &= \mathrm{Cov}\left(\frac{\hat{\mu}'_g}{\hat{\sigma}'}, \frac{\hat{\sigma}'_h}{\hat{\sigma}'}\right) \\ &\approx \frac{1}{\sigma'^2}z'_{gh} - \frac{\mu'_g}{\sigma'^3}\mathrm{Cov}(\hat{\sigma}', \hat{\sigma}'_h) - \frac{\sigma'_h}{\sigma'^3}\mathrm{Cov}(\hat{\mu}'_g, \hat{\sigma}') + \hat{\mu}'_g\hat{\sigma}'_h\mathrm{Var}\left(\frac{1}{\sigma'}\right) \\ &\approx \frac{1}{\sigma'^2}[z'_{gh} - \mu^*_g\mathrm{Cov}(\hat{\sigma}', \hat{\sigma}'_h) - \sigma^*_h\mathrm{Cov}(\hat{\mu}'_g, \hat{\sigma}') + \mu^*_g\sigma^*_h\mathrm{Var}(\hat{\sigma}')] \\ &= \frac{1}{\sigma'^2}[z'_{gh} - \mu^*_g t_h - \sigma^*_h r_g + \mu^*_g\sigma^*_h\mathrm{Var}(\hat{\sigma}')]. \end{aligned} \tag{B9}$$

We then have

$$\mathbf{Z}^* \approx \tfrac{1}{\sigma'^2}[\mathbf{Z}' - (\mathbf{M}^*\mathbf{T} + \mathbf{R}^t\boldsymbol{\Sigma}^{*t}) + \mathbf{M}^*\boldsymbol{\Sigma}^{*t}\mathrm{Var}(\hat{\sigma}')]. \tag{B10}$$

Equations B4, B8, and B10 require $\mathrm{Var}(\hat{\sigma}')$. Note first, that we can derive[12] the sampling variance of $\hat{\sigma}'^2$ as

$$\begin{aligned} \mathrm{Var}(\hat{\sigma}'^2) &= \mathrm{Var}(\mathbf{P}\widehat{\mathbf{M}}'^{\circ 2} + \mathbf{Q}\widehat{\boldsymbol{\Sigma}}'^{\circ 2}) \\ &= 4\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{V}'[\mathrm{diag}(\mathbf{M}')]\mathbf{P}^t + 4\mathbf{Q}[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{W}'[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{Q}^t \\ &\quad + 8\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{Z}'[\mathrm{diag}(\boldsymbol{\Sigma}')]\mathbf{Q}^t. \end{aligned} \tag{B11}$$

Then, by the Delta method,

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}') &\approx \frac{1}{4\sigma'^2}\mathrm{Var}(\hat{\sigma}'^2)\\ &\approx \frac{1}{\sigma'^2}[\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{V}'[\mathrm{diag}(\mathbf{M}')]\mathbf{P}^t + \mathbf{Q}[\mathrm{diag}(\mathbf{\Sigma}')]\mathbf{W}'[\mathrm{diag}(\mathbf{\Sigma}')]\mathbf{Q}^t \\ &+ 2\mathbf{P}[\mathrm{diag}(\mathbf{M}')]\mathbf{Z}'[\mathrm{diag}(\mathbf{\Sigma}')]\mathbf{Q}^t].\end{aligned} \tag{B12}$$

We substitute Equation B12 into Equations B4, B8, and B10 to obtain expressions for $\mathbf{V}^*$, $\mathbf{W}^*$, and $\mathbf{Z}^*$. To estimate $\mathbf{V}^*$, $\mathbf{W}^*$, and $\mathbf{Z}^*$, we replace the relevant terms in the resulting expressions by their estimated values.

Authors' Note

This article benefited from collaboration with Erin Fahle and thoughtful comments from J. R. Lockwood. Some of the data used in this article were provided by the National Center for Education Statistics (NCES). We thank Demetra Kalogrides for excellent research assistance, and Richard Williams for his responsiveness and assistance regarding our inquires about his -oglm- Stata package.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grants from the Institute of Education Sciences (#R305D110018 and #R305B090016) and a grant from the Spencer Foundation (#201500058). The opinions expressed here are ours and do not represent views of NCES, the Institute of Education Sciences, the Spencer Foundation, or the U.S. Department of Education.

Notes

1. The *V* statistic is a transformation-invariant metric quantifying the nonoverlap between two distributions and is equal to a Cohen's *d* standardized mean difference when both distributions are normal (Ho, 2009).
2. Note that we do not require that $p_g = n_g/N$; that is, the size of the sample in each group need not be proportional to the group's share of the population.
3. These specific constraints are not essential; other constraints will identify the parameters and may be preferable in some settings. The default in many software programs is to define some group *r* as the "reference group" and to constrain $\mu_r' = 0$ and $\gamma_r' = 0$. These constraints imply that the reference group has a mean of 0 and a standard deviation of 1, with the means and standard deviations of the other groups then interpreted relative to group *r*.

This is a reasonable default where there is a substantively important reference group and standardization is not needed. It is not the obvious default when there is no substantively important reference group and we would like to estimate each group's mean and standard deviation relative to the overall population distribution.

4. If we are using the default constraint of $\mathbf{P}\widehat{\boldsymbol{\Gamma}}' = 0$, then this together with the additional homoskedasticity constraint implies the single combined constraint $\hat{\gamma}'_1 = \hat{\gamma}'_2 = \; . \; . \; . \; = \hat{\gamma}'_G = 0$.
5. The modified program is a Stata ado-file called `-hetop-`; it can be downloaded from within Stata by typing "`ssc install hetop`" from the command line.
6. Even in cases where all groups have sufficient data to identify the model parameters, small sample sizes may slow or impede convergence of the maximum likelihood algorithm, because the likelihood function may be very flat over a wide range of the parameter space. In such cases, we have found that replacing the constraints $\mathbf{P}\widehat{\mathbf{M}}' = 0$ and $\mathbf{P}\widehat{\boldsymbol{\Gamma}}' = 0$ with a reference group constraint (i.e., constrain $\hat{\mu}'_r = 0$ and $\hat{\gamma}'_r = 0$ where r indicates a reference group) sometimes improved the speed of convergence. In such cases, convergence is improved when the reference group is one with a large sample size and a distribution of frequency counts that is similar to the population distribution. The speed of convergence can also be improved by providing the algorithm with feasible starting values, which can be obtained by using the two-group methods described in Ho and Reardon (2012) to separately estimate each group's mean and standard deviation relative to that of the selected reference group.
7. We used the median rather than the mean estimated standard error to reduce the impact of extreme standard error estimates, primarily in conditions with small sample sizes and wide or skewed cut scores.
8. The National Assessment of Educational Progress (NAEP) is administered to a sample of students in the nation, and special scoring and scaling techniques result in "plausible values" (e.g., Mislevy, Johnson, & Muraki, 1992) instead of individual scores. For each of the eight year-grade-subject combinations in the data set, we had five plausible values for each student. To generate a data set with a single score for each student that could be used to compare with HETOP estimates, we created a synthetic data set using the first set of plausible values for all students. In order to avoid complications from comparisons using multiple plausible values and sampling weights, we generated an artificial sample for each state using the following procedure. We drew a random sample with replacement from the set of nonmissing first plausible values, with probability of selection proportional to original sampling weights. This created a random sample from a population defined by the (weighted) observed first plausible values in each state. We sampled N_g values for each state, where N_g was the original number of unique students with nonmissing data in state g. Using NAEP's actual proficiency cut scores

for each subject/grade combination, we calculated the number of students in the synthetic sample scoring in each proficiency category.

9. We could have used more than 20 categories, but given the finite number of possible scale scores and size of the groups, additional categories add vanishingly little additional information.
10. Even under the assumption that the HETOP estimator provides unbiased estimates of $\mathbf{M}'$ and $\mathbf{\Sigma}'$, the assumption that $E[\hat{\sigma}'] = \sigma'$ is not strictly valid, given nonlinearities in Equations 9 and 10, but is a good approximation in practice.
11. Note that $\mathbf{Q}$ in these formulas depends upon whether a HETOP, HOMOP, or PHOP model is being used, as defined in Equation A16.
12. Note that if $\mathbf{A}$ and $\mathbf{B}$ are $1 \times G$ scalar vectors, $\mathbf{D}$ and $\mathbf{E}$ are $G \times G$ scalar matrices, and $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ are $G \times 1$ column vectors of random variables, then $\mathrm{Cov}(\mathbf{A}[\mathbf{D}\widehat{\mathbf{X}}]^{\circ 2}, \mathbf{B}[\mathbf{E}\widehat{\mathbf{Y}}]^{\circ 2}) \approx 4\mathbf{A}[\mathrm{diag}(\mathbf{X})][\mathbf{D}^t\mathbf{D}]\mathbf{C}[\mathbf{E}\mathbf{E}^t][\mathrm{diag}(\mathbf{Y})]\mathbf{B}^t$, where $\mathbf{C}$ is the $G \times G$ covariance matrix of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.

Alvarez, R. M., & Brehm, J. (1995). American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science*, *39*, 1055–1082.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*, 204–226.

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

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

Cox, C. (1995). Location—Scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine*, *14*, 1191–1203. Retrieved from http://doi.org/10.1002/sim.4780141105

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

Freeman, E., Keele, L., Park, D., Salzman, J., & Weickert, B. (2015, August 14). *The plateau problem in the heteroskedastic probit model*. Retrieved from http://arxiv.org/abs/1508.03262v1

Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. New York, NY: Cambridge University Press.

Gu, Y., Fiebig, D. G., Cripps, E., & Kohn, R. (2009). Bayesian estimation of a random effects heteroscedastic probit model. *Econometrics Journal*, *12*, 324–339. Retrieved from http://doi.org/10.1111/j.1368-423X.2009.00283.x

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*, 546–582. Retrieved from http://doi.org/10.1177/0193841X14554212

Hedeker, D., Demirtas, H., & Mermelstein, R. J. (2009). A mixed ordinal location scale model for analysis of ecological momentary assessment (EMA) data. *Statistics and Its Interface*, *2*, 391.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. Retrieved from http://doi.org/10.3102/0162373707299706

Ho, A. D. (2008). The problem with "Proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*, 351–360. Retrieved from http://doi.org/10.3102/0013189X08323842

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*, 201–228. Retrieved from http://doi.org/10.3102/107699860933275

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*, 489–517. Retrieved from http://doi.org/10.3102/1076998611411918

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*, 3–17. Retrieved from http://doi.org/10.3102/10769986027001003

Horowitz, J. L., Sparmann, J. M., & Daganzo, C. F. (1982). An investigation of the accuracy of the Clark approximation for the multinomial probit model. *Transportation Science*, *16*, 382–401. Retrieved from http://doi.org/10.1287/trsc.16.3.382

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2013). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis*. Retrieved from http://doi.org/10.3102/0162373713485814

Jennings, J. (2011). Open letter to the member states of PARCC and SBAC. *Center on Education Policy*. Retrieved from http://www.cep-dc.org/displayDocument.cfm?DocumentID=359

Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, *10*, 193–200. Retrieved from http://doi.org/10.1080/07350015.1992.10509898

Keele, L., & Park, D. K. (2006, March). *Difficult choices: An evaluation of heterogeneous choice models* (Working Paper). Retrieved from http://www3.nd.edu/̃rwilliam/oglm/ljk-021706.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, *42*, 109–142.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*, 131–154. Retrieved from http://doi.org/10.3102/10769986017002131

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.

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*, 158–189. Retrieved from http://doi.org/10.3102/1076998615570944

Shear, B. R., Castellano, K. E., & Lockwood, J. R. (2016, April). *Using the Fay-Herriot model to improve inferences from coarsened proficiency data*. Presented at the National Council on Measurement in Education 2016 Annual Meeting, Washington, DC.

StataCorp. (2013). *Stata statistical software: Release 13*. College Station, TX: StataCorp LP.

Tosteson, A. N. A., & Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making*, *8*, 204–215. Retrieved from http://doi.org/10.1177/0272989X8800800309

U.S. Department of Education. (2015). *State assessments in reading/language arts and mathematics: School year 2012-13 EDFacts Data Documentation*. Washington, DC. Retrieved from http://www.ed.gov/edfacts

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, *37*, 531–559. Retrieved from http://doi.org/10.1177/0049124109335735

Williams, R. (2010). Fitting heterogeneous choice models with oglm. *The Stata Journal*, *10*, 540–567.

Authors

SEAN F. REARDON is the Professor of Poverty and Inequality in Education and Professor (by courtesy) of Sociology at Stanford University, 520 Galvez Mall, #526, Stanford, CA, 94305; email: sean.reardon@stanford.edu. His research focuses on the causes, measurement, and consequences of educational and social inequality.

BENJAMIN R. SHEAR is a doctoral candidate at Stanford University, 520 Galvez Mall, Stanford, CA, 94305; email: bshear@stanford.edu. His research focuses on statistical issues in educational assessment and psychometrics, particularly those relevant for validity and validation.

KATHERINE E. CASTELLANO is a psychometrician at Educational Testing Service, 90 New Montgomery St., Suite 1500, San Francisco, CA 94105; email: kecastellano@ets.org. Her research interests include addressing educational policy issues with rigorous statistical and psychometric modeling, such as evaluating the use of student growth models in accountability systems.

ANDREW D. HO is Professor of Education at the Harvard Graduate School of Education, 455 Gutman Library, 6 Appian Way, Cambridge, MA 02138; email: andrew_ho@gse.harvard.edu. His research in educational measurement focuses on accountability metrics for proficiency, growth, college readiness, course completion, and value added.

Manuscript received January 22, 2016
Revision received June 10, 2016
Accepted July 31, 2016