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

Damian W. Betebenner
National Center for the Improvement of Education, Dover, NH
Yi Shang, Yun Xiang, Yan Zhao, and Xiaohui Yue
Department of Educational Research, Measurement and Evaluation
Lynch School of Education, Boston College

No Child Left Behind (NCLB) performance mandates, embedded within state accountability systems, focus school AYP (adequate yearly progress) 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 article, 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 (Kane & Staiger, 2002; Linn, Baker, & Betebenner, 2002). 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 adequate yearly progress (AYP) determination (Arce-Ferrer, Frisbie, & Kolen, 2002; Kane & Staiger, 2002; Linn & Haug, 2002; Rogosa, 2005; 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
Betebenner et al.

upon random sampling assumptions to insulate schools, particularly those with few students, from this volatility (Coladarci, 2003). Often discounted in the discussion of volatility is measurement error (Coladarci, 2003; Cronbach, Linn, Brennan, & Haertel, 1997; Hill, 2001). 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 and Novick (1973), Livingston and 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 an 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 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:

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

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 (Lee, Hanson, & Brennan, 2002; Livingston & Lewis, 1995; Martineau, in press; Noble, 2004; Rudner, 2001, 2005; Young and Yoon, 1998).
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 (Hanson & Brennan, 1990; Kupermintz, 2004; Rogosa, 1994). Following Coleman (1964), Wiggins (1973), Rogosa (1994), and Clauser, Margolis, and Case (2006), 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 matrix \( P \), consisting of conditional probabilities, is given by:

\[
P = \{p_{ij}\}_{1 \leq i, j \leq k} \quad \text{where} \quad p_{ij} = \Pr(A^* = j \mid A = i).
\]

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, \( \Pr(A^* = 1 \mid A = 2) \) and \( \Pr(A^* = 2 \mid A = 1) \), respectively. The complements of false-positive and false-negative error rates are the conditional correct classification rates, \( \Pr(A^* = 1 \mid A = 1) \) and \( \Pr(A^* = 2 \mid A = 2) \).

Numerous procedures exist to estimate conditional probabilities used to construct the misclassification matrix \( P \). 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:

\[
p_{ij} = \frac{\sum_{x \in \text{PL}_j} \sum_{\tau \in \text{PL}_i} \Pr(X = x, T = \tau)}{\sum_{\tau \in \text{PL}_i} \Pr(T = \tau)} \cdot g(\tau) / \sum_{\tau \in \text{PL}_i} \Pr(T = \tau).
\]

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 \( \text{PL}_i \) and \( \text{PL}_j \), and division by the marginal true score density of \( \text{PL}_i \) 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. The Hanson and Brennan procedure is also notable for distinguishing between a cut score on the true score scale and the corresponding “advancement” score on the observed score scale (Brennan, 1981), adding another layer of complexity to classification analyses. Rudner (2001, 2005) and Martineau (in press) simplify the estimation procedures by making strong normality assumptions for the conditional error distribution as well as the student true score density. Mislevy (as cited by Noble, 2004) employs simulation techniques to derive the misclassification
matrix, \( P \). Though determining which method yields the best estimates is a topic worthy of investigation, we take the misclassification matrix \( P \) 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), \( \Pr(X = x \mid T = \tau) \), 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 implicitly weight \( \Pr(X = x \mid T = \tau) \) according to this distribution. Indeed, conditional misclassification probabilities include information from both the measurement instrument and how individuals in the sample/population are situated relative to the cut scores used to define the performance levels. It is almost always the case that different populations—specifically, different schools or districts—have different true score densities, thus yielding different misclassification rates and matrices. 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.

The pooled state level misclassification matrix can be seen as an “average” of the school level misclassification matrices. As we demonstrate in the following sections, calculating school level misclassification matrices is desirable to investigate the impact of individual measurement error on the accuracy and precision of percent at performance level statistics commonly reported for schools. Deriving such school level misclassification rates allows for an understanding of the extent to which measurement error might infect aggregate summaries and permit measurement error adjustments if deemed necessary.

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 \( P \) is a transition matrix defining a mapping between latent and observed distributions (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 response or nonresponse. 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 (Chaudhuri & Mukerjee, 1988; van den Hout & van der Heijden, 2002). 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, he or she is able to recover an estimate of the true percentage of students who smoke marijuana at least once a week.

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 $P$ denote the misclassification matrix of equation (2) consisting of conditional misclassification probabilities on the latent and observed
categorical variables $A$ and $A^\ast$. Define the population density associated with $A$ to be $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$. That is, $\Pr(A = j) = \pi_j$. Similarly, let $\pi^\ast = (\pi^\ast_1, \pi^\ast_2, \ldots, \pi^\ast_k)$ denote the population proportions associated with the observed categorical variable $A^\ast$. $\Pr(A^\ast = j) = \pi^\ast_j$. Following common convention, let $\hat{\pi}$ and $\hat{\pi}^\ast$ denote statistics corresponding to parameters $\pi$ and $\pi^\ast$.

The misclassification matrix $P$ is a transition matrix—a matrix with nonnegative entries all of whose rows sum to 1. As such, $P$: $A \rightarrow A^\ast$ is a linear transformation from the set of latent category densities to the set of observed category densities defined by $\pi P = \pi^\ast$. 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 $P$. Under certain circumstances, this relationship yields an elegant means of estimating $\pi$.

**Proposition 1 (Moment Estimation of Latent Proportions).** If $P$ is nonsingular, then the unbiased moment estimator $\hat{\pi}$ of $\pi$ is given by

$$\hat{\pi} = \hat{\pi}^\ast P^{-1}$$

**Proof.** Since $\hat{\pi} P = \hat{\pi}^\ast$, $\Sigma(\hat{\pi} P) = \Sigma(\hat{\pi}^\ast) = \pi^\ast$. Thus, $\Sigma(\hat{\pi} P) = \pi^\ast \Longrightarrow \mathbb{E}(\hat{\pi}) = \pi^\ast P^{-1} = \pi$. ■

Multiplication by $P^{-1}$ disattenuates the observed proportions, $\hat{\pi}^\ast$, 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 $P$ 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 $P$ 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 $P$.

In some instances, estimation by moments is problematic since multiplication by $P^{-1}$ 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, $\hat{\pi}$, represents a valid probability distribution with components from the interior of the parameter space, then the moment estimator coincides with its maximum likelihood counterpart. The result, in many cases, simplifies the estimation of latent proportions.

**Proposition 2.** Let $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$ 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 $\hat{\pi}$ 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

\[ \text{Bias}(\pi^*) = \pi^* - \pi = \pi(P - I), \]

which is estimated by

\[ \text{Bias}(\hat{\pi}^*) = \hat{\pi}^* - \hat{\pi} = \hat{\pi}^*P^{-1} = \hat{\pi}^*(I - P^{-1}). \] (5)

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

The dependence of bias on the observed density complicates the way that measurement error/misclassification is approached at the individual level. Namely, given a single misclassification matrix, two different observed score densities associated with two different schools will result in different amounts of bias. Moreover, because misclassification is dependent upon the true score density associated with the students in the unit of aggregation, schools with different observed score densities might possess different misclassification matrices, yielding different amounts of bias. The practical effect of such differences is minor in most instances.

To investigate the practical effect, 2005 state assessment data was used to calculate school misclassification matrices which were used to estimate bias and precision of observed percent proficient statistics. The state assessments used to produce scale scores contain both multiple-choice and constructed-response items. For multiple-choice items, a three-parameter logistic model is used. For constructed-response items, the two-parameter partial credit model is used. The Livingston-Lewis procedure (Livingston & Lewis, 1995) was used to generate misclassification matrices. Though the test employed a vertical scale developed via a common item equating design, because misclassification matrices are based upon assessments within a given year, these equating procedures have no impact upon the result.

Using state assessment data, misclassification matrices were estimated for each of approximately 1,500 schools and used to derive measurement error adjusted percents proficient and nonproficient. For 75% of the schools, bias in percent proficient was less than 1.5%. Similarly, bias in reported non proficient students was less than 1.5% for approximately three in four schools. However, for 5% of the schools examined, bias in both percent proficient and non proficient exceeded 5%. For those schools demonstrating greater than 5% bias, conditional proficient classification rates, \( p_{11} \) ranged between 0.6 and 0.92. Those with greater than 5% bias in observed percent non proficient had conditional non proficient classification rates, \( p_{22} \) ranging between 0.5 and 0.7.

We have shown that one effect of measurement error on percent at performance level aggregates is introduce bias and thereby undermine the accuracy of the measures commonly used for accountability purposes. However, performance level
Betebenner et al.
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; Cronbach et al., 1997; Hill, 2001) 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 $\hat{\pi}$ and show the impact measurement error has on the precision of this statistic.

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

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

where $\text{Diag}(\pi^{*})$ is a $k \times k$ matrix with the entries of $\pi^{*}$ on the diagonal, $\pi^{*}$ is the transpose of $\pi^{*}$, and $n$ is the sample size on which $\hat{\pi}^{*}$ is based. Application of the multivariate delta method (Selén, 1986) to equation (4) yields

$$\text{Cov}(\hat{\pi}) = \text{Cov}(\hat{\pi}^{*}P^{-1}) = (P^{-1})' \text{Cov}(\hat{\pi}^{*}) P^{-1} = (P^{-1})'(\text{Diag}(\pi^{*}) - \pi^{*}\pi^{*}) P^{-1}/n.$$ 

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

$$(P^{-1})'(\text{Diag}(\hat{\pi}^{*}) - \hat{\pi}^{*}\hat{\pi}^{*}) P^{-1}/(n - 1). \quad (6)$$

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

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \quad \text{and} \quad \hat{\pi}^{*} = \begin{pmatrix} \hat{\pi}^{*}_{1} \\ 1 - \hat{\pi}^{*}_{1} \end{pmatrix}.$$
Then applying the expression for the unbiased estimator given in equation (6) yields:

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

Close examination of matrix 7 indicates how and the extent to which performance level misclassification increases the error variance associated with \(\hat{\pi}^*\). For perfect measurement with no misclassification, \(p_{11} = p_{22} = 1\), the error variance for \(\hat{\pi}^*\) equals the sampling error variance associated with the observed proportions: \(\frac{\hat{\pi}_1^*(1 - \hat{\pi}_1^*)}{n - 1}\). 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 \(\hat{\pi}^*\) rapidly increases.

In addition, because Matrix 7 has equal diagonal values, there is equal variability for the two levels of \(\hat{\pi}^*\). 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 \(\text{Cov}(\hat{\pi}^*)\) into the sum of two covariance matrices—one representing multinomial sampling and the other misclassification due to measurement. This follows from an expansion of \(\text{Cov}(\hat{\pi}^*)\):

\[
\text{Cov}(\hat{\pi}^*) = ((P^{-1})' (\text{Diag}(\pi^*) - (\pi P) \pi P^{-1})/n
= ((P^{-1})' (\text{Diag}(\pi^*) - P' \pi P \pi P^{-1})/n
= ((P^{-1})' \text{Diag}(\pi^*) P^{-1} - \pi P^{-1} \pi P^{-1})/n
= (\text{Diag}(\pi^*) - \pi P^{-1} \pi P^{-1} - \text{Diag}(\pi))/n
= (\text{Diag}(\pi^*) - \pi P^{-1} \pi P^{-1} - \text{Diag}(\pi))/n + ((P^{-1})' \text{Diag}(\pi^*) P^{-1} - \text{Diag}(\pi))/n \tag{8}
\]

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

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 an 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, 2001), this approach considers a universe of generalization consistent of two facets: measurement and students. Treating both universes as infinite in size yields a quantification of school quality similar to a student’s observed score:

The result for a school for 1 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 (Cronbach et al., 1997; Hill & DePascale, 2002; 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. Moreover, if such confidence intervals

128
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 nontrivial 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. Student 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 a hypothetical population.

This approach differs from Kane and Staiger (2002), Arce-Ferrer et al. (2002), and Linn and 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, 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 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, \( \hat{\pi}^* \), and the misclassification matrix \( P \). Due to dependencies in the components of \( \hat{\pi}^* \) and \( P \), 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 nonproficiency, and \( \hat{\pi}_1^* \), the observed percent proficient.\(^8\) 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 (\( \hat{\pi}_1^* - \hat{\pi}_1 = \text{observed - true} \)) as a function of the conditional correct classification rates, \( p_{11} \) and \( p_{22} \), across nine observed percent proficient levels ranging from 10% to 90% (represented by the nine panels). To aid interpretation, the panels use gray-scale depth to indicate the extent of bias. Portions of the figures which are white indicate conditions under which the observed percent proficient is an unbiased estimate of the true percent proficient. Conversely, portions shaded gray depict conditions under which the observed percent proficient is a biased estimate of true percent proficient. Lines in each panel are contours indicating the extent of bias.

In almost all cases—situations based upon different conditional classification rates \( p_{11} \) and \( p_{22} \)—where observed percent proficient is low, 10 or 20%, the result is an overestimate of the true percent proficient. That is, the true percentage of students rated as proficient in low-achieving schools is almost certainly less than the observed
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 nine panels represent nine different levels of observed percent proficient, $\hat{\pi}_1$, ranging from 10 to 90%. Areas shaded gray denote conditions where the observed percent proficient over/underestimates the true percent proficient. Areas in white denote the absence of bias. Contours and gray-scale depth represent the extent of bias.
percent proficient commonly reported. This is paradoxical when compared with how
measurement error affects individual student scores (S. Livingston, April 10, 2006
review of NCME presentation). Students with low-observed scores are likely to have
higher true scores, yet schools with high-observed percentages of such students are
likely to have higher true percentages. The findings are consistent after one considers
how a performance level threshold interacts with the distribution of student level
observed and true scores.

The extent of bias for schools with low-observed percent proficient 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. In high-performing schools, the observed results under-
estimate the true percent proficient with \( p_{11} \) being the dominant conditional classifi-
cation rate determining the extent of bias. For schools near parity with proficient/
nonproficient students, underestimation and overestimation are both possible de-
pending upon the amount of misclassification associated with both proficient and
nonproficient categories.

The consequences of underestimating percent proficient for high-performing
schools is a threat to the validity of accountability systems. As 2014 approaches
and NCLB AYP mandates move closer to 100%, schools close to meeting AYP tar-
gets using observed percent proficient would possibly reach the AYP 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 nonproficient students. However, failing to account for mis-
classification bias will attenuate the difference in nonproficient 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
cutoff is set. Based upon an examination of state assessment data, when the profi-
ciency threshold is low relative to the group tested (e.g., at the 20th percentile), \( p_{11} \)
for schools is generally quite high, \( .9 \leq p_{11} \leq 1 \), while \( p_{22} \) is less so, \( .7 \leq p_{22} \leq
.85 \). 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 and Lewis (1995, p. 183). In their
Table 2, as the proficient threshold increases, values of \( p_{11} \) decrease from .979 to
.758 while the values of \( p_{22} \) increase from .876 to .985.

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 perfor-
ance level misclassification, induces an “unsuspected bias” (Cochran, 1968, p. 637)
that can lead to erroneous judgments about AYP. For example, in a state with chal-
lenging proficiency standards, \( p_{11} \) is likely to range between .7 and .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, $p_{11} < .9$ can lead to a substantial underestimation (10–15%) of the percent proficient at the school. Clearly, this underestimation can have consequences regarding AYP 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% proficient students. However, for extreme schools with less than 15% and more than 85% 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 $\hat{\pi}_1$ is a function of the misclassification rates and the observed percent proficient at the school.

To quantify the impact of misclassification upon the total variability associated with reported percent proficient results, the ratio of total variability to sampling variability is represented as a function of three variables: the conditional correct classification rates, $p_{11}$ and $p_{22}$, and the observed percent proficient, $\hat{\pi}_1^*$. A ratio of 1, for example, indicates no measurement error. When observed percent proficient is low, the conditional correct classification rate associated with nonproficiency, $p_{22}$, displays greater relevance in terms of variability of $\hat{\pi}_1$ than does the conditional correct classification rate 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 nonproficient approach parity, the situation is more complicated, both $p_{11}$ and $p_{22}$ can impact the variability of $\hat{\pi}_1$. Combining the bias and variability there is a “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 .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 double error variability associated with sampling. For example, with 50% observed percent proficient, if $p_{11} = p_{22} = .85$, 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 .75 suggesting that in some schools with modest numbers of students ($n$ in equation 8), measurement error can lead to extremely imprecise
estimates. Future research will address the impact of these issues with regard to AYP 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 the impact of misclassification be understood as a source of bias and variability in those results.

Notes

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.

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.

3 Hanson and 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.

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

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

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

7 In the rare circumstance that the observed density is the steady-state distribution associated with the misclassification matrix $P$, 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”.

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

9 When the collection of students investigated is considered 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 and 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


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


**Authors**

DAMIAN W. BETEBENNER, National Center for the Improvement of Educational Assessment, PO Box 351, Dover, NH 03821; dbetebenner@nciea.org. Current research interests include longitudinal data analysis, particularly with regard to state accountability systems.

YI SHANG is currently a Doctoral Candidate in the Department of Educational Research, Measurement and Evaluation, Boston College, Chestnut Hill, MA 02467; yi.shang@bc.edu. Current research interests include applied statistics and educational policy.

YUN XIANG is currently a Doctoral Candidate in the Department of Educational Research, Measurement and Evaluation, Boston College, Chestnut Hill, MA 02467; yun.xiang@bc.edu.

YAN ZHAO is currently a Doctoral Candidate in the Department of Educational Research, Measurement and Evaluation, Boston College, Chestnut Hill, MA 02467; yan.zhao@bc.edu.

XIAOHUI YUE is currently a Research Assistant in the Educational Research and Evaluation Program, Department of Educational Leadership and Policy Studies, School of Education, Virginia Tech, Chestnut Hill, MA 02467; yuexi@vt.edu. Current research interests include psychometrics, IRT, and Computerized Adaptive Testing.