

Some Threats to Validity in Adequate Yearly Progress (AYP) Models under NCLB

Steve Hebbler
Office of Research and Statistics
Mississippi Department of Education

Handout to Accompany a Presentation in the Session
*Examining the Validity of State Accountability Systems:
Will Decisions Based on AYP Models Hold Up?*

34th Annual CCSSO Conference on Large Scale Assessment
Boston, Massachusetts
June 20-23, 2004

Figure A shows how student scores on statewide tests are generally converted to proficiency levels and then classified as "proficient" or "not proficient" for use in AYP models.

Figure A: Characteristics of the Scales

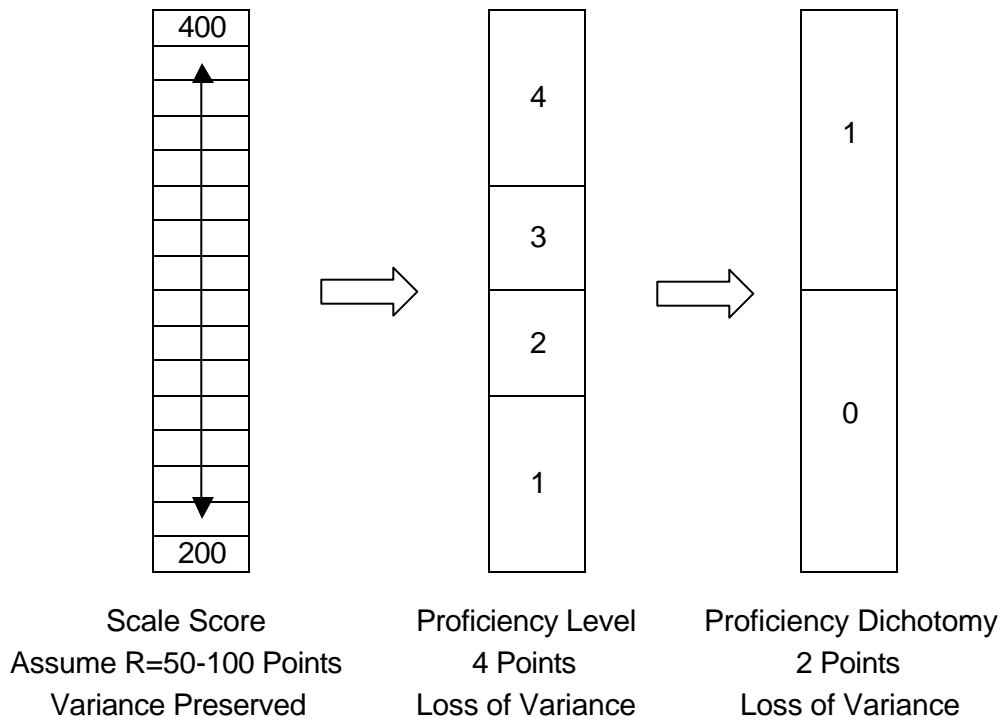
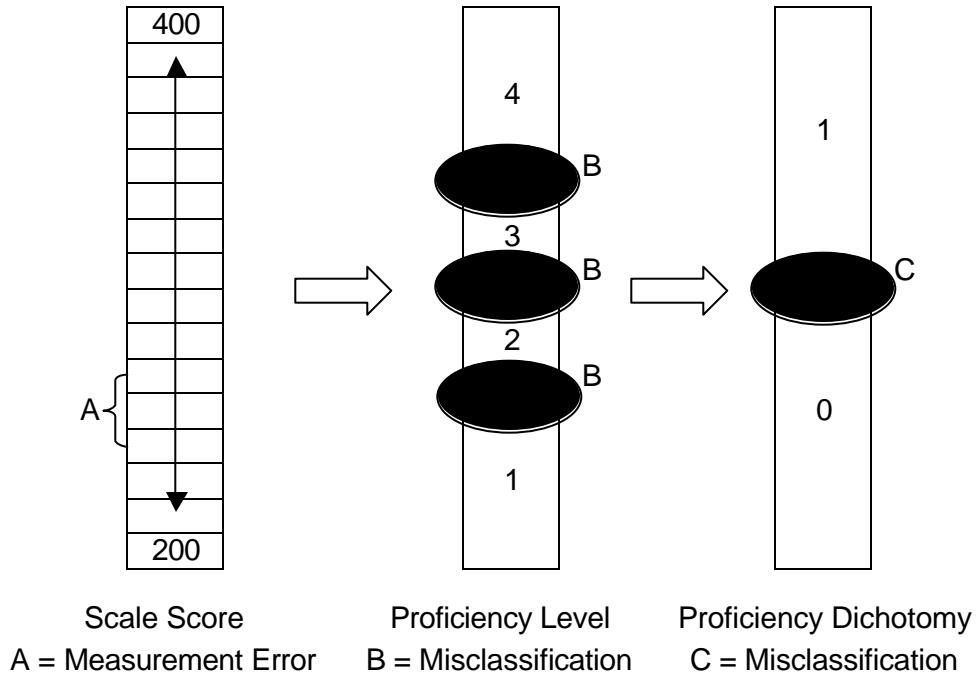
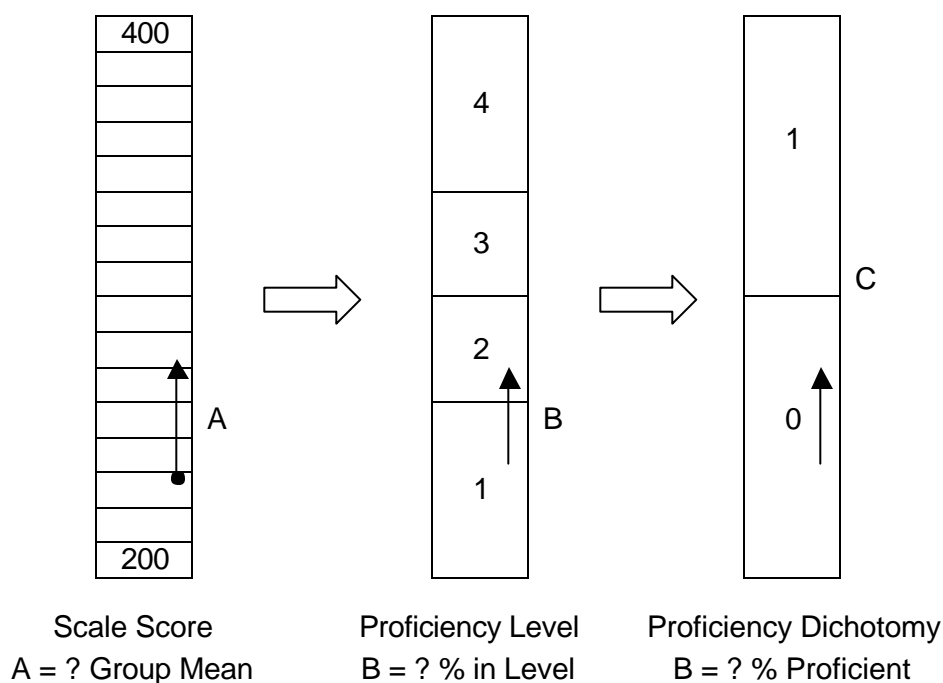


Figure B: Student Level Error Concerns



For reasonably reliable tests, measurement error (A) should be relatively small. The misclassification errors around B and C would also be relatively small. The sum of the positive and negative measurement errors around A would be close to 0 and the numbers of students misclassified "too high" and "too low" in B and C would split about evenly. With reasonably large groups, there should be no appreciable bias in the mean scale score or in the percentages of students assigned to proficiency levels or to the proficient/not proficient dichotomy.

Figure C: Sensitivity to Improvement in Achievement



A change in the group mean of (for example) 10 scale score points would be detectable within a true school improvement (or growth) model using the scale score as the unit of measure (see A). That same improvement might be detectable as a change in the percentage of students scoring in Level 2 (see B), while it might be considered "no improvement" if it did not increase the percentage of proficient students (see C).

The term Adequate Yearly Progress (AYP) implies that the model uses student progress (an increase in student achievement) to determine whether a school is effective. For low performing schools (or low performing subgroups within schools), however, there can be a large amount of improvement that will not be detected using a change in proficiency percentage as the unit of analysis. Even the "safe harbor" provision in NCLB, which was meant to keep improving low performance schools from failing AYP, exhibits this problem. Safe harbor is based on the change in the percentage of proficient students in a subgroup, not on actual changes in the students' achievement scores.

The assumption that "not making AYP" (as defined in NCLB and in subsequent USDE regulations) is an accurate generalization for ineffective instruction cannot be justified. Even if it is shown that the state assessments are reliable and are valid for measuring the state's challenging content standards and that there is no systematic bias in the proficiency classifications of students on those assessments, failure of a subgroup to meet AYP may represent a school level misclassification error. That is, we may be making an accurate assessment concerning the percentage of students scoring "proficient" and an accurate assessment concerning changes in that percentage, but a seriously inaccurate assessment of instructional effectiveness.

If there is a misclassification for even one subgroup at a school, the school will not meet AYP and may be subject to inappropriate sanctions. The use of conjunctive standards (AYP criteria) in the AYP model magnifies the school level misclassification problem.

The misclassification problem in this case is not focused on whether the calculations accurately determined whether the school met AYP when the model was applied under the state's approved plan. The more important question is whether the results of the AYP model (i.e., "met" or "did not meet" AYP) correspond with a "true" assessment of instructional effectiveness. The question can be answered, hypothetically, and the answer could be depicted as a four-fold truth table.

Table 1. AYP Truth Table

		AYP Determination (Results of the AYP Model)	
		Met AYP	Did Not Meet AYP
True School Effectiveness Status	Does Not Need Improvement	True Negative	False Positive (Error)
	Needs Improvement	False Negative (Error)	True Positive

If the AYP determinations were perfectly accurate, the true positive and true negative cells would contain the numbers of schools/LEAs actually needing or not needing improvement (as defined earlier). The false positive and false negative cells would both contain zeros.

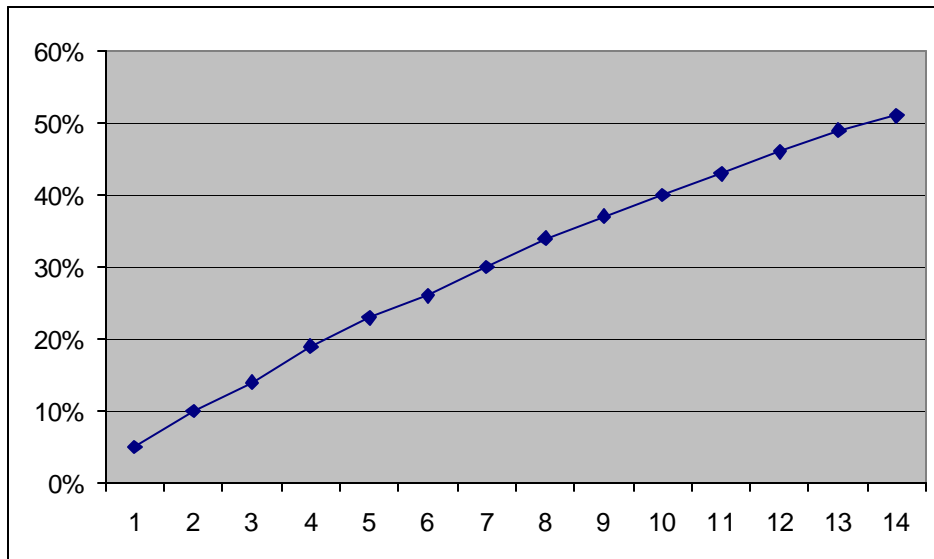
Table 2 shows actual results from two different accountability systems using the same data in 2003. The AAG model used student test scores and classified schools based on both achievement (status at the end of the school year) and growth (change in student scale scores from the previous school year). The AYP model was run consistent with the requirements in NCLB as approved by the U.S. Department of Education. While there was some agreement – 396 "Successful" schools met AYP and 129 "Below Successful" schools failed to meet AYP, the models disagreed on 296 schools. The differences are great enough to indicate that (1) the models are measuring different things, (2) one model is "correct" and one is not, or (3) both models are incorrect. We do not have enough information to determine which is true.

Table 2. Comparison of two Accountability Models

		Results of the AYP Model	
		Met AYP	Did Not Meet AYP
Results of the AAG Model	Successful (Levels 3-5)	396 Schools	283 Schools
	Below Successful (Levels 1-2)	13 Schools	129 Schools

The actual subgroup level classification error in an AYP model will vary depending on a many things – however, it certainly would be greater than zero. Assuming even a relatively small degree of error for each NCLB subgroup, the overall misclassification error for the school across the subgroups will be very large. This is because the overall AYP "met/not met" decision for a school is determined by the school failing to meet AYP for any of the subgroups. The probability that the overall school determination will reflect a misclassification is much greater than the probability of a misclassification for any of the separate subgroups. This is shown in Figure D.

Figure D. Probability of Misclassification



X Axis = Number of Conjunctive Standards Applied

Given a situation where the probability of a misclassification for any given group is 5% (this could be the case, for example, if the state's model applies a 95% confidence interval to the subgroup determinations to compensate for sampling error), the overall probability increases with the number of subgroups for which the school is held accountable. While each separate subgroup determination may be subject to only a 5% probability of error, the rate increases with each additional standard that is applied. Given that NCLB specifies nine separate subgroups, states that check AYP for subgroups across both subject areas (reading and mathematics), possibly at several separate grade spans, then add the AYP check for "other academic indicators" will quickly approach a misclassification probability of 50%. This means that the overall AYP determination for the school could as well have been predicted by flipping a coin. Even using 99% confidence intervals, the probability of error across 14 decisions is about 13%. Clearly, confidence intervals alone cannot mitigate the misclassification problem in a conjunctive standards model. That is why some states' AYP models use of a minimum N value in conjunction with confidence intervals.

Obviously, it is important to examine the various factors affecting reliability and validity in AYP models and ensure that large numbers of schools are not being misclassified.

The major threats to validity outlined in this presentation were:

1. The use of coarse measurement statistics in school accountability models
2. Using change in percent proficient to measure improvement in low performing schools
3. Large misclassification errors associated with a conjunctive standards model

Bibliography

- Center on Education Policy (2003). *From the Capitol to the classroom: State and federal efforts to implement the No Child Left Behind Act*. Washington, DC.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Daugman, J. (.). *Biometric Decision Landscapes*. Paper. Cambridge, England: The Computer Laboratory, University of Cambridge.
- Gong, B. (2002) *Designing school accountability systems: Towards a framework and process*. Washington, DC: Council of Chief State School Officers.
- Hill, R. (June, 1997). *Calculating and reducing errors associated with the evaluation of adequate yearly progress*. Paper presented at the CCSSO Annual Large Scale Assessment Conference. Colorado Springs, CO.
- Hill, R. (March, 2000). *Common problems with accountability systems*. Paper presented at the Conference on Policy and Measurement Issues in Large Scale Science and Mathematics Assessment. Washington, DC.
- Hill, R. (2001). *Issues related to the reliability of school accountability scores*. Dover, NH: National Center for the Improvement of Educational Assessment, Inc., Report on the reliability lecture from the 2000 Annual Edward F. Reidy Interactive Lecture Series.
- Hill, R. (April, 2002). *Examining the reliability of accountability systems*. Paper presented at the Annual Conference of the American Educational Research Association. New Orleans, LA.
- Hill, R. & DePascale, C. (2002). *Determining the reliability of school scores*. Dover, NH: National Center for the Improvement of Educational Assessment, Inc.
- Hill, R. & DePascale, C. (2003). *Reliability of No Child Left Behind Accountability Designs*. Dover, NH: National Center for the Improvement of Educational Assessment, Inc.
- Hoffman, R. G. & Wise, L. L. (2000). *School classification accuracy final analysis plan for the Commonwealth accountability and testing system*. Alexandria, VA: HumRRO.
- Kane, T. J. & Staiger, D. O. (2002) *Volatility in school test scores: Implications for test-based accountability systems*, Brookings Papers on Education Policy. Washington, DC: The Brookings Institution.
- Kane, T. J., Staiger, D. O., & Geppert, J. (2001). *Assessing the Definition of "Adequate Yearly Progress" in the House and Senate Education Bills*, UCLA Working Paper.
- 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.

Marion, S.F., White, C., Carlson, D., Erpenbach, W.J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

Forte-Fast, E. & Hebbler, S. (2004). *A framework for examining validity in state accountability systems*. Washington, DC: Council of Chief State School Officers.

Riddle, W. (2001). *Adequate yearly progress under the ESEA: Provisions, issues, and options regarding House and Senate versions of H.R. 1*. Congressional Research Service, The Library of Congress, CRS Report RL31035.

Shlyakhter, A., & Wilson, R. (1995, September). *Monte Carlo simulation of uncertainties in epidemiological studies: An example of false-positive findings due to misclassification*. Proceedings of the 3rd International Symposium on Uncertainty Modeling and Analysis, College Park, MD: University of Maryland.

Thum, Y. M. (2003). *No child left behind: Methodological challenges & recommendations for measuring adequate yearly progress*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, CSE Technical Report 590.