Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 19-05

Using Pooled Heteroskedastic Ordered Probit Models to Improve Small-Sample Estimates of Latent Test Score Distributions

AUTHORS

Benjamin R. Shear
University of Colorado Boulder

sean f. reardon
Stanford University

ABSTRACT

This paper describes an extension to the use of heteroskedastic ordered probit (HETOP) models to estimate latent distributional parameters from grouped, ordered-categorical data by pooling across multiple waves of data. We illustrate the method with aggregate proficiency data reporting the number of students in schools or districts scoring in each of a small number of ordered "proficiency" levels. HETOP models can be used to estimate means and standard deviations of the underlying (latent) test score distributions, but may yield biased or very imprecise estimates when group sample sizes are small. A simulation study demonstrates that the pooled HETOP models described here can reduce the bias and sampling error of standard deviation estimates when group sample sizes are small. Analyses of real test score data demonstrate use of the models and suggest the pooled models are likely to improve estimates in applied contexts.

VERSION

April 2020

Suggested citation: Shear, B.R., & Reardon, S.F. (2019). Using Pooled Heteroskedastic Ordered Probit Models to Improve Small-Sample Estimates of Latent Test Score Distributions (CEPA Working Paper No.19-05). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp19-05

Using Pooled Heteroskedastic Ordered Probit Models

to Improve Small-Sample Estimates of Latent Test Score Distributions

Benjamin R. Shear
University of Colorado Boulder

sean f. reardon
Stanford University

April, 2020

Note: this paper is based on work completed as part of the first author's doctoral dissertation. The research described here was supported by grants from the Institute of Education Sciences (R305D110018), the Spencer Foundation, the Russell Sage Foundation, the Bill and Melinda Gates Foundation, the Overdeck Family Foundation, and the William T. Grant Foundation. An earlier version of this paper was presented at the 2017 NCME Annual Meeting. Please direct comments and correspondence to Benjamin R. Shear, benjamin.shear@colorado.edu, School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309.

Using Pooled Heteroskedastic Ordered Probit Models

to Improve Small-Sample Estimates of Latent Test Score Distributions

Abstract

This paper describes an extension to the use of heteroskedastic ordered probit (HETOP) models to estimate latent distributional parameters from grouped, ordered-categorical data by pooling across multiple waves of data. We illustrate the method with aggregate proficiency data reporting the number of students in schools or districts scoring in each of a small number of ordered “proficiency” levels. HETOP models can be used to estimate means and standard deviations of the underlying (latent) test score distributions, but may yield biased or very imprecise estimates when group sample sizes are small. A simulation study demonstrates that the pooled HETOP models described here can reduce the bias and sampling error of standard deviation estimates when group sample sizes are small. Analyses of real test score data demonstrate use of the models and suggest the pooled models are likely to improve estimates in applied contexts.

Keywords: coarsened data; categorical data; heteroskedastic ordered probit; proficiency data

States administer millions of standardized assessments to public school students annually as part of their school accountability systems. The results of these assessments are often made publicly available only in highly coarsened form, and so are much less useful than they might be. Many states, for example, report the number students in a particular school or district scoring in each of a small number of ordered performance categories, such as “basic,” “proficient,” or “advanced,” rather than reporting the overall mean and standard deviation of students’ scores. These are referred to as “coarsened” test score data because they arise from coarsening continuous test scores according to a set of pre-determined cut scores. Such data have many widely recognized shortcomings (Ho, 2008; Ho & Reardon, 2012; Holland, 2002; Jacob et al., 2013), but continue to be a primary, and sometimes the only, publicly available source of state or district achievement test data. Having access to estimates of the mean and standard deviation of test scores can support a wider range of interpretations and analyses, ultimately leading to more accurate and useful interpretations about student achievement.

Reardon et al. (2017) described how heteroskedastic ordered probit (HETOP) models can be used to estimate the underlying means and standard deviations of the test score distributions based on coarsened test score data via maximum likelihood (ML), thus overcoming some limitations of the coarsening. In addition, because HETOP models use only ordinal information in the data, they do not rely on common interval scale assumptions. This fact provides some interpretational benefits and allows the models to be connected to other widely used ordinal statistics, as we describe in more detail below. Use of the HETOP model in this context does require that the coarsened scores in each group be based on a common test (or other measure) across groups that is coarsened using a common set of cut scores. At the same time, HETOP models can readily be applied to other contexts in which grouped, ordered-categorical scores are available and there is a need to summarize or compare the underlying distributions across groups. Examples include analyzing the aggregate responses to a Likert-style survey item across groups or across time, comparing aggregated Apgar scores (Apgar, 1953) across hospitals or regions, or analyzing continuous variables such as income that are reported in ordered categories in aggregate data sources such as the census.

The HETOP model described by Reardon et al. (2017) has some important limitations, however. When group sample sizes are small, the standard deviation estimates produced by the HETOP model are negatively biased and have large sampling variances (Reardon et al., 2017). Sparse data is the primary cause of this problem; when some groups have no observations in one or more categories, the coarse data provide limited information about the underlying distribution. In some cases finite ML estimates may not exist (Agresti, 2013). These sparse data problems can occur frequently, particularly in the context of analyzing coarsened test score data, where group sample sizes are often small and the cut scores used to coarsen the original test scores may be asymmetrically located throughout the distribution.

Researchers have proposed several methods to improve small-sample HETOP estimates. To illustrate how these approaches work, consider a case in which a HETOP model is used to estimate, from coarsened proficiency data, the distribution of mathematics achievement of third graders in each school across an entire state. As described in prior work, the HETOP model requires that all students complete the same test and that scores were coarsened using a common set of cut scores across all schools. To overcome small-sample problems, Reardon et al. (2017) proposed using models that constrain standard deviations to be equal across some or all schools in the sample. These constrained models attempt to improve standard deviation estimates for schools with small sample sizes by borrowing information from other small schools and estimating a single, common third grade mathematics standard deviation parameter for these small schools. In their most extreme form, the constrained models estimate only a single standard deviation parameter for all schools, regardless of size. Lockwood et al. (2018) describe Bayesian HETOP models that use a form of shrinkage estimators to improve small-sample estimates by borrowing information from other schools that are similar on observed covariates. Both of these approaches rely on borrowing information across groups (schools, in this case) to improve small-sample estimates, which can preclude the study of heterogeneity of within-group variances and rely on the potentially unrealistic assumption that the within-group variances are equal.

In this paper we propose a generalized version of the HETOP model, which we refer to as a pooled HETOP model, that can be used to estimate multiple latent distributions for each group simultaneously when coarsened data are available from multiple measures or time points. Returning to the case of achievement testing, analysts will often have access to additional sets of coarsened data for each school based on tests administered in other grades, years, or subjects. The pooled HETOP model allows these distributions to be estimated simultaneously, even when the tests and cut scores vary across grades, years, or subjects. Estimating these distributions simultaneously allows the model to use information from the same school in other grades, years, or subjects to improve estimates rather than borrowing information from different schools within the same grade, year, or subject. The intuition behind our approach is that when possible, it is preferable to pool information from the same group observed on different occasions rather than to pool information across different groups observed on the same occasion. This is partly an empirical question, and we analyze test score data from a national database to evaluate the tradeoff between pooling across versus within groups in the context of aggregate coarsened test score data.

The remainder of the paper is organized as follows. Section 1 provides an explanation of the HETOP model in the context of analyzing coarsened test score data, and describes an extension of the model to define what we refer to as the pooled HETOP model, which can be used to estimate distributions across multiple tests simultaneously. Section 2 analyzes test scores in a national database to evaluate the plausibility of assumptions made in the pooled HETOP model and to provide empirical evidence that placing constraints within rather than between groups is preferable. Section 3 uses a Monte Carlo simulation to evaluate how well the pooled HETOP model can recover parameters using small sample sizes under known conditions, and compares performance to the standard HETOP model and a constrained homoskedastic ordered probit model. Section 4 uses school-level coarsened proficiency data from a statewide mathematics assessment to illustrate the use of a pooled HETOP model in practice. Section 5 concludes with a brief discussion.

1. Statistical Models

To formalize discussion of the HETOP model, let there be a set of G groups (e.g., schools or districts). Students within each group take the same test and their scores are coarsened into one of K ordered proficiency categories using a common set of cut scores across all groups. We assume that there is an underlying, normally distributed latent variable y^* within each group that was coarsened into the set of K ordered categories based on a set of $K-1$ ordered cut scores denoted $c_1, \ldots, c_{K-1}$, where $c_{k-1} < c_k$ for all k. We define $c_0 = -\infty$ and $c_K = +\infty$. More formally, we assume that

$$y_{gi}^* \sim N(\mu_g, \sigma_g), \tag{1}$$

where y_{gi}^* represents an unobserved continuous score for student i in group g, and μ_g and σ_g are the mean and standard deviation, respectively, in group g. Let $\mathbf{N}$ be a G by K matrix, with elements n_{gk} equal to the number of students in group g scoring in category k.

We do not observe the values of y_{gi}^*, but rather observe the ordered categorical variable x_{gi}, $x \in \{1, \ldots, K\}$, for each student i in group g, where

$$x_{gi} = k \text{ if } c_{k-1} < y_{gi}^* \leq c_k. \tag{2}$$

The model-implied proportion of students in group g scoring in category k is

$$\pi_{gk} = \Phi\left(\frac{\mu_g - c_{k-1}}{\sigma_g}\right) - \Phi\left(\frac{\mu_g - c_k}{\sigma_g}\right) = \Pr(c_{k-1} < y_{gi}^* \leq c_k), \tag{3}$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function. This model is also sometimes referred to as a heterogeneous choice model (e.g., Alvarez & Brehm, 1995; Keele & Park, 2006; Williams, 2009), a rational model (McCullagh & Nelder, 1989), or a location-scale model (e.g., Cox, 1995; McCullagh, 1980). The use of HETOP models to estimate and interpret the means and standard deviations of y^* in each group is a generalization of the ML-based estimator of V, an ordinal method for estimating standardized achievement gaps between two groups described by Ho and Reardon (2012). The model can also be applied to the context of receiver operating characteristic curves (Dorfman & Alf, 1969; Tosteson & Begg, 1988).

Following the notation of Reardon et al. (2017), let n_{gk} be the number of students in group g scoring in category k and let $\mathbf{N}$ be the $G \times K$ matrix of observed n_{gk} values. The goal is to estimate the vectors $\mathbf{M} = [\mu_1, \ldots, \mu_G]^t$, $\mathbf{\Sigma} = [\sigma_1, \ldots, \sigma_G]^t$, and $\mathbf{C} = [c_1, \ldots, c_{K-1}]^t$.[1] In practice, $\mathbf{\Gamma} = [\gamma_1, \ldots, \gamma_G]^t$ is estimated in place of $\mathbf{\Sigma}$, where $\gamma_g = \ln(\sigma_g)$. This ensures that the estimates of σ_g will always be positive. Following estimation of $\mathbf{\Gamma}$, we have $\widehat{\mathbf{\Sigma}} = \left[e^{\hat{\gamma}_1}, \ldots, e^{\hat{\gamma}_G}\right]^t$. This is similar to the regression model with heterogeneous variances proposed by Harvey (1976). Reardon et al. (2017) describe how to estimate these parameters and their standard errors using ML. The estimation is based on expressing the log-likelihood function for the data as

$$l(\mathbf{N}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{C}) = A + \sum_{g=1}^{G} \sum_{k=1}^{K} n_{gk} \ln\left(\Phi\left(\frac{\mu_g - c_{(k-1)}}{e^{\gamma_g}}\right) - \Phi\left(\frac{\mu_g - c_k}{e^{\gamma_g}}\right)\right), \tag{4}$$

where A is a constant based on the multinomial distribution. The scale of the y^* variable is undefined and constraints must be placed on the model parameters to make the model identified. Reardon et al. (2017) describe different sets of equivalent constraints that can be used, as well as a process to linearly transform the resulting estimates of $\mathbf{M}$ and $\mathbf{\Sigma}$ to a scale such that the overall mean of y^* is 0 and the standard deviation is 1 (i.e., y^* is in a standardized metric).

---

[1] In some cases, researchers may know the operational cut scores used to coarsen the original test scores. However, because the original test score metric may not be the same as the latent, normal y^* metric, these cut scores cannot necessarily be used as fixed values when estimating the other model parameters. In cases where researchers believe the original scale score metric meets certain normality assumptions, then it would be possible to treat the cut scores as fixed values and estimate the remaining parameters relative to those cuts cores. Reardon et al. (2017) discuss this issue in greater detail.

Assumptions and Interpretation of the HETOP Model

The primary assumption of the HETOP model is that test score distributions are respectively normal and thus a probit link can adequately summarize the data (Albert & Chib, 1993; Ho & Haertel, 2006; Ho & Reardon, 2012; Reardon & Ho, 2015). Let y denote the latent test scores in their original, continuous metric. The scores are said to be respectively normal if there is a single, monotonic function $g(y) = y^*$ that can transform the original scale scores into the y^* metric, in which the within-group distributions are all normal. The HETOP model estimates the means and standard deviations of achievement expressed in the y^* metric, not necessarily the original test score metric.

It would be possible to use alternate within-group distributional forms, such as logistic distributions. In that case, the assumption would be that the latent distributions were respectively logistic. We elect to use normal distributions (i.e., a probit link function) due to their familiarity for many researchers and because analyses of real test score data by Reardon et al. (2017) suggest the respective normality assumption is reasonable and likely to be satisfied in practice when analyzing coarsened test score data. Prior research using similar methods in the two-group case to estimate achievement gaps suggest these models are likely to be robust to violations of respective normality and that the probit transformation may yield more accurate estimates than the logit transformation (Ho & Reardon, 2012).

The HETOP model parameters can be viewed as ordinal statistics, because they rely only on ordinal information in the data. That is, the y^* metric is invariant to monotonic transformations of the original score scale – any monotonic transformation to the original latent score scale (that also transforms the cut scores) will lead to identical parameter estimates in the y^* metric. Because there may be doubts about whether test score scales have meaningful interval properties (Ballou, 2009; Briggs, 2013; Domingue, 2014), the choice of a single metric such as y can be difficult to justify, making this a potential advantage of the HETOP model. While it would be possible to analyze coarsened proficiency data with other methods that rely only on ordinal information in the coarsened scores, we believe the

HETOP models are particularly useful in this context. Many ordinal statistics focus on making pairwise comparisons between distributions and may require adjustments for ties (repeat values), which occur frequently in coarsened data. The HETOP models naturally account for repeat values, while the model parameters allow one to both make pairwise comparisons between groups and summarize patterns across more than two groups. The HETOP model parameters, for example, can be used to estimate standardized mean differences or probability-based ordinal effect size measures between pairs of groups (Agresti & Kateri, 2017), to estimate intraclass correlation coefficients among groups (Reardon et al., 2017), or as outcome measures in regression models.

Problems with the HETOP Model

Although the HETOP model works well for recovering the means and standard deviations in the y^* metric when only $\mathbf{N}$ is observed, a number of problems can occur when attempting to estimate the parameters using ML with small samples. First, for some patterns of sampling zeros in $\mathbf{N}$, finite ML estimates may not exist for all groups. Second, even when there may be sufficient information for the ML estimates to be defined in theory, computer algorithms may not converge to a solution or may produce unstable estimates with extremely poor precision. Such issues are sometimes referred to as fragile identification (Freeman et al., 2015; Keane, 1992). Third, in cases where the ML estimates do exist and software can identify the estimates, the simulations in Reardon et al. (2017) show that there is negative bias and excessive sampling error in standard deviation estimates when group sample sizes are small (less than 50) and the cut scores are asymmetrically and/or widely spaced.

Reardon et al. (2017) considered two possible solutions to these challenges. The first was to fit a homoskedastic ordered probit (HOMOP) model that constrains all groups to have a common standard deviation. The second was a partially heteroskedastic ordered probit (PHOP) model that estimates a single, pooled standard deviation for all groups with sample sizes below a set threshold. The HOMOP model makes the potentially unrealistic assumption that all groups have equal standard deviations, precluding the study of heterogeneity of within-group variances. The PHOP model allows for the study of heterogeneity among some groups, but entails the arbitrary constraint that a subset of groups (here, those with sample sizes below some threshold) have a common standard deviation.

Lockwood et al. (2018) describe a Bayesian model that addresses these challenges by borrowing information from other groups and from covariates. As anticipated, the Bayesian model solves the identification and existence problems, and reduces sampling error of standard deviation estimates, but at the cost of additional bias and the requirement that analysts define or estimate appropriate prior distributions for the latent group parameters.

In the context of recovering achievement test score distributions for schools, each of these approaches borrows information from students in other schools taking the same test in the same year, because the models are defined assuming that the coarsened data are from a single test with a common set of cut scores. In the next section, we describe a generalized version of the HETOP model that can be used to estimate multiple latent distributions for each group simultaneously, even if the distributions are for different measures coarsened using different cut scores. Estimating the distributions simultaneously allows one to borrow information from students in the same school taking tests in these additional years, grades, and subjects. This approach will be preferable, in theory, if borrowing information from the same group provides better estimates than borrowing information from other groups. This could occur, for example, if there is more variability in the relative magnitude of parameters across schools (within time points) than within schools (across time points). This is an empirical question that we investigate in Section 2 with a national database of real test score data, where we find evidence that there is greater variability in standard deviations across districts than within districts over time.

The Pooled HETOP Model

When analysts have test score proficiency counts from multiple test administrations across years or grades for the same G groups, it is possible to pool information across administrations, resulting in a more general model that may also improve the estimates of some parameters, such as the estimates of σ_g. To define the model, suppose there are coarsened proficiency counts for a set of G schools across R grades. We now assume there is an underlying variable y^* that is normally distributed within each school g and grade r,

$$y^*_{gri} \sim N(\mu_{gr}, \sigma_{gr}), \tag{5}$$

and that the observed data, $x_{gri} \in \{1, \dots, K_r\}$, consist of ordered proficiency scores that arise from coarsening these y^* values with grade-specific cut scores such that

$$x_{gri} = k \text{ if } c_{r(k-1)} < y^*_{gri} \leq c_{rk}. \tag{6}$$

The goal is to estimate the school and grade-specific parameters μ_{gr} and σ_{gr} for each school in each grade simultaneously.

If we model the mean and standard deviation parameters with parametric functions of grade and group, with $\mu_{gr} = f(g, r)$ and $\gamma_{gr} = \ln(\sigma_{gr}) = h(g, r)$, the model-implied probability of student i in group g scoring in proficiency category k in grade r can be written as:

$$\begin{aligned} \pi_{grk} &= \Phi\left(\frac{f(g,r) - c_{r(k-1)}}{e^{h(g,r)}}\right) - \Phi\left(\frac{f(g,r) - c_{rk}}{e^{h(g,r)}}\right) \\ &= Pr\left(c_{r(k-1)} < y^*_{gri} \leq c_{rk}\right). \end{aligned} \tag{7}$$

Let n_{grk} be the number of students in group g scoring in proficiency category k in grade r, and let $n_{gr} = \Sigma_{k=1}^{K} n_{grk}$. We can write the log-likelihood of the model in terms of the parameters in $f(\ )$ and $h(\ )$ as

$$l(\mathbf{N}|\boldsymbol{f}, \boldsymbol{h}, \mathbf{C}) = A + \sum_{g=1}^{G} \sum_{r=1}^{R} \sum_{k=1}^{K} n_{grk} \ln\left(\Phi\left(\frac{f(g,r) - c_{r(k-1)}}{e^{h(g,r)}}\right) - \Phi\left(\frac{f(g,r) - c_{rk}}{e^{h(g,r)}}\right)\right). \tag{8}$$

For now we assume there are the same number of cut scores in each grade level (though they do not need to be equal across grades), but it is possible to relax this assumption.[2]

[2] If we restrict consideration to cases where K is equal across grades, the three-way $K \times G \times R$ table could be collapsed to a two-way $K \times (GR)$ table, thus making the model equivalent to the HETOP model described above.

To connect with the models above, fitting the HETOP model separately within each grade is equivalent to having fully nonparametric functions $f(g,r) = \mu_{gr}$ and $h(g,r) = \gamma_{gr}$. Fitting the HOMOP model separately within each grade, which constrains all groups in a given grade to have the same standard deviation, uses $h(g,r) = \gamma_r$. We consider two alternative forms for $h(\quad)$ that leverage information across grades but within groups. First, we define a model that estimates a single scale parameter for each group using

$$\gamma_{gr} = h(g,r) = \gamma_g. \tag{9}$$

Because this model estimates a single standard deviation parameter per group that is constant across grades, we refer to it as a "fully pooled HETOP model." Second, we define a model that estimates the scale parameter for each group with a group-specific linear function of grade using

$$\gamma_{gr} = h(g,r) = \beta_{0g} + \beta_{1g} * r. \tag{10}$$

We refer to this as the "linear trend pooled HETOP model." In Equation (10), β_{0g} is a unique scale parameter for each group corresponding to the grade level coded as 0, and β_{1g} is the rate of change in this scale parameter across grade levels.

Although there may be very little information with which to estimate a group's mean and standard deviation in a single year or grade, these models leverage additional data by pooling across multiple grades of data. While the model described here assumes data from multiple grades are available, the extension to additional dimensions (e.g., years or subjects) is straightforward. Pooled HETOP models can be applied most flexibly when pooling across time (e.g., grades or years) rather than subjects, although this will depend on both statistical and substantive considerations as we discuss below. In addition, while we focus on using the pooled model to improve small-sample standard deviation estimates, the models could be extended to have a functional form for the means or to include additional

However, this is only possible if K is equal across all grades and one constrains the locations of the cut scores across grades to be equal, neither of which are assumed for the pooled HETOP model in Equations 5 through 8.

covariates in the model that represent other group variables (such as school characteristics or aggregate student demographic information). We focus on the standard deviations because prior work suggests small-sample standard deviation estimates are more problematic than small-sample mean estimates.

The pooled HETOP models defined here treat the group parameters as fixed effects to be estimated individually. The data structure described above can also be conceptualized as a multi-level data structure, with repeated observations nested within groups. One could potentially treat the group-level parameters as random effects, estimating the distributions of random effects and using a second step to predict values for specific observations. However, because the tests and cut scores can vary across the different levels (i.e., repeated grades or years), the model allows for heteroskedasticity, and individual-level data are not available, estimating these models would likely not be possible using standard ordered mixed effects regression models. The Bayesian HETOP model described in Lockwood et al. (2018), for example, treats the group parameters as random variables, but was developed under the assumption that a single, common set of cut scores was used to coarsen all observed scores. Lockwood et al. discuss additional considerations when selecting between models that treat group parameters as fixed (i.e., directly estimated) or random effects.

Pooled HETOP Model Identification

Because the latent y^* metric is unobserved and indeterminate, constraints are needed to identify the scale of the estimates. In the standard HETOP model for a single grade, for example, two constraints are needed: one constraint to set the location of the latent y^* metric and one to define the scale of the y^* metric. To generalize this for the pooled HETOP models, let P_m be the number of parameters used per group to model the means, P_s be the number of parameters used per group to model the standard deviations, and K be the number of categories per grade (again assuming an equal number of cut scores in each grade, and assuming that $K \geq 3$ in each grade). In total the model defines $G(P_m + P_s) + R(K-1)$ total parameters, and requires at least $P_m + P_s$ constraints on these parameters to set the location and the scale of y^*. The fully pooled HETOP model, for example, uses R parameters per group to model the means (i.e., a separate mean estimated in each grade), but only one additional parameter per group for the standard deviations, and thus requires $R + 1$ constraints; R constraints to define the location for each grade, and one additional constraint to set the scale of the standard deviation parameters. The linear trend pooled HETOP model requires $R + 2$ constraints. Fitting the HETOP model separately within each grade requires $2R$ constraints, to set the location and scale of the estimates in each grade.

There are different ways to select constraints that satisfy these requirements and that result in statistically equivalent models, where parameters will be linear transformations of one another and the model log-likelihoods will be equal. One possibility, for example, would be to fix the first cut score in each grade level to a fixed value (e.g., to 0) and then constrain the second cut score for P_s of the grade levels to another fixed value (e.g., to 1). In the linear trend pooled HETOP model, another option is to constrain the weighted sum of group means to be 0 within each grade, and constrain the weighted sum of the β_{0g} and β_{1g} parameters to be 0 across groups.

These constraints assume that ML estimates exist for each relevant parameter. Certain patterns of sampling zeroes can prevent finite ML estimates from existing for some samples, even when the model specifications and data structure (e.g., number of grades, number of categories, and number of constraints) should, in theory, support model estimation. For example, if all observations in a single group are in the highest or lowest category in a given grade, a finite ML estimate will not exist for this group mean and hence for the model overall, despite having a sufficient number of grades, categories, and constraints to identify the model as described above. This problem arises due to patterns in some samples of data rather than due to the specification of the model. In the simulation section we describe an adjustment that can be made to sampled frequency counts to ensure the existence of finite ML estimates for all samples. Placing additional structure on the model, for example by modeling the group means with a linear trend in $f(\quad)$, is another potential option for overcoming problems caused by sparseness.

Pooled HETOP Model Assumptions and Standardization

The HETOP model assumes that the test score distributions are respectively normal and were coarsened with common cut scores within grades, years, and subjects. The fully pooled and linear trend pooled HETOP models place additional constraints on the relative magnitude of group standard deviations, which imply assumptions about the overall structure of group standard deviation parameters. To aid with the interpretation of results, once estimates of $\hat{\mu}_{gr}$ and $\hat{\sigma}_{gr} = \exp(\hat{\gamma}_{gr})$ have been obtained subject to necessary identification constraints, the estimates can be linearly transformed to a standardized within-grade metric in which the overall distribution of y^* has a marginal mean of 0 and a marginal standard deviation of 1 within each grade. Parameter estimates and standard errors on the within-grade standardized metric can be obtained by applying the formulas described in the Appendix of Reardon et al. (2017) to estimates from each grade separately. Letting $\hat{\mu}'_{gr}$ and $\ln(\hat{\sigma}'_{gr}) = \hat{\gamma}'_{gr}$ be the parameter estimates after standardizing within grades, standardization leads to the following relationships:

$$\begin{gathered} \hat{\mu}'_{gr} = \frac{\hat{\mu}_{gr} - \xi_r}{\exp(\Gamma_r)} \\ \ln(\hat{\sigma}'_{gr}) = \hat{\gamma}'_{gr} = \hat{\gamma}_{gr} - \Gamma_{\mathrm{r}}, \end{gathered} \tag{11}$$

where ξ_r is an estimate of the overall mean in grade r, and $\exp(\Gamma_r)$ is an estimate of the overall standard deviation in grade r, in the metric defined by the constraints used for identification.

Standardizing estimates within grades makes the assumptions of the pooled HETOP models slightly less restrictive. In the fully pooled HETOP model, for example, $\gamma_{gr} = \gamma_g$ is constant across grades, but this implies only that for a fixed pair of grades, r_1 and r_2, the ratio of any single group's standard deviations in the standardized metric will be constant, not that the standard deviations will be equal in absolute value. That is, the fully pooled HETOP model implies that

$$\frac{\sigma'_{gr_1}}{\sigma'_{gr_2}} = \frac{\exp\left(\gamma_g - \Gamma_{r_1}\right)}{\exp\left(\gamma_g - \Gamma_{r_2}\right)} = \exp\left(\Gamma_{r_2} - \Gamma_{r_1}\right) \tag{12}$$

will be constant across all schools (g) for a fixed pair of grades r_1 and r_2. The model also implies that the ratio of standard deviations for any pair of groups g_1 and g_2 will be constant across grades, meaning that $\sigma'_{g_1 r}/\sigma'_{g_2 r}$ will be constant for any choice of r. This implies that the rank ordering of group standard deviations remains constant across grades in the pooled model.

The linear trend pooled HETOP model instead implies that the ratio of any single group's (standardized) standard deviations across a pair of grades will depend on the group's slope, distance of the grades, and grade-specific standardization constants:

$$\begin{aligned} \frac{\sigma'_{gr_1}}{\sigma'_{gr_2}} &= \frac{\exp\left(\gamma_{gr_1} - \Gamma_{r_1}\right)}{\exp\left(\gamma_{gr_2} - \Gamma_{r_2}\right)} = \frac{\exp\left(\beta_{0g} + \beta_{1g} * r_1 - \Gamma_{r_1}\right)}{\exp\left(\beta_{0g} + \beta_{1g} * r_2 - \Gamma_{r_2}\right)} \\ &= \exp\left(\beta_{1g}(r_1 - r_2) + \left(\Gamma_{r_2} - \Gamma_{r_1}\right)\right). \end{aligned} \tag{13}$$

Likewise, the ratio of standard deviations for any pair of groups changes by a common factor across grades:

$$\frac{\sigma'_{g_1 r}}{\sigma'_{g_2 r}} = \frac{\exp\left(\beta_{0g_1} + \beta_{1g_1} * r - \Gamma_r\right)}{\exp\left(\beta_{0g_2} + \beta_{1g_2} * r - \Gamma_r\right)} = \exp\left(\left[\beta_{0g_1} - \beta_{0g_2}\right] + \left[\beta_{1g_1} - \beta_{1g_2}\right] * r\right). \tag{14}$$

Thus, the linear trend pooled HETOP model does not require that the rank ordering of group standard deviations remains constant across grades.

We have described the assumptions of the pooled HETOP models when pooling across grades here. The same assumptions would apply to other dimensions as well. If the model were used to pool across years, for example, the assumptions would apply to the relative magnitudes of group standard deviations across years; if the model were used to pool across subjects, the assumptions would apply to the relative magnitudes of group standard deviations across subjects. In addition to the statistical assumptions described here, one must also consider whether it makes sense substantively to pool across dimensions, something we discuss further below. In the next section we evaluate the plausibility of these assumptions about the relative magnitudes of group standard deviations within subjects across grades and years in an empirical dataset.

2. Empirical Test of Pooled Model Assumptions

This section analyzes district-level test score proficiency data from 40 states to evaluate whether there is evidence that the patterns among relative magnitudes of district-level standard deviations are consistent with the assumptions made by the pooled HETOP models introduced above. We use publicly available data from the Stanford Education Data Archive version 2.1 (SEDA; Reardon et al., 2018). SEDA contains estimated mathematics and English/Language Arts (ELA) grade 3-8 test score means and standard deviations for nearly every US public school district in the 2008-09 through 2014-15 school years.

The means and standard deviations in SEDA are estimated by fitting partially constrained HETOP models separately in each state, grade, year and subject using aggregate district-level proficiency counts obtained from the ED*Facts* database (Fahle et al., 2018). Because our goal is to study variation among group standard deviations, we exclude standard deviation estimates that were constrained during estimation, and focus only on freely estimated standard deviations. The exact sample restrictions are described in the Appendix. The final sample consists of 620,588 unique standard deviation estimates across 40 states and 9,266 unique districts. Each district has between 1 and 42 repeated observations (across six grades and seven years) in each subject, with an average of approximately 34 observations per district-subject. On average there are 231 districts per subject and state, ranging from 54 to 699.

Models

SEDA contains estimates of σ'_{grt} with an associated standard error for each district g in grade r and year t in each state and subject. These estimates are on a standardized metric such that within each state, subject, grade, and year, the weighted sum of the means is 0 and the total student-level variance is equal to 1. If the assumptions about the relative magnitudes of standard deviations for the fully pooled or trend models are met for a particular state by subject dataset, then the natural log of the standardized values should be related to the γ_{grt} values that would be obtained by fitting a fully pooled or trend HETOP model as

$$\ln(\sigma'_{grt}) = \gamma'_{grt} = \gamma_{grt} + \Gamma_{rt} = \beta_{0g} + \beta_{1g}r + \beta_{2g}t + \Gamma_{rt}. \quad (15)$$

This implies that if the fully pooled or trend HETOP model assumptions are valid, the γ'_{grt} parameters should follow a linear function of grade and year, net of grade-year specific fixed effects Γ_{rt}. In the case of the fully pooled HETOP model with constant γ_g parameters, $\beta_{1g} = \beta_{2g} = 0$ for all groups, and the γ'_{grt} parameters would be a group-specific constant plus a grade-year specific fixed effect.

We fit two precision-weighted hierarchical linear models (HLM; Raudenbush & Bryk, 2002) for each state-subject dataset, with estimates $\hat{\gamma}'_{grt} = \ln(\hat{\sigma}'_{grt})$ as outcomes. The first model includes grade-year fixed effects and a random intercept for each district, and represents the structure assumed by the fully pooled HETOP model in each state-subject dataset. The second model includes grade-year fixed effects as well as random intercepts, grade trends, and year trends for each district, and represents the structure assumed by the linear trend HETOP model with both year and grade trends for each district. These models are used to study two important patterns in the data. First, we use results from Model 1 to estimate the proportion of variance in γ_{grt} values that is between rather than within districts. If there is more variability between than within districts (net of grade and year fixed effects), this suggests that pooled or trend HETOP models are likely to be preferable to models that place constraints across districts. Second, we test whether adding grade and year trends in Model 2 explains a statistically and practically significant amount of the within-district variability of γ_{grt} values. If the grade and year trends explain a substantial proportion of within-district variability, it suggests that the trend HETOP model will be preferable to the fully pooled HETOP model.

Results

Across the 80 state-subject datasets, in Model 1 on average 65% of the total variance in γ_{grt} values in ELA (range 41% to 88%) and 64% in Math (range 43% to 89%) was between rather than within districts. The ratio was less than 50% in only 7 of the 80 models. This suggests that in nearly all cases the fully pooled HETOP model that places constraints within districts (across years and grades) would be preferable to the HOMOP model that places constraints across districts (within years and grades). In Model 2, the variance of year and grade trends was statistically significant in all but 1 of the 80 state-subject datasets, suggesting that a linear trend pooled HETOP model with district-specific grade and year trends would be preferable to a fully pooled model without these trends. Adding the grade and year trends reduced the unexplained within-district variability in γ_{grt} values by approximately 35% for ELA and 29% for Math, on average, relative to the fully pooled HETOP model. This suggests that including district-specific linear trends explains a substantial proportion of variability that is not explained by the fully pooled HETOP model.

We can also use the magnitude of the estimated variance components to quantify the anticipated gains in accuracy obtained by fitting one of the pooled HETOP models relative to a HOMOP model. Based on the magnitude of the estimated variance components across models, we would expect HOMOP standard deviation estimates to be within approximately ±14% of the true γ_{grt} values in ELA and ±17% in Math. Similar calculations suggest that on average estimates should be within approximately ±8% in ELA or ±10% in Math when using the fully pooled HETOP model, and within ±7% in ELA or ±9% in Math when using the linear trend pooled HETOP model. These results indicate that when placing constraints on standard deviation estimates within subjects, the linear trend pooled HETOP model should generally produce more accurate estimates of γ_{grt} than either the fully pooled HETOP model or HOMOP model. In the next section we use a computer simulation to evaluate model performance under known conditions.

3. Simulation

A Monte Carlo computer simulation was used to investigate the small (i.e., finite) sample performance of the fully pooled HETOP model and the linear trend pooled HETOP model (referred to in this section as the “trend HETOP” model) relative to the standard HETOP and HOMOP models when pooling data across repeated observations. Data were generated for a set of 25 groups observed across six occasions. This scenario could represent having data for 25 schools across six grades, and hence we refer to the occasions as “grades.” The simulation varied the true group standard deviation structure (either constant values or following group-specific linear trends across grades), group sample size (sizes of 10, 25, 50, 100, or 200), and cut score locations. The cut scores used to coarsen the data were placed at either the $20^{th}/50^{th}/80^{th}$ (mid), $5^{th}/30^{th}/55^{th}$ (skewed) or $5^{th}/50^{th}/95^{th}$ (wide) percentiles of the overall distribution within each grade, or were mixed such that scores in the first three grades were coarsened using the mid, skewed, and wide cut scores, respectively, with the same pattern for grades four through six. Overall there were (2 group structures) x (5 sample size conditions) x (4 cut score conditions) for a total of 40 simulation conditions. We generated and analyzed 1000 replications (i.e., samples) in each condition. All simulations and analyses were carried out using Stata 14.2 (StataCorp, 2015), with estimation of the HETOP models conducted using a custom program written by the authors and based on the Stata -ml- functions. All simulation code is available upon request from the authors.

Data Generation

For each group standard deviation structure by sample size condition we began by defining a population of 25 groups with fixed mean and standard deviation parameters at each grade level. Defining the true group mean and standard deviation parameters began by creating a 5-by-5 grid of β_{0g} and β_{1g} values, where the log standard deviation for group g in grade $r = \{0,1,\ldots,5\}$ is γ_{gr}:

$$\gamma_{gr} = \beta_{0g} + \beta_{1g} * r. \tag{16}$$

To determine the true values, we first assigned values of σ_g equal to 0.75, 0.85, 0.95, 1.05 or 1.15 to each group, and defined $\beta_{0g} = \ln(\sigma_g)$. These values were then re-centered such that $\Sigma_g \beta_{0g} = 0$. In the constant standard deviation condition, $\beta_{1g} = 0$ for all groups. In the linear trend condition, a grid with all possible combinations of the five β_{0g} values and the five β_{1g} values $\{-0.10, -0.05, 0.0, 0.05, 0.10\}$ was defined. These values are more extreme than the linear trends found in the national district-level data analyzed above, but were similar to those found in the school-level example below and are used in the simulation to evaluate model performance across a broader range of conditions that might be encountered in practice. The mean for each group was randomly sampled (with replacement) from the values $\{-0.6, -0.3, 0.0, 0.3, 0.6\}$ within each grade. These group means and standard deviations were standardized within each grade so that the marginal mean and standard deviation in each grade were 0 and 1, respectively. The standardized values were used to generate the random samples for each group in each grade, and are the target of recovery.

The standardized σ_{gr} and μ_{gr} values varied across grades based on the random assignment of group mean values, but produce approximately grid-like structures of group means and standard deviations within each grade of data in the standardized metric. The intraclass correlation coefficient (ICC) also depends on the randomly selected group parameters, and ranged from 0.1 to 0.18 (mean 0.14) within grades. The coefficient of variation (CV) among standardized group σ_g values ranged from approximately 0.15 to 0.39 (mean 0.20) across conditions. These values are similar to those found in prior analyses with real test score data (e.g., Fahle & Reardon, 2018; Hedges & Hedberg, 2007). In each replication of each sample size and standard deviation condition, a normally distributed random sample of size n (either 10, 25, 50, 100, or 200) was generated from each group for each of the grades and was coarsened using each set of cut scores (mid, skewed, wide, or mixed).

Parameter Estimation

For each of the four coarsened datasets in each condition, we fit the HETOP and HOMOP models separately within each grade and fit the fully pooled and trend HETOP models simultaneously to all grades. The fully pooled HETOP model was expected to perform best when the data generating model specified constant γ_g values across grades, while the trend HETOP model was anticipated to perform best in the linear trend condition. The HETOP model fit separately in each grade is correctly specified given the data generation process, but is less parsimonious than the fully pooled or trend models. The HOMOP model is incorrectly specified in all conditions. Although we anticipate the separate HETOP and HOMOP models will suffer from some of the problems described above, we include them in the simulation to compare their relative performance to that of the pooled and trend HETOP models.

We used the following procedure to ensure finite ML estimates exist for all samples. When a sampled count vector had only one non-zero count, had non-zero counts in only the top and bottom categories, or had non-zero counts in only two adjacent categories, we replaced the sampled counts for that group with

$$\hat{n}_{grk} = n_{gr} * \frac{n_{grk} + \alpha}{n_{gr} + K * \alpha}, \tag{17}$$

where $n_{gr} = \Sigma_{k=1}^{K} n_{grk}$ is the total group sample size for group g in grade r and $\alpha = \frac{1}{K} = \frac{1}{4}$. This process has been referred to as "flattening" (Fienberg & Holland, 1972) or "smoothing" (Simonoff, 1995) the observed frequency counts. The degree of smoothing depends upon the choice of α and the resulting proportions in each cell tend to get flattened towards a uniform distribution. The use of $\alpha = \frac{1}{K}$ was suggested by Perks (1947). This method is similar to the common technique of adding a small constant (often 0.5) to cells in sparse contingency tables (Agresti, 2013), but it has the desirable property that it leaves the total sample size for each group unaltered.

Outcomes

Evaluation of model performance is based on four outcomes. First, the convergence rate for each model was recorded, indicating whether the ML algorithm could reach a solution. We then evaluated the bias, root mean squared error (RMSE) and confidence interval (CI) coverage for the estimated group means and standard deviations (in the within-grade standardized metric). The bias, RMSE, and CI coverage was aggregated across all groups and grades for a particular condition (i.e., it is the average bias or pooled RMSE across groups and grades for a given condition). The CI coverage was evaluated by determining the proportion of individual estimates for which the estimated parameter value was within +/- 1.96 estimated standard errors of the true parameter value.

To compare the relative gain in efficiency when using a fully pooled HETOP model rather than separate HETOP models in each grade, we conducted one additional analysis. For each replication in the equal standard deviation condition, we fit a fully pooled model using only the first two, three, four, or five grades of data, in addition to the model using all six grades. We then compared the empirical sampling variance of the group standard deviation estimates in these pooled models relative to the separate HETOP models fit within each grade. The efficiency ratio between the fully pooled and separate HETOP models was defined as the ratio of the average observed sampling variance in the separate HETOP models relative to each of the fully pooled HETOP models, computed as:

$$Efficiency\ Ratio = \frac{\sum_{g=1}^{G} Var\left(\hat{\sigma}_{g,HETOP}\right)}{\sum_{g=1}^{G} Var\left(\hat{\sigma}_{g,pooled}\right)}. \qquad (18)$$

This ratio indicates how much smaller the sampling error would be if the group standard deviations remain constant and we pool across either two, three, four, five, or six grades rather than using only a single grade to estimate standard deviations. A ratio of 1 indicates that sampling error in the separate and pooled models are equal, ratios greater than 1 indicate the separate HETOP model estimates have larger sampling error, and ratios less than 1 indicate the pooled model has larger sampling error. A similar calculation was also made to compare the efficiency of the trend pooled HETOP model to the separate HETOP models.

Results

All models converged successfully. Table 1 summarizes the proportion of count vectors that were smoothed across simulation conditions. Across all conditions, approximately 6% of all sampled vectors were smoothed, and these were primarily concentrated in the wide and mixed cut score conditions with small sample sizes. In the wide cut score condition with n=10, for example, approximately 43% of vectors were smoothed, while in the mixed cut score condition with n=10 approximately 20% were smoothed. While smoothing the count vectors ensures existence of ML estimates, it may also lead to positive bias in standard deviation estimates by artificially adding variance to the observed count vectors, something we discuss below.

Table 1. Proportion of Smoothed Count Vectors Across Simulation Conditions.

	N=10		N=25		N=50		N=100		N=200	
Cut Scores	Equal	Trend	Equal	Trend	Equal	Trend	Equal	Trend	Equal	Trend
Mid	0.035	0.057	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000
Mixed	0.182	0.218	0.054	0.097	0.012	0.053	0.003	0.028	0.000	0.022
Skewed	0.107	0.137	0.013	0.030	0.003	0.005	0.000	0.001	0.000	0.000
Wide	0.412	0.455	0.139	0.224	0.040	0.116	0.007	0.058	0.000	0.040

Group Means. We do not show detailed results for the estimated means here because these were not the primary outcome of interest, and because there was little variation in the results across models. Average bias in estimated means was indistinguishable from 0 for all conditions. There was very little difference in the RMSE of means across models, and sample size was the primary factor influencing this outcome. CI coverage was generally good and converged towards the expected rate (95%) as sample sizes increased, with the following exceptions: coverage rates became too low for the HOMOP model as sample sizes increased, and with skewed cut scores rates were as low as 90% for the separate HETOP models with n=10 and for the pooled HETOP model with n=200 in the trend SD condition.

Group Standard Deviations. Figures 1 and 2 display the bias and RMSE for estimated standard deviations. Each panel displays results for a single cut score by group standard deviation structure condition; the x-axis depicts group sample sizes, the y-axis depicts the outcome of interest, and each line represents a different model. With n=10 and n=25 there was a reduction in bias for the fully pooled and linear trend models relative to the separate HETOP models. An exception was the wide cut score condition, in which all models slightly overestimated group standard deviations, on average, with very small sample sizes; as noted above this is likely due to the correction factor applied to ensure ML existence, which was applied most often in the wide cut score condition. The fully pooled and trend models tended to slightly overestimate standard deviation estimates with samples of size n=10, but this bias was smaller in magnitude than the negative bias in the separate HETOP model estimates and was reduced to near 0 with samples of size 25 or larger. The separate HOMOP models produced a small positive bias on average across nearly all conditions, which was larger when there were true trends in the standard deviations. This indicates that the single common standard deviation estimated in the HOMOP model was slightly larger than the true average within-group standard deviations, and is likely due to the misspecification of the HOMOP model.

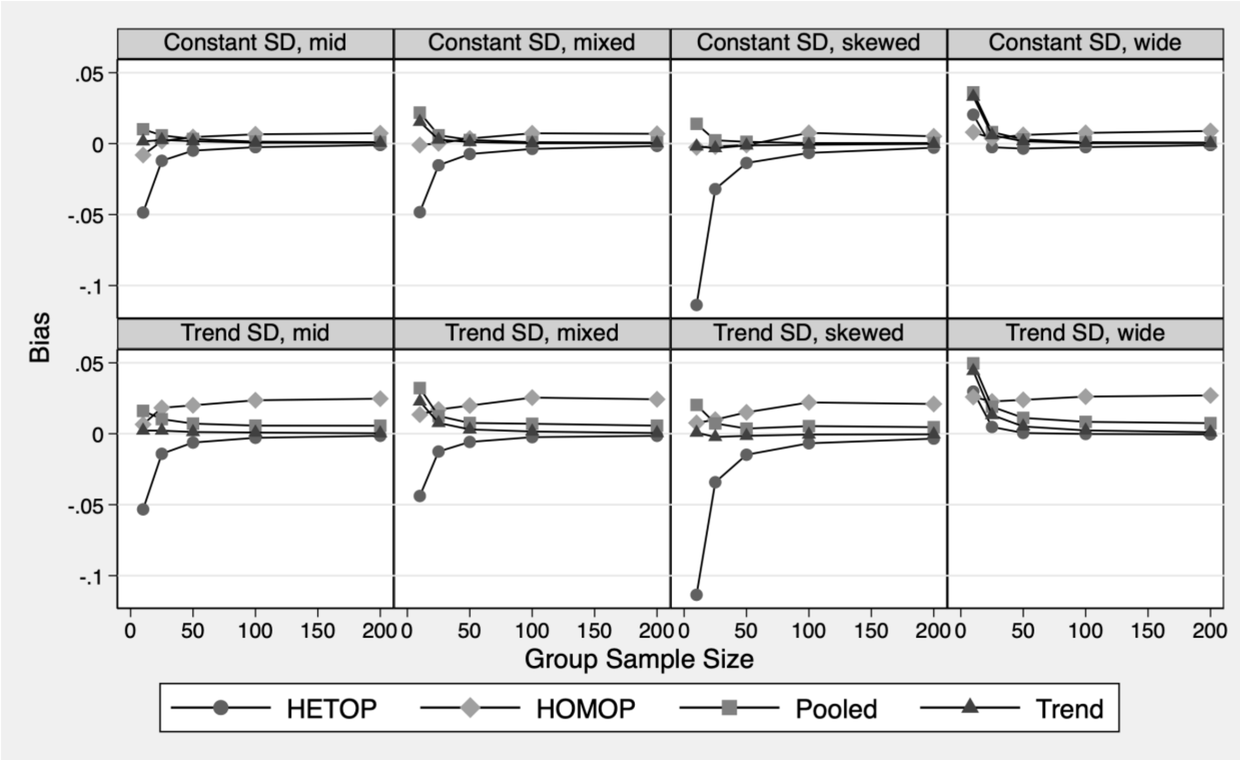

Figure 1. Bias in Estimated Standard Deviations by Standard Deviation Structure, Cut Score Type, and Sample Size for Each Model. HETOP=heteroskedastic ordered probit model; HOMOP=homoskedastic ordered probit model; Pooled=fully pooled HETOP model; Trend=linear trend pooled HETOP model. Constant SD and Trend SD refer to different patterns of true group standard deviations described in text. The mid, skewed, wide, and mixed headings refer to different cut score locations; mid=symmetric cut scores at approximately the 20/50/80 percentiles; skewed=asymmetric cut scores at approximately the 5/30/55 percentiles; wide=symmetric cut scores at approximately the 5/50/95 percentiles; mixed=mix of mid/skewed/wide cut score locations across grades.

Figure 2, depicting the RMSEs of the estimated standard deviations, is simpler to summarize. The separate HETOP models had the largest RMSEs when $n \leq 25$ across all conditions, except when there were true trends in the standard deviations, where the separate HOMOP models sometimes had the largest RMSE when $n = 25$. The difference was substantial for all conditions except the wide cut score condition; again the correction factor used for existence appears to have caused this difference. The separate HOMOP models had constant RMSE across different sample size conditions, with similar RMSE to the fully pooled model when $n = 10$, but larger RMSE than all other models when group sample sizes were greater than 50. In the constant SD conditions, the fully pooled HETOP model had the lowest RMSEs in all but the skewed cut score condition, where the separate HOMOP models had slightly lower RMSE when $n = 10$. While the fully pooled model RMSEs were only slightly lower than the trend model RMSEs, they were substantially smaller than the RMSEs for the separate HETOP models in all but the largest sample size conditions. In the trend SD conditions, the trend HETOP model had the lowest RMSEs in all conditions except when $n = 10$ with mid or mixed cut scores, when the pooled HETOP model had slightly smaller RMSEs. These results were anticipated; the pooled HETOP model is correctly specified (and most parsimonious) in the constant SD conditions, but is mis-specified in the trend SD conditions. In additional simulations using smaller trends in standard deviations (not reported), the pooled HETOP model often had lower RMSEs than the linear trend HETOP model, suggesting that whether the pooled or trend HETOP model achieves lower RMSEs will depend in part on the magnitude of the standard deviation trends.

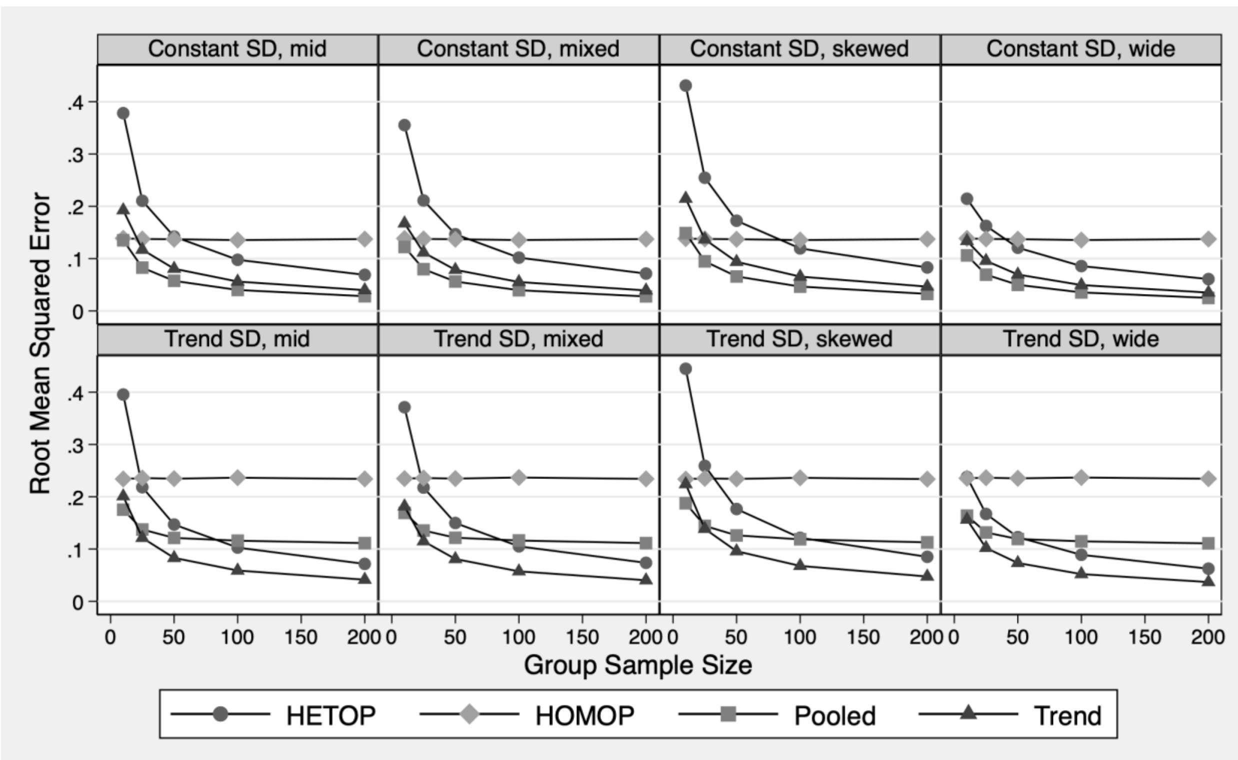

Figure 2. RMSE of Estimated Standard Deviations by Standard Deviation Structure, Cut Score Type, and Sample Size for Each Model. HETOP=heteroskedastic ordered probit model; HOMOP=homoskedastic ordered probit model; Pooled=fully pooled HETOP model; Trend=linear trend pooled HETOP model. Constant SD and Trend SD refer to different patterns of true group standard deviations described in text. The mid, skewed, wide, and mixed headings refer to different cut score locations; mid=symmetric cut scores at approximately the 20/50/80 percentiles; skewed=asymmetric cut scores at approximately the

5/30/55 percentiles; wide=symmetric cut scores at approximately the 5/50/95 percentiles; mixed=mix of mid/skewed/wide cut score locations across grades.

The CI coverage rates (not presented graphically) followed anticipated patterns. For the separate HETOP models, coverage rates were between 92.5% and 97.5% for all conditions when $n \geq 100$, and were too low in small sample size conditions (as low as 86% when n=10) except in the wide cut score condition where they were too high (99% when n=10), likely due to the smoothing correction. For the trend HETOP model, coverage rates were between 92.5% and 97.5% for all conditions except the wide cut score, constant standard deviation condition when n=10, where they were also too high. The trend HETOP model coverage rates were always more accurate than the separate HETOP coverage rates, except in the wide cut score condition where there were minor differences. For the fully pooled HETOP model, coverage rates were similar to the trend model for the constant SD condition, but became substantially less accurate in the trend SD condition as sample sizes increased due to the model mis-specification. Coverage rates for the HOMOP model were too low across all conditions due to model mis-specification (never higher than 25% in any condition) and were less accurate with larger sample sizes.

Figure 3 displays the efficiency ratio of the separate HETOP models relative to the pooled models when pooling across varying numbers of grades. Each panel represents a different cut score condition, and each line represents the efficiency ratio when pooling across a different number of grades. When using only 1 grade, the fully pooled model is equivalent to the separate HETOP models, indicated by the efficiency ratio of 1. In general, the efficiency ratios approach a value of p, the number of datasets being pooled, indicating that the mean squared error (MSE) of estimates using the fully pooled model is approximately $1/p$ times the MSE using the separate HETOP models, a substantial reduction.

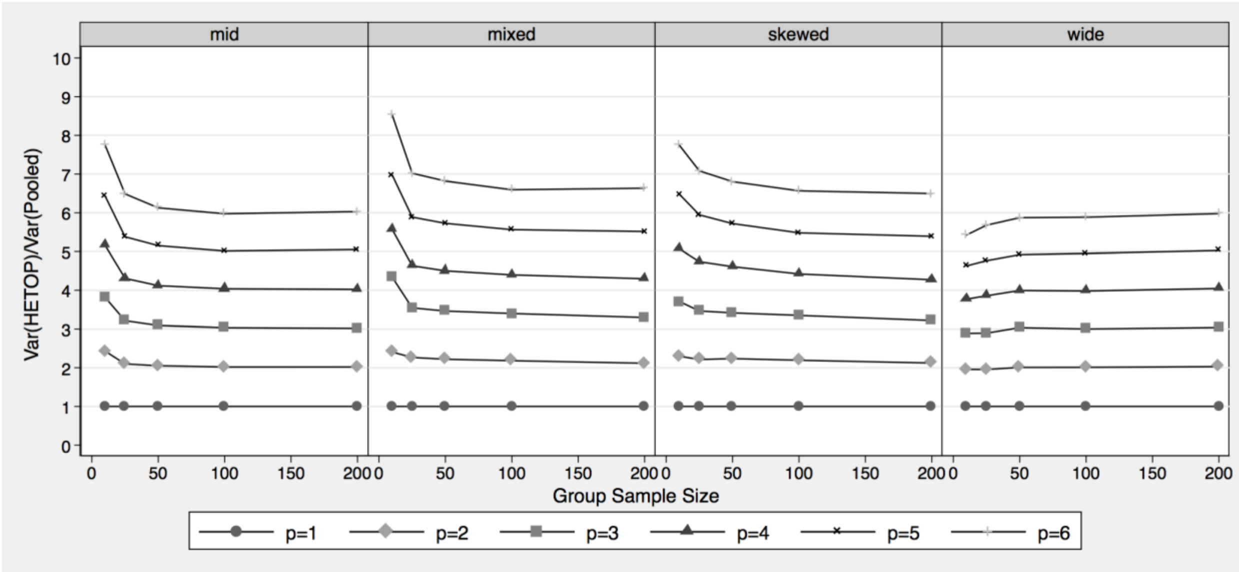

Figure 3. Efficiency Ratios between HETOP and Pooled HETOP Models by Cut Score Type, Sample Size, and Number of Pooled Grades in the Constant SD Condition. The "p" refers to the number of grades used to estimate the fully pooled HETOP model. The mid, skewed, wide, and mixed headings refer to different cut score locations; mid=symmetric cut scores at approximately the 20/50/80 percentiles; skewed=asymmetric cut scores at approximately the 5/30/55 percentiles; wide=symmetric cut scores at approximately the 5/50/95 percentiles; mixed=mix of mid/skewed/wide cut score locations across grades.

Figure 4 plots the observed efficiency ratios of the trend model estimates relative to the separate HETOP model estimates for the trend SD condition. Each panel represents a different cut score condition, and each line plots the efficiency ratio at a single grade level. The trend model has the greatest gains in efficiency for the middle grades (2 and 3), and the smallest efficiency gains for the extreme grades (0 and 5), in all but the mixed cut score condition (which we discuss below). This result is expected because the standard deviations are effectively predictions from a linear regression model, and regression predictions near the center of the predictor distribution will have smaller variance than predictions at the extremes. The estimated (or predicted) scale parameter in the trend model is $\hat{\gamma}_{gr} = \hat{\beta}_{0g} + \hat{\beta}_{1g}r$, where r is the grade level. In least squares (LS) regression, the sampling variance of the prediction is (Casella & Berger, 2002, pp. 557–558):

$$Var\left(\hat{\beta}_{0g} + \hat{\beta}_{1g}r\right) = \frac{\sigma^2}{n}\left(1 + \frac{(r-\bar{r})^2}{Var(r)}\right), \qquad (19)$$

where σ^2 is the residual error variance, n is the number of observations, $\bar{r}$ is the mean of r, and $Var(r)$ is the variance of r. The sampling error in LS thus depends on the specific value of r being considered – it will be σ^2/n at the mean of r and become larger as r gets further from the mean of r. If we assume that the sampling variance of the scale parameter estimates using the separate HETOP models represents σ^2, and the sampling variance of the trend model estimates can be approximated by the LS result in Equation (19), then the anticipated efficiency ratio of the trend model estimates for a model with $p = 6$ grades coded as $r = 0,1,\ldots,5$ would be approximately 1.91 (for $r = 0$ and 5), 3.39 (for $r = 1$ and 4), and 5.53 (for $r = 2$ and 3). The dashed horizontal lines in Figure 4 depict these anticipated efficiency ratios.

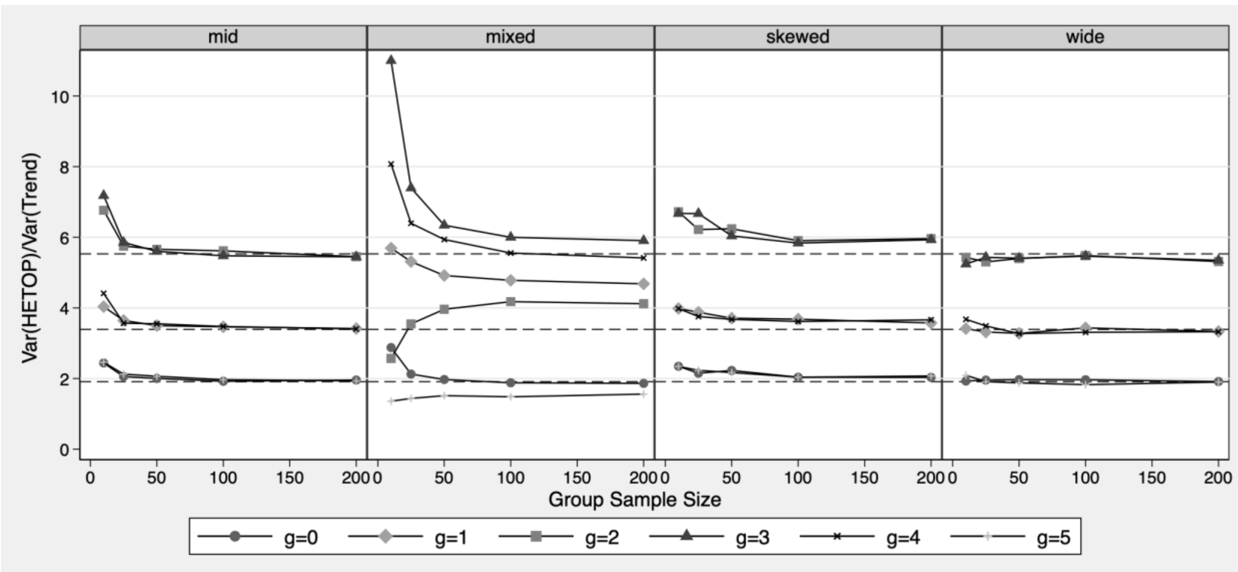

Figure 4. Efficiency Ratios between HETOP and Trend Pooled HETOP Models by Cut Score Type, Sample Size, and Grade in the Trend SD Condition. The "g" represents each of the six possible grade levels. The mid, skewed, wide, and mixed headings refer to different cut score locations; mid=symmetric cut scores at approximately the 20/50/80 percentiles; skewed=asymmetric cut scores at approximately the 5/30/55 percentiles; wide=symmetric cut scores at approximately the 5/50/95 percentiles; mixed=mix of mid/skewed/wide cut score locations across grades. The mixed cut score condition used mid cut scores for grades 0 and 3, skewed cut scores for grades 1 and 4, and wide cut scores for grades 2 and 5.

The approximations appear to work well for the mid, wide, and skewed cut score conditions, but are less accurate for the mixed cut score conditions. In the mixed cut score conditions the sampling variance of the separate HETOP estimates varies across grade levels depending upon the distribution of the cut scores, resulting in the equivalent of a heteroskedastic error term. These results suggest that the efficiency ratio of the trend model relative to separate HETOP models can be approximated using results from standard LS regression. When cut score locations vary substantially across grade levels the approximations may be less accurate, but substantial gains in efficiency remain. Hence, although the trend estimates are more efficient than the separate HETOP estimates, the gain in efficiency depends on factors such as the number and coding of the grades and the cut score locations.

Summary. These results suggest that when data for repeated test administrations are available, the fully pooled and trend HETOP models can substantially reduce bias and sampling error of standard deviation estimates relative to fitting separate HETOP models, particularly with very small sample sizes. The reduction in bias is smaller with larger samples or more equally spaced cut scores, but gains in efficiency remain across conditions. The fully pooled and trend models also had smaller sampling variance than the separate HOMOP models across nearly all conditions. Use of the smoothing correction did appear to induce some positive bias in standard deviation estimates, as anticipated. The results illustrate that the relative performance of the models depends on many factors, including the number of waves (grades) of data available, group sample sizes, cut score locations, and the true values of the standard deviations. In the next section we illustrate how analysts might go about selecting and estimating a pooled HETOP model with real data.

4. Real Data Example

Determining whether to use the fully pooled, linear trend, HOMOP or full HETOP model depends on a number of factors, including the type of data available, group sample sizes, location of the cut scores, average values of σ'_{gr}, and the true structure of the σ'_{gr} values. If all group sample sizes are large, full HETOP models that estimate a unique standard deviation for each group will likely be preferred. Often, however, the choice for estimating parameters of small groups will be either a model placing constraints across groups (e.g., a HOMOP model) or a model placing constraints within groups (e.g., a pooled HETOP model). If data are only available from a single measure or time point, then between-group constraints are the only option. When data are available from multiple measures or waves, the choice of model will depend on a combination of statistical and substantive factors. In this section we use a single year of publicly reported coarsened proficiency data from a statewide mathematics assessment administered in grades 3-8 to illustrate how analysts might go about selecting a pooled HETOP model in practice. Here we will consider models that pool information from students taking mathematics tests in the same school and year across different grades.

The data contain coarsened proficiency counts for 124 schools that enrolled at least 16 students in each grade (data for schools with smaller sample sizes were not reported publicly), resulting in $124 * 6 = 744$ school-grade cells. There are $K = 4$ proficiency categories in each grade, but the tests and cut scores vary across grades; the 3rd grade cut scores (approximately 7th/26th/65th percentiles) are similar to the skewed cut score simulation condition, while the 8th grade cut scores (approximately 19th/47th/75th percentiles) are more similar to the mid cut score condition. The goal is to estimate the mean and standard deviation of math achievement scores within each school-grade cell from the coarsened proficiency data. The within-grade sample sizes range from 16 to 310 (mean=67.9, median=58), suggesting that small sample bias and sampling error could be a concern for a large proportion of school-grade cells. There are also eight school-grade cells that do not have sufficient data to estimate both a mean and standard deviation without pooling or additional constraints.

The analyses of national district-level data above suggest that, a priori, when test score data are available across multiple grades, we would expect a linear trend HETOP model with linear grade trends to be optimal. We also use statistical criteria to select a HETOP model for this particular dataset. To do so, we fit a series of nested HETOP models that can be compared with likelihood ratio tests. Model 1 estimates a unique mean for each school in each grade while constraining the log standard deviation to be equal across schools within grades, and is equivalent to estimating a separate HOMOP model in each grade. Model 1 is identified by constraining the weighted sum of the means to be 0 within grades, and constraining the common scale parameter, $\ln(\sigma_r) = \gamma_r$, to be 0 in all grades. Model 2 is the fully pooled HETOP model introduced above that estimates a unique mean for each school in each grade and a single log standard deviation parameter for each school, pooled across grades. Model 2 is identified by constraining the weighted sum of the means to be 0 within grades and the weighted sum of the school-specific log standard deviations to be 0 across schools. Finally, Model 3 is the linear trend pooled HETOP model that estimates a unique mean for each school in each grade and a pooled log standard deviation with a linear grade trend for each school. Model 3 is identified by constraining the weighted sum of the means within each grade to be 0, and constraining both the weighted sum of the intercepts and the weighted sum of the linear trends to be 0. All three models allow the cut score locations to vary across grades.[3]

Table 2 summarizes the results across all three models. First, to determine whether between or within-school constraints on the log standard deviations are preferable for these data, we compare the fit of Models 1 and 2. A likelihood ratio test at $\alpha = 0.01$ suggests that Model 2 provides a statistically better fit to the data ($\chi^2 = 481.11$, $df = 123$, $p < 0.001$), indicating that constraints within schools (across grades) are preferable to constraints across schools (within grades). Substantively, this suggests that the relative variability in student performance tends to be more similar for students in different grades of the same school than it is for students across different schools within the same grade. Next, a likelihood ratio test comparing Models 2 and 3 suggests adding linear trends to the scale parameters for each school also leads to a statistically better fit ($\chi^2 = 171.47$, $df = 123$, $p = 0.0026$), implying there is enough systematic change in the relative variability of student mathematics performance across grades to include

[3] Because the tests differ across grades we do not expect the cut scores to be equal, but we also compared the fit of models that constrained the cut score locations to be equal across grades. For all three models, allowing cut scores to vary across grades provided statistically significantly better fit to the data. Therefore, we only report the results of models allowing cut scores to vary.

the additional parameters. Thus, for these data, we would select the linear trend pooled HETOP model to estimate means and standard deviations for each school in each grade.

Table 2 also summarizes the estimated log standard deviation intercepts and trends for Models 2 and 3. The summary statistics are not weighted by school sample size, in which case the average intercepts and trends would have been exactly 0 by construction. The average estimated intercepts were similar in Models 2 and 3, although there was slightly more variation in the Model 3 estimates. The linear trends in Model 3 ranged from -0.136 to 0.120 across schools (mean=0.002, SD=0.044), suggesting a level of heterogeneity in standard deviations that would lead the linear trend model to provide more accurate estimates than the fully pooled model based on the simulations. The table also summarizes the resulting means and standard deviations in the grade-standardized metric, $\hat{\mu}'_{gr}$ and $\hat{\sigma}'_{gr}$. The estimated means were similar across models, but as anticipated the estimated standard deviations differed. While average $\hat{\sigma}'_{gr}$ values were similar across models, Models 2 and 3 indicate substantial additional variability among these estimates (with estimates ranging from 0.475 to 1.413 across schools and grades in Model 3).

In addition to comparing the relative fit of models, different approaches might be used to assess overall goodness of model fit. Table 2 reports an overall chi-square goodness of fit statistic ("χ^2 GOF") based on the observed and expected frequency counts in each category in each school-grade cell. These may be of limited value because the large sample size could indicate statistically significant misfit that is not practically significant, and because these statistics do not indicate the nature of model misfit. As a descriptive measure of fit, Table 2 also reports the mean absolute difference between observed and expected proportions of students scoring in each category for each school-grade cell ("MAD P"). Across the 744 school-grade cells in Model 3, for example, the average difference was 0.030 (range 0.001 to 0.137; median=0.026), indicating that Model 3 appears to accurately characterize the observed proportions for most schools.

Table 2. Summary Statistics for Estimated HETOP Models.

	Statistic	Model 1: HOMOP	Model 2: Pooled	Model 3: Trend
	Log Likelihood	-58869.534	-58628.980	-58543.246
	Free Parameters	756	879	1002
	χ^2 GOF	2031.925	1556.515	1380.347
	MAD P	0.037	0.032	0.030
$\hat{\beta}_0$	Mean		-0.013	-0.018
	SD		0.115	0.151
	Range		[-0.386, 0.394]	[-0.344, 0.483]
$\hat{\beta}_1$	Mean			0.002
	SD			0.044
	Range			[-0.136, 0.120]
$\hat{\sigma}'$	Mean	0.856	0.852	0.850
	SD	0.013	0.098	0.117
	Range	[0.837, 0.877]	[0.569, 1.304]	[0.475, 1.413]
$\hat{\mu}'$	Mean	-0.019	-0.020	-0.020
	SD	0.526	0.518	0.519
	Range	[-1.249, 1.753]	[-1.254, 1.512]	[-1.266, 1.631]

Note: MAD P = mean absolute difference between predicted and observed proportions; SD=standard deviation. The rows corresponding to β_0 and β_1 represent 124 unique estimates across schools; the rows summarizing $\hat{\sigma}'$and $\hat{\mu}'$represent 744 estimates, although in Model 1 there are only 6 possible unique values of $\hat{\sigma}'$. Means and SDs are unweighted.

5. Discussion

This paper presented a generalization of the HETOP model described by Reardon et al. (2017) that can be used to analyze grouped, ordered-categorical data when there are multiple waves of data available for each group. The fully pooled HETOP model leverages the repeated observations by estimating a constant scale parameter for each group across datasets, while the linear trend pooled HETOP model is more flexible and allows each group's scale parameter to vary linearly across the datasets. The simulations and empirical analyses above document four primary reasons the pooled HETOP models might be preferred to standard HETOP models in practice. First, the pooled HETOP models can be estimated in some cases where there are not sufficient data to support estimation of full HETOP models. Second, the pooled HETOP models may better represent observed patterns in group standard deviations than do models placing constraints across groups, which provides another method to address sparse data problems. Third, the pooled models reduce bias in standard deviation estimates relative to full HETOP models when sample sizes are very small, particularly when cut scores are widely or asymmetrically placed as is common in coarsened proficiency data. And, fourth, the pooled HETOP models improve the precision (i.e., reduce RMSE) of estimated standard deviations beyond gains made through reductions in bias.

Whether these gains are realized in practice will depend on the nature of the data. When multiple waves of data are available, the pooled or trend HETOP models will preferable to models placing constraints across groups if there is more variability between groups than within. Our empirical analysis of national district-level data suggests that constraints within districts and subjects are likely to produce more accurate estimates than constraints across districts in the context of coarsened proficiency data, and that linear trends are likely to produce slightly more accurate estimates than fully pooled models. Analysts must also consider whether it is reasonable to expect greater heterogeneity between or within groups, and whether a linear trend is conceptually appropriate based on the nature of the data. It may not be reasonable, for example, to fit a linear trend across data from different subjects, where the repeated observations cannot be placed in a logical order as is possible with grades or years. However, it may still be reasonable to fit fully pooled models across subjects if the assumptions about the relative magnitudes of group standard deviations across subjects are plausible. The example in Section 4 demonstrated how analysts could select an appropriate model using theoretical and statistical criteria. The simulation results provide additional information about the conditions under which pooled HETOP models are expected to lead to the greatest reductions in bias or RMSE relative to full HETOP or HOMOP models. The anticipated reductions in sampling error, for example, can be approximated based on the number of pooled datasets and the coding of the linear predictors used in the trend models.

This paper also leaves important directions for future work. As with any simulation study, many additional factors could have been varied. These factors include additional structures for the standard deviations (including structures that do not conform to the linear trends) as well as violations of respective normality. Another avenue for additional work revolves around the problems caused by non-existence of ML estimates. Essentially, this is a problem of small samples containing limited information about the parameters of interest. In some simulation conditions, for example when sample sizes were n=10 for each group and cut scores were widely spaced, a substantial proportion of group count vectors needed to be adjusted to guarantee existence of the ML estimates when group means were freely estimated. A more complete proof of existence conditions for the ML estimates was not provided and would be a useful extension of the results here. It would also be worth testing models that place additional constraints (e.g., linear trends) on the estimated means as another method for overcoming sparse data problems, and evaluating additional model fit statistics.

As mentioned above, Bayesian and random effects models provide an alternative approach to addressing existence and small-sample problems, but were beyond the scope of the present investigation. These models rely on specifying or estimating prior distributions, rather than attempting to estimate each term individually (e.g., Hedeker et al., 2009; Kapur et al., 2015). Recent work pursuing a Bayesian HETOP model (Lockwood et al., 2018) is similar to the framework described here with an additional random component. However, these Bayesian models have not yet been extended to simultaneously model data from multiple measures with potentially varying cut scores. While Bayesian approaches can overcome problems with the non-existence of the ML estimates and potentially produce estimates with smaller RMSE, they can increase the bias in estimates for individual groups, require appropriate specifications or estimates of prior distributions, and as with the HOMOP and PHOP models, they have so far relied on constraints across rather than within groups. Under certain conditions, including when estimates might be used in secondary analyses, ML estimates may be preferable, and in those cases the models described here are a useful alternative. Pursuing extensions to these models that incorporate multiple sets of data would be a useful area for further study.

Finally, we note that the models described in this paper can be applied to a wide range of ordered-categorical data beyond coarsened test scores. The pooled HETOP models described here are applicable any time analysts have multiple sets of grouped, ordered-categorical data for a common set of groups and wish to estimate distributional parameters of an underlying continuous variable. These data could arise from test scores reported only on ordinal scales such as Advanced Placement (AP) scores, from responses to Likert survey items, or from continuous variables such as income that are often reported in a coarsened form.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Agresti, A., & Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, *73*(1), 214–219. https://doi.org/10.1111/biom.12565

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.1080/01621459.1993.10476321

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*(4), 1055–1082. https://doi.org/10.2307/2111669

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia*, *32*(4), 260–267.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383. https://doi.org/10.1162/edfp.2009.4.4.351

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226. https://doi.org/10.1111/jedm.12011

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

Cox, C. (1995). Location—scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine*, *14*(11), 1191–1203. https://doi.org/10.1002/sim.4780141105

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19. https://doi.org/10.1007/s11336-013-9342-4

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology*, *6*(3), 487–496. https://doi.org/10.1016/0022-2496(69)90019-4

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009–2015. *Educational Researcher*, *47*(4), 221–234. https://doi.org/10.3102/0013189X18759524

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., DiSalvo, R., & Ho, A. D. (2018). *Stanford Education Data Archive: Technical documentation (Version 2.1)*. https://stacks.stanford.edu/file/druid:db586ns4974/SEDA_documentation_v21.pdf

Fienberg, S. E., & Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis*, *2*(1), 127–134. https://doi.org/10.1016/0047-259X(72)90014-0

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

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*(3), 461. https://doi.org/10.2307/1913974

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*(4), 391. https://doi.org/10.4310/SII.2009.v2.n4.a1

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://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*(6), 351–360. https://doi.org/10.3102/0013189X08323842

Ho, A. D., & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples* (No. 665; CSE Report). National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "Proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489–517. https://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*(1), 3–17. https://doi.org/10.3102/10769986027001003

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*. https://doi.org/10.3102/0162373713485814

Kapur, K., Li, X., Blood, E. A., & Hedeker, D. (2015). Bayesian mixed-effects location and scale models for multivariate longitudinal outcomes: An application to ecological momentary assessment data. *Statistics in Medicine*, *34*(4), 630–651. https://doi.org/10.1002/sim.6345

Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, *10*(2), 193–200. https://doi.org/10.1080/07350015.1992.10509898

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

Lockwood, J. R., Castellano, K. E., & Shear, B. R. (2018). Flexible Bayesian models for inferences from coarsened, group-level achievement data. *Journal of Educational and Behavioral Statistics*, *43*(6), 663–692. https://doi.org/10.3102/1076998618795124

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, *42*(2), 109–142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall/CRC.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, *73*(2), 285–334.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications, Inc.

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158–189. https://doi.org/10.3102/1076998615570944

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & DiSalvo, R. (2018). *Stanford Education Data Archive* (Version 2.1) [Computer software]. https://cepa.stanford.edu/seda/overview

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3–45. https://doi.org/10.3102/1076998616666279

Simonoff, J. S. (1995). Smoothing categorical data. *Journal of Statistical Planning and Inference*, *47*(1–2), 41–69. https://doi.org/10.1016/0378-3758(94)00121-B

StataCorp. (2015). *Stata statistical software: Release 14* (Version 14) [Computer software]. StataCorp LP. http://www.stata.com/

Tosteson, A. N. A., & Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making*, *8*(3), 204–215. https://doi.org/10.1177/0272989X8800800309

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, *37*(4), 531–559. https://doi.org/10.1177/0049124109335735

Appendix

This Appendix provides more details about the data and statistical models used to analyze the empirical data from the Stanford Education Data Archive (SEDA; Reardon et al., 2018) data in Section 2.

Data

As noted in the main text, the data analyzed here are from SEDA version 2.1 (Fahle et al., 2018). We make the following sample restrictions to the SEDA version 2.1 database for the purposes of our analyses. First, some states reported proficiency data in only two categories during some years, requiring that HOMOP models were fit to these datasets. We drop these observations because all districts within that particular state, grade, subject, and year were constrained to have equal standard deviations. Second, in cases where data were reported in three or more proficiency categories (the majority of data), PHOP models were fit by constraining the logged standard deviation for districts with fewer than 50 students to be equal to the average logged standard deviation of all districts with more than 50 students in the same state, grade, year, and subject. We therefore drop all district observations with estimates based on fewer than 50 students. These restrictions ensure that the remaining standard deviation estimates were estimated without constraints. After these restrictions, we drop all states with estimates for fewer than 50 districts. Final sample sizes are presented in the main text of the paper.

Statistical Models

SEDA contains estimates of σ'_{grt} (standardized within states, grades, years, and subjects) with an associated standard error for each district g in grade r and year t in each state and subject. The SEDA data also contain estimated standard errors of the $\hat{\sigma}'_{grt}$'s. We use the delta method to estimate the standard error of $\hat{\gamma}'_{grt} = \ln\left(\hat{\sigma}'_{grt}\right)$ as:

$$SE\left(\hat{\gamma}'_{grt}\right) = \sqrt{\frac{1}{\hat{\sigma}'^{2}_{grt}} SE\left(\hat{\sigma}'_{grt}\right)^2} = \frac{1}{\hat{\sigma}'_{grt}} SE\left(\hat{\sigma}'_{grt}\right). \quad \text{(A20)}$$

We use the estimated sampling variances of the $\hat{\gamma}'_{grt}$ values in a variance-known model (Raudenbush & Bryk, 2002) that accounts for the sampling error in the estimates. For each state-subject dataset, the general form of the model begins with an equation for the estimated $\hat{\gamma}'_{grt}$ values

$$\begin{gathered}\hat{\gamma}'_{grt} = \gamma'_{grt} + \epsilon_{grt} \\ \epsilon_{grt} \sim N\left(0, \hat{V}_{grt}\right)\end{gathered} \tag{A21}$$

where $\hat{V}_{grt}$ is the square of the estimated standard error of $\hat{\gamma}'_{grt}$. We then fit two models for the γ'_{grt} values in each state-subject dataset:

Model 1: $\gamma'_{grt} = \beta_{0g} + \Gamma_{rt} + e_{grt}$, where $e_{grt} \sim N(0, \omega_1^2)$ and $\beta_{0g} \sim N(0, \nu_{00})$;

Model 2: $\gamma'_{grt} = \beta_{0g} + \beta_{1g} r + \beta_{2g} t + \Gamma_{rt} + e_{grt}$, where $e_{grt} \sim N(0, \omega_2^2)$ and (A22)

$$\begin{bmatrix}\beta_{0g} \\ \beta_{1g} \\ \beta_{2g}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}, \mathbf{T} = \begin{bmatrix}\tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22}\end{bmatrix}\right).$$

The Γ_{rt} are grade by year fixed effects. Grade and year variables were centered at the mean value within each state-subject dataset. All models were fit using the software HLM 7 (Raudenbush, Bryk, & Congdon, 2013).

Model 1 includes a random intercept for each district and represents the fully pooled HETOP model structure in each state-subject dataset. In this model, ν_{00} is the variance between districts in γ_{grt} values, while ω_1^2 is the variance within districts across grades and years. Model 2 includes random grade and year linear trends for each district and represents the linear trend pooled HETOP model structure in each state-subject dataset; ω_2^2 is the unexplained within-district variance in γ_{grt} values and the elements of $\mathbf{T}$ indicate the variance of district-specific intercepts and linear trends. We can use estimates from Model 1 to estimate the ratio $\rho = \hat{\nu}_{00}/(\hat{\nu}_{00} + \hat{\omega}_1^2)$ to quantify the proportion of variation in γ_{grt} that is between districts. We then calculate the quantity $\Delta_{12} = 1 - \hat{\omega}_2^2/\hat{\omega}_1^2$, the percent of unexplained, within-district variance in γ_{grt} values that can be explained by adding district-specific linear grade and year trends. Table A1 summarizes the average values across the 80 state-subject models. Each row represents 40 models estimated using ELA test score data and 40 models estimated using Math test score data. Interpretation of the results is presented in the main text.

References

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., DiSalvo, R., & Ho, A. D. (2018). *Stanford Education Data Archive: Technical documentation (Version 2.1)*. Retrieved from https://stacks.stanford.edu/file/druid:db586ns4974/SEDA_documentation_v21.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2013). *HLM 7.01 for Windows*. Skokie, IL: Scientific Software International, Inc.

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & DiSalvo, R. (2018). Stanford Education Data Archive (Version 2.1). Retrieved from https://cepa.stanford.edu/seda/overview

Table A1. Summary of HLM Model Estimates by Subject.

Statistic	ELA			Math		
	Min	Mean	Max	Min	Mean	Max
ω_1^2	0.0004	0.0018	0.0045	0.0008	0.0028	0.0064
ν_{00}	0.0015	0.0033	0.0058	0.0015	0.0049	0.0103
ρ	0.4072	0.6519	0.8814	0.4280	0.6374	0.8846
ω_2^2	0.0000	0.0012	0.0036	0.0005	0.0020	0.0051
τ_{00}	0.0015	0.0033	0.0059	0.0015	0.0049	0.0104
τ_{11}	0.0000	0.0001	0.0004	0.0001	0.0002	0.0004
τ_{22}	0.0000	0.0001	0.0002	0.0000	0.0001	0.0002
Δ_{12}	0.1544	0.3463	0.9997	0.2071	0.2935	0.4260

Note: This table summarizes estimates from Models 1 and 2 estimated in each state-subject dataset. ω_1^2 = residual variance in γ_{grt} values in Model 1; ν_{00} = between-district variance in γ_{grt} values in Model 1; $\rho = \nu_{00}/(\nu_{00} + \omega_1^2)$ is the percent of total variance between rather than within districts; ω_2^2 = residual variance in γ_{grt} values in Model 2; τ_{00} = between-district variance in γ_{grt} values in Model 2; τ_{11} and τ_{22} are between-district variances in grade and year trends, respectively; $\Delta_{12} = (\omega_1^2 - \omega_2^2)/\omega_1^2$ is the percent of unexplained variance in Model 1 that is explained by including linear grade and year trends in Model 2.