SCORING EXAMINATIONS:

EQUATING AND SCALING

by Lee Schroeder, Ph.D.

Editor's note: This discussion is intended to illustrate the concepts of equating and scaling using simplified examples. While the concepts themselves are rather simple, the actual calculations done to adjust raw scores are more complex than represented here and involve working with a number of statistics in addition to mean scores.

Many of us tend to think of examinations in terms of the tests we took in school. When we took an examination in high school or college, our teacher might have given us a test consisting of 50 multiple-choice items. All students in the class would receive the same examination. Our score was simply the number of questions correct, and sometimes this would be represented as a simple percentage. If we got 40 of the 50 items correct, our score would be 80 percent. The process of scoring these tests was simple and easy to understand.

For professional regulatory examinations, however, calculations are not the same, nor are they as easy. Examinations used for high-stakes decision making must follow more rigorous standards than do teacher-made examinations from high school, college, or even law school.

One examination used early for high-stakes decision making is the Scholastic Achievement Test (SAT). Those of us that took the SAT may remember that it is composed of two examinations, the Verbal and Mathematical Examinations. For each part of the examination we received a score ranging from 200 to 800 points. Obviously, the scores were not percentages. In fact, the scores are from a reporting scale that is different from, though related to, the raw score or number of questions correct. What we may not have noticed during the exam was that the candidate sitting next to us received an entirely different set of questions than we did.

Developers of these types of tests create multiple forms for numerous administrations. Yet despite the best efforts of professional test developers, no two forms of a particular examination are exactly the same in terms of difficulty. Thus, without adjustment, some candidates could be advantaged by being assigned easier forms, while other candidates may be disadvantaged by being assigned more difficult forms. This is when equating and scaling become essential to fairness.

The use of equating and scaling in the preparation of professional regulatory examinations has been supported in the courts. For example, a lawsuit involving a client of mine was heard recently where a failing candidate complained about the unfairness of an examination score. The candidate blamed this unfairness on calculations associated with equating and scaling. When these processes were explained by

an expert witness from my firm, the trial judge found no merit in the candidate's complaint and ruled in favor of the client. This is often the result of such litigation.

Hypothetical Testing Situation

The processes of equating and scaling are complicated and somewhat abstract. In view of this, the following example explains these processes in terms that should be easy to understand.

Suppose that two different groups of candidates (Group 1 and Group 2) took two different forms (Form A and Form B) of an examination on different dates. Perhaps one group of candidates took a given test form in January and a group composed of different candidates took another form of the examination in February.

If the average test score for the two groups is different, what conclusions can be drawn about the two groups or the two forms? Do both groups have the same level of knowledge on the two examinations, or is one group more knowledgeable than the other? Are both examinations of the same level of difficulty, or is one examination more difficult than the other?

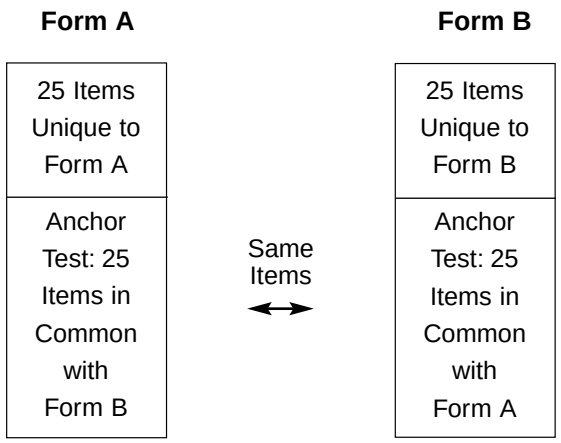
Suppose, for example, that the average score for Group 1 was 38 and that the average score for Group 2 was 33. (Assume that both Form A and Form B are 50 items in length.) The following is a list of several possible situations that could have contributed to this 5-point difference in averages:

- Form A and Form B are equally difficult, but Group 1 is, on average, more knowledgeable than Group 2. (The entire 5-point average difference is due to ***group differences***.)
- Form A is easier than Form B, but Group 1 and Group 2 have the same average level of knowledge. (The entire 5-point average difference is due to ***form difficulty differences***.)
- Form A is easier than Form B, and Group 1 is more able than Group 2. (Part of the 5-point average difference is due to differences in ***form difficulty*** and the other part is due to ***group differences***.)

From the scant data available, we do not know very much about the relative difficulty of the two forms. We are also unaware of the relative levels of knowledge in the two groups when each group takes two different forms of an examination.

Equating

A common technique used to help understand form and group differences is to include a common set of items in both forms of the examination. These common items are sometimes referred to collectively as an anchor test; the individual items in this set of common items are known as equators. Suppose in the previous example that 25 equators appeared on both Form A and Form B. This could be represented by the following tables.



In these figures, both Group 1 and Group 2 took the anchor test. If the anchor test is sufficiently broad, we can determine the average scores on the anchor test, and these averages tell us how Group 1 and

Group 2 compare in terms of knowledge of the material being tested.

In addition, from the difference between the two groups on the anchor test, we can determine what portion of the difference in average scores in either examination is due to group differences and what portion is due to form differences. The process of making these calculations is called ***equating***.

A COMMON TECHNIQUE USED TO HELP UNDERSTAND FORM AND GROUP DIFFERENCES IS TO INCLUDE A COMMON SET OF ITEMS IN BOTH FORMS OF THE EXAMINATION. THESE COMMON ITEMS ARE SOMETIMES REFERRED TO COLLECTIVELY AS AN ANCHOR TEST; THE INDIVIDUAL ITEMS IN THIS SET OF COMMON ITEMS ARE KNOWN AS EQUATORS.

To further explain the process of equating, consider the following example:

Suppose two forms of a 50-question examination are administered, Form A to Group 1 and Form B to Group 2. Suppose the average of Group 1 on Form A is 38, and the average for Group 2 on Form B is 33. Also suppose that an anchor test of 25 equators is part of both Form A and Form B and that both Group 1 and Group 2 have an average score of 15 on the anchor test. This data is shown in the following table.

Group 1 **Form A**		**Group 2** **Form B**
25 Items Unique to Form A		25 Items Unique to Form B
Anchor Test: Average is 15 out of 25	Same Items ↔	Anchor Test: Average is 15 out of 25
Average Score: 38 out of 50		**Average Score: 33 out of 50**

Because both groups have the same average score on the anchor test, we can say that the groups are similarly knowledgeable of the material in the examination. Thus, the difference in the averages for Form A (Average=38) and Form B (Average=33) is due to differences in difficulty between the forms.

In this case, if there were no adjustment to the scores, candidates in Group 1 would receive an average score of 38, while candidates in Group 2, who on average have a level of knowledge equal to those in Group 1, would receive an average score of 33. This would be unfair to all candidates in Group 2.

Further, if the minimum passing score on the test was set at 70 percent, many candidates would pass if they took Form A, but fail if they took Form B. This would be extremely unfair to those candidates who took Form B.

A simple solution to this problem would be to add 5 points to the scores of candidates who took Form B. This would make a correct answer on Form B have more weight or a higher value than a correct answer on Form A. This formula would convert a raw score of 33 on Form B to a score of 38, making it have an equivalent meaning to scores on Form A.

The above provided scoring adjustment is an example of ***equating***. Equating determines how raw scores from one test may be weighted so as to have equal meaning with scores from another test. This process eliminates the effects of differences in test difficulty. Since test forms do differ in difficulty,

equating is important to ensure fairness to all candidates who are tested.

Scaling

Given that equating is necessary, we must also know how to report scores on equated examinations. In the example above, a candidate taking Form B and earning a raw score of 33 has the same level of knowledge as a candidate with a raw score of 38 on Form A. This could be represented in various ways, such as:

- Add 5 points to all Form B scores, thus reporting an earned score of 38 for candidates who get 33 questions correct. In this case, how are the scores reported? Do candidates who take Form A wonder why their scores are not adjusted? What do we tell them?

- Subtract 5 points from all Form A scores, thus reporting an earned score of 33 for candidates who get 38 questions correct. Do candidates who take Form A wonder why their scores are adjusted? What do we tell them?

Actually, there is no way to report equal raw or percent scores on equated examinations without creating some confusion. To prevent confusion, the process of ***scaling*** is used to report scores from equated examinations. This process begins with the adoption of an arbitrary scale. To further explain the process of scaling we could, in our example, create a scale that may run from 5 to 15 with the cut score set at 12. A score of 38 on Form A may be set at 13 on this scale. Further, all scores on future forms that are equal to 38, after equating, would also be set at 13. Therefore, in this example, a score of 33 on Form B would have a scaled score of 13 as well.

SUMMARY

This article was written to explain why the process of equating and scaling are necessary to ensure fairness for high-stakes examinations. Equating helps us understand whether differences in test scores are due to form difficulty or group differences. Scaling provides a means of representing test scores from test forms of different levels of difficulty. Both equating and scaling assure candidates the highest level of fairness.

DR. LEE L. SCHROEDER is the founder and President of Schroeder Measurement Technologies (SMT) in Dunedin, Florida. SMT is a full-service testing company that examines more than 100,000 candidates each year in the United States and Canada as well as Europe and Asia. Prior to his work at SMT, Dr. Schroeder served as President of National Assessment Institute in Clearwater, Florida; he was also the founder and President of Applied Measurement Services in New Jersey.

Dr. Schroeder received his doctorate in Statistics and Measurement from Rutgers University in New Jersey. He is a frequent contributor to the CLEAR Exam Review where he provides a regular column on software related to the measurement process.