WAKE COUNTY PUBLIC SCHOOLS EVALUATION AND RESEARCH DEPARTMENT
FOR GOOD MEASURE...

Measurement Error

Authors: Stephen Johnson, Chuck Dulaney, and Karen Banks (850-1903)

Overview

Educators give tests to determine what students know and are able to do in a particular content area. If we have confidence in a test, we believe that a student who scores high on the test knows more in that area than a student who scores low. Likewise, two students whose scores are similar probably have roughly similar achievement in the area being tested.

How much confidence should we have in any particular test? How should our level of confidence in the test affect the way we think about student scores? Teachers and parents are actually interested in how well students read, not how well they do on a test. A reading test is a convenient way to estimate how well students read. If the test is a good one, we can confidently interpret students' scores on the test as a sound estimate of their true reading achievement. Yet no test, however well designed, can measure a student's true reading achievement because there are numerous factors that interfere with our ability to measure it. Those factors are sources of "measurement error." Every score we obtain has some amount of error in it. The goal is to create tests that have as little measurement error in them as possible.

Where does error come from?

What sorts of things create measurement error? Error can result from the way the test is designed, factors related to the individual students, the testing situation, and many other sources. Some students may know the answers, but fatigue, distractions, and nervousness affect their ability to concentrate. Students may know correct answers but accidentally mark wrong answers on an answer sheet. Students may misunderstand the instructions on a test or misinterpret a single question. Scores can also be an overestimate of true achievement. Students may make random guesses and get some questions right.

There are also test-specific sources of error. Suppose the test uses reading selections as the basis for some questions. If a class happened to have previously studied the text passage being used, that class will probably do better than a class of students who have never seen the text before. For some tests, we know that changing the order of the items on the test leads to higher or lower scores. This means the order of the items is causing measurement error. Some test items may be biased in favor of or against particular groups of students. For example, if the reading passage contains a story that takes place on a farm, students from the inner city may be at a systematic disadvantage in making inferences based on the story.

We always have "error" in our tools and can therefore only obtain an approximation of the "truth" plus or minus some error. Measurement error means that we can underestimate or overestimate a student's true ability or achievement.

The Standard Error of Measurement

We can assume that the student's true score lies within some range around his/her reported score. The Standard Error of Measurement (SE_m) is used to determine the range of certainty around a student's reported score. Technical manuals published by test developers usually provide information about the SE_m, and a formula for calculating SE_m can be found at the end of this report.

The SE_m estimates the range of scores individual students might get if they were to take the same test over and over again (assuming no benefit from the repeated practice). Due to measurement error, students who take a test over and over are unlikely to obtain the same score each time. The error range represents limits around a test score within which we would expect to find the individual's true score. The SE_m makes it possible to determine how reliable a particular test is and how much confidence we can place in the scores it yields.

If one SE_m is added to an observed score and one SE_m is subtracted from it, we can be 68% sure the true score falls within the range we created. Similarly, if two SE_m are added to the score and two SE_m are subtracted from it, a wider interval is created, and we can be 95% certain that the true score falls within this wider range. For example, if a third grade student has a North Carolina End-of-Grade reading score of 150 (the SE_m is approximately 3), we can be 68% certain that the student's true reading score is between 147 and 153 (150 plus or minus 3). Likewise, we can be 95% certain that the true score is between 144 and 156 (150 ± 6).

We cannot assume, however, that all scores on a particular test have the same levels of error. The likelihood that we have measured more error in an individual student's score increases the further away that score is from the mean (average score) of the group. Sources of error (e.g., fatigue, health, luck, etc.) are more likely to lead to very high or very low scores than they are to affect average scores.

When we realize how much room there is for error in student scores, we realize how careful we must be in making decisions about students based on their test scores. For example, third grade students who score 141 in reading (the cut-off score between Level II and Level III) are 95% likely to have true scores between 137 and 145, making them potentially Level II or III. On some EOG tests, students can receive scores in Level III when their true scores are 95% likely to fall anywhere from Level II to Level IV.

Measurement Error and Group Scores

Schools, districts, states, and the nation use test scores to assess the effectiveness of educational systems and programs and to make many large and small policy decisions. So what happens when the local school, district, or state uses those error-filled observed scores to make

decisions about the effectiveness of a teacher, a school, or a district? As shown below, error becomes less important as the amount of reported data grows.

A "true" score is the hypothetical score a student obtains when no error enters an assessment. We can never know a "true" score with certainty because there are always forms of error present in any testing situation. For well-designed tests, the error for individuals is randomly distributed, and the more scores you group together, the more likely it is that one student's error will cancel out another's. Remembering that error can result in an underestimation or overestimation of true achievement, let's look at a small set of students and their reported scores compared to their hypothetical true scores.

Student	Reported Score	"True" Score	Error
Bill	153	147	+6
Rosa	150	149	+1
Kelly	148	152	-4
Latoya	147	147	0
Michael	151	153	-2
Ryan	153	150	+3
Taylor	150	148	+2
Means *(sum/N)*	150.3	149.4	0.9

The means (averages) for each of the columns is the sum of all the scores in the column divided by the total number of students in the group. We can see from this table that the errors for individual students run from –4 to +6, but the average error for the group is only +0.9. The pluses and minuses of individual students cancel each other. Grouping individual student scores into classes, schools, or districts provides a closer approximation of true achievement than individual scores can provide. For an individual student, multiple scores will be more reliable than any one score.

In order to estimate the error for a score reported for a group, we use another statistic called the Standard Error of the Mean (S_e). Using the formula found at the end of this report for the group of students above, the S_e is 0.9.

The bigger the group, the more likely it is that errors will cancel or offset each other, and the more likely that the estimate of the group error will approach zero. The error levels for the mean score of a group of students will be smaller than the error for an individual student, and we can be more confident that a group score is an accurate reflection of the group's true ability.

Standard Error for North Carolina End-of-Grade Tests

The following table shows SE_m for the North Carolina End-of-Grade Reading tests. A similar table can be found in the EOG technical report showing SE_m for the EOG math tests. As noted earlier in this article, sixty-eight percent of the time students' true scores should be within plus or minus these values of their reported scores, and 95% of the time within plus or minus two times these values. Note that standard error of measurement increases as scores move away from the middle of the scoring range. Students at the extremes with very high scores or very low

scores are more likely to have misleading scores that were the result of sources of error such as guessing, testing conditions, or the match between test content and student interests.

Reported	SE_m for the North Carolina End-of-Grade Reading Tests					
Score	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
190						
180						4
170			3	3	3	2
160	3	2	2	2	2	2
150	3	2	2	2	3	3
140	2	3	4	5	5	
130	4	5				
120						
110						

Taken from Sanford, E. (1996). Technical Report #1 North Carolina End of Grade Tests. NC Department of Education. Page 46.

Using the chart, we can say with 95% confidence that a student scoring 150 on the fifth grade reading test probably has a true score between 146 and 154. However, the 95% confidence range for a student scoring 140 would be wider and range from 132 to 148.

As noted earlier, standard errors for group averages have less error than individual scores. The formulas shown below and the information in the table shown above were used to determine the 95% confidence range on the fifth grade reading test for a typical WCPSS school in 1999. The range within which a "true" score probably lies is approximately plus or minus:

5 points for an individual,
3 points for a class of 20 students, and
2 points for a school of 100 students.

In summary, we can be much more confident about our interpretation of the scores of groups of students on the End-of-Grade tests than we can be about interpreting the score of an individual student. End-of-Grade tests were designed to provide information about school accountability and they provide reliable measurement of groups of students.

Formulas

Standard Error of Measurement

$$SE_m = S_x \times \sqrt{1 - r_x}$$

where S_x is the standard deviation of the test, and r_x is the reliability coefficient of the test. Test user handbooks should report the r_x and S_x.

Standard Error of the Mean

$$S_e = S_x \div \sqrt{N}$$

where S_x is the standard deviation of the test, and N is the number of cases/students.