TEST EQUATING

Randall E. Schumacker, Ph.D.

INTRODUCTION

Multiple forms of a certification exam are desirable for a variety of reasons. However, the problem of comparability among test scores using different test forms must be addressed in order to insure fairness and consistency in each testing situation. Test scores must be independent of examinee characteristics and the particular test form used. Test forms must be interchangeable across test administrations. Psychometric procedures known as equating methods can be utilized to produce comparable (equated) scores. Equating procedures consist of (1) a design for collecting test data for equating, (2) a clearly defined level of expected correspondence among test scores, and (3) specific statistical procedures that are used to estimate score correspondence.

EQUATING DESIGNS

The choice of an equating design involves both practical and statistical issues. Three commonly used designs are the single group design, random groups design, and common item nonequivalent groups design. In the single group design, one group is administered each form to be equated. While the design is useful, it suffers from several practical and technical problems. Random group designs utilize a spiraling process where alternate examinees are administered alternate forms of the exam. The effect is to produce randomly equivalent groups taking each form of the test. From a practical point of view, the random groups design is often preferable to the single group design. However, much larger samples are necessary to obtain stable parameter estimates. This requirement limits their use in many situations. Common item nonequivalent groups designs utilize different groups of examinees and a common set of items for each test form. Two variations of the design are employed depending on whether the common items are administered internally or externally. The major disadvantage of the common item nonequivalent groups design is the stringent statistical analysis underlying the technique.

EQUATING ERROR

When applying equating methods, different types of equating error influence the interpretation of the results. The first type - random equating error - is always present because samples are used to estimate parameters such as means, standard deviations, and percentile ranks. However, random error can be reduced by using large samples of examinees and by the choice of equating design. Systematic error is more difficult because it results from the violation of assumptions and conditions unique to the particular equating methodology used. Unlike random error, which can be quantified using standard error calculations, systematic error is more difficult to estimate.

© 2010 Applied Measurement Associates LLC

CONCLUSION

Equating designs and equating methods should be carefully chosen to reduce equating error. Fairness in testing requires assurance that comparable score results are obtained across testing situations. Fortunately, equating procedures exist that can greatly improve the quality of measurement and assessment. We are available to work with new and existing clients to help create equivalent forms that produce comparable test score results.

© 2010 Applied Measurement Associates LLC