IDENTIFICATION AND ESTIMATION OF BOUNDS ON SCHOOL PERFORMANCE MEASURES: A NONPARAMETRIC ANALYSIS OF A MIXTURE MODEL WITH VERIFICATION

By Jeff Dominitz* and Robert P. Sherman† [1]

*Carnegie Mellon University †California Institute of Technology

April, 2005

Abstract

This paper identifies and nonparametrically estimates sharp bounds on school performance measures based on test scores that may not be valid for all students. A mixture model with verification is developed to handle this problem. This is a mixture model for data that can be partitioned into two sets, one of which (the so-called verified set) is more likely to be from the distribution of interest than the other. An administrative classification of each student as English proficient or limited English proficient determines these sets. An analysis of performance measures for some California public schools reveals how verification information and plausible monotonicity restrictions can bound the range of disagreement about school performance based on observed scores.

1. Introduction

School administrators have long used standardized test scores to assess the performance of schools and school districts. The *No Child Left Behind Act* (NCLB, P.L. 107-119, H.R.1) now

[1]We thank William A. Bibbiani, former Director of Research, Evaluation, and Testing for the Pasadena Unified School District, for providing the data analyzed in this paper and for sharing his knowledge and insights. We also thank Valentina Bali for help in obtaining the data. Finally, we thank Arie Kapteyn, Chuck Manski, and the seminar participants at Carnegie Mellon, Johns Hopkins, Northwestern, Princeton, Rutgers, Tilburg, and UCLA for helpful comments and suggestions.

federally mandates this practice. This law requires that each state set challenging performance standards in reading and math, and annually test every child's progress in these subjects in third through eighth grades and at least once during 10th through 12th grades. By school year 2007-2008, testing in science will also be required. Schools and school districts failing to meet these statewide performance goals can face serious consequences including loss of jobs, paid relocation of students, loss of accreditation, and even takeover by private companies.

Fair and effective implementation of NCLB will require proper evaluation of school performance measures. Kane and Staiger (2002) show that questions of statistical precision must be considered when evaluating performance criteria based on student test scores. In this paper, we are principally concerned with the more fundamental issue of identification that arises when a significant fraction of test scores may be invalid. In particular, we are concerned about the validity of the scores of students who are classified as limited English proficient (LEP), but who must, nonetheless, take tests written and administered in English.

Concerns about the validity of test scores of LEP students have been the subject of national attention. Last year, U.S. Secretary of Education Rod Paige announced revised NCLB procedures for handling such scores (see "Secretary Paige Announces New Policies to Help English Language Learners," U.S. Department of Education press release, 2/19/04). For example, LEP students in the first year of U.S. enrollment are now allowed to "take the mathematics assessment, with accommodations as appropriate". Such accommodations could include tests administered in native languages, tests administered in simplified English, and the provision of extra time for completion of the test.[2]

The validity of LEP scores has also been the focus of much debate and litigation in California,

[2]Abedi (2002) reviews evidence on the effects of accommodations on LEP student test scores.

where each year public school students in grades 2 through 11 take the Stanford 9 standardized tests in reading, mathematics, language, science, and social science. Of the over 4 million public school students in California who take the tests each year, approximately one-fourth are classified by school officials as LEP. The remaining are classified as EP, or English proficient.

Prior to 1998, the scores of LEP students in California were reported separately from the scores of EP students, and only the latter scores were used to evaluate school performance. In addition, school comparisons were only made within school districts. The policy changed in 1998, when English immersion programs were mandated statewide in accordance with Proposition 227. Since then, state law has required that the scores of all students, regardless of English proficiency, be used to evaluate school performance at all levels: district, county, and state. Moreover, the California tests are written and administered in English and do not explicitly allow for the types of accomodations mentioned previously with respect to the implementation of NCLB.

These requirements have been the source of controversy among school and public officials. According to one account of litigation to prevent the release of scores, school officials supported the practice of using only the scores of EP students to assess school performance, arguing that tests taken by students "in a language they do not understand would unfairly make urban school districts with large numbers of immigrants look worse than they are."[3] On the other hand, the reporting of all scores was supported by public officials, led by then-Governor Pete Wilson, who stated: "It is deplorable that one judge and a number of education bureaucrats are so fearful of accountability for how poorly education is being provided in parts of our state." See "Judge Blocks Release of Test Scores", Los Angeles Times, 6/28/98.

In evaluating student test scores, it is important to distinguish between being classified as EP

[3]See also Thompson et al. (2002) for a critique of the validity of Stanford 9 tests scores of LEP students.

and being truly English proficient. We say that students are truly English proficient if their English language scores equal their native language scores, where native language scores are scores they would get if they took the tests in their native languages. We say that native language scores are valid scores. Certainly, native English speakers are truly English proficient in the sense just defined. However, as we shall see, some students classified as EP may not be truly English proficient. Conversely, some students classified as LEP may be truly English proficient.[4]

We focus on two distributions of interest: the distribution of valid scores for truly English proficient students and the distribution of valid scores for all students. The former distribution is implicitly favored by the school officials mentioned above, whereas the latter is implicitly favored by the public officials.

We do not take a stand on which distribution is the more appropriate one to study. Rather, we seek to bound the range of disagreement about important characteristics (e.g., mean valid test scores) of either distribution by clarifying the assumptions needed to make valid inferences based on the empirical evidence. Central to our methods is the search for plausible restrictions that tighten the bounds and so strengthen the conclusions that can be drawn. In our analysis of Stanford 9 scores, we utilize three important types of information: verification information in an EP/LEP administrative classification, expert assessment of bounds on misclassification probabilities, and various plausible monotonicity restrictions. In fact, we show, among other things, that under certain sets of plausible restrictions, both the mean valid score of truly English proficient students and the mean valid score of all students are bounded below by the mean score of all students and

[4]In California, students are classified as LEP if they fail to pass an English proficiency test taken after an initial enrollment survey indicates that a language other than English is spoken at home. Until recently, individual transitions from LEP to EP status would occur only rarely from passage of retests taken at the request of the student or a parent. For this reason and others, the administrative classification may not accurately reflect a student's ability to obtain a valid score on a given test.

above by the mean score of EP students. In other words, the polar measures endorsed by the public and school officials are the bounds for these distribution characteristics. In this way, our results provide a step in the direction of resolving the contentious issue of how to make valid inferences from observed test scores.

Empirical researchers are often faced with the type of situation described above, where a flawed data generating mechanism produces data that are not always representative of a population of interest. Often, such data can be viewed as observations from a mixture model. According to such a model, each observation is generated from either a distribution of interest, say F, or another, potentially spurious, distribution. Unless untestable assumptions about the data generating process hold, it is not possible to identify characteristics of F such as moments, probabilities, and quantiles. However, given a lower bound on the probability of generation from F, Horowitz and Manski (1995) identify and nonparametrically estimate sharp bounds on such characteristics.

Sometimes there is more information than simply a lower bound on the probability of generation from F. Sometimes data generated from a mixture model can be partitioned into two sets, and it is reasonable to assume that observations from one set are more likely to be from F than observations from the other set. We call the former set the *verified* set, and say that data generated in this way come from a mixture model with verification.

It is natural to model student test scores with a mixture model where a score is valid when a student is truly English proficient, and invalid otherwise. While the observed EP classification is not a perfect indicator of being truly English proficient, it is reasonable to assume that students classified as EP are more likely to be truly English proficient than students classified as LEP. Thus, we have a mixture model with verification, where the EP classification acts as an imperfect verification indicator. As mentioned above, we are interested in the distribution of valid scores for truly English

proficient students, as well as the distribution of valid scores for all students. As we will show, verification information can be used to develop sharp bounds on characteristics of both of these two distributions that can be substantially tighter than the corresponding bounds of Horowitz and Manski (1995). We also show how to tighten bounds by imposing natural monotonicity conditions.

We construct sample analogs of the population bounds for characteristics of these distributions, and show that they are $\sqrt{n}$-consistent and asymptotically normally distributed, where n is the sample size. Extensions to allow for discrete covariates are immediate. The establishment of the limiting normal distribution for the sample bounds for characteristics of the distribution of scores conditional on being truly English proficient depends on a parametrization that induces convexity and concavity in the functions used to define these bounds. This makes short work of a problem that would otherwise be difficult to solve. Convexity and concavity in the sample functions can also be exploited to significantly reduce computations in various settings, as will be explained.

Mixture models with verification apply to a wide range of other interesting data problems. Consider, for example, self-reported data on income. Self-reports by some respondents can be verified by administrative records, as was done in the Survey of Income and Program Participation Record Check Study (U.S. Bureau of the Census, 1998, Section 6.3.4). Alternatively, some respondents may report that they consulted pay stubs when reporting income, as is routinely done in the Family Expenditure Survey conducted by the United Kingdom's Office for National Statistics (http://www.mimas.ac.uk/surveys/fes/). In both cases, it is reasonable to assume that the verified data are more likely to be from the distribution of interest than the unverified data. Some models of survey nonresponse (Horowitz and Manski, 1998) and treatment effects (Molinari, 2002) can also be viewed as mixture models with verification. See also Dominitz and Sherman (2004) for a related analysis of environmental pollutant data previously studied by Lambert and Tierney (1997).

The rest of the paper is organized as follows. In Section 2, we formally define the mixture model with verification, and present and discuss the assumptions of the model as it applies to test score data. In Section 3, we derive sharp bounds on characteristics of the distributions of interest. Section 4 shows how to tighten these bounds by imposing various monotonicity conditions. Section 5 defines sample analogs of the population bounds derived in Sections 3 and 4. We also establish convexity and concavity of the functions defining the sample bounds for characteristics of one distribution, and discuss the consequent computational and asymptotic benefits. In Section 6, we estimate bounds on measures of achievement in math based on test scores of ninth graders from the five high schools in the Pasadena Unified School District in California. We compare our verification bounds to the bounds of Horowitz and Manski (1995). Unlike the bounds of Horowitz and Manski (1995), the verification bounds for truly English proficient students are tight enough to yield informative comparisons of performance across schools. When plausible monotonicity conditions are incorporated, we also make informative comparisons of performance for all students, and, as mentioned above, show that various performance measures of interest are bounded by those endorsed by opposing school and public officials, providing a basis for resolving their dispute. Section 7 summarizes. Proofs of some theorems are given in an appendix.

2. Mixture Models with Verification

In this section, we formally define a mixture model with verification in the context of school performance measures based on student test scores. However, it should be borne in mind that the model, as well as the subsequent bounds and estimation procedures developed, are more generally applicable. We also define the distributional characteristics of interest and state and discuss the basic assumptions under which we derive sharp bounds on these characteristics.

Recall that native language scores are scores students would get if they took the Stanford 9 tests in their native languages. We view these scores as valid test scores. For concreteness, we focus on math scores. Define Y_1 to be the student's native language math score. We say students are truly English proficient in math if their English language math scores equal their native language math scores. Define $Z = 1$ if a student is truly English proficient and $Z = 0$ otherwise. Define Y_0 to be the score a student obtains on the Stanford 9 math test when $Z = 0$. We observe Y, a mixture of Y_1 and Y_0. That is,

$$Y = Y_1 Z + Y_0(1 - Z)\,. \tag{1}$$

Finally, define the verification indicator $V = 1$ if a student is classified as EP and $V = 0$ if the student is classified as LEP. As mentioned in the introduction, it is important to realize that V is subject to the following misclassification errors: (i) $V = 1$ and $Z = 0$ and (ii) $V = 0$ and $Z = 1$. For example, during registration at a California public school, parents must fill out a home language survey indicating whether or not a language other than English is spoken at home. Only students whose parents indicate that a non-English language is spoken at home are tested to see if they merit the LEP designation. All others are classified as EP. Parents who speak a non-English language at home may not indicate this fact on the survey. In addition, judgement errors of type (i) can occur in testing for limited English proficiency. According to William Bibbiani, former Director of Research and Testing for the Pasadena Unified School District (PUSD), it is reasonable to assume that no more than 5% of PUSD students are designated as EP when they should be classified as LEP. Error (ii) may be more common, and can occur, for example, when children who are initially classified as LEP become truly English proficient, but, for various reasons, are not reclassified. According to Bibbiani, at least one-third of the PUSD high school students who are classified as LEP should be reclassified as EP, yet such reclassifications were rare during the time period leading

up to the tests we analyze.[5] Judgement errors in testing for limited English proficiency can also contribute to error (ii).

To summarize, in this mixture model with verification, each member of the sampling distribution is characterized by a vector (Y, V, Z, Y_1, Y_0), where Y and V are observed.[6]

Let M denote a known, real-valued function on the support of Y_1. Assuming a mixture model with verification, we develop and estimate sharp bounds on $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$ for $M(t) = t$ and $M(t) = \{t \geq 50\}$. Thus, we are interested in the mean and the probability of exceeding 50 for each of the distributions of interest. The threshold score of 50 corresponds to the national median score for the Stanford 9 tests, and the proportion of students exceeding 50 in a given test is reported in the School Accountability Report Card issued for each school.

We now state and discuss the basic assumptions we make to establish sharp bounds on the quantities $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$.

A1. The data are draws from a mixture model with verification where $I\!P\{V = 1\} > 0$.

A2. $I\!P\{Z = 1 \mid V = 1\} \geq I\!P\{Z = 1 \mid V = 0\}$.

A3. There exists a known constant $d_0 \geq 0$ for which $I\!P\{Z = 1 \mid V = 0\} \geq d_0$.

A4. There exists a known constant $d_1 > 0$ for which $I\!P\{Z = 1 \mid V = 1\} \geq d_1$.

Assumptions A1 and A2 are plausible assumptions for the PUSD data analyzed in Section 6.

[5] An annual English proficiency test, the California English Language Development Test (CELDT), was initiated during the 2001-02 school year. The results of this test, required for "students whose primary language is other than English," could be used for reclassification. See http://www.cde.ca.gov/ta/tg/el/. According to Bibbiani, prior to CELDT, retesting for possible reclassification was rare and would occur only in response to individual requests.

[6] We abstract from two other sources of identification problems in school performance assessment: censored test scores due to absenteeism and measurement errors due to guessing. The NCLB calls for a censoring rate not to exceed 5 percent of enrolled students. If this rate is known, it is easy to incorporate into the verification bounds, which would become wider. As for measurement errors, the bounds we derive below hold when the errors for truly English proficient students satisfy a plausible unbiasedness condition. To account for censoring and measurement errors would entail additional notation without adding much to the substance of the analysis. We therefore focus on the test scores of those who take the test.

Note that assumption A2 states that students classified as EP are more likely to be truly English proficient than students classified as LEP. Assumptions A3 and A4 say that there exist known lower bounds on the probability of being truly English proficient given that one is classified as LEP or EP, respectively. As discussed above, it may be reasonable to take $d_0 = .33$ and $d_1 = .95$ for the PUSD data. Note, however, that officials on both sides of the litigation mentioned in the introduction endorsed the use of EP students scores for school assessment, and therefore may be prepared to assume that $d_1 = 1$.

Assumptions A1 and A2 are the main assumptions that distinguish mixture models with verification from the mixture models studied in Horowitz and Manski (1995) (hereafter, HM95), where verification information is not available. The key aspect of A1 is that the verification indicator, V, is observed for each member of the sample. This extra information can be used to develop bounds that are tighter than the corresponding HM95 bounds.

In order to construct estimable HM95 bounds on characteristics of either distribution of interest, a positive lower bound on $I\!P\{Z = 1\}$ must be known or estimable from the data. It follows from assumptions A1, A3, and A4 that $I\!P\{Z = 1\} \geq d_1 I\!P\{V = 1\} + d_0 I\!P\{V = 0\} > 0$. Thus, if A3 and A4 hold and $I\!P\{V = 1\}$ is known or estimable, then it is possible to construct estimable HM95 bounds. However, the verification bounds developed in this paper exploit the verification status of individual observations, whereas the HM95 bounds do not. Because of this, the verification bounds are always contained in the HM95 bounds.

3. Sharp Bounds

In this section, we derive sharp bounds on $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$ for $M(t) = t$ and $M(t) = \{t \geq 50\}$ under assumptions A1 through A4 described in the last section. We also derive the corresponding HM95 bounds and compare them to the verification bounds.

We begin by noting that the observed scores variable Y in (1) is discrete, taking integer values between 1 and 99. Define a continuous analogue $\mathcal{Y} = Y + U$ where U is distributed uniformly on $(-1, 0]$. We introduce $\mathcal{Y}$ for notational convenience. Bounds on all quantities of interest mentioned above can be easily expressed in terms of various quantiles of $\mathcal{Y}$.

We now develop sharp bounds on $I\!E[M(Y_1) \mid Z = 1]$ under assumptions A1 through A4.

Note that $I\!E[M(Y_1) \mid Z = 1] = I\!E[M(Y) \mid Z = 1]$. Write p_1 for $I\!P\{V = 1 \mid Z = 1\}$. Then

$$I\!E[M(Y_1) \mid Z = 1] = I\!E[M(Y) \mid Z = 1, V = 1]p_1 + I\!E[M(Y) \mid Z = 1, V = 0](1 - p_1)\,. \tag{2}$$

Write δ_1 for $I\!P\{Z = 1 \mid V = 1\}$ and Q_i for the quantile function of $\mathcal{Y}$ given $V = i$, $i = 0, 1$. By Proposition 4 in HM95, the interval

$$[I\!E[M(Y) \mid \mathcal{Y} \leq Q_1(\delta_1), V = 1],\ I\!E[M(Y) \mid \mathcal{Y} > Q_1(1 - \delta_1), V = 1]] \tag{3}$$

contains $I\!E[M(Y) \mid Z = 1, V = 1]$.

Define $v_1 = I\!P\{V = 1\}$. Bayes' Rule implies that $I\!P\{Z = 1 \mid V = 0\} = [(1-p_1)v_1\delta_1]/[p_1(1-v_1)]$. Write $\pi(p_1, \delta_1)$ for this quantity. By Proposition 4 in HM95, the interval

$$[I\!E[M(Y) \mid \mathcal{Y} \leq Q_0(\pi(p_1, \delta_1)), V = 0],\ I\!E[M(Y) \mid \mathcal{Y} > Q_0(1 - \pi(p_1, \delta_1)), V = 0]] \tag{4}$$

contains $I\!E[M(Y) \mid Z = 1, V = 0]$. Combine (2), (3), and (4) to bound $I\!E[M(Y_1) \mid Z = 1]$. Note, however, that these bounds are infeasible since p_1 and δ_1 are unknown.

To develop feasible bounds, assume that $d_1 \geq d_0$ so that A4 is binding. Thus, $\delta_1 \in [d_1, 1]$.[7]

[7] If $d_1 < d_0$, then A4 is not binding since A2 and A3 imply that $\delta_1 \in [d_0, 1]$. We assume $d_1 \geq d_0$, since this is the more interesting case. A similar argument works for the case $d_1 < d_0$.

Apply Bayes' rule once again to get $p_1 = \delta_1 v_1/[\delta_1 v_1 + \pi(p_1, \delta_1)(1 - v_1)]$. Apply A2 and A3 to get $v_1 \leq p_1 \leq \delta_1 v_1/[\delta_1 v_1 + d_0(1 - v_1)]$. For each $\delta \in [d_1, 1]$, define $\sigma(\delta) = \delta v_1/[\delta v_1 + d_0(1 - v_1)]$. For each $\delta \in [d_1, 1]$ and $p \in [v_1, \sigma(\delta)]$ define $\pi(p, \delta) = (1 - p)v_1\delta/[p(1 - v_1)]$. Define lower and upper bound functions

$$\begin{aligned} L(p,\delta) &= p\mathbb{E}[M(Y) \mid \mathcal{Y} \leq Q_1(\delta), V = 1] + (1-p)\mathbb{E}[M(Y) \mid \mathcal{Y} \leq Q_0(\pi(p,\delta)), V = 0] \\ U(p,\delta) &= p\mathbb{E}[M(Y) \mid \mathcal{Y} > Q_1(1-\delta), V = 1] + (1-p)\mathbb{E}[M(Y) \mid \mathcal{Y} > Q_0(1-\pi(p,\delta)), V = 0]\,. \end{aligned}$$

Note that $\sigma(\delta)$ is strictly increasing on $[d_1, 1]$. For each $p \in (\sigma(d_1), \sigma(1)]$ define the inverse function $\sigma^{-1}(p) = [p(1 - v_1)d_0]/[(1 - p)v_1]$. Define $\delta(p) = d_1\{v_1 \leq p \leq \sigma(d_1)\} + \sigma^{-1}(p)\{\sigma(d_1) < p \leq \sigma(1)\}$. Note that for each $p \in [v_1, \sigma(1)]$, the function $L(p, \delta)$ is increasing in δ, and so is minimized over $\delta \in [\delta(p), 1]$ at $\delta = \delta(p)$. Similarly, for each $p \in [v_1, \sigma(1)]$, the function $U(p, \delta)$ is decreasing in δ, and so is maximized over $\delta \in [\delta(p), 1]$ at $\delta = \delta(p)$. This leads to the following result.

THEOREM 1. *If A1 through A4 hold, then* $\lambda_1 \leq \mathbb{E}[M(Y_1) \mid Z = 1] \leq u_1$, *where*

$$\begin{aligned} \lambda_1 &= \inf_{p \in [v_1, \sigma(1)]} L(p, \delta(p)) \\ u_1 &= \sup_{p \in [v_1, \sigma(1)]} U(p, \delta(p))\,. \end{aligned}$$

Moreover, these bounds are sharp.

REMARK 1. Under assumptions A1, A3, and A4, $\mathbb{P}\{Z = 1\} \geq d_1 v_1 + d_0(1 - v_1)$. Since v_1 is estimable from the data, this lower bound on $\mathbb{P}\{Z = 1\}$ is sufficient to construct estimable HM95 bounds on $\mathbb{E}[M(Y_1) \mid Z = 1]$. Write Q for the unconditional quantile function for $\mathcal{Y}$. By

Proposition 4 in HM95, the HM95 bounds on $I\!E[M(Y_1) \mid Z = 1]$ are the endpoints of the interval

$$[I\!E[M(Y) \mid \mathcal{Y} \leq Q(d_1 v_1 + d_0(1 - v_1))], I\!E[M(Y) \mid \mathcal{Y} > Q(1 - d_1 v_1 - d_0(1 - v_1))]] \ .$$

It is easy to show that this interval must contain $[\lambda_1, u_1]$. It is interesting to note that if assumption A2 is dropped from the conditions of Theorem 1, then sharp bounds weakly wider than the verification bounds in Theorem 1 and weakly narrower than the HM95 bounds on $I\!E[M(Y_1) \mid Z = 1]$ can be obtained by an argument similar to that preceding the statement of Theorem 1.

It is also interesting to note that if it is reasonable to make the additional assumption that Y_1 is independent of Z (or the weaker mean independence assumption that $I\!E[M(Y_1) \mid Z = 1] = I\!EM(Y_1)$), then the bounds in Theorem 1 are sharp bounds on $I\!EM(Y_1)$. This independence assumption is among the defining properties of contaminated mixture models (see HM95, for example). As mentioned previously, for the test scores application analyzed in Section 6, this assumption is implausible: valid test scores (Y_1) are not likely to be independent of true English proficiency ($Z = 1$). However, if, in addition to assumptions A1 through A4, this type of independence assumption holds, then Theorem 1 provides sharp bounds on $I\!EM(Y_1)$ for all contaminated mixture model with verification applications.

Next, we develop sharp bounds on $I\!EM(Y_1)$ under assumptions A1 through A4.

Let M be a known, real-valued function on $I\!R$, and suppose the support of $M(Y_1)$ is contained in the closed interval $[a, b]$, where a and b are known. In the scores application, when $M(t) = t$, then $[a, b] = [1, 99]$; when $M(t) = \{t \geq 50\}$, then $[a, b] = [0, 1]$.

THEOREM 2. *If A1 through A4 hold, then* $\lambda_2 \leq I\!E M(Y_1) \leq u_2$, *where*

$$\begin{aligned} \lambda_2 &= L(\sigma(d_1), d_1)[d_1 v_1 + d_0(1 - v_1)] + a[1 - d_1 v_1 - d_0(1 - v_1)] \\ u_2 &= U(\sigma(d_1), d_1)[d_1 v_1 + d_0(1 - v_1)] + b[1 - d_1 v_1 - d_0(1 - v_1)] \,. \end{aligned}$$

Moreover, these bounds are sharp.

REMARK 2. Using the results from Remark 1, it is easy to show that under assumptions A1 through A4, the HM95 bounds on $I\!E M(Y_1)$ are given by

$$\begin{aligned} &I\!E[M(Y) \mid \mathcal{Y} \leq Q(d_1 v_1 + d_0(1 - v_1))][d_1 v_1 + d_0(1 - v_1)] + a[1 - d_1 v_1 - d_0(1 - v_1)] \\ &I\!E[M(Y) \mid \mathcal{Y} > Q(1 - d_1 v_1 - d_0(1 - v_1))][d_1 v_1 + d_0(1 - v_1)] + b[1 - d_1 v_1 - d_0(1 - v_1)] \,. \end{aligned}$$

The interval with these endpoints must contain $[\lambda_2, u_2]$. Also, if A2 is dropped, then sharp bounds weakly wider than the verification bounds in Theorem 2 and weakly narrower than the HM95 bounds on $I\!E M(Y_1)$ can be obtained.

4. Sharp Bounds under Monotonicity

In this section, we derive sharp bounds on $I\!E[M(Y_1) \mid Z = 1]$ and $I\!E M(Y_1)$ under additional monotonicity assumptions that can considerably tighten the bounds derived in Section 3. We also derive the corresponding HM95 bounds and compare them to the verification bounds. We restrict attention to a set of restrictions that make sense in the test scores application. In other applications, other plausible monotonicity restrictions may be adopted to tighten bounds (see, for example, the analyses of environmental pollutant data in Lambert and Tierney, 1997, and Dominitz and Sherman, 2004).

We consider the following monotonicity assumptions:

A5. $I\!E[M(Y_1) \mid Z = 1] \geq I\!E[M(Y_0) \mid Z = 0]$.

A6. $I\!E[M(Y_1) \mid Z = 1, V = 1] \geq I\!E[M(Y_1) \mid Z = 1, V = 0]$.

A7. $I\!E[M(Y_1) \mid Z = 0] \geq I\!E[M(Y_0) \mid Z = 0]$.

A8. $I\!E[M(Y_1) \mid Z = 1] \geq I\!E[M(Y_1) \mid Z = 0]$.

Suppose $M(t) = t$. In this case, assumption A5 says that the average observed score of students who are truly English proficient is at least as high as the average observed score of students who are not truly English proficient. If the EP/LEP classification were a perfect indicator of true proficiency, then this condition would be directly testable and, in fact, would be found to hold in our math scores data. Misclassification, of course, complicates the matter, but the restriction seems natural enough, given the obstacles to learning and to demonstrating mathematical achievement on these tests faced by students who are not truly English proficient. Assumption A6 says that the average observed score of students who are truly English proficient and are classified as EP is at least as high as the average observed score of students who are truly English proficient but are classified as LEP. To the extent that expected mathemental achievement is positively related to verification, this assumption also seems quite plausible. Adding assumption A5 can raise the lower bound, while adding A6 can lower the upper bound, on $I\!E[M(Y_1) \mid Z = 1]$.

The next pair of assumptions is perhaps more controversial than the previous pair. Assumption A7 says that for students who are not truly English proficient, their average valid score is at least as high as their average invalid score. Assumption A8 says that the average valid score of students who are truly English proficient is at least as high as the average valid score of students who are not truly English proficient. In fact, evidence exists that instruction in English can lead LEP students

to perform better on a test administered in English, because important terms are only familar in English (Abedi, 2002). This finding could violate A7 (e.g., if $V = 0$ implies $Z = 0$), but redefining valid scores would take care of this problem (see Remark 4, below). It is also easy to imagine the presence of well-educated immigrant groups whose performance on a native language math test would exceed the average performance of their truly English proficient counterparts, in violation of A8. By and large, however, we believe that A8 is a plausible restriction for the math scores we analyze. Adding assumption A7 can raise the lower bound, while adding A8 can lower the upper bound, on $I\!EM(Y_1)$.

THEOREM 3. *If A1 through A6 hold, then* $\lambda_3 \leq I\!E[M(Y_1) \mid Z = 1] \leq u_3$, *where*

$$\begin{aligned} \lambda_3 &= I\!EM(Y) \\ u_3 &= \min\{I\!E[M(Y) \mid \mathcal{Y} > Q_1(1 - d_1), V = 1], u_1\}\,. \end{aligned}$$

If A1 through A4, A7, and A8 hold, then $\lambda_3 \leq I\!EM(Y_1) \leq u_3$. *Moreover, these bounds are sharp.*

REMARK 3. It is easy to show that under the corresponding assumptions in Theorem 3, the HM95 bounds on both $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$ are the endpoints of the interval

$$[I\!EM(Y), I\!E[M(Y) \mid \mathcal{Y} > Q(1 - d_1 v_1 - d_0(1 - v_1))]]\,.$$

Note that the HM95 upper bounds on $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$ under monotonicity are the same as the HM95 upper bound on $I\!E[M(Y_1) \mid Z = 1]$ under assumptions A1 through A4 (see Remark 1 after the statement of Theorem 1 in Section 3). As before, the HM95 upper bound under monotonicity must be at least as large as u_3.

It is interesting to note that in the special case of A3 and A4 when $d_0 = 0$ and $d_1 = 1$, $u_3 = I\!E[M(Y) \mid V = 1]$. Recall from the introduction that the lower bound, $\lambda_3 = I\!EM(Y)$, was viewed as the better measure of educational achievement by the political officials, while the upper bound, $I\!E[M(Y) \mid V = 1]$, was preferred by the school officials.

REMARK 4. Under certain circumstances, it may be reasonable to (i) say that a student is truly English proficient ($Z = 1$) if the student's English language score is at least as high as the student's native language score and (ii) define a valid score (Y_1) to be the greater of a student's English language score and the student's native language score. This would allow for the possibility that some students who are not native English speakers may, over time, acquire English language skills that exceed their native language skills. If definitions (i) and (ii) are adopted, then assumption A7 is automatically satisfied, and so λ_2 in Theorem 2 is equal to λ_3 in Theorem 3.

REMARK 5. It is possible to impose tighter restrictions on the correspondence between Y_1 and Y_0 based on additional information that would serve to further tighten the bounds of interest or even obtain point identification. For example, during the first year of enrollment, LEP students who speak Spanish at home are required to take the Spanish Assessment of Basic Education, 2nd edition (SABE/2). This assessment includes a math test, for which "reference group" percentile ranks are reported, where the reference group is "Spanish-speaking students enrolled in bilingual education programs across the nation" (CTB MacMillan/McGraw-Hill, 1994). Suppose there exists a one-to-one mapping from Stanford 9 math scores to SABE/2 math scores. In addition, suppose the EP/LEP classification is a perfect indicator of proficiency ($Z = V$) and the native language is Spanish for all LEP students. Applying the mapping to all observed LEP scores would point identify $I\!E[M(Y_1) \mid Z = 0]$, leading to point indentification of $I\!EM(Y_1)$. If this type of evidence

is used instead to obtain weaker restrictions on the correspondence between Y_1 and Y_0, such as a bound on the size of the discrepancy between valid and observed scores when $Z = 0$, then sharp bounds on the performance measures of interest could be derived in a manner similar to that presented here.

REMARK 6. In addition to combined reports, the NCLB requires that schools report scores separately by EP/LEP classification. In California, the School Accountability Report Cards may also include such a comparison. With monotonicity restrictions, one may obtain informative bounds on valid performance measures for each group. Consider, in particular, LEP students. With $d_0 = 0$ and no monotonicity restrictions, informative bounds for LEP students cannot be obtained. Under A7, the observed LEP student scores determine a lower bound on valid performance. Under the monotonicity restriction $I\!E[M(Y_1) \mid V = 1] \geq I\!E[M(Y_1) \mid V = 0]$, the observed scores of EP students determine an upper bound.

5. ESTIMATION

We begin by developing sample analogs of the population bounds on $I\!E[M(Y_1) \mid Z = 1]$ given in Theorem 1 in Section 3. We establish convexity and concavity of the sample functions defining these bounds, and discuss the consequent computational and asymptotic benefits.

Let $(Y_i, V_i, Z_i, Y_{1i}, Y_{0i})$, $i = 1, \ldots, n$, be independent draws from from the mixture model with verification defined in Section 2. Define $n_1 = \sum_{i=1}^{n} V_i$, $n_0 = n - n_1$, and $\hat{v}_1 = n_1/n$. Recall the definition of $\mathcal{Y}$ given at the beginning of Section 3. Define $\mathcal{Y}_i = Y_i + U_i$ where the U_i's are independent $U(-1, 0]$ random variables. Let $\hat{Q}_i$ denote the empirical quantile function of $\mathcal{Y}$ given $V = i$, $i = 0, 1$. For each $\delta \in [d_1, 1]$, define $\hat{\sigma}(\delta) = \delta\hat{v}_1/[\delta\hat{v}_1 + d_0(1 - \hat{v}_1)]$. For each $\delta \in [d_1, 1]$ and $p \in [\hat{v}_1, \hat{\sigma}(\delta)]$ define $\hat{\pi}(p, \delta) = (1 - p)\hat{v}_1\delta/[p(1 - \hat{v}_1)]$.

Define the sample lower and upper bound functions

$$\begin{aligned}
\hat{L}(p,\delta) &= p\sum_{i=1}^{n} M(Y_i)V_i\{\mathcal{Y}_i \le \hat{Q}_1(\delta)\}/[\delta n_1] \\
&+ (1-p)\sum_{i=1}^{n} M(Y_i)(1-V_i)\{\mathcal{Y}_i \le \hat{Q}_0(\hat{\pi}(p,\delta))\}/[\hat{\pi}(p,\delta)n_0] \\
\hat{U}(p,\delta) &= p\sum_{i=1}^{n} M(Y_i)V_i\{\mathcal{Y}_i > \hat{Q}_1(1-\delta)\}/[\delta n_1] \\
&+ (1-p)\sum_{i=1}^{n} M(Y_i)(1-V_i)\{\mathcal{Y}_i > \hat{Q}_0(1-\hat{\pi}(p,\delta))\}/[\hat{\pi}(p,\delta)n_0]\,.
\end{aligned}$$

For each $p \in (\hat{\sigma}(d_1), \hat{\sigma}(1)]$ define the inverse function $\hat{\sigma}^{-1}(p) = [p(1-\hat{v}_1)d_0]/[(1-p)\hat{v}_1]$. Define $\hat{\delta}(p) = d_1\{\hat{v}_1 \le p \le \hat{\sigma}(d_1)\} + \hat{\sigma}^{-1}(p)\{\hat{\sigma}(d_1) < p \le \hat{\sigma}(1)\}$. Finally, define the extreme value estimators

$$\begin{aligned}
\hat{\lambda}_1 &= \inf_{p\in[\hat{v}_1,\hat{\sigma}(1)]} \hat{L}(p,\hat{\delta}(p)) \\
\hat{u}_1 &= \sup_{p\in[\hat{v}_1,\hat{\sigma}(1)]} \hat{U}(p,\hat{\delta}(p))\,.
\end{aligned}$$

One can show that $\hat{L}(p,\hat{\delta}(p))$ is a piecewise linear convex function and $\hat{U}(p,\hat{\delta}(p))$ is a piecewise linear concave function for $p \in [\hat{v}_1, \hat{\sigma}(1)]$. This result holds whether Y is discrete or continuous, and for any feasible values of d_0 and d_1. Below, we state the result for a discrete Y taking values $y_1 < y_2 < \cdots < y_m$. (For the scores application, we take $m = 99$ and $y_k = k$, $k = 1, 2, \ldots, 99$.) We prove the result in the appendix, where, for simplicity, we treat the special case $d_0 = 0$ and $d_1 = 1$, so that $\hat{\sigma}(1) = 1$ and $\hat{\delta}(p) = 1\{\hat{v}_1 \le p \le 1\}$.

THEOREM 4. *Suppose Y is a discrete random variable, taking values $y_1 < y_2 < \cdots < y_m$. Then $\hat{L}(p,\hat{\delta}(p))$ is a piecewise linear convex function and $\hat{U}(p,\hat{\delta}(p))$ is a piecewise linear concave function on $[\hat{v}_1, \hat{\sigma}(1)]$.*

Theorem 4 can be useful computationally. For example, in the scores application, $M(t) = t$ or $M(t) = \{t \geq 50\}$. From the proof of Theorem 4 in the appendix, we see that a search to find $\hat{\lambda}_1$ when $M(t) = t$ can be limited to the potential kink point ordinates $\hat{\pi}^{-1}(\hat{H}_0(k-1))$, $k = 1, 2, \ldots, 99$. Similarly, a search to find $\hat{u}_1$ can be limited to the points $\hat{\pi}^{-1}(1 - \hat{H}_0(k-1))$, $k = 1, 2, \ldots, 99$. Moreover, convexity and concavity make binary searches over these points possible. When $M(t) = \{t \geq 50\}$, $d_0 = 0$, and $d_1 = 1$, it is easy to show that $\hat{\lambda}_1 = \hat{L}(\hat{\pi}^{-1}(\hat{H}_0(49)), 1)$ and $\hat{u}_1 = \hat{U}(\hat{\pi}^{-1}(1 - \hat{H}_0(49)), 1)$. That is, only a single evaluation of the sample functions is needed to find the extreme value estimators.

The computational shortcuts described above can result in substantial savings in computation time when n is large, when bootstrap estimates of the distribution of the extreme value estimators are desired, or when it is of interest to compute the estimators for many discrete covariate values. For example, in the scores application, it may be of interest to compute the extreme value estimators conditional on student gender, parental marital status, or level of parental income.

Theorem 4 also confers asymptotic benefits. Asymptotic distribution theory not only enables asymptotic inference but also provides information about the quality of an estimator by revealing its exact rate of convergence. Deriving the asymptotic distribution of an extreme value estimator is, in general, a very difficult problem when the estimator is defined as an extreme value of a complicated, nonsmooth sample function (as is the case for $\hat{\lambda}_1$ and $\hat{u}_1$). However, when such an estimator is the infimum of a piecewise linear convex function or the supremum of a piecewise linear concave function, then it can be relatively straightforward to determine the limiting distribution. We illustrate this in the appendix by showing that $\hat{\lambda}_1$ is $\sqrt{n}$-consistent for λ_1 and asymptotically normally distributed.

6. Empirical Results

The state of California requires that all public school students in grades 2 through 11 take the Stanford 9 standardized tests in reading, mathematics, language, science, and social science. These tests are written and administered in English, with no explicit allowance for accomodations based on limited English proficiency. Prior to 1998, only the scores of EP students were used to evaluate school performance, and then, only within school districts. Since then, California law has required that the scores of all students, both EP and LEP, be used to evaluate educational performance at all levels: district, county, and state.

As discussed in Section 2, data of this sort can be modeled with a mixture model with verification, where the indicator of English proficiency (1 if a student is classified as EP and 0 if the student is classified as LEP) serves as a verification indicator. In this section, using the results developed in Sections 3, 4, and 5, we construct verification bounds on math scores of ninth-graders in the Pasadena Unified School District (PUSD) who took the Stanford 9 tests in the year 2000. We demonstrate the identifying power of the verification information by comparing these bounds to the bounds of Horowitz and Manski (1995). The verification bounds yield informative comparisons of performance measures across schools for truly English proficient students. Further, when we incorporate the plausible monotonicity restrictions defined in Section 4, we are also able to make informative comparisons of performance measures for all students.

The data we analyze are norm-referenced data. That is, a representative national sample of 9th grade students took the Stanford 9 math test and generated a distribution of test scores. The score of each 9th grade student in PUSD who took this math test in 2000 is compared to this national distribution of scores. A score of 50 corresponds to the mean and median of the national distribution of scores.[8]

[8]More specifically, each 9th grader in the national sample obtained a raw score on the math test. This is simply the number of correct answers on the test. Each correct answer was reweighted to account for differences in questions

Each PUSD high school reports results from the Stanford 9 tests in its annual School Accountability Report Card. A key statistic reported is the percentage of students who score at or above the 50th percentile of the national distribution of adjusted scores. For further details, see California Department of Education Score Explanations, at http://star.cde.ca.gov/star2000f/.

Table 1 describes the math test score data for ninth graders at each of the five high schools in the school district, as well as for PUSD as a whole. The share of students who are classified as LEP ranges from 15% to 17% for schools 64, 80, 82, and 84. For school 90, a continuation high school serving students with behavioral or academic problems, LEP students constitute 35% of the student body. Overall, 18% of PUSD high school students are officially classified as LEP. The mean math score is lower for the LEP students than EP students. While the difference ranges from about 9 to 14 points at the other schools, the difference in means at School 90 is under 4 points. The latter differential is difficult to interpret, given the process of selective enrollment in this continuation school.

The analysis presented in this section is based on the scores of students who took the test, and yet about 6% of enrolled students did not take the test. The proportion missing varies across schools and within schools by EP status. The bounds derived here can easily be revised to account for this censoring, but this would entail additional notation and revisions to the theorems in the previous sections without adding much to the analysis, apart from a widening of the bounds. Of course, this censoring problem is of substantive interest, especially with provisions of the NCLB

(e.g., more difficult questions received greater weight) and the new weights were added to form an adjusted score. The 1st through 99th percentiles of the adjusted scores were then computed. Each PUSD 9th grader who took the math test in 2000 generated a raw score that was adjusted and then mapped to the nearest percentile of the national distribution of adjusted scores. These percentiles were then mapped to the 1st through 99th percentiles of a normal distribution with mean 50 and standard deviation 21.06. These scores are called NCE, or normal curve equivalent, scores. (Note that an NCE score of 50 corresponds to the 50th percentile of the national distribution of adjusted scores.) These NCE scores were then rounded to the nearest integer. The rounded NCE scores of the PUSD students are the data we received from the Pasadena Unified School District and which we analyze in this section. These data take values in the set of integers from 1 to 99.

requiring at least a 95% testing rate. To illustrate the effect of censoring on the bounds, Figures A1 and A2 in the appendix display sharp bounds on mean math scores for all enrolled students, not just those students who took the test.

We begin by presenting estimated bounds on $I\!E[M(Y_1)|Z = 1]$ for $M(Y_1) = Y_1$ and $M(Y_1) = \{Y_1 \geq 50\}$. When $M(Y_1) = Y_1$, the quantity of interest is the mean valid math score of truly English proficient students. When $M(Y_1) = \{Y_1 \geq 50\}$, the quantity of interest is the probability that the valid scores of truly English proficient students exceed the national median score. The corresponding population bounds are given in Theorem 1 in Section 3 and Theorem 3 in Section 4.

The appendix describes in detail our estimates of the lower and upper bound functions $\hat{L}(p, \hat{\delta}(p))$ and $\hat{U}(p, \hat{\delta}(p))$. Examples of these functions, for students at all PUSD high schools, are plotted in Figures 1 and 2. Inspection of these figures reveals the relative sensitivity of the bound functions and, by implication, the bounds on $I\!E[M(Y_1) \mid Z = 1]$, to the values d_0 and d_1. In particular, the main effects of an increase in d_0 from 0.00 to 0.33 are to introduce a kink in each bound function and to restrict the feasible values of the share of truly English proficient students who are classified as EP. The effect on the bounds of interest is rather modest: the upper bound decreases a bit and the lower bound is unchanged. In contrast, what seems at face value to be a modest decrease in d_1 from 1.00 to 0.95 causes the bound functions to shift out considerably and yields relatively large increases in the width of the bounds. When the monotonicity restrictions are imposed, the figures show that the given variation in d_1 still considerably increases the upper bound whereas the given variation in d_0 has little or no impact on either bound. Thus, for this application, it appears crucial to obtain a credible and tight lower bound d_1 on the fraction of EP students who are truly English proficient. If d_1 is understated, then the estimated bounds on a given performance measure may be too wide to be useful. Conversely, if d_1 is overstated, then the estimated bounds may not cover

the given performance measure.

Tables 2 and 3 report point estimates and estimated confidence intervals for HM95 bounds, verification bounds, and verification bounds under monotonicity restrictions on both $I\!E[M(Y_1) \mid Z = 1]$ and $I\!EM(Y_1)$ when $M(Y_1) = Y_1$ and $M(Y_1) = \{Y_1 \geq 50\}$ for PUSD as well as for each school in the district. This is done for the cases $(d_1, d_0) = (1, 0)$, $(1, .33)$, and $(.95, .33)$. Figures 3 through 6 graphically present point estimates for easy comparison across types of bounds and schools. What becomes clear from these comparisons is the usefulness of verification information and monotonicity restrictions, as well as how this usefulness depends on the existence of tight but credible bounds on the share of truly English proficient students within the verified set and, to a lesser extent, the unverified set.

Consider first Table 2, which treats the case $M(Y_1) = Y_1$. We present joint 95% confidence interval estimates using the bootstrap. For each school and for PUSD as a whole, we draw 1000 bootstrap samples from the original sample, with each bootstrap sample size equal to the original sample size. Each bootstrap sample produces a bootstrap estimate of a given pair of population bounds. The estimates are denoted (λ^*, u^*). Following Horowitz and Manski (2000), we use the empirical distribution of (λ^*, u^*) pairs to find the smallest value z^* such that 95% of the pairs $(\lambda^* - z^*, u^* + z^*)$ contain (λ, u), the original point estimate of the given population bounds. The joint 95% confidence interval estimate that we report is then $(\lambda - z^*, u + z^*)$.[9]

Next, consider the point estimates of bounds depicted in Figures 3 thru 6. It is worth noting

[9]HM95 bounds always contain the corresponding verification bounds. However, the method of computing bootstrap confidence intervals described above can produce verification confidence intervals that contain the corresponding HM95 confidence intervals. For instance, consider the lower bound under monotonicity for PUSD, with $(d_1, d_0) = (.95, .33)$ (bottom-right corner of Table 2). HM95 and verification lower bounds are equal here. However, the verification upper bound is more variable (yet never larger) than the HM95 upper bound. This variability leads to a larger value of z^* for the verification bounds and, hence, a slightly smaller lower limit for the confidence interval (45.4 vs. 45.5). Slight anomalies of this sort occur several times in Tables 2 and 3. Also, for some cases in Table 3, estimated confidence intervals extend beyond the unit interval. When this happens, we enforce the restriction that the support of $M(Y_1)$ is $[0, 1]$.

that school performance evaluations are typically based solely on point estimates. See Kane and Staiger (2002) for a review and critique of such practices. We focus here on Figures 5 and 6, which depict the bounds on the fraction of students whose performance exceeds the national median. The school report cards in Pasadena focus attention on comparisons of this point estimate for each school to the point estimates for the district, county, and state.

The estimated bounds for PUSD are reported in the right-most panel of each figure. The PUSD bounds in Figure 5 include the bounds in Figure 2 with $(d_1, d_0) = (1, 0)$ and are reported in the sixth row of Table 3. They illustrate the findings in Sections 3 and 4 that the verification bounds on $I\!E\{Y_1 \geq 50\}$ (thin line in column marked "V") contain the bounds on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$ (black-filled rectangle in column V), which, in turn, contain the monotonicity bounds $I\!E\{Y_1 \geq 50\}$ and $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$ (limits of gray-filled rectangle in column marked "M"). Further, note that the estimated HM95 bounds on $I\!E\{Y_1 \geq 50\}$ (thin line in column marked "HM") contain the verification bounds on $I\!E\{Y_1 \geq 50\}$, while the HM95 bounds on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$ (black-filled rectangle in column HM) contain the verification bounds on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$. The extent to which the verification bounds are tighter than the HM95 bounds quantifies the identifying power of the verification information. For example, the estimated HM95 bounds on $I\!E\{Y_1 \geq 50\}$ are twice as wide as the corresponding verification bounds (35.5 versus 17.7 percentage points wide), and the HM95 bounds on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$ are about two-and-one-half times as wide as the verification bounds on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$ (21.6 versus 8.8 percentage points wide). The corresponding comparisons of 95% confidence intervals (Table 3, row 6) are 41.6 versus 22.9 and 27.9 versus 14.2, respectively.

Now consider what can be learned about the performance of truly English proficient students. The point estimates in Figure 5 of both the HM95 and verification upper bounds fall below the

benchmark value of 0.50 for three of the five high schools and for the district as a whole. The 95% confidence intervals (Table 3, top panel) for the verification bounds are also below this benchmark for two schools and for PUSD, whereas the HM95 intervals are completely below the benchmark for just one school. Further, the points estimates of the HM95 bounds are overlapping for all schools but the continuation school (90), whereas the verification bounds for School 80 fall completely below the verification bounds for Schools 63, 82, and 84.

The qualitative relationships among the estimated bounds reported in Figure 5 continue to hold when $(d_1, d_0) = (.95, .33)$ in Figure 6, and when $M(t) = t$ with $(d_1, d_0) = (1, 0)$ and $(.95, .33)$ in Figures 3 and 4, respectively. Note, however, that HM95 and verification upper bounds are closer when $(d_1, d_0) = (.95, .33)$. This change seems to be driven by the relaxation of the bound d_1 on the share of truly English proficient students among those classified as EP. When a maximum of 5%, as opposed to none, of the EP students are posited to not be truly proficient, the student-specific verification information becomes less useful. Yet the verification bounds for School 80 in Figure 6 still fall completely below the bounds for Schools 64 and 84.

Next, consider the relationships among these bounds and the verification bounds obtained under monotonicity restrictions, as depicted by the gray-filled rectangle in column M. Recall from Theorem 3 that, when $(d_1, d_0) = (1, 0)$, the observed performance measures $I\!E\{Y_1 \geq 50\}$ and $I\!E[\{Y_1 \geq 50\} \mid V = 1]$ are the lower and upper verification bounds, respectively, for $I\!E[Y_1 \mid Z = 1]$ when A5 and A6 hold. They are also the lower and upper bounds for $I\!E\{Y_1 \geq 50\}$ when A7 and A8 hold. As noted in Remark 3, under the corresponding assumptions, the HM95 lower bounds under monotonicity are identical to the verification lower bounds, but the HM95 upper bounds under monotonicity are equal to the unrestricted HM95 upper bound on $I\!E[\{Y_1 \geq 50\} \mid Z = 1]$. We see in Figure 5 for PUSD (and row 6 of Table 3) that the estimated verification bounds under mono-

tonicity are $[0.402, 0.458]$, whereas the estimated HM95 bounds under monotonicity are about 50 percent wider–$[0.402, 0.489]$. The estimated confidence intervals are $[0.375, 0.485]$ and $[0.375, 0.516]$, respectively. Thus, the confidence intervals for the HM95 bounds contain the benchmark value of 0.50 whereas the verification bounds confidence intervals do not. Note, however, that HM95 bounds and verification bounds are nearly identical when $(d_1, d_0) = (0.95, 0.33)$, reported in the bottom panel of Table 3.

To get a better sense of the valid range of disagreement about school performance based on the empirical evidence, suppose that officials or parents wish to focus on the scores of all students. The lower limits of the gray-filled rectangles in Figure 5 correspond to the school report card reports of the proportion of students who scored at least 50. Note that these values are below .50 for each school and, hence, for the school district. This finding suggests that the scores of more than half of the students at any school are below the national median score, indicating subpar performance at all schools. However, the verification upper bounds (thin lines in column V) for three schools and for PUSD exceed 0.50, indicating that performance at these schools and for the district as a whole may be "better than average."

By incorporating the monotonicity restrictions, we may reduce these upper bounds considerably. In particular, note that the upper bound in column M with $(d_1, d_0) = (1, 0)$, which is the fraction of EP scores exceeding the national median, exceeds 0.50 for just School 84. When $(d_1, d_0) = (.95, .33)$ (Figure 6), the upper bound for School 64 also exceeds the benchmark.

Moreover, note that the verification bounds under monotonicity are rather tight. Suppose we accept the notion that test scores are valid for all students classified as EP (i.e., $d_1 = 1$), as one could infer from both opposing arguments described in the introductory discussion of California test score litigation. Then, with the weakest bound on the share of valid scores among students

classified as LEP (i.e., $d_0 = 0$), the bounds on any school's performance ranges from 3.2 to 6.4 percentage points wide. Should we instead assume that at least one-third of LEP scores are valid (Table 3, middle panel), then the bounds range from just 2.6 to 6.0 percentage points wide. Given such tight bounds on the range of disagreement, it seems possible that much acrimony could be avoided by trying to come to an agreement on what conditions hold and then reporting sharp bounds based on the implied restrictions.

7. Summary

This paper undertakes a nonparametric analysis of mixture models with verification. These are mixture models for data that can be partitioned into two sets: a verified set and an unverified set. According to these models, observations from the verified set are more likely to be from the distribution of interest than observations from the unverified set. As indicated in the introduction, these models apply to a wide range of interesting data problems.

Sharp bounds are derived on characteristics of distributions of interest in these models, allowing for misclassification in the verification indicator. Sharp bounds under additional monotonicity conditions are also derived. For a certain distribution of interest, the functions optimized to produce the lower and upper bounds are shown to be piecewise linear convex and piecewise linear concave, respectively. These results lead to computational and asymptotic benefits. In particular, convexity and concavity can be used to establish the limiting distribution of the extremum estimators.

The identifying power of verification information is revealed through an analysis of math test scores of ninth graders in a California public school district, where an indicator of English proficiency plays the role of a verification indicator. The new methods yield informative comparisons of schools in this district with respect to various performance measures of interest, such as mean test scores and proportion of students exceeding the national median score. The conclusions drawn would

not be possible using previous methodology. In addition, the analysis calls attention to the need for tight and credible bounds on misclassification probabilities. Finally, the results can be used to resolve the contentious issue of how to use all the observed data to make valid inferences about school performance measures even when some test scores may not be valid.

APPENDIX

PROOF OF THEOREM 1.

Recall the definition of $L(p, \delta)$ and $U(p, \delta)$ given prior to the statement of Theorem 1. From (2), (3), and (4) we obtain the infeasible bounds

$$L(p_1, \delta_1) \leq I\!E[M(Y_1) \mid Z = 1] \leq U(p_1, \delta_1)\,.$$

It follows from this and the definition of λ_1 and u_1 that

$$\lambda_1 \leq L(p_1, \delta_1) \leq I\!E[M(Y_1) \mid Z = 1] \leq U(p_1, \delta_1) \leq u_1\,.$$

That is, λ_1 and u_1 are bounds for $I\!E[M(Y_1) \mid Z = 1]$ under A1 through A4. We want to prove that they are sharp bounds under these assumptions.

Assume, in addition to A1 through A4, that p_1 and δ_1 are known. Then $L(p_1, \delta_1)$ and $U(p_1, \delta_1)$ are sharp bounds for $I\!E[M(Y_1) \mid Z = 1]$ since they are based on simultaneously attainable sharp HM95 bounds for $I\!E[M(Y) \mid Z = 1, V = 1]$ and $I\!E[M(Y)|Z = 1, V = 0]$. The lower bounds (upper bounds) on these expectations are simultaneously attainable since the expectations are over disjoint subsets of the sample space.

Now drop the assumption that p_1 and δ_1 are known. Write Θ for the compact set $[v_1, \sigma(\delta)] \times$

$[d_1, 1]$. Since $\mathcal{Y}$ is a continuous random variable, both $L(p,\delta)$ and $U(p,\delta)$ are continuous on Θ. Since any pair $(p,\delta) \in \Theta$ is feasible under assumptions A1 through A4, so are the pairs that either minimize $L(p,\delta)$ or maximize $U(p,\delta)$. This proves sharpness. □

Proof of Theorem 2.

We shall prove the lower bound result. The proof of the upper bound result is similar. Note that $\mathbb{E}[M(Y_1) \mid Z = 1] = \mathbb{E}[M(Y) \mid Z = 1]$. It follows that

$$\mathbb{E}M(Y_1) = \mathbb{E}[M(Y) \mid Z = 1]\mathbb{P}\{Z = 1\} + \mathbb{E}[M(Y_1) \mid Z = 0]\mathbb{P}\{Z = 0\}\,. \tag{5}$$

Recall the following definitions: $p_1 = \mathbb{P}\{V = 1 \mid Z = 1\}$, $\delta_1 = \mathbb{P}\{Z = 1 \mid V = 1\}$, and $\pi(p_1, \delta_1) = \mathbb{P}\{Z = 1 \mid V = 0\}$, where $\pi(p,\delta) = (1-p)v_1\delta/[p(1-v_1)]$ for each $(p,\delta) \in \Theta \equiv [v_1, \sigma(\delta)] \times [d_1, 1]$ with $\sigma(\delta) = \delta v_1/[\delta v_1 + d_0(1-v_1)]$. Recall the definition of $L(p,\delta)$ given before the statement of Theorem 1.

Note that $\mathbb{P}\{Z = 1\} = \delta_1 v_1 + \pi(p_1,\delta_1)(1-v_1)$. For each $(p,\delta) \in \Theta$, write $w(p,\delta)$ for the weight function $\delta v_1 + \pi(p,\delta)(1-v_1)$. Deduce from (5), the argument preceding the statement of Theorem 1, and the lower bound on $M(Y_1)$ that

$$\begin{aligned} \mathbb{E}M(Y_1) &\geq L(p_1,\delta_1)w(p_1,\delta_1) + a[1 - w(p_1,\delta_1)] \\ &\geq \inf_{(p,\delta)\in\Theta} \left[L(p,\delta)w(p,\delta) + a[1 - w(p,\delta)]\right]\,. \end{aligned}$$

Since $L(p,\delta) \geq a$ for each $(p,\delta) \in \Theta$, the last bound is minimized when $w(p,\delta)$ is minimized over $(p,\delta) \in \Theta$. This occurs at $\delta = d_1$ and $p = \sigma(d_1)$ since $\pi(\sigma(d_1), d_1) = d_0$. This yields the stated lower bound.

To prove that the lower bound is sharp, consider a distribution for the data satisfying the following conditions: $I\!P\{Z = 1 \mid V = 1\} = d_1$, $I\!P\{Z = 1 \mid V = 0\} = d_0$, $M(Y_1) = a$ whenever $Z = 0$, $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid \mathcal{Y} \leq Q_1(d_1), V = 1]$, $I\!E[M(Y) \mid Z = 1, V = 0] = I\!E[M(Y) \mid \mathcal{Y} \leq Q_0(d_0), V = 1]$. Note that the first two conditions imply that $p_1 = \sigma(d_1)$. Thus, the lower bound λ_2 is attained for this distribution. Moreover, this distribution is consistent with A1 through A4 and the assumptions about the support of $M(Y_1)$. □

Proof of Theorem 3.

First, we show that u_3 is the sharp upper bound for $I\!E[M(Y_1) \mid Z = 1]$ under assumptions A1 through A4 and A6.

Temporarily, assume only that A1, A4, and A6 hold. Recall $p_1 = I\!P\{V = 1 \mid Z = 1\}$. Apply A4, the law of total probability, and A6 to get

$$\begin{aligned} I\!E[M(Y_1) \mid Z = 1] &= I\!E[M(Y) \mid Z = 1] \\ &= I\!E[M(Y) \mid Z = 1, V = 1]p_1 + I\!E[M(Y) \mid Z = 1, V = 0](1 - p_1) \\ &\leq I\!E[M(Y) \mid Z = 1, V = 1]\,. \end{aligned}$$

Apply A4 and Proposition 4 in HM95 to get

$$I\!E[M(Y) \mid Z = 1, V = 1] \leq I\!E[M(Y) \mid \mathcal{Y} > Q_1(1 - d_1), V = 1]\,. \tag{6}$$

Consider a distribution for the data for which $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid Z = 1, V = 0]$ and $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid \mathcal{Y} > Q_1(1 - d_1), V = 1]$. The upper bound in (6) is attained for this feasible distribution, proving that it is sharp under assumptions A1, A4, and A6.

Now add assumptions A2 and A3 and argue as in the proof of Theorem 1 to show that u_3 is sharp under assumptions A1 through A4 and A6.

Next, apply A5, A4, and the fact that $I\!E[M(Y_0) \mid Z = 0] = I\!E[M(Y) \mid Z = 0]$ to get

$$\begin{aligned} I\!E[M(Y_1) \mid Z = 1] &\geq I\!E[M(Y_1) \mid Z = 1]I\!P\{Z = 1\} + I\!E[M(Y_0) \mid Z = 0]I\!P\{Z = 0\} \\ &= I\!E[M(Y) \mid Z = 1]I\!P\{Z = 1\} + I\!E[M(Y_0) \mid Z = 0]I\!P\{Z = 0\} \\ &= I\!E[M(Y) \mid Z = 1]I\!P\{Z = 1\} + I\!E[M(Y) \mid Z = 0]I\!P\{Z = 0\} = I\!EM(Y)\,. \end{aligned}$$

Consider a distribution for the data for which $I\!E[M(Y_1) \mid Z = 1] = I\!E[M(Y_0) \mid Z = 0]$. The lower bound λ_3 is attained for this feasible distribution, proving that it is sharp under assumptions A1 through A5.

Next, we show that u_3 is the sharp upper bound on $I\!EM(Y_1)$ under assumptions A1 through A4 and A8. Temporarily, assume only that A1, A4, and A8 hold. Apply the law of total probability and A8 to get

$$\begin{aligned} I\!EM(Y_1) &= I\!E[M(Y_1) \mid Z = 1]I\!P\{Z = 1\} + I\!E[M(Y_1) \mid Z = 0]I\!P\{Z = 0\} \\ &\leq I\!E[M(Y_1) \mid Z = 1]\,. \end{aligned}$$

Now use (6) to get an upper bound on $I\!EM(Y_1)$. Consider a distribution for the data for which (i) $I\!E[M(Y_1) \mid Z = 1] = I\!E[M(Y_1) \mid Z = 0]$ (ii) $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid Z = 1, V = 0]$ and (iii) $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid \mathcal{Y} > Q_1(1 - d_1), V = 1]$. The upper bound in (6) is attained for this feasible distribution, proving that it is sharp under assumptions A1, A4, and A8. As before, add assumptions A2 and A3 and argue as in the proof of Theorem 1 to show that u_3 is sharp under assumptions A1 through A4 and A8.

Next, apply the law of total probability, A7, A4, and $\mathbb{E}[M(Y_0) \mid Z = 0] = \mathbb{E}[M(Y) \mid Z = 0]$ to get

$$\begin{aligned}
\mathbb{E}M(Y_1) &= \mathbb{E}[M(Y_1) \mid Z = 1]\mathbb{P}\{Z = 1\} + \mathbb{E}[M(Y_1) \mid Z = 0]\mathbb{P}\{Z = 0\} \\
&\geq \mathbb{E}[M(Y_1) \mid Z = 1]\mathbb{P}\{Z = 1\} + \mathbb{E}[M(Y_0) \mid Z = 0]\mathbb{P}\{Z = 0\} \\
&= \mathbb{E}[M(Y) \mid Z = 1]\mathbb{P}\{Z = 1\} + \mathbb{E}[M(Y_0) \mid Z = 0]\mathbb{P}\{Z = 0\} \\
&= \mathbb{E}[M(Y) \mid Z = 1]\mathbb{P}\{Z = 1\} + \mathbb{E}[M(Y) \mid Z = 0]\mathbb{P}\{Z = 0\} = \mathbb{E}M(Y)\,.
\end{aligned}$$

Consider a distribution for the data for which $\mathbb{E}[M(Y_1) \mid Z = 0] = \mathbb{E}[M(Y_0) \mid Z = 0]$. The lower bound λ_3 is attained for this feasible distribution, proving that it is sharp under assumptions A1 through A4 and A7. This completes the proof. □

Proof of Theorem 4 when $d_0 = 0$ and $d_1 = 1$.

Start with $\hat{L}(p, 1)$. Write $\hat{\gamma}$ for $\sum_{i=1}^{n} M(Y_i)V_i/n_1$ and $\hat{c}$ for $(1 - \hat{v}_1)/\hat{v}_1$. We have that

$$\hat{L}(p, 1) = p\left[\hat{\gamma} + \hat{c}\sum_{i=1}^{n} M(Y_i)(1 - V_i)\{\mathcal{Y}_i \leq \hat{Q}_0(\hat{\pi}(p, 1)\}/n_0\right].$$

Define $y_0 = -\infty$. Define $\tau_0 = 0$, $\tau_k = M(y_k)$, $k = 1, 2, \ldots, m$, and $\tau_{m+1} = \tau_m$. Finally, define $\hat{H}_0$ to be the empirical cumulative distribution function for $\mathcal{Y}$ given $V = 0$. Straightforward calculations show that

$$\hat{L}(p, 1) = \sum_{k=m}^{1} (\hat{\beta}_k p + \tau_k)\{\hat{\pi}^{-1}(\hat{H}_0(y_k)) \leq p < \hat{\pi}^{-1}(\hat{H}_0(y_{k-1}))\} \tag{7}$$

where

$$\hat{\beta}_k = \hat{\gamma} - \tau_k - \hat{c}\sum_{j=1}^{k}(\tau_j - \tau_{j-1})\hat{H}_0(y_{j-1})$$

and, for $p \in [0,1]$,

$$\hat{\pi}^{-1}(p) = \hat{v}_1/[\hat{v}_1 + p(1-\hat{v}_1)]\,.$$

Note that the summation in (7) runs from m to 1, not from 1 to m. It is easy to show that $\hat{\beta}_k$ is nondecreasing as k decreases, and that $\hat{L}(p,1)$ is continuous on $[\hat{v}_1,1]$. Deduce that $\hat{L}(p,d)$ is a piecewise linear convex function on $[\hat{v}_1,1]$.

Similar calculations show that

$$\hat{U}(p,1) = \sum_{k=1}^{m} (\hat{\beta}_k p + \tau_k)\{\hat{\pi}^{-1}(1-\hat{H}_0(y_{k-1})) < p \le \hat{\pi}^{-1}(1-\hat{H}_0(y_k))\} \tag{8}$$

where

$$\hat{\beta}_k = \hat{\gamma} - (\hat{c}+1)\tau_k + \hat{c}\tau_m - \hat{c}\sum_{j=k}^{m}(\tau_{j+1}-\tau_j)\hat{H}_0(y_j)\,.$$

It is easy to show that β_k is nonincreasing as k increases, and that $\hat{U}(p,1)$ is continuous on $[\hat{v}_1,1]$. Deduce that $\hat{U}(p,1)$ is a piecewise linear concave function on $[\hat{v}_1,1]$. □

The Limiting Distribution of λ_1 when $d_0 = 0$ and $d_1 = 1$.

Recall the definitions of y_k and τ_k, $k = 0,1,\ldots,m$. Define H_0 to be the cumulative distribution function for $\mathcal{Y}$ given $V = 0$. Define $\gamma = I\!E[M(Y_1) \mid V = 1]$ and $c = (1-v_1)/v_1$. Using the proof technique of Theorem 4, it is straightforward to show that

$$L(p,1) = \sum_{k=m}^{1} (\beta_k p + \tau_k)\{\pi^{-1}(H_0(y_k)) \le p < \pi^{-1}(H_0(y_{k-1}))\} \tag{9}$$

where

$$\beta_k = \gamma - \tau_k - c\sum_{j=1}^{k}(\tau_j - \tau_{j-1})H_0(y_{j-1})$$

and, for $p \in [0,1]$,

$$\pi^{-1}(p) = v_1/[v_1 + p(1 - v_1)]\,.$$

Likewise, it is easy to check that $L(p,1)$ is a piecewise linear convex function on $[v_1, 1]$. It follows that there exists a kink point ordinate $\pi^{-1}(H_0(y_k)) \in [v_1, 1]$ such that $\lambda_1 = L(\pi^{-1}(H_0(y_k)), 1)$.

Extend $\hat{L}(p,1)$ and $L(p,1)$ in the obvious way so that $\hat{L}(p,1)$ and $L(p,1)$ are defined for all $p \in [0,1]$ and are piecewise linear convex functions on $[0,1]$. By weak laws of large numbers, $\hat{\gamma}$ converges in probability to γ, $\hat{v}_1$ converges in probability to v_1, and $\hat{H}_0(y_j)$ converges in probability to $H_0(y_j)$, $j = 1, 2, \ldots, m$. These facts, together with (7) and (9), imply that (i) $\hat{\pi}^{-1}(\hat{H}_0(y_j))$ converges in probability to $\pi^{-1}(H_0(y_j))$, $j = 1, 2, \ldots, m$ and (ii) for each $p \in [0,1]$, $\hat{L}(p,1)$ converges in probability to $L(p,1)$. It easily follows from (i), (ii), $\lambda_1 = L(\pi^{-1}(H_0(y_k)), 1)$, and piecewise linear convexity of $\hat{L}(p,1)$ and $L(p,1)$, that with probability tending to one as $n \to \infty$, $\hat{\lambda}_1 = \hat{L}(\hat{\pi}^{-1}(\hat{H}_0(y_k)), 1)$.

In sum, with probability tending to one as $n \to \infty$, there exists a fixed point y_k in the support of Y such that $\hat{\lambda}_1 = \hat{L}(\hat{\pi}^{-1}(\hat{H}_0(y_k)), 1)$ and $\lambda_1 = L(\pi^{-1}(H_0(y_k)), 1)$. When this happens, it follows from (7) and (9) that

$$\hat{\lambda}_1 - \lambda_1 = f(\hat{\theta}) - f(\theta_0)$$

where $\hat{\theta} = (\hat{\gamma}, \hat{v}_1, \hat{H}_0(y_1), \ldots, \hat{H}_0(y_k))'$, $\theta_0 = (\gamma, v_1, H_0(y_1), \ldots, H_0(y_k))'$, and, for each parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_{k+2})' \in I\!R^{k+2}$,

$$f(\theta) = \left[\theta_1 - \tau_k - \frac{1 - \theta_2}{\theta_2} \sum_{j=2}^{k} (\tau_j - \tau_{j-1})\theta_{j+1}\right] \left[\frac{\theta_2}{\theta_2 + \theta_{k+2}(1 - \theta_2)}\right]\,.$$

Let $n^{-1} \sum_{i=1}^{n} \Delta_{1i}$ denote the zero-mean average $\hat{\gamma} - \gamma$. Similarly, let $n^{-1} \sum_{i=1}^{n} \Delta_{2i}$ denote $\hat{v}_1 - v_1$ and let $n^{-1} \sum_{i=1}^{n} \Delta_{2+j,i}$ denote $\hat{H}_0(y_j) - H_0(y_j)$, $j = 1, 2, \ldots, k$. Let the symbol $\Longrightarrow$ denote

convergence in distribution. By the Multivariate Central Limit Theorem,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0, \Sigma)$$

where Σ is a $(k+2)\times(k+2)$ matrix with stth entry equal to the probability limit of $n^{-1}\sum_{i=1}^{n} \Delta_{si}\Delta_{ti}$. The next result follows by an application of the delta method.

THEOREM 5. *Under the assumptions of Theorem 4,* $\sqrt{n}(\hat{\lambda}_1 - \lambda_1) \Longrightarrow N(0, \frac{\partial f}{\partial \theta}' \Sigma \frac{\partial f}{\partial \theta})$.

The asymptotic variance in Theorem 5 can be estimated from the data by replacing population quantities with sample analogues. An argument similar to the one used to prove Theorem 5 can be used to show that $\sqrt{n}(\hat{u}_1 - u_1)$ is asymptotically normal. These results, in turn, can be used to develop an asymptotic confidence interval for $I\!E[M(Y_1) \mid Z = 1]$ using Bonferroni's inequality, as is done in Section 3.3 of Horowitz and Manski (1997).

Finally, we note that sample analogs of the population bounds derived in Theorems 2 and 3 in Sections 3 and 4 can be easily constructed by replacing expectations with sample averages and population quantile functions with sample quantile functions. Standard asymptotic methods can be applied to prove $\sqrt{n}$-consistency and asymptotic normality of these estimators. Then asymptotic confidence intervals can be developed as discussed in the last paragraph.

BOUNDS ON $I\!E[M(Y_1)|Z = 1]$ FOR $M(Y_1) = Y_1$ AND $M(Y_1) = \{Y_1 \geq 50\}$.

When $M(Y_1) = Y_1$, the quantity of interest is the mean math score of truly English proficient students. When $M(Y_1) = \{Y_1 \geq 50\}$, the quantity of interest is the probability that the scores of truly English proficient students exceed the national median score. The corresponding population bounds are given in Theorem 1 in Section 3 and Theorem 3 in Section 4.

A brief word on notation. For notational simplicity, in previous sections, we have suppressed the dependence of $\hat{\sigma}(\delta)$ on d_0 and $\hat{\delta}(p)$, the $\hat{\lambda}_i$'s, and the $\hat{u}_i$'s on d_1 and d_0. For the purpose of better understanding the figures, it will be convenient to make these dependencies explicit. Thus, we will write $\hat{\sigma}(\delta, d_0)$ for $\hat{\sigma}(\delta)$, $\hat{\delta}(p, d_1, d_0)$ for $\hat{\delta}(p)$, $\hat{\lambda}_i(d_1, d_0)$ for $\hat{\lambda}_i$ and $\hat{u}_i(d_1, d_0)$ for $\hat{u}_i$.

Figure 1 depicts the bounds for $M(Y_1) = Y_1$, along with the sample lower and upper bound functions $\hat{L}(p, \hat{\delta}(p, d_1, d_0))$ and $\hat{U}(p, \hat{\delta}(p, d_1, d_0))$ for $(d_1, d_0) = (1, 0)$, $(.95, 0)$, and $(.95, .33)$. The bounds are based on the data for all 1543 ninth-grade math test takers in PUSD in 2000. Of these, 82% were classified as EP. That is, $\hat{v}_1 = .82$.

First, consider the functions $\hat{L}(p, \hat{\delta}(p, 1, 0))$ and $\hat{U}(p, \hat{\delta}(p, 1, 0))$ in Figure 1. These functions are constructed under the assumption that all students classified as EP are truly English proficient ($d_1 = 1$), while possibly none of the students classified as LEP are truly English proficient ($d_0 = 0$). These functions are optimized over $p \in [\hat{v}_1, \hat{\sigma}(1, 0)] = [.82, 1]$. Starting from $p = \hat{v}_1$, we see that $\hat{L}(\hat{v}_1, \hat{\delta}(\hat{v}_1, 1, 0)) = \hat{U}(\hat{v}_1, \hat{\delta}(\hat{v}_1, 1, 0)) = \hat{I\!E}Y$, the sample mean of the observed scores of all students. Note that $\hat{L}(p, \hat{\delta}(p, 1, 0))$ is minimized at $p = .85$ where it takes the value 46.05. That is, $\hat{\lambda}_1(1, 0) = 46.05$. Next, at $p = 1$, we see that $\hat{L}(1, \hat{\delta}(1, 1, 0)) = \hat{U}(1, \hat{\delta}(1, 1, 0)) = \hat{I\!E}[Y|V = 1]$, the sample mean of the observed scores of EP students. Note that $\hat{I\!E}Y$ and $\hat{I\!E}[Y|V = 1]$ are the sample analogues of the Theorem 3 lower and upper bounds, λ_3 and u_3, under monotonicity restrictions A5 and A6 when $d_1 = 1$ and $d_0 = 0$. Finally, note that the upper bound function $\hat{U}(p, \hat{\delta}(p, 1, 0))$ is maximized at $p = .97$ where it takes the value 48.89. That is, $\hat{u}_1(1, 0) = 48.89$.

Next, consider the functions $\hat{L}(p, \hat{\delta}(p, .95, 0))$ and $\hat{U}(p, \hat{\delta}(p, .95, 0))$ in Figure 1. These functions illustrate changes that occur when misclassification of EP students is allowed ($d_1 = .95$). As before, they are optimized over $p \in [\hat{v}_1, \hat{\sigma}(1, 0)] = [.82, 1]$. The most prominent change is the introduction of gaps between the functions at $p = \hat{v}_1$ and $p = 1$. Otherwise, the shapes of the functions

are similar to those just described, except with lower infimum ($\hat{\lambda}_1(.95, 0) = 44.18$) and higher supremum ($\hat{u}_1(.95, 0) = 50.43$). Also, note that the lower bound, under monotonicity conditions A5 and A6 when $d_1 = .95$ and $d_0 = 0$, is still $\hat{\lambda}_3 = \hat{I\!E}Y$, whereas the upper bound increases to $\hat{u}_3(.95, 0) = \hat{U}(1, \hat{\delta}(1, .95, 0))$.

Next, consider the functions $\hat{L}(p, \hat{\delta}(p, .95, .33))$ and $\hat{U}(p, \hat{\delta}(p, .95, .33))$ in Figure 1. These functions are computed under the assumption that at least 95% of EP students are truly English proficient ($d_1 = .95$) and at least 33% of LEP students are truly English proficient ($d_0 = .33$); they are optimized over $p \in [\hat{v}_1, \hat{\sigma}(1, .33)] = [.82, .93]$. Note that $\hat{\delta}(p, d_1, d_0)$ is increasing in d_0 while $\hat{L}(p, \hat{\delta})$ is increasing in $\hat{\delta}$ and $\hat{U}(p, \hat{\delta})$ is decreasing in $\hat{\delta}$. It follows that when $d_0^* \geq d_0$, $\hat{L}(p, \hat{\delta}(p, d_1, d_0^*)) \geq \hat{L}(p, \hat{\delta}(p, d_1, d_0))$ and $\hat{U}(p, \hat{\delta}(p, d_1, d_0^*)) \leq \hat{U}(p, \hat{\delta}(p, d_1, d_0))$ for $p \in [\hat{v}_1, \hat{\sigma}(1, d_0^*)]$. Deduce that when $d_0^* \geq d_0$, $\hat{\lambda}_1(d_1, d_0^*) \geq \hat{\lambda}_1(d_1, d_0)$ and $\hat{u}_1(d_1, d_0^*) \leq \hat{u}_1(d_1, d_0)$. Note that for the scores data, $\hat{\lambda}_1(.95, .33) = \hat{\lambda}_1(.95, 0)$ while $\hat{u}_1(.95, .33) < \hat{u}_1(.95, 0)$.

Finally, note that all the sample lower bound functions in Figure 1 are piecewise linear convex functions while all the sample upper bound functions are piecewise linear concave functions on their respective domains, as guaranteed by Theorem 4 and the paragraph immediately preceding it in Section 5.

Results parallel to those presented in Figure 1 are presented in Figure 2 for the case $M(Y_1) = \{Y_1 \geq 50\}$. As in Figure 1, the piecewise linear convexity of the sample lower bound functions and the piecewise linear concavity of the upper bound functions are evident. We also see the gaps at $p = \hat{v}_1$ and $p = 1$ that are introduced when d_1 decreases from 1 to .95, as well as the sharp rise in the sample lower bound function and the sharp drop in the sample upper bound function at $\hat{\sigma}(d_1 = .95, .33)$ when d_0 increases from 0 to .33.

REFERENCES

Abedi, J. (2002): "Assessment and Accomodations of English Language: Issues, Concerns, and Recommendations," *Journal of School Improvement*, Volume 3, Issue 1.

CBT Macmillan/McGraw-Hill. (1994): *Spanish Assessment of Basic Education, Second Edition: Norms Book*, Monterey, CA: CTB.

Dominitz, J. and R. P. Sherman (2004): "Sharp Bounds under Contaminated or Corrupted Sampling with Verification, with an Application to Environmental Pollutant Data," *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 319–338.

Horowitz, J., and C. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Horowitz, J., and C. Manski (1997): "What Can Be Learned about Population Parameters when the Data are Contaminated", in *The Handbook of Statistics: Robust Inference*, Vol.15, 439–466. C.R. Rao and G.S. Maddala, eds., North Holland, Amsterdam.

Horowitz, J., and C. Manski (1998): "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations", *Journal of Econometrics*, 84, 37–58.

Horowitz, J., and C. Manski (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", *Journal of the American Statistical Association*, 95, No. 449, 77–84.

Kane, T. J., and D. O. Staiger (2002): "The Promise and Pitfalls of Using Imprecise School Accountability Measures", *Journal of Economic Perspectives*, 16, 91–114.

Lambert, D. and L. Tierney (1997): "Nonparametric Maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias," *Journal of the American Statistical Association*, 92, 937–944.

Molinari, F. (2002): "Missing Treatments," Manuscript, Northwestern University, July 16.

Thompson, M. S., K. E. DiCerbo, K. Mahoney, and J. MacSwan (2002): "Exito en California: A Validity Critique of Language Program Evaluations and Analysis of English Learner Test Scores," *Education Analysis Policy Archives*, Volume 10, No. 7.

U.S. Bureau of the Census (1998): "SIPP Quality Profile: 1998," SIPP Working Paper Number 230, available at http://www.bls.census.gov/sipp/workpapr/wp230.pdf