Article

An Evaluation of Different Statistical Targets for Assembling Parallel Forms in Item Response Theory

Applied Psychological Measurement
2016, Vol. 40(3) 163–179

DOI: 10.1177/0146621615613308
apm.sagepub.com
SAGE

Usama S. Ali[1,2] and Peter W. van Rijn[3]

Abstract

Assembly of parallel forms is an important step in the test development process. Therefore, choosing a suitable theoretical framework to generate well-defined test specifications is critical. The performance of different statistical targets of test specifications using the test characteristic curve (TCC) and the test information function (TIF) was investigated. Test length, the number of test forms, and content specifications are considered as well. The TCC target results in forms that are parallel in difficulty, but not necessarily in terms of precision. Vice versa, test forms created using a TIF target are parallel in terms of precision, but not necessarily in terms of difficulty. As sometimes the focus is either on TIF or TCC, differences in either difficulty or precision can arise. Differences in difficulty can be mitigated by equating, but differences in precision cannot. In a series of simulations using a real item bank, the two-parameter logistic model, and mixed integer linear programming for automated test assembly, these differences were found to be quite substantial. When both TIF and TCC are combined into one target with manipulation to relative importance, these differences can be made to disappear.

Keywords

test assembly, test information function, test characteristic curve, item response theory, statistical target, mixed integer linear programming

In many large-scale assessments, parallel test forms are needed so that the test can be taken at different sessions. Such forms need to meet a set of statistical and content requirements so that the test scores are comparable. The use of algorithmic approaches for creating test forms subject to a set of requirements is commonly referred to as automated test assembly (van der Linden, 2005). Both classical test theory (CTT) and item response theory (IRT) methods are extensively used in test assembly for the purpose of creating test forms that are parallel from a psychometric perspective (Sanders & Verschoor, 1998; van der Linden, 1998). The key in creating parallel

[1]Educational Testing Service, Princeton, NJ, USA
[2]South Valley University, Qena, Egypt
[3]ETS Global, Amsterdam, The Netherlands

Corresponding Author:
Usama S. Ali, Educational Testing Service, 660 Rosedale Rd, MS-02P, Princeton, NJ 08451, USA.
Email: uali@ets.org

forms lies in the definition of parallelism in terms of both statistical and content specifications (see, for example, McDonald, 1999). Typically, this definition consists of a statistical target function combined with a set of constraints. For example, the test characteristic curve (TCC) can be used as a statistical target to make sure that the test forms are of equal difficulty. In addition, examples of constraints are the distribution of items from each of the content domains, the distribution of selected- and constructed-response items, and test length. Although automated test assembly can be used for both linear and adaptive testing (see, for example, Diao & van der Linden, 2011), the focus of the study lies on creating parallel forms for linear tests. In this article, the effects of different statistical targets on the outcomes of the test assembly process are studied.

In test assembly, test developers try to satisfy both content and statistical specifications to the extent possible. Due to the different nature of testing programs, there exist different types of guidelines and rules to assemble and evaluate new test forms. An important choice seems to be needed between CTT and IRT methods. Davey and Hendrickson (2010) compared CTT and IRT statistical specifications by means of simulations under realistic operational conditions. They found, however, that both methods are quite capable of producing test forms that resemble a specified target form and that, therefore, choosing one method over the other is a matter of operational convenience. Similar results were found by other researchers in comparing CTT and IRT methods in assembling test forms (e.g., Ali, Guo, & Puhan, 2012; Fan, 1998; Lin, 2008). The focus of the study will be on IRT because it is used in many large-scale testing programs.

IRT methods have been applied extensively in automated test assembly using both unidimensional and multidimensional IRT models (Theunissen, 1985; Veldkamp, 2002). In many cases, the methods need to be automated because of the size of the problem. Mixed integer linear programming (MILP) has been used a lot for test assembly problems in which IRT is the psychometric framework (van der Linden, 2005). The problem is formalized by specifying an objective function subject to a number of constraints. For example, the objective is to create parallel forms of which the test information function (TIF) is as close as possible to a specified target. In contrast, the objective can be to create equally difficult forms, where the objective is to minimize the difference in TCCs.

Although there are examples where differences in either TIF or TCC are minimized (e.g., Armstrong, Belov, & Weissman, 2005; Armstrong, Jones, & Kunce, 1998), the TIF seems to be more commonly used as the statistical target function (Ackerman, 1989; Chyn, Tang, & Way, 1994; Swanson & Stocking, 1993; van der Linden & Adema, 1998; Veldkamp, Matteucci, & de Jong, 2013). The main reason for the focus on the TIF seems to be that it is believed that test difficulty is implied when the TIF is used (see, for example, Ariel, Veldkamp, & Breithaupt, 2006). However, in a comparison of three different test assembly methods, Chen, Chang, and Wu (2012) found that constraining the TIFs to be equal does not necessarily guarantee similarity of the TCCs.[1] Obviously, the opposite can occur as well: Differences in TIFs can be found when the objective is to minimize the difference in TCCs. Choosing either one as the statistical target can thus lead to either differences in difficulty or differences in information. Therefore, it is important to study the effect of selecting either the TIF or TCC as a statistical target on the outcomes of the test assembly process and the possible consequences for the purpose of the test.

The aim of the current study is to investigate the interplay between two different IRT targets of statistical test specifications: the TCC and the TIF. More specifically, the authors are interested in the extent to which test forms differ in terms of their TIFs when they are assembled based on a target TCC. In addition, they want to determine the differences in TCCs when test forms are assembled based on a target TIF. Furthermore, they investigate the case where both the differences in TCCs and TIFs are minimized. This can be performed in at least three different ways. First, the TIF can be used as the objective function, and differences in TCC are taken

care by means of a constraint. Second, obviously, this can be reversed as well. Third, targets are specified for both the TCC and TIF, and the objective is to minimize the sum of the differences. Both the TCCs and TIFs of different forms can be evaluated not only in an absolute manner (i.e., the distance to the target) but also in a relative manner (i.e., the distance between parallel forms).

MILP is selected to perform the assembly of the parallel forms for several reasons. First, MILP generally works well for creating parallel forms (see, for example, Chen et al., 2012). Second, MILP is nowadays available through standard statistical software packages such as R (see Diao & van der Linden, 2011). Finally, for the purpose of this article, the exact assembly method is actually not that relevant, as long as the assembly method can reasonably solve the stated combinatorial optimization problem (i.e., reach the objective under the given constraints).

The article is organized as follows: The two statistical targets in the context of IRT are described first. Next, the design, methods, and results of a simulation study for comparing the two targets are discussed. Finally, a method is introduced that combines both targets. The article ends with a discussion of the results.

Statistical Targets for Parallel Test Forms in IRT

For test forms to be parallel, they have to be equivalent in terms of their statistical and content-related properties. In CTT, test forms are referred to as strictly parallel if the true score and the conditional error variance are the same for everyone (Lord, 1980). However, in most practical settings, test forms are considered roughly statistically equivalent when their summary statistics (e.g., mean, variance, and reliability) are similar (Dorans, Pommerich, & Holland, 2007).

In the context of IRT, test forms are said to be strictly parallel if the items have equal parameters in a joint unidimensional (or multidimensional) IRT model (McDonald, 1999).[2] If test forms are strictly parallel in this sense, the TCCs and TIFs will be the same. However, other forms of parallelism are also entertained, and there is no agreement in the IRT literature on terminology with respect to parallel forms. The authors distinguish between strictly parallel, TCC&TIF-parallel, TCC-parallel, and TIF-parallel. The task of creating strictly parallel forms is feasible for small problems only, and generally this type of parallelism seems hard to realize. Therefore, the focus of the study is on the other three notions of IRT parallelism. TIF-parallel is also called weakly parallel in an IRT framework, and this definition does not require the number of items or score categories to be the same (Samejima, 1977).

Test assembly studies that focus on TIF-parallelism restrict the TIFs of the created forms within a certain range of a specified target TIF (van der Linden, 2005). For a new testing program, generating test forms that target a given shape of the TIF is useful in terms of managing the conditional standard error of measurement (CSEM). McDonald (1999) argued, however, that it would not make sense to match test forms on either TCC or TIF. For instance, if test forms are only TCC-parallel, differences in error variance can randomly affect the test scores. In contrast, if test forms are TIF-parallel, differences in TCCs can systematically affect the test scores. In both cases, issues of test fairness are at stake, although differences in TCC are less problematic if the item parameters of different test forms can be put on the same scale. Nevertheless, it makes sense to investigate the extent to which such differences can arise and how they can be mitigated.

The choice of IRT model is the unidimensional two-parameter logistic model (2PLM; Birnbaum, 1968). The item response function, which is the probability of a correct response $Y_i = 1$ on item i, of the 2PLM can be given by

$$P_i(\theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}, \tag{1}$$

where a_i is a discrimination parameter, b_i a difficulty parameter, and θ person ability. It is assumed that items are independent conditional on θ, so that the probability of response vector $\mathbf{Y} = \mathbf{y}$ can be written as the product over items:

$$\Pr(\mathbf{Y} = \mathbf{y}|\theta) = \prod_{i=1}^{n} P_i(\theta)^{y_i}(1 - P_i(\theta))^{1-y_i}. \tag{2}$$

This is generally referred to as the assumption of local independence (Hambleton, Swaminathan, & Rogers, 1991).

TCC

The TCC of a test is the sum of the item response functions because of the assumption of local independence. The TCC, denoted by $T(\theta)$, is given by the following equation:

$$T(\theta) = \sum_{i=1}^{n} P_i(\theta). \tag{3}$$

The TCC can be used as the basis for determining statistical equivalence in test equating. A useful feature is that the sum in Equation 3 can be used to predict the observed scores defined by the sum of observed item scores of test takers at given ability levels. If the test is made up of test items that are relatively difficult, then the TCC is shifted to the right and test takers tend to have lower expected scores on the test than if easier test items are included. In addition, the TCC can be used to explain how test takers with a fixed ability can perform differently on two tests measuring the same construct. When the IRT model holds, the TCC connects ability scores in IRT to true scores in CTT because a test taker's expected test score at a given ability level is by definition the test taker's true score on that set of test items (Hambleton & Jones, 1993).

TIF

The item information function gives the amount of information that an item can provide at specific ability values. Depending on the IRT model and the item type (i.e., dichotomous or polytomous), the item information function can differ in its shape and scale. Item information is additive as well because of the local independence assumption of IRT models. That is, the test information is simply a summation over all item information functions. Using the 2PLM, the TIF, denoted by $I(\theta)$, can be obtained by

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} a_i^2 P_i(\theta)(1 - P_i(\theta)), \tag{4}$$

where $I_i(\theta)$ is the item information.

Observed Score Distribution

Van der Linden and Luecht (1996) described an assembly method with a target for the observed-score distribution. To link the TCC and TIF to the observed-score distribution, let S

denote the sum score of a test with n items. Then, the conditional mean and variance of S are equal to (Lord, 1977)

$$\mathrm{E}(S|\theta) = T(\theta),$$

$$\mathrm{Var}(S|\theta) = \sum_{i=1}^{n} P_i(\theta)(1 - P_i(\theta)) = \sum_{i=1}^{n} \frac{1}{a_i^2} I_i(\theta).$$

According to van der Linden (2005), for any population, the observed score distributions for two test scores S_1 and S_2 are identical if and only if

$$\sum_{i}^{n} P_{i1}(\theta)^r = \sum_{i}^{n} P_{i2}(\theta)^r, \quad r = 1, \ldots, n, \tag{5}$$

where $-\infty < \theta < \infty$. It is easily seen that when r is equal to 1, Equation 5 means that the TCCs of the two tests are the same. In addition, if Equation 5 holds for $r \leq 2$, the conditional mean and variance of S as given above are the same. As the focus of this study is on IRT methods and not on observed-score methods, TCC and TIF targets are only considered.

Automated Test Assembly Using MILP

A test assembly problem translated into a MILP problem consists of two key elements: an objective function and a set of constraints (Theunissen, 1985). Following Diao and van der Linden (2011), the objective function can be written as

$$\min \mathbf{c}'\mathbf{x}, \tag{6}$$

where $\mathbf{x}$ is the vector of decision variables and $\mathbf{c}$ contains known coefficients. The set of constraints is given by

$$\mathbf{A}\mathbf{x} \leq \mathbf{d}, \tag{7}$$

with $\mathbf{A}$ a coefficient matrix with one row for each constraint and $\mathbf{d}$ a known numeric vector for the constraints.

In this case, m parallel forms of length n have to be created. Then, $\mathbf{x}$ consists of $n \times m$ variables, with x_{ik} equal to 1 if item i is allocated to test form k, and 0 otherwise. If forms that are TIF-parallel have to be created, z is minimized, subject to

$$\sum_{i=1}^{n} I_i(\theta) x_{ik} \leq \mathcal{I}(\theta) + z, \quad \text{and} \tag{8}$$

$$\sum_{i=1}^{n} I_i(\theta) x_{ik} \geq \mathcal{I}(\theta) - z, \quad \text{for } k = 1, 2, \ldots, m, \tag{9}$$

where $\mathcal{I}(\theta)$ is the target TIF and $z \geq 0$. Typically, the use of three to five ability levels to constrain the TIFs is enough to obtain good results, especially for longer tests consisting of dichotomous items (van der Linden, 2005). Other constraints relating to, for example, test length, overlap (i.e., common items), item format, and content domain can be easily added (see Diao & van der Linden, 2011). For example, if forms without overlap are to be created, the constraint $\sum_{k=1}^{m} x_{ik} \leq 1$ can be used for all items. Other constraints can be added in a similar

fashion. If TCC-parallel forms are to be created, the item information function and target TIF are simply replaced by the item response function and target TCC. The latter is denoted by $\mathcal{T}(\theta)$.

If test forms with minimal differences in both TCC and TIF are to be created, an assembly problem with multiple objectives is available.[3] Several methods for multiobjective test assembly are available (see, for example, van der Linden, 2005, § 3.3.4), but weighting of the two objectives is the most straightforward method and is known to work well (Veldkamp, 1999). If weights w_T and w_I for the relative importance of the TCC and TIF, respectively, are specified, z can be minimized again, but now subject to

$$\sum_{i=1}^{n} P_i(\theta)x_{ik} \leq \mathcal{T}(\theta) + w_T z, \tag{10}$$

$$\sum_{i=1}^{n} P_i(\theta)x_{ik} \geq \mathcal{T}(\theta) - w_T z, \tag{11}$$

$$\sum_{i=1}^{n} I_i(\theta)x_{ik} \leq \mathcal{I}(\theta) + w_I z, \tag{12}$$

$$\sum_{i=1}^{n} I_i(\theta)x_{ik} \geq \mathcal{I}(\theta) - w_I z, \quad \text{for } k = 1, 2, \ldots, m. \tag{13}$$

With the weighting method, a single objective function is retained. This is convenient because many solvers can only deal with a single objective function. The equivalence of either equal TIFs or equal TCCs can be considered more important depending on the application, but specifying the weights is not straightforward because the TCC and TIF do not have the same metric. The weighting method can be extended to include different weights for different values of θ. For example, a somewhat looser target may be used for extreme values of θ. This extension is relatively straightforward as only a limited number of θ values is used in practice.

Another way to address such relative importance is to minimize differences in the one considered most relevant as part of the objective function and set a maximum allowable difference for the other as constraint (see, for example, Debeer, Ali, & van Rijn, 2015). The maximum allowable difference can be specified in two ways. First, it can be specified with respect to the particular target (see, for example, van der Linden, 2005, § 6.2.2). A downside of using tolerances on targets through constraints is that they may not be feasible. For example, if the target is a flat TIF for some values of θ, one does not necessarily know the shape of the TCC (and vice versa). A second option then is to specify a tolerance on the differences between the TCCs without using a target TCC. A downside of both approaches is that the differences are not actually minimized. This can mean that for larger tolerances, there may exist better solutions than the one found by the solver, and for smaller tolerances, there may not exist a solution at all.

Another option is to use an absolute target for one and a relative target for the other.[4] In case of an absolute TIF target and relative TCC target, the objective is to minimize z with the regular constraint on the TIFs

$$\sum_{i=1}^{n} I_i(\theta)x_{ik} \leq \mathcal{I}(\theta) + w_I z, \tag{14}$$

$$\sum_{i=1}^{n} I_i(\theta)x_{ik} \geq \mathcal{I}(\theta) - w_I z, \quad \text{for } k = 1, 2, \ldots, m, \tag{15}$$

and the following constraint on the TCCs

$$\sum_{i=1}^{n} P_i(\theta)x_{ik} - \sum_{i=1}^{n} P_i(\theta)x_{ik'} \leq w_T z, \tag{16}$$

$$-\sum_{i=1}^{n} P_i(\theta)x_{ik} + \sum_{i=1}^{n} P_i(\theta)x_{ik'} \leq w_T z, \quad \text{for } 1 \leq k < k' \leq m. \tag{17}$$

In contrast to minimizing the relative difference, one can set a maximum allowable difference for the TCCs (ε_T). In specifying this tolerance, one can use, for example, the so-called difference that matters (DTM) criterion proposed by Dorans and Feigenbaum (1994) in the context of equating, which would amount to 0.5 if the unit of the score scale is 1 point. A TIF version can be found in an analogous fashion. In specifying the TIF tolerance, Luecht (2006) described procedures to control decision accuracy for the case of mastery testing. If the focus is not on mastery testing, one can also choose to minimize the TIF differences in terms of their relative efficiency as follows:

$$\frac{\sum_{i=1}^{n} I_i(\theta)x_{ik}}{\sum_{i=1}^{n} I_i(\theta)x_{ik'}} \leq 1 + z, \tag{18}$$

$$\frac{\sum_{i=1}^{n} I_i(\theta)x_{ik'}}{\sum_{i=1}^{n} I_i(\theta)x_{ik}} \leq 1 + z, \quad \text{for } 1 \leq k < k' \leq m. \tag{19}$$

Again, a tolerance can be chosen here (ε_I) instead of minimizing z. For example, $\varepsilon_I = 1.05$ means that the TIFs of all forms differ by less than 5% at evaluated values of θ. This can be directly related to the CSEM, because the variance of the maximum likelihood estimate of θ is the inverse of TIF. This constraint can be used in combination with a minimum value for the TIF, so that a certain precision is ensured. Obviously, such a minimum needs to be set realistically given the test length.

A disadvantage of the relative targets is that the constraints in Equations 16 to 19 grow exponentially with the number of test forms. In contrast, the constraints for the absolute targets in Equations 10 to 13 grow linearly with the number of forms to be assembled. This can have substantial impact on the amount of solving time, up to the point that it basically becomes prohibitive for practical purposes (i.e., multiple days). For example, for 10 forms, five θ points, and a relative TCC target, there are $\frac{1}{2} \times 10 \times 9 \times 5 \times 2 = 450$ constraints on the TCC (Equations 16 and 17). For an absolute target, there are $10 \times 5 \times 2 = 100$ constraints. For this reason, the absolute targets are the main focus in this study.

Method

In this study, a real item bank from a large-scale testing program is used. In addition to the statistical target (TIF vs. TCC), other factors are considered such as test length (n; 30 items to

represent short tests vs. 60 items to represent long tests) and number of test forms (m; 5 vs. 10 forms). The item parameters were obtained using the 2PLM.

As noted, the particular test assembly methodology is not critical for the authors' message. The method of MILP as described in a clear and accessible manner by Diao and van der Linden (2011) is chosen. With this method, different objective functions can be specified, and content and other constraints in the assembly of parallel forms can be included. The aim of the study is to investigate how severe the differences in TCCs (or TIFs) can be for parallel forms with nearly identical TIFs (or TCCs). The lpSolveAPI library of statistical software package R to assemble the tests is used. The source code for test assembly is available upon request.

Three different objective functions with absolute targets are compared to obtain TIF-parallel, TCC-parallel, and TIF&TCC-parallel test forms, respectively. The objective functions were described in the previous section. The assembled test forms are evaluated by making use of a graphical depiction of TIFs and TCCs. In addition, the means of both item parameters are computed. Finally, a mean square deviation (MSD) statistic is used to quantify the closeness between the assembled tests and the target form for both the TIFs and TCCs. The value of these MSD statistics is determined by

$$\mathrm{MSD}_{\mathrm{TIF}} = \frac{\sum_{j=1}^{q} \left[I(\theta_j) - \mathcal{I}(\theta_j) \right]^2}{q}, \tag{20}$$

$$\mathrm{MSD}_{\mathrm{TCC}} = \frac{\sum_{j=1}^{q} \left[T(\theta_j) - \mathcal{T}(\theta_j) \right]^2}{q}, \tag{21}$$

where q is the number of ability levels at which the functions are evaluated. If relative targets were used, the MSD between the minimum and maximum TCC or TIF could be used.

Data

An item bank of 392 multiple-choice items was used for assembling parallel test forms. The characteristics of the item bank according to the 2PLM are reported in Table 1. There are four content domains, and the content constraints are that each test form needs to have, respectively, 10%, 30%, 40%, and 20% items from each domain. The item bank information and response curves along with the five ability levels are displayed in Figure 1. In this study, the minimization at five ability levels is applied (i.e., -2.5, -1.5, -0.5, 0.5, 1.5). While we evaluated the assembled test forms over 41 points on the ability scale from -3 to 3, hence $q = 61$ in Equations 20 and 21. Although many different absolute and relative targets can be set, the ratio of the test length and the number of items in the bank multiplied by the TCC of the item bank as the statistical target are chosen to be used. With this TCC target, 10 forms were created, and the average TIF was taken as the TIF target. This method is suggested by van der Linden (2005). The target values for the tests of length 30 are reported in Table 2. For the assembly of the tests of length 60, the values of the statistical targets are simply doubled.

Results

The means and standard deviations of the item parameters and the MSD statistics for the five test forms ($m = 5$) assembled according to each of three statistical targets presented in Table 3

Table 1. Mean, Standard Deviation, Minimum, and Maximum of CTT and IRT Parameters for 392-Item Bank.

Statistic	*p*	*r*-biserial	*a*	*b*
M	.60	.54	0.81	−0.51
SD	.22	.12	0.23	1.07
Minimum	.13	.26	0.26	−3.74
Maximum	.97	.79	1.54	2.29

Note. CTT = classical test theory; IRT = item response theory.

Table 2. Target Values for TIF and TCC in a 30-Item Test With Five Ability Levels.

θ	$\mathcal{I}(\theta)$	$\mathcal{T}(\theta)$
−2.5	4.81	3.57
−1.5	9.60	8.27
−0.5	11.77	15.30
0.5	8.37	22.12
1.5	3.88	26.43

Note. TIF = test information function; TCC = test characteristic curve.

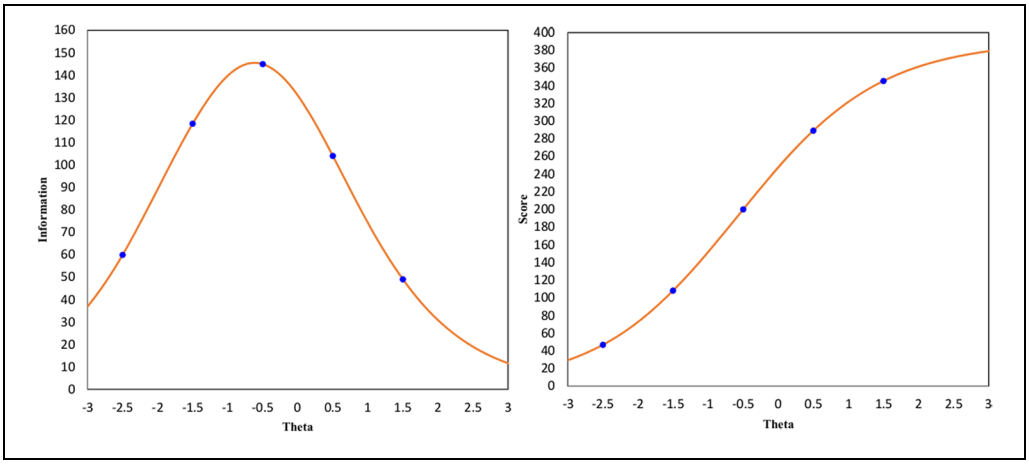

Figure 1. TIF and TCC of item bank.
Note. TIF = test information function; TCC = test characteristic curve.

for a test length of 30 items. It can be seen that the MSD statistics are small for the statistical target function that was used in assembling the test forms and large for the statistical function that was not used as a target. For example, MSD_{TIF} values are smaller (0.005 to 0.034) in the case of TIF-parallel test forms than those (0.437 to 3.928) in the case of TCC-parallel forms. In terms of the distribution of item parameters *a* and *b*, the TIF-parallel forms were consistent in terms of *a* parameters but not for *b* parameters. The TCC-based forms were consistent in both parameters, although more consistent for *b* than *a*. Highly similar results were found for the case of $n = 30$ and $m = 10$, so the authors refrain from showing them. They note that the content constraints were met in all cases.

Table 3. Mean and Standard Deviation of Item Parameters and MSD Statistics of Assembled Test Forms With Different Absolute Statistical Targets ($n = 30$, $m = 5$).

		a		*b*		MSD	
Target	Form	*M*	*SD*	*M*	*SD*	TIF	TCC
TIF	1	0.83	0.40	−0.46	1.42	0.005	0.837
	2	0.83	0.34	−0.19	1.34	0.034	2.089
	3	0.83	0.28	−0.34	1.24	0.013	0.614
	4	0.84	0.29	−0.64	1.36	0.011	0.447
	5	0.83	0.28	−0.76	1.23	0.015	0.977
TCC	1	1.01	0.31	−0.51	1.44	3.928	0.002
	2	0.95	0.25	−0.48	1.37	1.821	0.005
	3	0.95	0.23	−0.51	1.32	1.878	0.002
	4	0.90	0.22	−0.53	1.26	0.791	0.004
	5	0.87	0.23	−0.53	1.26	0.437	0.002
TIF&TCC	1	0.82	0.29	−0.51	1.09	0.002	0.003
	2	0.80	0.27	−0.52	1.22	0.023	0.006
	3	0.81	0.28	−0.50	1.01	0.024	0.001
	4	0.82	0.24	−0.51	1.18	0.003	0.002
	5	0.81	0.25	−0.51	1.02	0.006	0.002

Note. MSD = mean square deviation; TIF = test information function; TCC = test characteristic curve.

Table 4. Mean MSDs of TIF and TCC for Combined Absolute Target With Different Weights ($n = 30$, $m = 5$).

Weights[a]		MSD	
w_I	w_T	TIF	TCC
1	1	0.011	0.003
10	1	0.012	0.002
1	10	0.004	0.023
20	1	0.493	0.005
1	20	0.004	0.405
∞	1	1.771	0.003
1	∞	0.013	0.993

Note. MSD = mean square deviation; TIF = test information function; TCC = test characteristic curve.
[a]Lower means more important.

Table 4 shows the means of the MSDs for the combined absolute target with different weights for the TIF and TCC. It can be seen that the larger the weight, the larger the MSD is of the associated function. The values obtained with an infinite weight correspond to the values found in Table 3 for the single targets.

Figure 2 shows the results for creating TIF-parallel test forms of 30 items for both five and 10 forms. The left panel shows the TIFs, and the right panel shows the TCCs. From the graphical representation, the TIF-parallel forms generally had TCCs within 5% of the TCC target. However, such differences in difficulty are not small and might be highly unwanted. As noted, the results for both five and 10 forms are basically the same.

Figure 3 displays the results for five and 10 TCC-parallel forms of 30 items. Here, the left panel shows the TCCs and the right panel the TIFs. It can be seen that major differences are

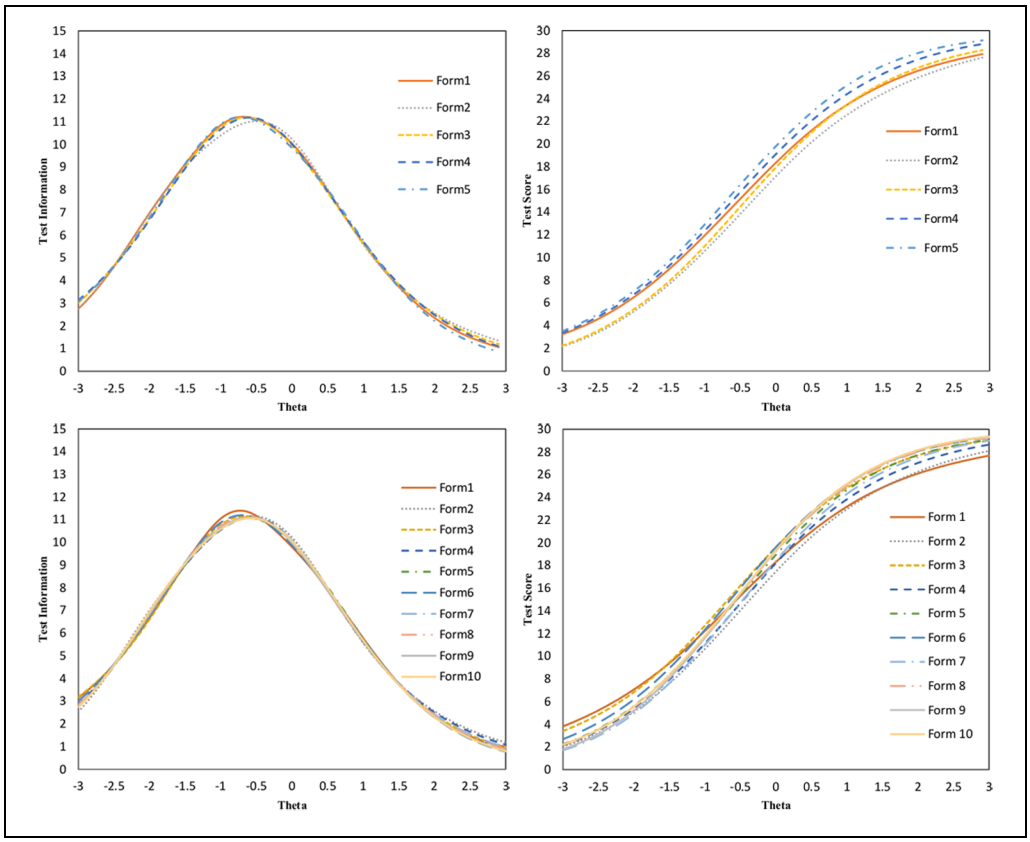

Figure 2. TIFs for five and 10 forms (top and bottom left) and their corresponding TCCs (top and bottom right) with TIF target for 30-item tests.
Note. TIF = test information function; TCC = test characteristic curve.

found between the TIFs for the assembled forms. This information gap is clearly seen in the ability levels ranging from −1.5 to 0.0, that is, the range were the test forms provide the most information. The maximum difference in TIF for these cases is almost 40% for $m = 10$.

The results of the test assembly for a test length of 60 items are similar to the results for test forms with 30 items. The means and standard deviations of the item parameters and the MSD statistics for the resulting five forms assembled according to each statistical target are presented in Table 5. Again, the MSD statistics are relatively small for the statistical target function that was used in assembling the test forms and relatively large for the statistical function that was not used as a target. For example, MSD_{TCC} values are smaller (0.004 to 0.014) in the case of TCC-parallel test forms than those (0.036 to 3.132) in the case of TIF-parallel test forms. In terms of the distribution of *a* and *b*, the TIF-based forms were only consistent in terms of *a* parameters but not for *b* parameters. A difference with the test length of 30 items is that the TCC-parallel forms of 60 items have less consistent *a* parameters (0.76 to 0.97 for 60 items vs. 0.87 to 1.01 for 30 items). In addition, the *SD* of the *b* parameters seem to vary somewhat more.

Figure 4 shows the results for creating TIF-parallel and TCC-parallel test forms of 60 items. The left panels show the TIFs, and the right panels the TCCs. From the graphical representation, the TIF-based assembled forms had parallel TCCs within 5% of the target (see the top right

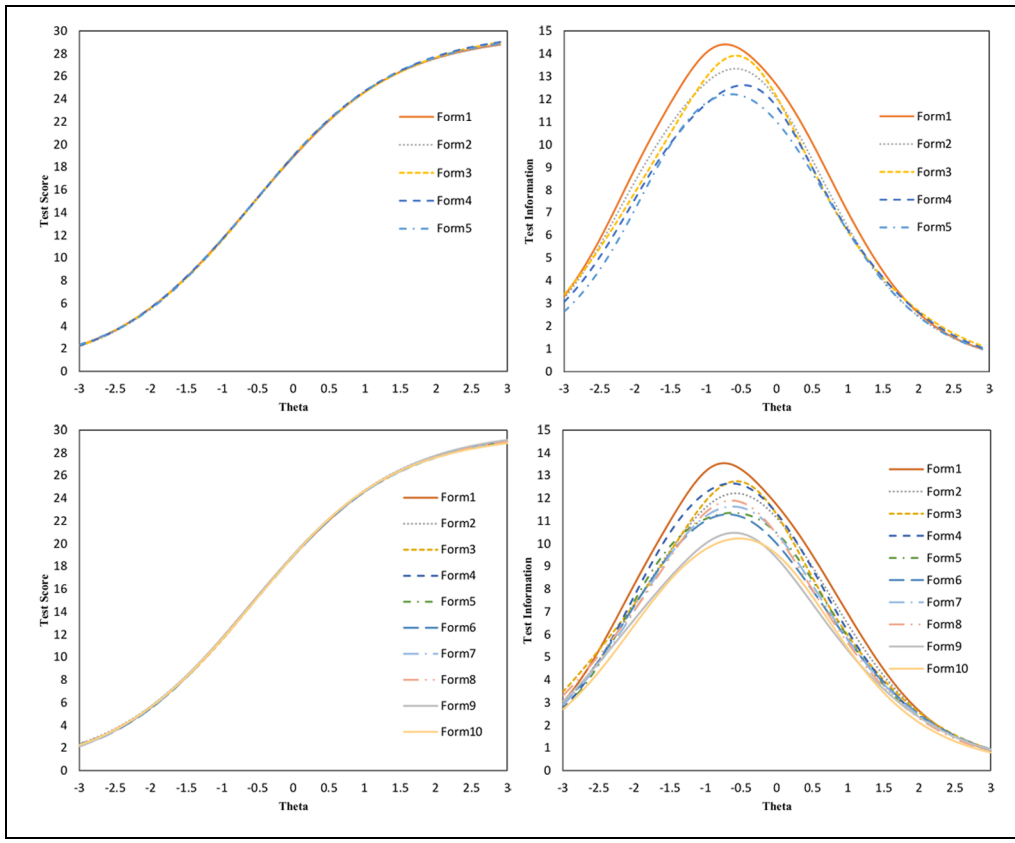

Figure 3. TCCs for five and 10 forms (top and bottom left) and their corresponding TIFs (top and bottom right) with TCC target for 30-item tests.
Note. TCC = test characteristic curve; TIF = test information function.

Table 5. Mean and Standard Deviation of Item Parameters and MSD Statistics of Assembled Test Forms With Different Absolute Statistical Targets ($n = 60$, $m = 5$).

		a		*b*		MSD	
Target	Form	*M*	*SD*	*M*	*SD*	TIF	TCC
TIF	1	0.83	0.37	−0.39	1.35	0.014	3.132
	2	0.83	0.27	−0.51	1.35	0.031	1.005
	3	0.81	0.20	−0.49	1.01	0.006	0.036
	4	0.80	0.20	−0.52	0.94	0.002	0.208
	5	0.80	0.17	−0.59	0.95	0.011	0.729
TCC	1	0.97	0.29	−0.52	1.41	10.377	0.011
	2	0.91	0.23	−0.51	1.23	3.010	0.011
	3	0.84	0.20	−0.50	1.14	0.219	0.005
	4	0.79	0.21	−0.49	1.06	0.240	0.014
	5	0.76	0.19	−0.51	0.97	1.071	0.004
TIF&TCC	1	0.81	0.33	−0.52	0.93	0.010	0.004
	2	0.82	0.24	−0.50	1.17	0.014	0.008
	3	0.81	0.24	−0.50	1.09	0.008	0.007
	4	0.81	0.22	−0.51	1.10	0.005	0.004
	5	0.82	0.20	−0.51	1.04	0.005	0.003

Note. MSD = mean square deviation; TIF = test information function; TCC = test characteristic curve.

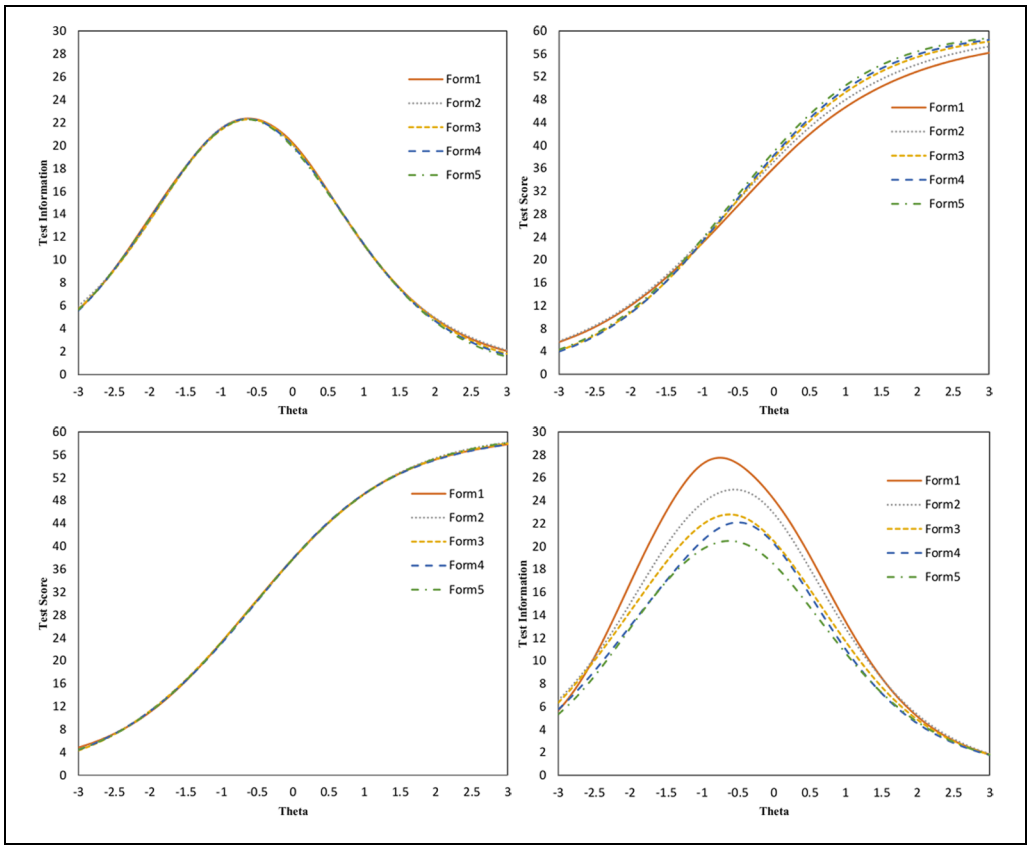

Figure 4. TIFs and TCCs for five forms (top left and right) with TIF target and TCCs and TIFs for five forms (bottom left and right) with TCC target for 60-item tests.
Note. TIF = test information function; TCC = test characteristic curve.

panel of Figure 4). Again, a major difference existed between the TIFs for the forms assembled according to the TCC target (see the bottom right panel of Figure 4).

Figure 5 shows the results of a combined TIF and TCC target for five test forms of 30 items. The differences in both TIFs and TCCs between the test forms are very small.

Discussion

Test assembly is a crucial step in the test development process (Wendler & Walker, 2006). Therefore, choosing a suitable theoretical framework to generate well-defined test specifications is critical. In test assembly, test developers try to satisfy both content and statistical specifications to the extent possible. There exist different types of guidelines to assemble and evaluate a new test form due to the different nature of testing programs. The focus of this study was to investigate the performance of three different IRT targets of statistical test specifications. These targets were aligned to different types of parallelism: TIF-, TCC-, and TIF&TCC-parallelism.

It is shown that there can be substantial differences in parallel forms if either the TIF or TCC is used. This is in contrast to the belief that test difficulty is implied when tests are TIF-parallel. Findings can be summarized as follows: If the TIF was used as the statistical target for test assembly, substantial differences in TCC between the resulting test forms are found. This

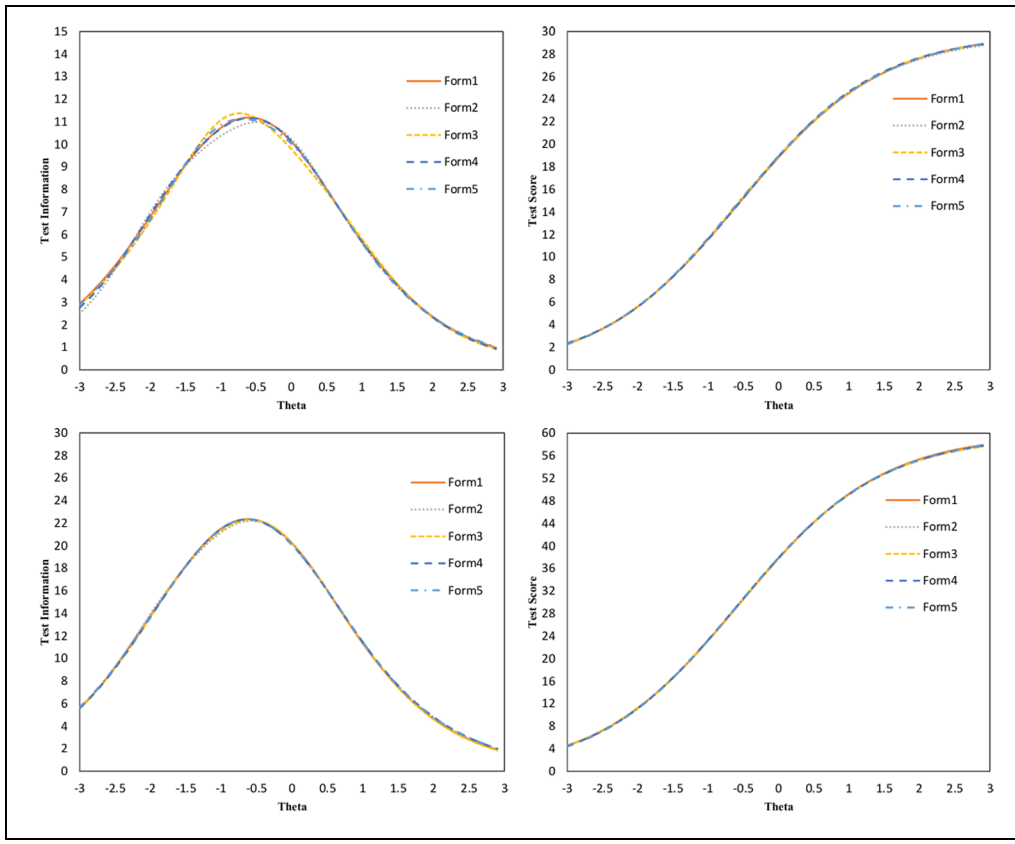

Figure 5. Information curves and their corresponding response curves for five 30-item test forms (top left and right panels) and same for five 60-item test forms (bottom left and right panels) assembled according to a combined information and response target.

difference was found with both shorter and longer tests, and with different numbers of test forms to be assembled. In most high stakes applications, differences in difficulty are acceptable only to a certain extent. With the TCC as the statistical target, the differences in TIF between the resulting test forms were much larger. Generally, it can be considered inappropriate to administer test forms with highly different precisions. This is analogous to the case of having two tests with different reliabilities.

Results show that when both TIF and TCC are combined as the statistical target, the differences observed with a single target disappear. This is a hybrid way to benefit from the features of both IRT targets and avoids their individual shortcomings. This relatively straightforward extension of the statistical target mitigates the problems found with each individual target. That is, both the differences in difficulty and precision can be controlled with the combined statistical target. However, an absolute target function proportional to that of the item bank is only used. The feasibility of this approach remains to be determined for other targets. If the combination of both TCC and TIF targets fails (i.e., the model is infeasible), there are other options as well. For example, one can specify the problem as a truly multiobjective optimization problem with the TCC and TIF targets as separate objective functions (Veldkamp, 1999). Another approach would be to use MILP with one target (e.g., TIF). Then, if a solution is found, build a second

model with the other target (TCC) and add the solution to the first problem as a constraint. This makes it possible to use different bounds on the solutions for different targets. A third approach would be to solve the MILP problem with one target (e.g., TIF) and determine the average of the other (e.g., the average of TCCs taken over all forms) with the solution. Then, add this average with some bound as a constraint, and run the MILP problem with the original target again. This approach can be sensible if one of the targets is more important than the other or has a particular shape (e.g., a flat TIF). In addition, the study investigated different weighting schemes for the purpose of relative importance of the two objectives within the combined target approach. The procedure used for setting these weights was used to overcome the different metrics of TCC and TIF. These weights are inversely related to the importance to meet a certain target.

A limitation of this study is that a relatively small item bank is used (some real item banks have thousands of items). From one aspect, the study results can be generalizable to a larger item pool given that the item pool is a big test in terms of test specifications (i.e., a test form is a mini-item pool), and these test specifications are proportional to the resources available in the item pool. From another aspect, this choice was made because there is a limit to how many items can fit in a unidimensional IRT model. Although there are assembly methods available using MIRT methods (e.g., Veldkamp, 1999, 2002). However, if the item bank is really large, so is the assembly problem. In that case, it might be wise to take a ''divide-and-conquer'' approach by using multiple assembly stages. In the first stage, a number of smaller, parallel item pools is created from the main item pool. Then, the parallel forms are created from each smaller item pool. This approach is described by van der Linden (2005) in the context of larger problems for adaptive testing. Also, a freely available solver is used (with known limitations). If the assembly problem really is large, it might be necessary to switch to commercial solvers (see, for example, Donoghue, 2015).

For better understanding of the implications and limitations of the findings, it is critical to consider the context in which different statistical targets can perform well. If a new testing program is about to be established, these contextual findings will be helpful. Therefore, it is recommended for future research to consider other factors that might affect the choice of an IRT statistical target in item selection to build new test forms. For example, only one IRT model for dichotomous items is considered. In addition, the impact of uncertainty in item parameter estimates on both TIF and TCC is not addressed (Veldkamp et al., 2013; Zhang, 2012). Also, the authors only focused on assembling test forms to be used for linear testing. It would be interesting to see what the differences between the statistical targets would look like for different test delivery methods. For instance, one could compare the targets in the assembly of blocks that are used in multistage testing (Zheng & Chang, 2015). Finally, one could compare the impact of the different targets on different scoring methods (number-correct vs. maximum likelihood estimators; Yen, 1984; Yen & Candell, 1991). Notwithstanding that these factors can have substantial impact on the results, it is expected that a combined TIF and TCC statistical target is likely to outperform each single one.

Acknowledgment

The authors thank Shelby Haberman, Frederic Robin, the editor, and the anonymous reviewers for their comments that improved this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Van der Linden and Luecht (1996) noted that a test characteristic curve (TCC) target implicitly constrains the test information function (TIF), because the slope of the TCC is related to the value of the TIF. However, this relationship is not one-to-one for most item response theory (IRT) models.
2. This is also referred to as item-parallel.
3. If the differences in either TCC or TIF are only minimized, it is actually already a multiobjective assembly problem because the minimization is performed at different values of θ (see van der Linden, 2005).
4. The notion of a relative target is different from that in van der Linden (2005, §5.1.1), because the amount of information is relative to the test forms and not relative to θ points.

References

Ackerman, T. A. (1989, March). *An alternative methodology for creating parallel test forms using the IRT information function*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Ali, U. S., Guo, H., & Puhan, G. (2012, July). *The correspondence of the classical and IRT methods in statistical test specifications*. Paper presented at the 77th annual meeting of Psychometric Society, Lincoln, NE.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement*, *30*, 204-215.

Armstrong, R. D., Belov, D. I., & Weissman, A. (2005). Developing and assembling the Law School Admission Test. *Interfaces*, *35*, 140-151.

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*, 237-247.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Chen, P.-H., Chang, H.-H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement*, *72*, 933-953.

Chyn, S., Tang, K. L., & Way, W. D. (1994). *An investigation of IRT-based assembly of the TOEFL Test* (ETS RR-94-38). Princeton, NJ: Education Testing Service.

Davey, T., & Hendrickson, A. (2010, April). *Classical versus IRT statistical test specifications for building test forms*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Debeer, D., Ali, U. S., & van Rijn, P. W. (2015, April). *Evaluating statistical targets for assembling parallel mixed-format test forms*. Paper presented at the annual meeting of National Council on Measurement in Education, Chicago, IL.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_Solve Version 5.5 in R. *Applied Psychological Measurement*, *35*, 398-499.

Donoghue, J. R. (2015). *Comparison of integer programming (IP) solvers for automated test assembly (ATA)* (ETS RR-15-05). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS RM-94-10) (pp. 91-122). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Pommerich, J., & Holland, P. (2007). *Linking and aligning scores and scales*. New York, NY: Springer-Verlag.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/persons statistics. *Educational and Psychological Measurement*, *58*, 357-381.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Lin, C.-J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *Journal of Technology, Learning, and Assessment*, *6*(8). Available from http://www.jtla.org

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575-596). Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*, 193-198.

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement*, *22*, 212-223.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151-166.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika*, *50*, 411-420.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195-211.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer-Verlag.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*, 185-198.

van der Linden, W. J., & Luecht, R. M. (1996). An optimization model for test assembly to match observed-score distributions. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 405-418). Norwood, NJ: Ablex Publishing.

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*, 253-266.

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement*, *26*, 133-146.

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*, 123-139.

Wendler, C. L., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445-467). Mahwah, NJ: Lawrence Erlbaum.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*, 93-111.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education*, *4*, 209-228.

Zhang, J. (2012). The impact of variability of item parameter estimators on test information function. *Journal of Educational and Behavioral Statistics*, *37*, 737-757.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*, 104-118.