The Impact of Performance Level Misclassification on the Accuracy and Precision of Percent at Performance Level Measures

Damian W. Betebenner, Yi Shang, Yun Xiang, Yan Zhao & Xiaohui Yue

Department of Educational Research, Measurement and Evaluation
Lynch School of Education, Boston College

October 2, 2006

Abstract

NCLB performance mandates, embedded within state accountability systems, focus school AYP/AMO compliance squarely on the percentage of students at or above proficient. The singular importance of this quantity for decision making purposes has initiated extensive research into percent proficient as a measure of school quality. In particular, technical discussions have scrutinized the impact of sampling, measurement, and other sources of error on percent proficient statistics. In this paper we challenge the received orthodoxy that measurement error associated with individual students' scores is inconsequential for aggregate percent proficient statistics. Synthesizing current classification accuracy research with techniques from randomized response designs, we establish results which specify the extent to which measurement error—manifest as performance level misclassifications—produces bias and increases error variability for percent at performance level statistics. The results have direct relevance for the design of coherent and fair accountability systems based upon assessment outcomes.

Background

As a result of No Child Left Behind (2001), there is acute interest regarding the (in)adequacy of using percentage of examinees scoring at or above proficient as a measure of school quality (Linn, Baker, & Betebenner, 2002; Kane & Staiger, 2002).[1] Beyond the skepticism of whether assessment outcomes alone adequately capture school quality, measurement and policy specialists have raised more technical concerns about year-to-year volatility in school results used for AYP determination (Rogosa, 2005; Kane & Staiger, 2002; Linn & Haug, 2002; Arce-Ferrer, Frisbie, & Kolen, 2002; Yen, 1997). If fluctuations in percent at performance level statistics are so large as to obscure changes resulting from institutional improvement, then the ability of such measures to function as useful auditing mechanisms, providing feedback about achievement over time, is seriously impaired.

Annual changes in percent at performance level reflect numerous factors unrelated to changes in school effectiveness including sampling error, measurement error, differences in cohorts, and equating error (Arce-Ferrer et al., 2002). Many states, recognizing the variability accompanying such percentages, use confidence intervals based upon random sampling assumptions to insulate schools, particularly those with few students, from this volatility (Coladarci, 2003). Often discounted in the

[1] To simplify wording we use *percent at performance level* generically to describe an aggregate percentage of students performing at some arbitrary performance level (e.g., proficient). To further simplify, we use school to denote an arbitrary aggregation unit used for reporting purposes. However, the analyses herein apply to any aggregation of students—for example, districts or states.

discussion of volatility is measurement error (Coladarci, 2003; Hill, 2001; Cronbach, Linn, Brennan, & Haertel, 1997). In this paper we challenge the contention that measurement error is of little or no concern in percent at performance level aggregates. Indeed, in some very common circumstances we demonstrate that measurement error can seriously undermine the accuracy and precision of percent at performance level statistics and, consequently, jeopardize inferences one may wish to draw about school quality.[2] We establish results that quantify both the bias and variability in percent at performance level statistics attributable to measurement error and use these results to clarify discussions about accountability systems based upon percent at performance level statistics.

Performance Level Misclassification: Consistency and Accuracy

Going back nearly 40 years, an extensive body of research exists into measurement error and criterion referenced testing (see, for example, Hambleton & Novick (1973), Livingston & Lewis (1995) and references contained therein). Gradually, two somewhat overlapping research approaches toward measurement error and classification emerged: consistency and accuracy. Classification consistency quantifies the extent to which two observed categorizations coincide based upon two independent examinations whereas classification accuracy relates the extent to which a examinee's observed categorization matches his or her true status. In a review of state assessment documentation and related research articles, we find classification consistency to be the predominant method used to quantify the impact of measurement error on performance level categorization. This is likely due to its affinity with classical test reliability. Deviating from this norm, in what follows we establish the utility of concepts associated with classification accuracy.

The literature on classification accuracy, unfortunately, is plagued by inconsistencies in both notation and terminology. To better situate our own discussion, we begin by reviewing different approaches toward the concept. The majority of researchers develop classification accuracy *vis-à-vis* the estimation of the *joint* distribution of observed and true classifications. That is, based upon the administration of a some assessment, if A denotes the latent (i.e., true) categorical variable and A^* its observed counterpart, each consisting of k performance levels, then classification accuracy is defined as the probability of correct classification across all performance levels:

$$\sum_{i=1}^{k} \Pr(A^* = i,\, A = i). \tag{1}$$

The probability of correct classification (i.e., *accuracy* at the individual level) given by Equation 1 is used by many authors to describe the relationship between measurement error and performance level (mis)classification (Martineau, 2006; Rudner, 2005; Noble, 2004; Lee, Hanson, & Brennan, 2002; Rudner, 2001; Young & Yoon, 1998; Livingston & Lewis, 1995).

A closely related approach begins with the *conditional* classification probabilities instead of the joint probabilities of Equation 1. Though conditional probabilities are easily derivable from the joint distribution, there is less research using this approach, likely due to researchers seeking to summarize classification in terms of the accuracy statistic (Kupermintz, 2004; Rogosa, 1994; Hanson & Brennan, 1990). Following Coleman (1964), Wiggins (1973), and Rogosa (1994), we use conditional probabilities to quantify performance level misclassification and represent the collection of all such probabilities as a stochastic matrix called a *misclassification matrix*. Formally, the misclassification

[2] The terms accuracy and precision are often used incorrectly. Yen (1997), for example, refers to the accuracy of percent above cutpoints when it is precision that is investigated. Accuracy is a quantification of veracity whereas precision quantifies reproducibility. See Stallings and Gillmore (1971) for a concise and very readable comparison of accuracy/precision with validity/reliability.

matrix $\mathbf{P}$, consisting of conditional probabilities, is given by:

$$\mathbf{P} = \{p_{ij}\}_{1 \le i,j \le k} \quad \text{where} \quad p_{ij} = \Pr(A^* = j \,|\, A = i). \tag{2}$$

Frequently, in applications, there are $k = 2$ levels for the latent and observed variables, A and A^*. For example, current state accountability systems based upon NCLB commonly employ the proficient/not proficient dichotomy. Accordingly, if $i = 1$ refers to proficient and $i = 2$ refers to not proficient, then two frequently encountered quantities are the *false-positive* and *false-negative* error rates, $\Pr(A^* = 1 \,|\, A = 2)$ and $\Pr(A^* = 2 \,|\, A = 1)$, respectively.[3] The complements of false-positive and false-negative error rates are the conditional correct classification rates, $\Pr(A^* = 1 \,|\, A = 1)$ and $\Pr(A^* = 2 \,|\, A = 2)$.

Numerous procedures exist to estimate conditional probabilities used to construct the misclassification matrix $\mathbf{P}$. For scale score based exams with performance levels determined by predefined cut-scores, all the methods share a common foundation, estimation of the joint distribution for observed and true scores:

$$\begin{aligned} p_{ij} &= \sum_{x \in \mathsf{PLj}} \sum_{\tau \in \mathsf{PLi}} \Pr(X = x, T = \tau) \Big/ \sum_{\tau \in \mathsf{PLi}} \Pr(T = \tau) \\ &= \sum_{x \in \mathsf{PLj}} \sum_{\tau \in \mathsf{PLi}} \Pr(X = x \,|\, T = \tau) \cdot g(\tau) \Big/ \sum_{\tau \in \mathsf{PLi}} \Pr(T = \tau). \end{aligned} \tag{3}$$

Here, X represents observed scores, T true scores, and $g(\tau)$, the true score density. Summation (or integration for continuous score distributions) across appropriate scores in performance levels PLi and PLj, and division by the marginal true score density of PLi yields the corresponding conditional probability, p_{ij}. Various methods are available to estimate the quantities of Equation 3. The well known procedure of Livingston and Lewis (1995), building upon the work of Hanson and Brennan (1990), invokes a strong true score model—a four-parameter beta true score distribution combined with a binomial/compound binomial error distribution—to estimate classification accuracy. Rudner (2001, 2005) and Martineau (2006) simplify the estimation procedures by making strong normality assumptions for both the conditional error distribution as well as the student true score density. Mislevy (Noble, 2004) employs simulation techniques to derive the misclassification matrix, $\mathbf{P}$. Though determining which method yields the best estimates is a topic worthy of investigation, we take the misclassification matrix $\mathbf{P}$ as given and proceed to demonstrate its utility regarding bias and variability of percent at performance level statistics.

Current uses of classification statistics (e.g., accuracy) focus primarily on individual level inferences with regard to the measurement process. For example, the standard frequentist interpretation of accuracy indicates the chances that an individual, selected at random, has of being correctly classified according to his/her true performance level. This interpretation takes account of only half of the numerator of Equation 3. The multiplier of Equation 3, $\Pr(X = x \,|\, T = \tau)$, is indicative of the measurement process, namely the conditional standard error of measurement, while the multiplicand, $g(\tau)$, weights the multiplier based upon the sample/population under consideration. Currently, most accuracy analyses use all available test takers as the sample/population.

By assuming this underlying reference population, accuracy and other classification statistics implicitely weight $\Pr(X = x \,|\, T = \tau)$ according to this distribution. Indeed, conditional misclassification probabilities include information from *both* the measurement instrument as well as how individuals

[3] Hanson & Brennan (1990, p. 348) discuss false-positive and false-negative error rates with regard to classification accuracy, however their rates are unconventional in that they represent joint instead of conditional probabilities.

in the sample/population are situated relative to the cut scores used to define the performance levels. It is entirely likely that different populations—specifically, different schools or districts—have different true score densities thus yielding different misclassification rates. This is not surprising: A school having students lying close to performance level cutpoints would be expected to have higher misclassification rates (i.e., lower accuracy rates) than a school with students well away from those thresholds.

A large body of research exists relating latent and observed distributions using the misclassification probabilities when misclassification rates are specified. In the situation with 2 levels, the impact of known misclassification upon accuracy and precision of estimates was first investigated by Bross (1954). These results were extended to more than two levels by Mote and Anderson (1965). At present, the approach has been generalized such that the conditional misclassification probabilities of Equation 3 represent a latent class model composed of a mixture of distributions (Agresti, 2002). As such, the misclassification matrix $\mathbf{P}$ is a transition matrix defining a mapping between latent and observed distributions—a Markov process between latent and observed categories (Gnedenko, 1967). In the next section, drawing upon techniques developed for randomized response designs, we elaborate on the relationship between latent and observed score distributions and derive distributional measures that specify percent at performance level statistics.

Randomized Response

Many areas of research exist where categorical data are misclassified subject to known misclassification rates. (Kuha & Skinner, 1997). The techniques utilized are derived from sampling designs in which data are *purposely* misclassified subject to known probabilities. As such, the sampling designs specify the exact relationship between latent and observed characteristics. Such designs are often employed when honest answers to sensitive questions might not be forthcoming. In such situations, assuring respondent anonymity is critical to minimize the answer bias resulting from false or non-response. An elegant and simple way to protect respondent identity and thereby address answer bias is the Randomized Response (RR) design (Warner, 1965).

In RR, respondents answer question(s) randomly assigned from a set of questions. The researcher is unaware exactly what question each respondent has answered but knows the frequency with which each of the questions are asked. With this information, response rates based upon the desirable latent characteristics are calculated (van den Hout & van der Heijden, 2002; Chaudhuri & Mukerjee, 1988).[4] Consider the following scenario associated with an exercise given in an elementary statistics text:

> Suppose that we want to determine what percentage of the students at a large university smoke marijuana at least once a week. We construct 20 flash cards, write "I smoke marijuana at least once a week" on 12 of the cards, where 12 is an arbitrary choice, and "I do not smoke marijuana at least once a week" on the others. Then we let each student (in the sample interviewed) select one of the cards at random, and respond "yes" or "no" without divulging the question (Freund & Perles, 2007, p. 156).

Here, the researcher is unaware which question the student answers making a *yes* or a *no* response uninformative at the individual level. However, because the researcher knows the probability with which the questions were asked, (s)he is able to recover an estimate of the true percentage of students who smoke marijuana at least once a week.

[4] Another area where misclassification is used to protect individual identity include Post Randomization Method (PRAM). Both RR and PRAM assume known misclassification rates that are used to recover statistics related to the latent variable(s).

In RR, the interest is in recovering quantities associated with the true (i.e. latent) categories using available observed categorical information. This is analogous to the situation with measurement error and criterion referenced testing. In the following, using techniques associated with RR designs, we establish results relating proportions of students in observed and latent categories. This forms the basis for the derivation of estimates of central tendency and dispersion for the multinomially distributed latent categorical variable. We thus establish a distributional foundation to understand how sampling and measurement error combine to undermine the precision and accuracy of school percent at performance level aggregates. By separating variability attributable to sampling and measurement, the results inform recent discussions of whether a school should be treated as a sample or a population.

Latent Parameter Estimation: Central Tendency

As above, let $\mathbf{P}$ denote the misclassification matrix of Equation 2 consisting of conditional misclassification probabilities on the latent and observed categorical variables A and A^*. Define the population density associated with A to be $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$. That is, $\Pr(A = j) = \pi_j$. Similarly, let $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ denote the population proportions associated with the observed categorical variable A^*, $\Pr(A^* = j) = \pi_j^*$. Following common convention, let $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}^*$ denote statistics corresponding to parameters $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$.

The misclassification matrix $\mathbf{P}$ is a transition matrix—a matrix with non-negative entries all of whose rows sum to 1. As such, $\mathbf{P} : A \longrightarrow A^*$ is a linear transformation from the set of latent category densities to the set of observed category densities defined by $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}^*$. Viewed in this way it is possible to consider the measurement procedure as a Markov process between latent and observed categories governed by the conditional misclassification probabilities given in $\mathbf{P}$. Under certain circumstances, this relationship yields an elegant means of estimating $\boldsymbol{\pi}$.

Proposition 1 (Moment Estimation of Latent Proportions) *If $\mathbf{P}$ is non-singular, then the unbiased moment estimator $\hat{\boldsymbol{\pi}}$ of $\boldsymbol{\pi}$ is given by*

$$\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}^* \mathbf{P}^{-1} \tag{4}$$

Proof. Since $\hat{\boldsymbol{\pi}}\mathbf{P} = \hat{\boldsymbol{\pi}}^*$, $\mathbb{E}(\hat{\boldsymbol{\pi}}\mathbf{P}) = \mathbb{E}(\hat{\boldsymbol{\pi}}^*) = \boldsymbol{\pi}^*$. Thus, $\mathbb{E}(\hat{\boldsymbol{\pi}}\mathbf{P}) = \boldsymbol{\pi}^* \Longrightarrow \mathbb{E}(\hat{\boldsymbol{\pi}}) = \boldsymbol{\pi}^*\mathbf{P}^{-1} = \boldsymbol{\pi}$. ∎

Multiplication by $\mathbf{P}^{-1}$ disattenuates the observed proportions, $\hat{\boldsymbol{\pi}}^*$, with regard to measurement error. Depending upon the misclassification probabilities, the estimated true proportions, $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k$, will be smaller/larger than their observed counterparts. The requirement that $\mathbf{P}$ be invertible is not burdensome in measurement situations. One can show that any diagonally dominant matrix is invertible (Strang, 1980, p. 304). And because the diagonal elements of $\mathbf{P}$ represent the conditional probabilities of correct classification, it is rarely ever the case (Selén, 1986, p. 77) in measurement situations that the diagonal elements do not dominate $\mathbf{P}$.[5]

In some instances, estimation by moments is problematic since multiplication by $\mathbf{P}^{-1}$ results in values lying outside the parameter space. In such cases, maximum likelihood estimation yields solutions on the boundary of the parameter space. However, if the moment estimator, $\hat{\boldsymbol{\pi}}$, represents a valid probability distribution with components from the interior of the parameter space, then the moment estimator coincides with its maximum likelihood counterpart.[6] The result, in many cases, simplifies the estimation of latent proportions.

[5] Note that diagonal dominance is sufficient but not necessary for invertibility.

[6] See Appendix B of van den Hout & van der Heijden (2002) for a proof of the proposition.

Proposition 2 *Let $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$ denote the moment estimator from Equation 4. If $0 < \hat{\pi}_j < 1$ for $j \in \{1, 2, \ldots, k\}$, and $\sum_j \hat{\pi}_j = 1$, then the moment estimator $\hat{\boldsymbol{\pi}}$ equals the maximum likelihood estimator.*

In situations where moment estimators are unacceptable, estimation of parameters via maximum likelihood can be performed using various software including R (R Development Core Team, 2006) with the poLCA library (Linzer & Lewis, 2006), LEM (Vermunt, 1997), and M-Plus. The two former programs are freely available.

An immediate and important consequence of Equation 4 is that the observed school percent at performance levels used for accountability are biased estimators of the true proportions. That is, observed percentages are not accurate estimates. The extent of bias is equal to

$$\mathsf{Bias}(\boldsymbol{\pi}^*) = \boldsymbol{\pi}^* - \boldsymbol{\pi} = \boldsymbol{\pi}(\mathbf{P} - \mathbf{I}).$$

Which is estimated by

$$\mathsf{Bias}(\hat{\boldsymbol{\pi}}^*) = \hat{\boldsymbol{\pi}}^* - \hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}^* - \hat{\boldsymbol{\pi}}^*\mathbf{P}^{-1} = \hat{\boldsymbol{\pi}}^*(\mathbf{I} - \mathbf{P}^{-1}). \tag{5}$$

Clearly, when measurement is perfect, the misclassification matrix is the identity matrix and there is no bias.[7] However, close examination of Equation 5 shows that bias depends upon *two* factors: the observed density and the misclassification matrix, $\hat{\boldsymbol{\pi}}^*$ of A^*. Thus, a single misclassification matrix can yield different bias for two different observed densities (e.g., two different schools).

We have shown that one effect of measurement error on percent at performance level aggregates is to introduce bias and thereby undermine the accuracy of the measures commonly used for accountability purposes. However, performance level misclassification affects more than just the accuracy of school level aggregates. Measurement error results in a "double whammy" (Carroll, Ruppert, & Stefanski, 1995, p. 22), eroding both the accuracy *and* precision of estimates. In the next section we illustrate the extent to which precision is compromised by establishing error variances and covariances for the estimates of Equation 4. The central tendency and dispersion results together provide the theoretical basis from which to discuss volatility in school percent at performance level results.

Latent Parameter Estimation: Dispersion

Despite the possibility of leading to incorrect AYP decisions, bias in percent at performance level statistics due to measurement error has garnered little to no research interest. Some researchers have incorrectly reported that no bias exists (Hebbler, 2004). Instead, researchers considering measurement error and its impact on percent at performance level measures (Coladarci, 2003; Hill, 2001; Cronbach et al., 1997) looked primarily at the extent to which the variability (i.e., the precision) of school aggregates was impacted. This research suggests that measurement error minimally impacts the variability associated with percent proficient statistics. This, we argue, is too simple an answer. Having demonstrated that measurement error introduces bias into percent at performance level statistics, in this section we derive results for the variance of $\hat{\boldsymbol{\pi}}$ and show the impact measurement error has on the precision of this statistic.

Following Chaudhuri & Mukerjee (1988, Section 3.3) and Greenland (1988), assume the observed density, $\hat{\boldsymbol{\pi}}^*$, is distributed multinomially with parameter $\boldsymbol{\pi}^*$. It follows that the variance-covariance matrix of $\hat{\boldsymbol{\pi}}^*$, denoted $\mathsf{Cov}(\hat{\boldsymbol{\pi}}^*)$, is given by

[7] In the rare circumstance that the observed density is the steady state distribution associated with the misclassification matrix $\mathbf{P}$, there will be no bias in the estimate even though misclassification occurs. In such an instance, the individual level misclassifications between performance levels "balance out".

$$\mathsf{Cov}(\hat{\boldsymbol{\pi}}^*) = \frac{\mathsf{Diag}(\boldsymbol{\pi}^*) - \boldsymbol{\pi}^{*'}\boldsymbol{\pi}^*}{n},$$

where $\mathsf{Diag}(\boldsymbol{\pi}^*)$ is a $k \times k$ matrix with the entries of $\boldsymbol{\pi}^*$ on the diagonal, $\boldsymbol{\pi}^{*'}$ is the transpose of $\boldsymbol{\pi}^*$, and n is the sample size on which $\hat{\boldsymbol{\pi}}^*$ is based. Application of the multivariate delta method (Rao, 1973, p. 388) to Equation 4 yields

$$\begin{aligned}
\mathsf{Cov}(\hat{\boldsymbol{\pi}}) &= \mathsf{Cov}\left(\hat{\boldsymbol{\pi}}^*\mathbf{P}^{-1}\right) \\
&= \left(\mathbf{P}^{-1}\right)'\mathsf{Cov}\left(\hat{\boldsymbol{\pi}}^*\right)\mathbf{P}^{-1} \\
&= \left(\mathbf{P}^{-1}\right)'\left(\mathsf{Diag}(\boldsymbol{\pi}^*) - \boldsymbol{\pi}^{*'}\boldsymbol{\pi}^*\right)\mathbf{P}^{-1}\Big/n.
\end{aligned}$$

An unbiased estimator of $\mathsf{Cov}(\hat{\boldsymbol{\pi}})$ is given by

$$\left(\mathbf{P}^{-1}\right)'\left(\mathsf{Diag}(\hat{\boldsymbol{\pi}}^*) - \hat{\boldsymbol{\pi}}^{*'}\hat{\boldsymbol{\pi}}^*\right)\mathbf{P}^{-1}\Big/(n-1). \tag{6}$$

To illustrate, consider a measurement instrument associated with $k = 2$ levels. Let the misclassification matrix, $\mathbf{P}$, and the observed proportions, $\hat{\boldsymbol{\pi}}^*$, be given by

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\pi}}^{*'} = \begin{pmatrix} \hat{\pi}_1^* \\ 1 - \hat{\pi}_1^* \end{pmatrix}.$$

Then applying the expression for the unbiased estimator given in (6) yields:

$$\frac{\hat{\pi}_1^*(1-\hat{\pi}_1^*)}{(n-1)(p_{11}+p_{22}-1)^2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{7}$$

Close examination of Matrix 7 indicates how and the extent to which performance level misclassification increases the error variance associated with $\hat{\boldsymbol{\pi}}$. For perfect measurement with no misclassification, $p_{11} = p_{22} = 1$, the error variance for $\hat{\boldsymbol{\pi}}$ equals the *sampling* error variance associated with the observed proportions: $\hat{\pi}_1^*(1-\hat{\pi}_1^*)/(n-1)$. In common measurement situations, $0.5 < p_{11}, p_{22} < 1$, the sampling error variance is multiplied by a factor that accounts for the misclassification, $(p_{11}+p_{22}-1)^2$. When the conditional probabilities of correct classification both approach 0.5—the worst case measurement scenario—error variance of $\hat{\boldsymbol{\pi}}$ rapidly increases.

In addition, because Matrix 7 has equal diagonal values, there is equal variability for the two levels of $\hat{\boldsymbol{\pi}}$. This is somewhat counterintuitive. When different misclassification rates exist for the two levels, one might suspect that the different levels would be measured with different degrees of precision. However, as Matrix 7 shows, even with different misclassification rates for the two levels, the error variances for both $\hat{\pi}_1$ and $\hat{\pi}_2$ are identical. However, in instances with more than two levels, it is rarely the case that the covariance matrix has equal diagonal elements. Thus different performance level proportions would have different error variability. This, in turn, could lead to complex inferences involving proportions in $k > 2$ categories.

It is instructive to represent the amount of variability contributed by misclassification by decomposing $\mathsf{Cov}(\hat{\boldsymbol{\pi}})$ into the sum of two covariance matrices—one representing multinomial sampling and the other misclassification due to measurement. This follows from an expansion of $\mathsf{Cov}(\hat{\boldsymbol{\pi}})$:

$$
\begin{aligned}
\mathsf{Cov}(\hat{\boldsymbol{\pi}}) &= \Big(\left(\mathbf{P}^{-1}\right)'\left(\mathsf{Diag}(\boldsymbol{\pi}^*) - \left(\boldsymbol{\pi}\mathbf{P}\right)'\boldsymbol{\pi}\mathbf{P}\right)\mathbf{P}^{-1}\Big)\Big/n \\
&= \Big(\left(\mathbf{P}^{-1}\right)'\left(\mathsf{Diag}(\boldsymbol{\pi}^*) - \mathbf{P}'\boldsymbol{\pi}'\boldsymbol{\pi}\mathbf{P}\right)\mathbf{P}^{-1}\Big)\Big/n \\
&= \Big(\left(\mathbf{P}^{-1}\right)'\mathsf{Diag}(\boldsymbol{\pi}^*)\mathbf{P}^{-1} - \boldsymbol{\pi}'\boldsymbol{\pi}\Big)\Big/n \\
&= \Big(\mathsf{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}'\boldsymbol{\pi} + \left(\mathbf{P}^{-1}\right)'\mathsf{Diag}(\boldsymbol{\pi}^*)\mathbf{P}^{-1} - \mathsf{Diag}(\boldsymbol{\pi})\Big)\Big/n \\
&= \underbrace{\Big(\mathsf{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}'\boldsymbol{\pi}\Big)\Big/n}_{Sampling} + \underbrace{\Big(\left(\mathbf{P}^{-1}\right)'\mathsf{Diag}(\boldsymbol{\pi}^*)\mathbf{P}^{-1} - \mathsf{Diag}(\boldsymbol{\pi})\Big)\Big/n}_{Misclassification} \qquad (8)
\end{aligned}
$$

Substituting statistics for parameters and dividing by $n - 1$ instead of n provides an unbiased estimator of $\mathsf{Cov}(\hat{\boldsymbol{\pi}})$. Note that whereas Matrix 7 specifies sampling error variance with regard to the observed proportions, $\hat{\boldsymbol{\pi}}^*$, the sampling component of Equation 8 is defined with regard to the estimated latent proportions, $\hat{\boldsymbol{\pi}}$.

The covariance decomposition of Equation 8 can be used to separate contributions based upon sampling assumptions and contributions based upon measurement error. This assists greatly in discussions of whether to treat a school as a sample drawn randomly from a idealized population or as a population unto itself. As Equation 8 implies, even if a school is treated as a population, the percent at performance level observations are biased estimates of the latent performance level percentages and are subject to variability based upon measurement error. In the next section we apply these results to typical scenarios within current accountability systems and demonstrate the impact that performance level misclassification has upon the bias and variability of percent at performance level measures.

Discussion

In the previous section we established, based upon a known misclassification matrix and observed proportions in each of k mutually exclusive categories, the mean and covariance of the multinomial distribution of true proportions. In the context of percent at performance level aggregates for schools, our result provides the mean and variance of the underlying distribution from which the observed proportions are theoretically drawn. Using terminology from generalizability theory (Brennan, 1992), this approach distinguishes two universes of admissible observations: those based upon measurement and those based upon students. Treating both of these universes as infinite in size yields a quantification of school quality similar to a student's observed score:

> The result for a school for one year is just one observation from which to infer a school's true score—what the school's average would be if we could test an infinite number of students from the school's catchment area an infinite number of times on all the test questions that might be asked (Hill, 2002, p. 2)

Imprecision in school estimates, considering infinite admissible universes of students and tests, derives from both sampling and measurement, with sampling receiving the majority of concern. However, it is relevant to define a "school's true score" based upon a finite number of students and an infinite number of testing occasions. Doing so, brings to the fore the question of whether to treat students at a school as a sample or a population.

There has been considerable debate about whether the assessment results of a school's cohort of students represents a sample or a population—that is, whether the universe of admissible students

extends beyond those in the current student body (Hill & DePascale, 2002; Cronbach et al., 1997; Yen, 1997). The results of those discussions manifest themselves in decisions about whether to place sample based confidence intervals about the percent at performance level estimates. Many authors advocating the use of such confidence intervals cite Cronbach et al. (1997, p. 393) in their defense:

> To conclude on the basis of an assessment that a school is effective as an institution requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population.

Though correct, the statement is a red herring. Inferring effectiveness requires more than placing a confidence interval about a statistic. Indeed, one of the most challenging issues in growth modeling using student assessment data is in trying to make effectiveness claims based upon observational data. Unless certain design issues are met, judging a school to be effective based upon percent of proficient students is not defensible, with or without confidence bands.[8] Moreover, if such confidence intervals embolden users into believing that it is safe to make school effectiveness claims, then perhaps their use should be avoided.

To be clear, we are not arguing against the use of sample based confidence intervals about school percent proficient aggregates in every instance. However, it is important that their use align with a defensible purpose. Describing how well a fixed student body performs on a test in the presence of measurement error is both non-trivial and relevant. The latent class approach developed in this paper is useful in this regard because it concisely separates the contributions of measurement and sampling:

1. Performance level misclassification affects *both* the accuracy and precision of percent at performance level estimates. Bias, defined in Equation 5, quantifies the extent to which measurement error affects accuracy. The impact of measurement error on precision is given in Equation 8.

2. Sampling error (based upon an assumption of random sampling) affects *only* the precision of percent at performance level estimates.

Measurement error is present regardless of whether students are treated as a sample or as a population. In this sense, it is the more pervasive of the two sources of error. Adjustment for bias provides a more accurate estimate of percent proficient for either the sample or population. Error variability due to measurement (Equation 8) comprises imprecision in the latent percent proficient estimate due to measurement error in either the sample or population. And lastly, sampling variability indicates additional imprecision in the latent percent proficient estimate based upon the desire to consider the current student body as a random sample from an abstract population.

This approach differs from Kane & Staiger (2002), Arce-Ferrer et al. (2002), and Linn & Haug (2002) who examine year-to-year differences. This, we think, complicates the issue because of the changing nature of cohorts, tests, and schools over time. In order to facilitate year-to-year comparisons, it is imperative to first understand performance, as Hill (2002) suggests, in a given year. In that regard, the results in this paper establish the precision and accuracy of percent at performance level aggregates for a single year. Based upon the mean and variance of the latent distribution, it is possible to examine subsequent years and determine whether these results are likely to have come from the same distribution. *However*, if a significant difference in percent proficient occurs,

[8] In "defense" of NCLB, if one assumes, for example, that student cohorts from year to year, for a given school, represent random samples from an unchanging population, then using percent proficient to make inferences concerning increasing/decreasing effectiveness for the school are quite defensible. Of course, this random sampling assumption is debatable.

it is problematic to infer a change in school effectiveness without assuming that the two cohorts of students taught in subsequent years represent random draws from the same population—a questionable assumption. If no significant changes occur, then all that can be argued is that the results are consistent with random variation that would be expected based upon a random draws from a single population.

To illustrate these issues, in the following sections we discuss how accuracy and precision of percent proficient statistics for schools is impacted based upon measurement and sampling considerations. The results are pertinent to NCLB based accountability systems where school AYP/AMO compliance is contingent upon these values. The results show that measurement error should not be dismissed as inconsequential to decisions made based upon percentages of student at a given performance level.

Bias/Accuracy

Bias in a school's observed percent proficient, as defined in Equation 5, is a function of *both* the observed percent proficient, $\hat{\boldsymbol{\pi}}^*$, and the misclassification matrix $\mathbf{P}$. Due to dependencies in the components of $\hat{\boldsymbol{\pi}}^*$ and $\mathbf{P}$, bias in the observed percent proficient can be expressed as function of, at a minimum, three variables: p_{11}, the conditional correct classification rate for proficiency, p_{22}, the conditional correct classification rate for non-proficiency, and $\hat{\pi}_1^*$, the observed percent proficient.[9] It is straightforward to show that the bias in the observed percent *not*-proficient is the additive inverse of the bias in the observed percent proficient.

Figure 1 illustrates bias in observed percent proficient ($\hat{\pi}_1^* - \hat{\pi}_1 =$ observed - true) as a function of the conditional correct classification rates, p_{11} and p_{22}, across nine observed percent proficient levels ranging from 10 percent to 90 percent (represented by the 9 panels). To aid interpretation, the panels use color to indicate both the direction and extent of bias. Portions of the figures shaded magenta indicate conditions under which the observed percent proficient *over-estimates* the true percent proficient. Conversely, portions shaded blue depict conditions under which the observed percent proficient *under-estimates* the true percent proficient. Areas that are white indicate no bias. Lines in each panel are contours indicating the extent of bias, which is concurrently shown with depth of color (darker color indicates more bias).

For low and high levels of percent proficient (e.g., 10, 20, 80, and 90 percent proficient) the figures quantify the attenuation in observed percent proficient due to measurement error. In almost all cases where observed percent proficient is low, 10 or 20 percent, the result is an overestimate of the true percent proficient. Moreover, the extent of bias is almost exclusively dependent upon the conditional classification rate associated with non-proficiency, p_{22}. This is consistent with the fact that most students at such schools reside in that category. Therefore, misclassification with regard to that category is the primary determinant of bias. Conversely, in high performing schools, the observed results under estimate the true percent proficient with p_{11} being the dominant conditional classification rate determining the extent of bias. For schools near parity with proficient/non-proficient students, under-estimation and over-estimation are both possible depending upon the amount of misclassification associated with both proficient and non-proficient categories.

The consequences of underestimating percent proficient for high performing schools is a threat to the validity of accountability systems. As 2014 approaches NCLB AYP/AMO mandates move closer to 100 percent, schools close to meeting AYP/AMO targets using observed percent proficient would possibly reach the AYP/AMO if the unbiased estimate was used to determine compliance. The result is equally problematic with regard to safe harbor criteria. NCLB compliance via safe harbor is predicated upon an annual decrease of non-proficient students. However, failing to account

[9] Other choices of three variables are possible, for example, p_{12}, p_{21}, and $\hat{\pi}_2^*$

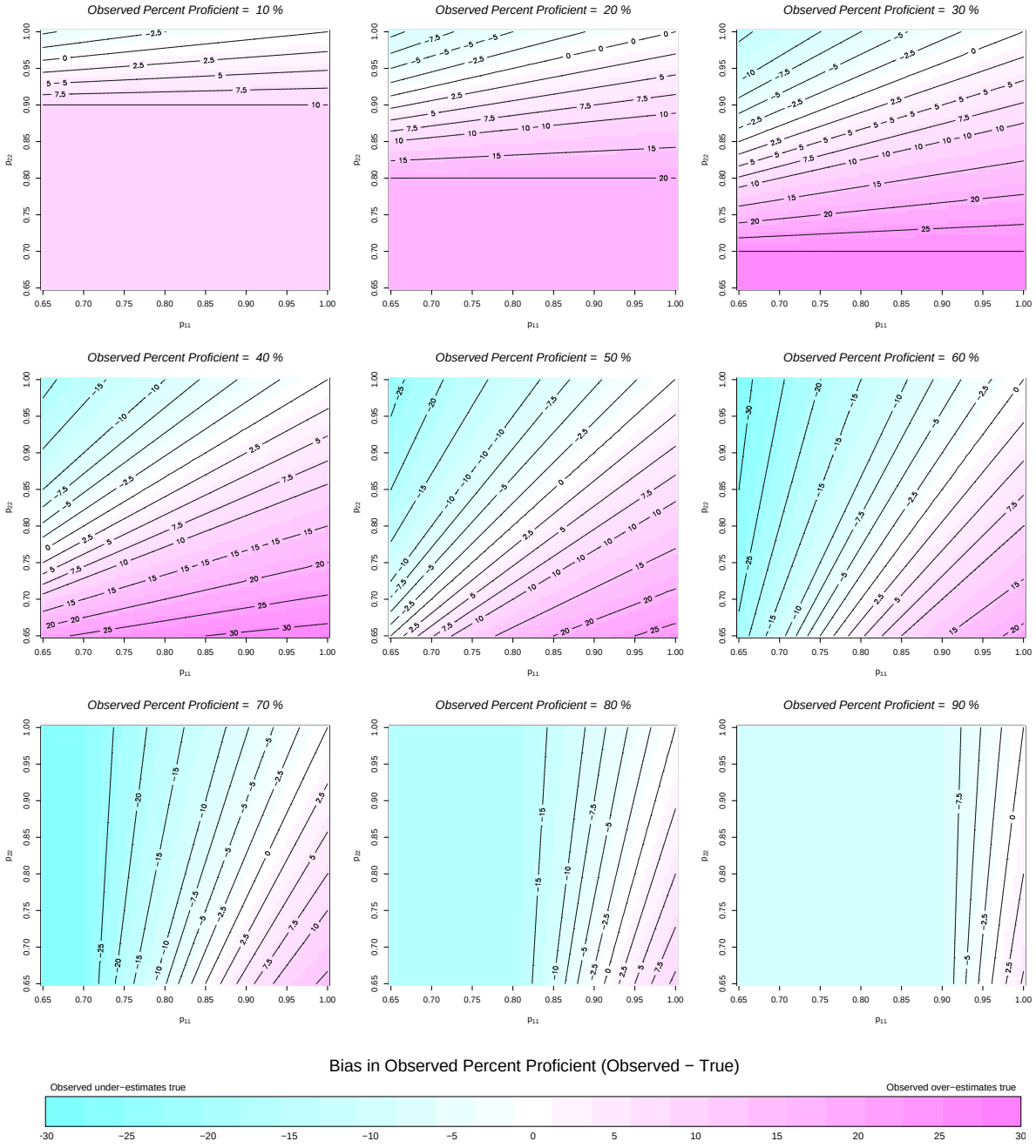

Figure 1: Bias in observed percent proficient as a function of correct conditional classification rates, p_{11} and p_{22}, and observed percent proficient, $\hat{\pi}_1^*$ (See Equation 5). The 9 panels represent 9 different levels of observed percent proficient, $\hat{\pi}_1^*$, ranging from 10 to 90 percent. Areas shaded blue denote conditions where the observed percent proficient under-estimates the true percent proficient whereas areas shaded magenta denote where observed percent proficient over-estimates true percent proficient. Areas in white denote the absence of bias. Contours and color depth represent the extent of bias.

for misclassification bias will attenuate the different in non-proficient students toward zero leading to a consistent underestimate of any decrease that might occur (Schwartz, 1985).

An important factor impacting the misclassification rates, is the stringency level of the state's proficiency threshold. That is, differences in the conditional correct classification rates, p_{11} and p_{22}, depend, in some part, upon where the proficiency cutpoint is set. Based upon an examination of state assessment data, when the proficiency threshold is low relative to the group tested (e.g., at the 20th percentile), p_{11} for schools is generally quite high, $0.9 \leq p_{11} \leq 1$, while p_{22} is less so, $0.7 \leq p_{22} \leq 0.85$. The opposite holds when the proficiency threshold is high. In such cases, the conditional probability of correctly classifying a student proficient, p_{11} is generally substantially less than p_{22}. This relationship between p_{11} and p_{22} and the cut-score level is consistent with the findings of Livingston & Lewis (1995, p. 183). In their Table 2, as the proficient threshold increases, values of p_{11} decrease from 0.984 to 0.876 while the values of p_{22} increase from 0.757 to 0.979.

The stringency of a state's proficiency threshold can differentially impact high and low performing schools due to differential bias resulting from misclassification. This result contradicts the claims of some authors that "reliability" has no bearing upon accountability decisions (Rogosa, 2004). Lack of reliability, manifest as performance level misclassification, induces an "unsuspected bias" (Cochran, 1968, p. 637) that can lead to erroneous judgments about AYP/AMO. For example, in a state with challenging proficiency standards, p_{11} is likely to range between 0.7 and 0.9 for schools. Differences in p_{11} have little impact on low performing schools since most of their students are not-proficient and bias results from misclassification of those students, reflected in p_{22}. However, for schools with higher percentages of proficient students, $p_{11} < 0.9$ can lead to a substantial underestimation (10 to 15 percent) of the percent proficient at the school. Clearly, this underestimation can have consequences regarding AYP/AMO determination. Future research will address issues along these lines.

Variance/Precision

As mentioned previously, research into measurement error and its impact upon percent proficient statistics has focused exclusively upon precision. Based upon generalizability theory analyses, the results suggest that measurement error contributes substantially less to overall variability than does sampling error. Broadly, this result coincides with our findings for a majority of schools with between 20 and 80 percent proficient students. However, for extreme schools with less than 15 percent and more than 85 percent proficient, due to the small sampling error at the extremes, measurement error dominates sampling error in eroding precision. As decomposition of variance in Equation 8 demonstrates, for each school, the extent to which measurement error, manifest as performance level misclassification, adds to the variability of $\hat{\boldsymbol{\pi}}$ is a function of the misclassification rates and the observed percent proficient at the school.

To demonstrate the contribution to total variability of estimated true percent proficient, $\hat{\pi}_1$, Figure 2 depicts the total variability (sampling + measurement) relative to that supplied by sampling. Formation of the ratio eliminates sample size as a consideration. Following the depiction of bias in Figure 1, the ratio of total variability to sampling variability is represented as a function of 3 variables: the conditional correct classification rates, p_{11} and p_{22}, and the observed percent proficient, $\hat{\pi}_1^*$. The 9 panels represent nine different observed percent proficient levels ranging from 10 to 90 percent. The panels are shaded to indicate significant measurement error contribution to the variability of $\hat{\pi}_1$ above and beyond sampling. A ratio of 1 (white portions of the panels) indicates no measurement error.

The panels demonstrate similarities to those associated with bias. When observed percent proficient is low, the conditional correct classification rate associated with non-proficiency, p_{22}, displays greater relevance in terms of variability of $\hat{\pi}_1$ than does the conditional correct classification rate

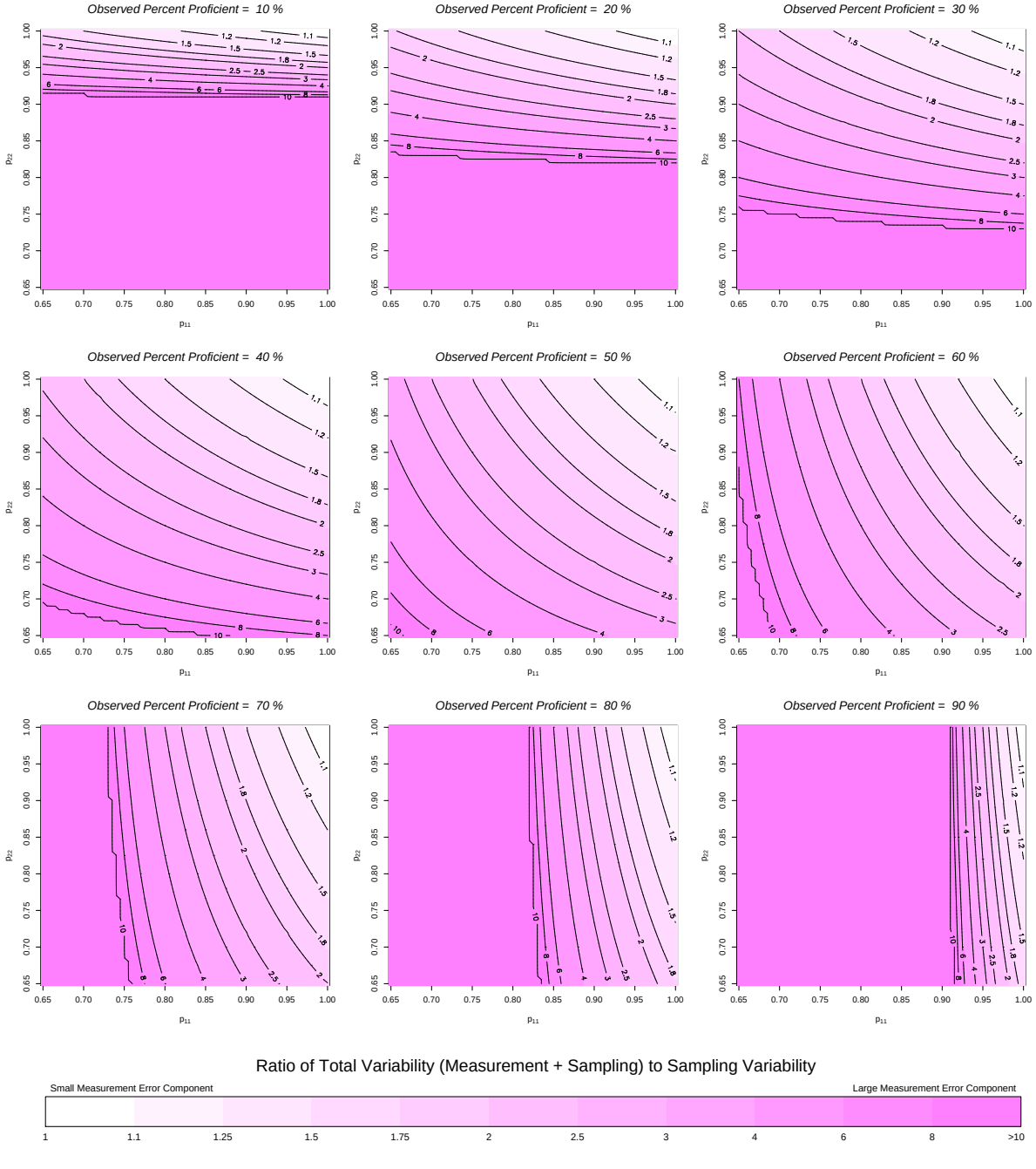

Figure 2: Ratio of total variability (measurement + sampling) to sampling variability of estimated true percent proficient, $\hat{\pi}_1$, as a function of correct classification rates, p_{11} and p_{22}, observed percent proficient, $\hat{\pi}_1^*$, and sample/population size (See Equation 8). The 9 panels represent 9 different levels of observed percent proficient, $\hat{\pi}_1^*$ ranging from 10 to 90 percent. Areas shaded magenta indicate conditions with a large measurement error component, with deeper shades reflecting greater contribution.

associated with proficient, p_{11}. For schools with high observed percent proficient the converse is true, p_{11} dominates instead of p_{22}. As observed percent proficient and non-proficient approach parity, the situation is more complicated, both p_{11} and p_{22} can impact the variability of $\hat{\pi}_1$. The panels also illustrate the "bias versus variance tradeoff" (Carroll et al., 1995, p. 32)—as reduction in bias yields an increase in estimate variability.

For schools with extremely high or low observed percent proficient, sampling error associated with $\hat{\pi}_1$ is small. Thus, even with excellent conditional classification rates above 0.9, measurement error variability can be more than 10 times as large as sampling variability. The square root of the ratio provides the total standard error relative to the sampling standard error. In schools with moderate observed percent proficient, modest conditional classification rates can doubling error variability associated with sampling. For example, with 50 percent observed percent proficient, if $p_{11} = p_{22} = 0.85$, then then measurement error approximately doubles the ratio. Hence, error variability from measurement equals that of sampling. Our empirical investigations of schools have shown conditional classification probabilities as low at 0.75 suggesting that in some schools with modest numbers of students (n in Equation 8), measurement error can lead to extremely imprecise estimates.[10] Future research will address the impact of these issues with regard to AYP/AMO decision making.

Conclusion

By synthesizing classification accuracy research with techniques associated with randomized response designs, this paper establishes results that quantify the amount of bias and variability due to measurement error in percent at performance level aggregates commonly used in state accountability systems. As such, the results provide a means to quantify the amount of inaccuracy and imprecision due to performance level misclassification in percent at performance level aggregates. The results provide a link between imprecision at the individual level and how that imprecision manifests as bias and imprecision at the group level. Because accountability systems rely so heavily upon aggregates of student performance, it is imperative that misclassification be considered as a significant source of bias and variability in those results.

The techniques presented in the paper raise a number of questions including:

- Should percent proficient reported for schools be adjusted for measurement error?
- Should confidence intervals be placed about observed or adjusted percent proficient scores?
- Should confidence intervals about percent proficient include contributions associated with misclassification?
- Should a school's cohort of students be considered a sample or a population?

Answers to each of these questions is critical for the implementation of valid and fair accountability systems. In future research, we will employ the techniques derived herein to investigate the practical outcomes that misclassification has on accountability systems.

[10] When the collection of students investigated is consider a sample, variance estimates like those of Equation 8 assume the observations are independent of one another. It is well documented, that in education settings, the assumption of independent observations is false yielding smaller confidence intervals than is likely the case. Thus, the n used in the denominator is likely an overestimate of the effective sample size (Snijders & Bosker, 1999). Confidence bands currently employed to account for sampling variability implicitly assume independent observations and thus infer greater precision than is actually the case.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, New Jersey: Wiley.

Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment*, *8*(1), 59–75.

Brennan, R. L. (1992). *Elements of generalizability theory* (Revised ed.). Iowa City: ACT Publications.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics*, *10*, 478–486.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models*. London: Chapman & Hall.

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized response: Theory and techniques*. New York: Marcel Dekker, Inc.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*(4), 637–666.

Coladarci, T. (2003, Oct). *Gallup goes to school: The importance of confidence intervals for evaluating "adequate yearly progress" in small schools* (Policy Brief). Washington D.C.: The Rural School and Community Trust.

Coleman, J. S. (1964). *Models of change and response uncertainty*. Englewood Cliffs, NJ: Prentice-Hall.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399.

Freund, J. E., & Perles, B. M. (2007). *Modern elementary statistics* (12th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Gnedenko, B. V. (1967). *The theory of probability* (4th ed.; B. D. Seckler, Trans.). New York: Chelsea.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, *7*, 745–757.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(1), 159–170.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345–359.

Hebbler, S. (2004, June). *Some threats to validity in adequate yearly progress (AYP) models under NCLB.* (Paper Presented at the 2004 CCSSO Conference on Large-Scale Assessment, Boston, MA. Retrieved August 14, 2006 from http://www.ccsso.org/content/pdfs/LargeScale2004Hebbler.pdf#search=%22bias%20in%20percent%20proficient%22)

Hill, R. K. (2001). *Issues related to the reliability of school accountability systems* (Tech. Rep.). Dover, NH: The National Center for the Improvement of Educational Assessment.

Hill, R. K. (2002, April). *Examining the reliability of accountability systems.* (Paper presented at the 2002 meeting of the American Educational Research Association, New Orleans (Retrived August 10th, 2006 from http://www.nciea.org/publications/NCME_RHCD03.pdf)

Hill, R. K., & DePascale, C. (2002, December). *Determining the reliability of school scores* (Tech. Rep.). Washington, D.C.: Council of Chief State School Officers (CCSSO). (Retrived July 10th, 2006 from http://www.ccsso.org/content/pdfs/DeterminingReliability.pdf)

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 235–283). Washington, D.C.: Brookings Institution.

Kuha, J., & Skinner, C. (1997). Categorical data and misclassification. In L. Lyberg et al. (Eds.), *Survey measurement and process quality* (pp. 633–670). New York: John Wiley & Sons.

Kupermintz, H. (2004). On the reliability of categorically scored examinations. *Journal of Educational Measurement*, *41*(3), 193–204.

Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*(4), 412–432.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29–36.

Linzer, D. A., & Lewis, J. (2006). *poLCA: Polytomous variable latent class analysis* [Computer program, version 0.9]. (Available online from http://dlinzer.bol.ucla.edu/poLCA)

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Martineau, J. A. (2006). *An expansion and practical evaluation of expected classification accuracy.* (To appear in Applied Psychological Measurement)

Mote, V. L., & Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. *Biometrika*, *52*, 95–109.

Noble, A. J. (2004, May). *Testing: Not an exact science* (Vol. 16; Education Policy Brief). Newark, Delaware: University of Delaware, Education Research & Development Center.

No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425.

R Development Core Team. (2006). *R: A language and environment for statistical computing.* Vienna, Austria. (3-900051-07-0)

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.

Rogosa, D. R. (1994). *Misclassification in student performance categories* (CLAS Technical Report). Monterey, CA: CTB/McGraw-Hill. (Appendix to CLAS draft technical report)

Rogosa, D. R. (2004, Oct). *Irrelevance of reliability coefficient to accountability systems: Statistical disconnect in Kane-Staiger "Volatility in School Test Scores".* (Retrieved from the World Wide Web, August 2, 2006, from http://www-stat.stanford.edu/~rag/api/kscresst.pdf)

Rogosa, D. R. (2005). Statistical misunderstandings of the properties of school scores and school accountability. In J. L. Herman & E. H. Haertel (Eds.), *Yearbook of the National Society for the Study of Education.* Chicago: National Society for the Study of Education.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(4). (Retrieved July 15, 2006 from http://PAREonline.net/getvn.asp?v=7&n=14)

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, *10*(13). (Retrieved July 15, 2006 from http://PAREonline.net/getvn.asp?v=10&n=13)

Schwartz, J. E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research*, *13*(4), 435–466.

Selén, J. (1986, March). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, *81*(393), 75–81.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage Publications.

Stallings, W. M., & Gillmore, G. M. (1971). A note on "accuracy" and "precision". *Journal of Educational Measurement*, *8*(2), 127–129.

Strang, G. (1980). *Linear algebra and its applications* (2nd ed.). New York: Harcourt Brace Jovanovich.

van den Hout, A., & van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, *70*(2), 269–288.

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data* [Computer program, version 1.0]. Tilburg, The Netherlands: Department of Methodology and Statistics, Tilburg University.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating answer bias. *Journal of the American Statistical Society*, *60*, 63–69.

Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. San Francisco: Jossey-Bass/Elsevier.

Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement: Issues and Practice*, *16*, 5–15.

Young, M. J., & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (Tech. Rep.). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. (CSE Technical Report 475)