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 can not 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-
viewed. Then, the set of conditions on item response function to ensure two test forms to have identical observed-score distributions is derived. The set replaces an earlier approximate condition given in van der Linden and Luecht (1996). It will then be shown how these conditions can be included in a linear-programming (LP) model for test assembly that optimizes the composition of the new test subject to a set of constraints representing the test specifications. Next, it will be indicated how the method can be generalized to deal with item pools dependent on more than one ability variable as well as other scoring systems than number correct. Use of the test assembly model is empirically illustrated for an item pool from the Law School Admission Test (LSAT).

Equating Transformations

The following notation is needed to present the equating transformations. Index \( j = 1, \ldots, n \) will be used to denote the items in the old test form, whereas \( i \) is used to denote the items in the new test form \( (i = 1, \ldots, n) \) or in the pool from which the form is assembled \( (i = 1, \ldots, I) \). Responses by examinee \( a \) to item \( i \) or \( j \) will be represented by random variables \( U_{ai} = u_{ai} \) and \( U_{aj} = u_{aj} \), respectively. Number-correct scores for examinee \( a \) on the two tests are defined as \( X_a = \sum_{i=1}^{n} U_{ai} \) and \( Y_a = \sum_{j=1}^{I} U_{aj} \), with true scores \( \tau_{X_a} = E(X_a) \) and \( \tau_{Y_a} = E(Y_a) \), respectively. Finally, it is assumed that \( X \) and \( Y \) have cumulative distribution functions \( F(x) \) and \( G(y) \).

In equipercentile equating, both test forms are administered to a single sample or two independent random samples from the population of examinees to estimate the transformation, \( e(x) \), that maps score \( X \) on the scale of score \( Y \):

\[
e(x) = G^{-1}(F(x)).
\]

(1)

The first step in this transformation identifies \( x \) as a percentile under the distribution of \( X \); the second step equates \( x \) to the same percentile under the distribution of \( Y \).

To discuss the mathematical equations involved in IRT equating, it is assumed that the responses to the items in the two test forms fit the 3-parameter logistic (3-PL) model. The model gives the probability of a correct response \( U_{ai} = 1 \) as

\[
P_r(\theta) = c_i + (1 - c_i)[1 + \exp(-a_i(\theta - b_i))]^{-1},
\]

(2)

where \( \theta \in (-\infty, \infty) \) is a parameter for the ability of the examinee, \( b_i \in (-\infty, \infty) \) and \( a_i \in [0, \infty) \) are parameters for the difficulty and discriminating power of item \( i \), respectively, and \( c_i \in [0, 1] \) is the guessing parameter of the item. The 3-PL model is chosen because it was used to calibrate the LSAT item pool in the empirical example at the end of this paper. If \( h(\theta) \) is the density of the ability distribution in the population of examinees, the probability functions of \( X \) and \( Y \) are given by

\[
f(x) = \int_{-\infty}^{\infty} p_X(x|\theta)h(\theta) \, d\theta \tag{3}
\]

and

\[
g(y) = \int_{-\infty}^{\infty} p_Y(y|\theta)h(\theta) \, d\theta, \tag{4}
\]

where \( p_X(x|\theta) \) and \( p_Y(x|\theta) \) are the probability functions of the conditional distributions of \( X \) and \( Y \) given \( \theta \), respectively. These conditional distributions are generalized binomial (Lord, 1980, sec. 4.1; Kendall & Stuart, 1977, sec. 5.10).
In IRT observed-score equating, the probability functions $f(x)$ and $g(y)$ are estimated from a random sample of examinees, estimates of the cumulative distribution functions $F(x)$ and $G(x)$ are calculated, and (1) is used to estimate the transformation from $X$ to $Y$. Two alternative methods to implement IRT observed-score equating are discussed in Zeng and Kolen (1995).

In IRT true-score equating, the fact is used that the true scores of $X$ and $Y$ are equal to:

$$
\tau_X = \tau_X(\theta) = \sum_{i=1}^{n} P_i(\theta)
$$

(5)

$$
\tau_Y = \tau_Y(\theta) = \sum_{i=1}^{n} P_i(\theta).
$$

(6)

These two equations, known as the test characteristic functions of form $X$ and $Y$, define the (parametric) relation between $\tau_X$ and $\tau_Y$ that can be used to equate the true score on $X$ to the one on $Y$.

To apply an equating method, a data collection design has to be chosen. Standard designs in equating practice are the single-group, random-groups, and common-item designs. Descriptions of these designs are given in Braun and Holland (1982) and Kolen and Brennan (1995). Though observed-score equating is the ideal, for large tests true-score equating is sometimes used as an alternative because the equations in (5) and (6) are simpler to apply than the procedure based on (3) and (4). Also, experience shows that for some equating designs true-score equating may produce more robust results than observed-score equating.

Conditions for Observed-Score Distributions to be Identical

From the probability functions in (3) and (4) it is clear that since the ability distribution is common, the observed-score distributions of $X$ and $Y$ are identical if the conditional distributions of $X$ and $Y$ given $\theta$ are. This fact is used in the proof of the following proposition.

**Proposition 1.** For any $h(\theta)$, the distributions of observed scores $X$ and $Y$ are identical if and only if

$$
\sum_{i=1}^{n} P_i(\theta) = \sum_{j=1}^{n} P_j(\theta), -\infty < \theta < \infty,
$$

(7)

for $r = 1, \ldots, n$.

**Proof.** The distributions of observed test scores given $\theta$ do not have a probability function in closed form but their probabilities can be obtained via the generating function $P_n(\theta, P)$. In addition, this probability generating function is known to have the following exact expansion in the powers of $P_1 - \xi, P_2 - \xi, \ldots, P_n - \xi$:

$$
\text{Prob}(X = x) = p_n(x) + \frac{n}{2} V_2 C_2(x) + \frac{n}{3} V_3 C_3(x)
$$

$$
+ \left(\frac{n}{4} V_4 - \frac{n^2}{8} V_2^2\right) C_4(x)
$$

$$
+ \left(\frac{n}{5} V_5 - \frac{5n^2}{6} V_2 V_3\right) C_5(x) + \ldots, x = 0, 1, \ldots, n,
$$

(8)
with

\[ p_n(x) = \binom{n}{x} \xi^{1-x} \xi^{-x}, \quad (9) \]

\[ C_r(x) = \sum_{v=0}^{r} (-1)^{r+v+1} \binom{r}{v} p_{n-r}(x-v), \quad r = 2, \ldots, n, \quad (10) \]

\[ V_r = n^{-1} \sum_{i=1}^{n} (P_i - \xi)^r, \quad r = 2, \ldots, n, \quad (11) \]

where \( \xi \) is defined as \( n^{-1} \sum_{i=1}^{n} P_i \) (Lord & Novick, 1968, sec. 23.10; Walsh, 1953, 1963).

Since (8) is an exact identity, the distributions of \( X \) and \( Y \) given \( \theta \) are equated if the expressions in (9) through (11) are. For (9) through (10) this condition is realized if \( \xi \) is equal for both tests, that is, if (7) is true for \( r = 1 \). In addition, (11) can be written as

\[ V_r = n^{-1} \sum_{i=1}^{n} \sum_{v=0}^{r} (-1)^{r-v} \binom{r}{v} P_i^{-v}, \quad r = 2, \ldots, n. \quad (12) \]

Substitution of \( r = 2, \ldots, n \) into (12) shows that these expressions are equal for both tests if (7) holds for \( r = 1, \ldots, n \). These two conclusions establish the proposition.

The following two propositions formulate useful relations between (7) and (8).

**Proposition 2.** For \( r = 1, \ldots, R \leq n \), the equalities in (7) equate the first \( R \) terms of the series in (8).

The truth of this proposition follows immediately from the substitution at the end of the proof of Proposition 1.

**Proposition 3.** For \( n \rightarrow \infty \), the contributions of the terms in (8) of the order \( r \geq 2 \) vanish.

**Proof.** For \( r = 1 \), the condition in (7) formulates that the true scores associated with \( X \) and \( Y \) in (5)-(6) be equal for all possible values of \( \theta \). From Proposition 2 it follows that equal true scores imply that the distributions of \( X \) and \( Y \) have identical first terms for (8). Thus, true-score equating can be viewed as a first-order approximation to observed-score equating. From the central limit theorem for sums of (nonidentical) independent variables (e.g., Shiryaev, 1996, chap. 4) it follows that the observed scores converge to their true scores if the test length increases. Hence, the first-order approximation improves with test length, and the contributions of the higher-order terms in (8) vanish for long tests.

The practical implication of the proposition is that the series in (8) can be truncated after a few terms. The same can be done for the test assembly model below imposing (7) on the selection of the items only for small values of \( r \). In the empirical example later in this paper only the first three terms were used.

An interesting consequence, however, is obtained if all \( n \) equalities in (7) are imposed. The set of conditions is then equivalent to the one of the response functions of the two tests being pairwise identical. This property as well as its proof were suggested by N. D. Verhelst (personal communication, November 1, 1996):
**Proposition 4.** The conditions in (7) hold simultaneously for all values of $r$ if and only if the two tests have pairwise identical response functions.

**Proof.** If the two tests have pairwise identical response functions, then the conditions in (7) hold trivially. The proof of the reverse implication is based on the idea to define probability spaces over the two sets of response functions and to invoke the principle of moments (e.g., Kendall & Stuart, 1977, sec. 4.22). Thus let be $(X_\theta, \mathcal{P}(X_\theta), p)$ a (finite) probability space, where $X$ is the set of response probabilities in test $X$ for a fixed value of $\theta$, $\mathcal{P}(X_\theta)$ is the power set of $X_\theta$ and $p$ is the (uniform) probability function $p(i) = 1/n$. In addition, a random variable $X_\theta(i) = P_i(\theta)$ is defined. An analogous probability space and random variable is defined for test $Y$. The conditions in (7) stipulate that the first $n$ moments of the distributions induced by the two random variables are identical. Therefore, the two distributions are identical; that is, for each value of $X_\theta(i) = P_i(\theta)$ there exist a value $Y_\theta(j) = P_j(\theta)$, and vice versa. Because the argument holds for an arbitrary value of $\theta$, the pair of functions $P_i(\theta)$ and $P_j(\theta)$ have more than two points in common and are identical.

As already noted, test forms with pairwise identical response functions yield parallel measurements. Proposition 4 thus shows the conditions in IRT under which the classical definition of parallel measurements holds. It also shows why a method of item matching could be an alternative to the model presented below. However, Propositions 2 and 3 imply that identity beyond the first few conditions in (7) is not necessary. The conditions in Proposition 4 are therefore too stringent. Also, item-matching models are generally based on a two-stage process that is computationally more complicated than the convenient one-stage approach implied by the model below. The latter is therefore preferred.

**Test Assembly Model**

Because the conditions in (7) are linear in the items, they can be used as objective function and/or constraints in an LP model for optimal test assembly. Such models have been proposed earlier, for example, to assemble tests to match a target information function (Swanson & Stocking, 1993; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1989), to assemble sets of parallel test forms (Adema, 1992; Armstrong, Jones & Wu, 1992; Boekkooi-Timminga, 1987, 1990), to maximize classical test reliability (Adema & van der Linden, 1989; Armstrong, Jones & Wang, 1994), to match tests item by item (Armstrong & Jones, 1992; van der Linden & Boekkooi-Timminga, 1988), to assemble multidimensional tests (van der Linden, 1996), or to implement constrained adaptive testing (van der Linden & Reese, 1998). In addition, these models allow for all other test specifications typically constraining the selection of items in a testing program.

Following is the model proposed to select a new test form from a pool of items with an observed-score distribution optimally equated to the distribution of an old form. Let $z_i$, $i = 1, \ldots, I$, be the decision variables to denote whether ($z_i = 1$) or not ($z_i = 0$) item $i$ is included in the new test form. Because the first terms in (8) are most important, the idea is to choose the values of $z_i$ such that the differences between the two left-hand and right-hand sums in (7) are minimized for $r = 1, \ldots, R \leq n$ at a series of values $\theta_k$, $k = 1, \ldots, K$. As already noted $R$ can be small. Also, since the item response functions in (7) are well-behaved continuous functions, only a few points are necessary to control their shapes over the range of $\theta$ values considered. However, there are no limitations as to the number of values and their spacing, and test assemblers are free to select the set best fitting their needs.

To illustrate how the model is able to deal with possible additional test specifications,
it is assumed that the test form has to meet constraints on several quantitative and categorical item attributes. For example, if the total length of the test, as measured by its number of lines, should be smaller than a given number, the number of lines in item $i$ can be formulated as a quantitative attribute and a constraint on the values of this attribute can be added to the model to guarantee the result. Likewise, if some subsets of items in the pool differ in content, the assembly process can be constrained to produce a form with a desired distribution of the items over the subsets. Various other types of constraints are possible; for a review see van der Linden (1998) or van der Linden and Boekkooi-Timminga (1989). The set of quantitative constraints in the model is denoted by parameters $q_{is}$, $i = 1, \ldots, S$, whereas the subsets of items representing the values of the categorical attributes is denoted as $V_t$, $t = 1, \ldots, T$.

The model is as follows:

\begin{align*}
\text{minimize } y & \tag{13} \\
\text{subject to} & \\
\sum_{i=1}^{l} P_i(\theta_k)z_i - \sum_{j=1}^{n} P_j(\theta_k) & \leq y, \quad k = 1, \ldots, K; \quad r = 1, \ldots, R \tag{14} \\
\sum_{i=1}^{l} P_i(\theta_k)z_i - \sum_{j=1}^{n} P_j(\theta_k) & \geq -y, \quad k = 1, \ldots, K; \quad r = 1, \ldots, R \tag{15} \\
\sum_{i=1}^{l} z_i & = n, \tag{16} \\
\sum_{i=1}^{l} q_{is}z_i & \leq r_s, \quad s = 1, \ldots, S, \tag{17} \\
\sum_{i\in V_t} z_i & = n_t, \quad t = 1, \ldots, T, \tag{18} \\
z_i & \in 0, 1, \quad i = 1, \ldots, I, \tag{19} \\
y & \geq 0, \tag{20}
\end{align*}

The constraints in (14)–(15) require the difference between $\sum_{i=1}^{l} P_i(\theta_k)z_i$ and $\sum_{j=1}^{n} P_j(\theta_k)$ to be in the interval $[-y, y]$, $y \geq 0$, for $k = 1, \ldots, K$ and $r = 1, \ldots, R$ whereas the objective function in (13) minimizes $y$. The model therefore effectively minimizes the largest difference between these sums over the $\theta$ values selected and is thus of the minimax type. The constraint in (16) sets the length of the new test form equal to $n$. The constraints in (17) and (18) deal with all possible additional test specifications. The sums of values of quantitative attributes $q_{is}$ are required to be not larger than bound $r_s$, $s = 1, \ldots, S$, whereas from the subset of items representing categorical attribute value $t$ exactly $n_t$ items are selected, $t = 1, \ldots, T$. Finally, the constraints in (19) and (20) define the range of the decision variables.

As already noted, the series in (8) approximates the distribution generally good for only a few terms but that the precision of the results goes up if the upper bound $R$ in (14) and (15) is increased. This result is only analytical, however; the actual problem is one of combinatorial optimization. In practice, item pools are finite and not all possible combinations of values for the item parameters are available. As a consequence, imposing too
many of the conditions in (7) may occasionally lead to combinations of items compromis-
ing between the conditions with results slightly worse than those for a case of fewer
conditions. Though weights could be added to the right-hand side variables in the indi-
vidual constraints in (14) and (15), it is hard to base the choice of their values on a
theoretic argument. In the empirical example below, it proved best to include constraints
for the first two or three terms in (8) in the model and to apply no weighing.

The above model can be solved for optimal values of \( z_i \) and \( y \) using standard LP
software or the test assembly software package ConTEST (Timminga, van der Linden &
Schweizer, 1996). For models with a special structure, heuristics as in Luecht and Hirsch
(1992) are convenient. The choice of algorithm to solve the model is further addressed in
the presentation of the empirical example below.

Multidimensional Ability

A potential danger to IRT-based equating is lack of fit of the response model to the
data. Such lack of fit is most likely due to the fact that success on the item pool can be
dependent on more than one ability. An obvious remedy is to use a multidimensional
response model. A well-known model is the following extension of the 2-PL model, that is,
the model in (2) with \( c_i = 1 \) for \( i = 1, \ldots, n \):

\[
P_i(\theta) = P(U_i = 1|\theta_1, \ldots, \theta_D, a_{i1}, \ldots, a_{iD}, b_i) = \frac{\exp \left( \sum_{d=1}^{D} a_{id} \theta_d - b_i \right)}{1 + \exp \left( \sum_{d=1}^{D} a_{id} \theta_d - b_i \right)},
\]

(21)

where \( \theta_d, d = 1, \ldots, D, \) are the ability variables, \( a_{id} \) is the parameter for the discrimi-
nating power of item \( i \) along \( \theta_d, \) and \( b_i \) is a parameter for the composite difficulty of item
\( i. \) Detailed information about (a different parameterization of) the model is given in
McKinley and Reckase (1983), Reckase (1985, 1997), and Samejima (1974). To equate
tests measuring possible multidimensional abilities, Glas (1992) uses a multidimensional
Rasch or 1-PL model. The model is equivalent to the one in (21) with \( a_{id} = 1 \) for all \( i \) and
\( d, \) but assumes that the items display a "simple structure" with respect to their dependen-
cies on the ability variables; that is, the success on disjoint subsets of items in the pool is
modeled as being dependent on different unidimensional ability variables. In addition, the
individual abilities are linked by the assumption of a multivariate normal distribution for
the population of examinees. To assess if a set of response data fits the assumption of a
simple ability structure, goodness-of-fit tests for the 1-PL model can be applied to the
subsets of items or the hypothesis of a simple structure can be formulated in the multi-
dimensional model by Kelderman (1997) which has statistical tests for testing the hypoth-
esis.

Test assembly from a multidimensional item pool requires a slight adaptation of the
model. The only changes necessary are substituting the multidimensional response func-
tions into the conditions in (7) and specifying a multidimensional rather than a unidimen-
sional grid of ability values for the constraints. That is, (14) and (15) have to be replaced by

\[
\sum_{i=1}^{l} P_i(\theta_{dk}) z_i - \sum_{j=1}^{n} P_j(\theta_{dk}) \leq y, \quad d = 1, \ldots, D; k = 1, \ldots, K; r = 1, \ldots, R,
\]

(22)
In the model by Glas, the grids remain unidimensional but different grids have to be specified for the separate ability variables for the different subsets of items in the pool.

Other Scores than Number Correct

Test results are often reported as a conversion of the number-correct score. If the conversion is a monotonically increasing function, the test assembly model in this paper can just be applied to the number-correct scores which are then converted to the desired scale afterwards. Examples of conversions to which this principle applies are changes of origin and/or unit of number-correct scores and "formula scoring" to correct for possible random guessing on multiple-choice items.

Empirical Example

The test assembly model was applied to a former pool of 753 items from the Law School Admission Test (LSAT) program. The items in the pool were calibrated using the 3-PL model in (2). The pool consisted of items falling into three different content categories, labeled SA, SB, and IA here. In addition, items varied in (sub)type, gender and minority orientation, answer key, and word count. Finally, a portion of the item pool had a set structure with items in the same set sharing a common stimulus. The type of stimulus varied in content description. All existing specifications for the LSAT were modeled as linear constraints following the general format in (17) and (18). To model the inclusion of item sets in the test, a second type of decision variables was needed in addition to the variables $z_i$ in (13) and (20). In all, the model had 729 variables and 433 constraints. An old test assembled by hand to meet several specifications of the LSAT was known to the authors. The model in (13) and (20) was used to assemble new tests of 75 items with observed-score distribution optimally equated to the distribution on the old test.

The model was solved using the First Acceptable Integer Solution heuristic as implemented in the ConTEST program. The heuristic first calculates an upper bound to the value of the objective function in (13) relaxing the other decision variables in the model and then performs a branch-and-bound search for the optimal solution that is stopped when the first integer solution with objective function value within a small tolerance from the upper bound is found. The search is speeded up using the optimal reduced costs in the solution to the relaxed model to fix some of the decision variables. For further details, see Timminga, van der Linden and Schweizer (1996, sec. 6.6.5). In the current application, the search for a 0-1 solution was stopped as soon as the value of the objective function, $y$, was smaller than .01. The observed-score distributions for the old and new tests were generated according to (3) and (4), with $\theta$ distributed as $N(0, 1)$, using a recursive algorithm introduced in Lord and Wingersky (1984).

Two sets of results were calculated, both for $r = 1, \ldots, 3$ but one with the response functions controlled only at $\theta = 0$ and the other at $\theta_1 = -1.0$ and $\theta_2 = 1.0$. The probability functions of the observed-score distributions for the first set are plotted in Figures 1 through 3. The figures also show the extent to which the sums of the $r$th powers of the response functions for the old and new forms in (7) are equal. The general impression is that the probability functions of the old and new tests are nearly identical in all four cases, with perfect results for $r = 1, 2, 3$. In this case, the sums of the three powers of the response functions are also identical for all practical purposes. The sets of curves for
Figure 1
Probability functions of observed-score distributions (upper panel) and sums of the response functions (lower panel) for the two forms for \( \theta = 0 \) (solid line: new form; dashed line: old form).
Figure 2
Probability functions of observed-score distributions (upper panel) and sums of $r$th powers of the response functions (lower panel) for the two forms for $r = 1, 2$ and $\theta = 0$ (solid line: new form; dashed line: old form).
Figure 3
Probability functions of observed-score distributions (upper panel) and sums of $r$th powers of the response functions (lower panel) for the two forms for $r = 1, 2, 3$ and $\theta = 0$ (solid line: new form; dashed line: old form).
second case are given in Figures 4–6. For \( r = 1 \) the fit is comparable to the one obtained
for the previous case. The only change is a shift of the distribution slightly to the left. For
\( r = 1, 2 \) the results are nearly perfect again. The slight decrease in fit for \( r = 1, 2, 3 \) is due
to the size of the item pool. As noted earlier, if the item pool does not have all possible
combinations of parameter values, the goal to find a good compromise between all con-
ditions may lead to worse result, in particular if the new condition \( (r = 3) \) only has a slight
impact on the shape of the observed-score distribution.

As a standard practice, it is recommended that these graphs be supplemented with
one of the equating transformations for the set of values of \( r \) studied. The value with an
equating transformation "closest" to the identity line is the best choice. In the present case,
the differences between the transformations are too small to reveal their differences if
plotted in the same figure.

Discussion

The success of the test assembly model proposed in this paper is predicated on the
quality of the item pool. If the pool is small relative to the size of the form or not well
designed, the observed-score distribution on the form assembled may fit the distribution on
the target form not as well as in the empirical example for the LSAT above. It is hard to
give general rules for the size of a quality item pool. In principle, a small set of items with
distributions on the item attributes required by the constraints in the model could already
produce a perfect result. On the other hand, since it is hard to predict the values of the
items on statistical attributes, in practice a larger item pool has to written to produce such
a set. However, experience with several professionally developed item pools have shown
that item pools of 500 or more items generally give excellent solutions to test assembly
models with complicated sets of constraints.

If the composition of the item pool does not allow for a good solution, the use of the
test assembly model is still recommended. Its results guarantee that the additional trans-
formation necessary to equate the two test forms exactly involves a minimal distortion of
scale over all possible test forms from the pool. Whether or not additional (equipercentile)
equating is necessary can immediately be inferred from such output as in Figures 1 through
6.

The quality of the solution is also determined by the fit of the response data to the
IRT model. A large variety of statistical methods for testing the goodness of the model are
available in the literature. In addition, the accuracy of the item parameter estimates has to
be high. If sound statistical procedures for parameter estimation are used, the quality of
item calibration depends only on the sample size. Sample sizes for IRT equating have been
studied in Tang, Way and Carey (1993). For the method proposed in this paper the
requirements are expected not to be much different.

It seems tempting to conduct a study in which application of the method in this paper
for a well-designed item pool is followed by administration of the new and old forms and
estimation of the equipercentile equating transformation in (1) from their empirical dis-
tribution functions. As already noted, theory predicts this transformation to be the identity
line. However, if the identity line is not obtained, the results are rather inconclusive. If the
item pool is well designed and the response data for the two forms fit the IRT model, the
result would point at a possible problem with the equating design (e.g., differences in
motivation between groups, order effects, or groups sampled from populations with dif-
ferent ability distributions). Such validity threats are known to exist for current equating
methods (Kolen & Brennan, 1995, chap. 1) but are absent for the method of pre-equating
in this paper. On the other hand, if the response data do not fit the IRT model, the first
reaction should be to distrust equating at all and diagnose the quality of the items. For
Probability functions of observed-score distributions (upper panel) and sums of the response functions (lower panel) for the two forms for $\theta_1 = -1.0$ and $\theta_2 = 1.0$ (solid line: new form; dashed line: old form).

**Figure 4**

Probability functions of observed-score distributions (upper panel) and sums of the response functions (lower panel) for the two forms for $\theta_1 = -1.0$ and $\theta_2 = 1.0$ (solid line: new form; dashed line: old form).
FIGURE 5
Probability functions of observed-score distributions (upper panel) and sums of $r$th powers of the response functions (lower panel) for the two forms for $r = 1, 2$ and $\theta_1 = -1.0$ and $\theta_2 = 1.0$ (solid line: new form; dashed line: old form).
FIGURE 6
Probability functions of observed-score distributions (upper panel) and sums of rth powers of response functions (lower panel) for the two forms for $r = 1, 2, 3$ and $\theta_1 = -1.0$ and $\theta_2 = 1.0$ (solid line: new form; dashed line: old form).
example, if the lack of fit appears to be due to the fact that some of the items show differential item functioning (DIF), the test forms do not order examinees consistently and any equating method would produce meaningless results.

References


*Manuscript received 9/10/97
Final version received 2/12/98*