OBSERVED-SCORE EQUATING AS A TEST ASSEMBLY PROBLEM

Wim J. van der Linden
University of Twente

Richard M. Luecht
National Board of Medical Examiners

A set of linear conditions on item response functions is derived that guarantees identical observed-score distributions on two test forms. The conditions can be added as constraints to a linear programming model for test assembly that assembles a new test form to have an observed-score distribution optimally equated to the distribution on an old form. For a well-designed item pool and items fitting the IRT model, use of the model results into observed-score pre-equating and prevents the necessity of post hoc equating by a conventional observed-score equating method. An empirical example illustrates the use of the model for an item pool from the Law School Admission Test.

Key words: item response theory, test equating, test assembly, generalized binomial distribution, 0-1 linear programming.

A well-known method of observed-score equating is equipercentile equating. The method assumes that estimates of the observed-score distributions on the old and new test forms are available and equates an observed scores on the new form to the score on the old form that estimates the same percentile in the population of examinees. With the advent of item response theory (IRT; Hambleton & Swaminathan, 1985; Lord, 1980; Rasch, 1960; van der Linden & Hambleton, 1997), new methods of equating have become available. These methods assume that the items in the two test forms have been calibrated on the same scale for the ability parameter in the IRT model. In one method, the response functions are used to generate the observed-score distributions on both test forms, and the equipercentile method is employed to find the transformation that equates the two distributions. Another method uses the test characteristic functions of the two tests as a system of parametric equations that equates the true scores on the two forms. If the two tests have high reliability, true-score equating is often used as an approximation to observed-score equating. An introduction to equipercentile and IRT equating is given in Braun and Holland (1982) and Kolen and Brennan (1995); IRT equating is also discussed in Lord (1980).

Under IRT, tests from the same pool are automatically scored on the same scale for the ability parameter. From a theoretical point of view, it seems therefore superfluous to equate the observed-score scale as well. Nevertheless, practical reasons for this additional equating exists. Many testing programs had already fixed their score scales before IRT was introduced and replacing them by ability estimates with a more complicated relation to the response vectors than number-right scores might have been difficult to explain to their examinees. In addition, since the ability scale has a nonlinear relation to the observed-
score scale, the sudden change of score distributions could have confused these examinees too. It is therefore not uncommon to find testing programs using IRT for such routines as item parameter estimation, screening of item quality, test assembly, and test equating but reporting their scores still on a traditional scale.

This paper identifies a set of conditions on item response functions that ensure identical observed-score distributions on two test forms. These conditions can be used to assemble a new test form from an item pool to have an observed-score distribution identical to the distribution on an old form to which the new form has to be equated. If the item pool is well designed and the items fit the IRT model, a test assembly procedure realizing the conditions would make observed-score pre-equating possible. Realization of this ideal has several advantages over the current practice of post hoc equating, that is, after the new form has been administered operationally. The advantages include:

1. The results hold for any population of students for which the calibration of the item pool is valid no matter its ability distribution. There is no need to know the distribution of the population but, as in any other IRT observed score equating method, the assumption is that the distribution remains stable.

2. No resources are lost running separate equating studies;

3. Scores on the new test form can be reported immediately after its administration;

4. Unlike current equating practice, the scale of the observed scores on the new test form is not distorted by a (nonlinear) score transformation. As a consequence, the scores keep their interpretation as number-right scores. Also, it is not necessary to find (arbitrary) solutions for transformations that map scores on new test forms outside the scale of the old form. Finally, there is no need to resort to interpolation methods to deal with the discreteness of the number-right scores.

5. The scores on the two forms are automatically equitable because the procedure ensures identity of the conditional distributions of observed scores on the two forms for each possible ability level. Lord (1980, sec. 13.2) proved that this condition cannot be met when equating an existing form. However, a new form may very well be assembled to meet this important condition on equating.

This paper also presents a test assembly model that can be used to realize the conditions on the response functions. Since the conditions are linear in the items, it is proposed to build the conditions as constraints in a 0-1 linear programming (LP) model for test assembly. 0-1 LP models have been developed earlier for a variety of other test assembly problems and have proven to yield practical results in many applications. A favorable feature of these models is that they allow for large sets (e.g., several hundreds) of additional test specifications to build into the model. These models are discussed more in detail below.

An alternative to a test assembly model with conditions on the response functions to guarantee identical observed-score distributions would be to assemble the new test form such that it is matched item by item to the old form. A 0-1 linear programming method for matching items on classical parameters was given earlier in van der Linden and Boekkooi-Timminga (1988; see also Armstrong & Jones, 1992). The method can easily be adapted to match items on their response functions. Tests forms with pairwise identical response functions have equal true scores and observed-score variances for each examinee in the population for which the IRT model holds and are therefore parallel (Lord & Novick, 1968, definition 2.13.1). Consequently, they have identical observed-score distributions. However, a method of pairwise item matching would impose conditions on the test assembly process that are more stringent than necessary. We will return to this topic when a proposition has been presented that clarifies this issue (Proposition 4).

In the rest of the paper, first the theory of equipercentile and IRT equating is re-