JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
*Journal of Educational Measurement*
*Summer 2017, Vol. 54, No. 2, pp. 218–242*

# Evaluating Statistical Targets for Assembling Parallel Mixed-Format Test Forms

**Dries Debeer**
*University of Zurich*
**Usama S. Ali**
*Educational Testing Service*
**Peter W. van Rijn**
*ETS Global*

*Test assembly is the process of selecting items from an item pool to form one or more new test forms. Often new test forms are constructed to be parallel with an existing (or an ideal) test. Within the context of item response theory, the test information function (TIF) or the test characteristic curve (TCC) are commonly used as statistical targets to obtain this parallelism. In a recent study, Ali and van Rijn proposed combining the TIF and TCC as statistical targets, rather than using only a single statistical target. In this article, we propose two new methods using this combined approach, and compare these methods with single statistical targets for the assembly of mixed-format tests. In addition, we introduce new criteria to evaluate the parallelism of multiple forms. The results show that single statistical targets can be problematic, while the combined targets perform better, especially in situations with increasing numbers of polytomous items. Implications of using the combined target are discussed.*

In many large-scale educational testing programs, different test forms are assembled with minimal or no overlap to secure the test and maintain its validity. At the same time, it should be a matter of indifference for test takers which test form is administered (Lord, 1980). Thus, test forms that are to be used in parallel should be equivalent in terms of their statistical and content-related properties (Chen, Chang, & Wu, 2012). The automated selection of items from an item pool to assemble one or more new test forms, commonly referred to as automated test assembly (ATA), plays a key role in the creation of parallel tests forms (van der Linden, 2005).

When applying ATA, the parallelism requirement can be operationalized in different ways. In the context of classical test theory (CTT; Lord, 1980, p. 6), statistical parallelism means that test forms produce equal true scores and that the observed-score variances, conditional on the true scores, are the same in each test form (van der Linden & Luecht, 1998). In the context of item response theory (IRT), strict parallelism is the case when test forms have items with equal parameters in an equal IRT model (i.e., item-parallel forms; McDonald, 1999). However, given the limitations of a realistic item bank, the requirement for test forms to be item-parallel may often be too stringent. Therefore, rather than attempting to develop parallel items, test forms are assembled to have equal test characteristics as represented by the test information function (TIF; van der Linden, 2005), or the test characteristic curve (TCC; Armstrong, Belov, & Weissman, 2005). In this article, the statistic that is

 

used to attain parallelism across test forms in ATA will be referred to as a *statistical target*.

The choice for a specific statistical target (i.e., TIF or TCC) has its consequences. Although it is often implicitly assumed that test forms assembled to be TIF-parallel are TCC-parallel (and vice versa), this is not necessarily the case (Chen et al., 2012). On the one hand, when test forms are assembled to be TCC- but not TIF-parallel, the conditional measurement error at certain ability values can differ between the test forms. On the other hand, when tests forms are assembled to be TIF- but not TCC-parallel, there can be systematic differences in the expected test score of the test forms (i.e., the tests are not equally difficult for certain ability values). In this article, differences in the conditional measurement precision across test forms that should be parallel (i.e., different TIF values at certain ability values) will be referred to as *information gaps*. Differences in the conditional expected test score (i.e., TCC) will be referred to as *score gaps*. When equated scores are reported, score gaps are not necessarily problematic, but they are problematic when only sum scores are reported. However, information gaps are problematic in both situations because they lead to different measurement errors in reporting.

Recently, Ali and van Rijn (2016) showed that for the two-parameter logistic model (2PLM; Birnbaum, 1968), the use of either the TIF or the TCC as the statistical target to attain parallelism can give rise to either score gaps or information gaps, respectively. As a solution they proposed a combined approach where both the TCC and TIF are considered as the statistical target. Ali and van Rijn (2016) obtained promising results that mitigate the problems produced by using either the TIF or the TCC as a statistical target. Moreover, their combined approach has been applied successfully to a reasoning test with 2PLM items (Weeks, 2015).

The findings of Ali and van Rijn (2016) are limited to the case of 2PLM items. In practice, however, many assessments combine both dichotomously scored items (e.g., multiple choice items) and polytomously scored items (e.g., constructed response items), because combining different item formats tends to increase the validity of test scores (Ercikan et al., 1998). There exist several examples of high-stakes assessments that use items of mixed formats and administer multiple parallel forms such as TOEFL®, SAT®, and GMAT®. In this article, we aim to extend the research of Ali and van Rijn (2016) to item pools that contain polytomous items besides dichotomous items (i.e., mixed-format test forms), as well as to other item response models. More specifically, we consider the one-parameter logistic model (1PLM; Rasch, 1993), the 2PLM, and the three-parameter logistic model (3PLM; Birnbaum, 1968) with respect to dichotomously scored items; and the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992) with respect to polytomously scored items. In addition, we introduce new criteria to evaluate the parallelism of test forms that are directly interpretable and relevant for practice.

In the next sections, we first introduce the response and information functions for polytomous items by the GPCM, and explain why adding polytomous items is expected to complicate the assembly of parallel test forms. Then, we discuss ATA problems as formulated by mixed integer linear programming models. In addition to

the combined approach of Ali and van Rijn (2016), we propose two new methods to combine statistical targets in ATA.

## Automated Assembly of Mixed-Format Tests

### Statistical Targets in the GPCM

Different IRT models for polytomously scored items with successively ordered response categories have been proposed, such as the graded response model (Samejima, 1969) and the rating scale model (Andrich, 1978). In this article, however, we focus on the PCM (Masters, 1982) and the GPCM (Muraki, 1992) because of their analogy with the 1PLM and the 2PLM. Both polytomous models are based on the assumption that the probability of selecting the $k$th category over the $k-1$th category is governed by a dichotomous response model. For an item $i$ with possible response categories $Y_i = 0, 1, \ldots, m_i$ (with $m_i$ being the highest possible category for that item), the probability of selecting the $k$th category $P_{ik}(\theta)$ given the latent ability $\theta$ is formulated by the GPCM as

$$P_{ik}(\theta) = \Pr(Y_i = k|\theta) = \frac{\exp\left[\sum_{v=0}^{k} Da_i(\theta - (b_i + d_{iv}))\right]}{\sum_{g=0}^{m_i} \exp\left[\sum_{v=0}^{g} Da_i(\theta - (b_i + d_{iv}))\right]}, \tag{1}$$

for $k = 0, 1, \ldots, m_i$, and with $d_{i0} = 0$ and $\sum_{v=0}^{k} d_{iv} = 0$ for all items $i$, and where $D$ is a scaling constant — typically 1.7. Equation 1 is called the item category characteristic curve. The slope parameter $a_i$ is analogous to the item discrimination parameter in the 2PLM. The $b_i$ and the $d_{ik}$ are the item location parameter and the $m_i$ threshold distances, respectively. If the index $i$ of $a_i$ is dropped, the GPCM reduces to the PCM. If $m_i = 1$, the model in Equation 1 reduces to the 2PLM. The combination of these two restrictions results in the 1PLM.

In the GPCM, the item characteristic curve (ICC) or response function, denoted $P_i$, is the conditional mean of item scores at a given $\theta$ (Muraki, 1993):

$$P_i(\theta) = \sum_{k=0}^{m_i} k\, P_{ik}(\theta). \tag{2}$$

Although there are many possible shapes for the ICC in Equation 2, unless the slope parameter $a_i$ in Equation 1 is negative, the ICC a strictly increasing function of $\theta$.

The TCC for a certain test models the probability of a test score given the latent ability $\theta$ of a person. Because of the local independence assumption, the TCC, denoted $T(\theta)$, is the sum of the ICCs of all $n$ items in a test:

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

The item information function (IIF), denoted $I_i(\theta)$, represents the expected information contributed by a specific item $i$ across the $\theta$ range. The IIF in the GPCM (Donoghue, 1994) can be formulated as

$$I_i(\theta) = D^2 a_i^2 \left[ \sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left( \sum_{k=0}^{m_i} k P_{ik}(\theta) \right)^2 \right]. \quad (4)$$

Because of the local independence assumption, the TIF, denoted $I(\theta)$, is obtained by the sum of the IIFs of all items in the test

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta). \quad (5)$$

**Impact of Polytomous Items on Test Assembly**

For dichotomous items modeled by the 1PLM, the shapes of the ICCs and the IIFs are always the same, but their locations can differ according to the item difficulty parameter. In addition to these location differences, when using the 2PLM, the steepness of the ICCs and the peakedness of the IIFs can differ depending on the item discrimination parameter. On top of these effects, the guessing parameter within the 3PLM sets the lower bound for the ICC and mainly decreases the height of the IIF. Despite such effects of parameters on the ICCs and IIFs of dichotomous items, their shapes are similar and their differences are expected.

With respect to polytomous items, these curves (i.e., ICCs and IIFs) are more diverse. For instance, Muraki (1993) showed that for the GPCM there can be more than one peak in the IIF, depending on the threshold distances $d_{ik}$. The more categories included, the greater the possible variety in ICC and IIF shapes. As an illustration, Figure 1 presents the ICCs and IIFs of two five-category items following the GPCM. Although both items have the same slope and location parameters, the curves are dissimilar.

The greater variety in ICCs and IIFs for polytomous items is likely to cause a greater variety in the shapes of the TCC and TIF of test forms that contain more polytomous items. Hence, we expect that the discrepancies demonstrated by Ali and van Rijn (2016) in tests with only 2PLM items will be even more pronounced in test forms that include polytomous items, such as mixed-format tests, but less pronounced in test forms that include only 1PLM items.

**ATA via Mixed Integer Linear Programming**

ATA is the automated process of selecting specific items from an item pool to form new test forms that satisfy a series of specifications. Different methods are available. Some are based on sampling-and-classification methods such as the Cell Only and the Cell and Cube methods (Chen et al., 2012), while others rely on constrained combinatoral optimization techniques (e.g., Finkelman, Kim, and Roussos, 2009; Luecht, 1998; van der Linden, 2005). The most commonly used optimization technique involves translating the ATA problem to a mixed integer linear programming (MILP) model (Diao and van der Linden, 2011; Theunissen, 1985; van der Linden, 2005).

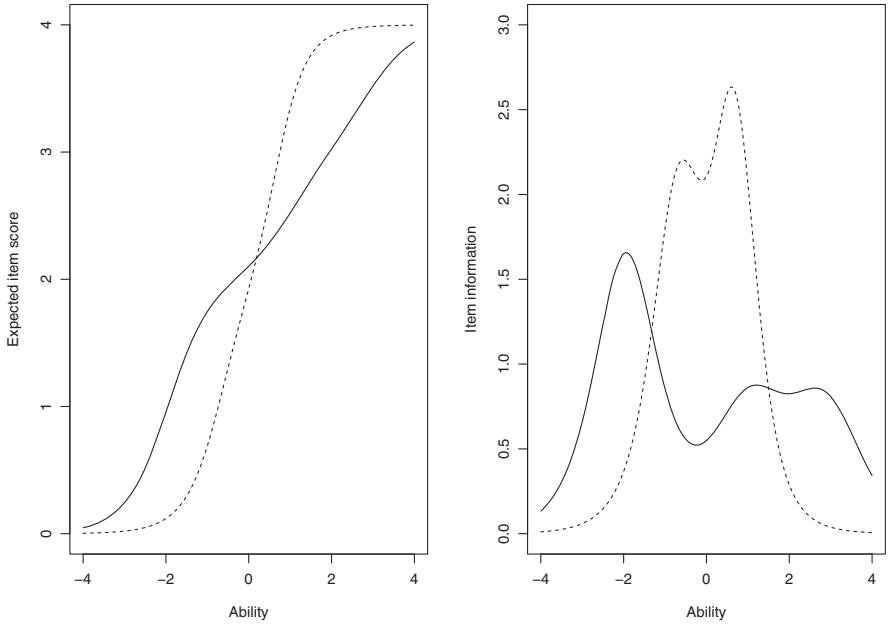


*Figure 1.* The item characteristic curves (left-hand panel) and item information curves (right-hand panel) of two GPCM items with five response categories. The threshold distances $d_{ik}$ for the solid and the dashed lines are, respectively, $-2.2$, $-1.7$, 1.0, and 2.9; and $-0.7$, $-0.7$, 1.0, and 0.4. The slope and location parameters for both items are 1 and 0, respectively.

Because MILP and linear programming models are widely used in various industries, efficient algorithms have been formulated, and many different software packages (usually referred to as "solvers") are available. Some of the commercial solvers, such as XPRESS (FICO, 2016), CPLEX (IBM, 2016) and GUROBI (Gurobi Optimization, 2016), are extremely powerful. However, there are also free open source solvers available such as lp_solve (Berkelaar, Eikland, & Notebaert, 2004) and GLPK (*GNU Linear Programming Kit*, 2016). (For an overview of the solvers that can be used for ATA problems see Donoghue, 2015). In this article we focus on the MILP approach for ATA problems.

MILP models are linear programming models for which (some of) the variables are restricted to integer values. When variables in a MILP model are further restricted to be binary, they can be interpreted as decision variables. In the context of ATA, each decision variable relates to whether or not a specific item from the pool is selected to be in a specific test form. In general, a MILP model includes two main parts (Ali and van Rijn, 2016; Diao and van der Linden, 2011): (1) the objective function,

$$\min \ \mathbf{c}^T \mathbf{x}; \tag{6}$$

and (2) a set of constraints,

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

where **x** is the vector of variables that MILP needs to solve for, and **c** is a numeric vector of known coefficients for the objective function. **A** is a known coefficient matrix with one row for each constraint and **d** is a vector with the corresponding right-hand values of the constraints.

The specifications that assembled test forms should satisfy, such as the number of items from a specific content domain, can also be divided into objectives and constraints (van der Linden, 2005). Constraints require a test or item attribute to satisfy an upper or lower bound. Objectives require an attribute to take a minimum or maximum possible value. Translating the constraints and the objectives from the ATA problem to a MILP model is generally straightforward.

One of the most commonly used MILP approaches for ATA of parallel test forms is the Minimax model (van der Linden, 2005). After deciding the desired shape of the statistical target for the assembled test form, the Minimax model minimizes the maximum distance between the statistical target of the assembled test forms and the statistical target of the reference test forms (Chen et al., 2012). Although there are other MILP approaches available for the ATA of parallel test forms, in this article we focus on the Minimax approach. We discuss examples of constraints for test assembly in the Minimax model using single and combined statistical targets.

**Minimax models with single statistical targets.** If $F$ new test forms are assembled from an item pool of $J$ items to be parallel with existing or ideal reference test form $t$, then we have $J \times F$ binary decision variables $x_{jf}$, $j = 1, 2, \ldots, J$, and $f = 1, 2, \ldots, F$ in **x** that indicate if item $j$ is selected for test form $f$. If the TCC is used as the statistical target to attain parallelism, the Minimax model minimizes the maximum absolute distance between the TCC of the assembled test forms and the TCC of the reference test form (also referred to as the target TCC) at $H$ prespecified ability values (or θ-points), using the following constraints:

$$\begin{aligned}
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} - u_h z \leq T^{(t)}(\theta_h), \\
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} + u_h z \geq T^{(t)}(\theta_h), \quad \text{for } h = 1, 2, \ldots, H, \\
&\qquad \text{and for } f = 1, 2, \ldots, F,
\end{aligned} \tag{8}$$

where $P_j(\theta_h)$ is the ICC of item $j$ at θ-point $h$, $T^{(t)}(\theta_h)$ is the target TCC at θ-point $h$, and $z$ is an additional positive real-valued variable (i.e., $z \geq 0$) in **x** that should be minimized:

$$\min(z). \tag{9}$$

Further, $u_h$ is the weight for $z$ at θ-point $h$. In this case, the constraints in Equation 8 translate to two times $H \times F$ rows in **A** with the corresponding right-hand values in **d**, and **c** contains zeros for all $x_{jf}$ and 1 for $z$, so that Equation 6 is equal to Equation 9. When the weights $u_h$ are 1 for all θ-points $h$, $z$ corresponds to the maximum distance between the TCC of a new test form and the target TCC at one of the prespecified θ-points. Hence, the maximum score gap between two new test forms at

one of the θ-points is $2z$. If score gaps are considered more problematic at certain θ-points, the weights $u_h$ can be adjusted accordingly. Note that the absolute distance is used, rather than the Euclidean distance. The latter would result in a nonlinear programming model, and although many solvers can deal with quadratic constraints, in general, linear problems are easier to solve.

Other constraints related to test length, item content, item type, and so on, can be added to **A** and **d** in a similar way. For instance, if item overlap is not allowed among operational test forms:

$$\sum_{f=1}^{F} x_{jf} \leq 1, \quad \text{for } j = 1, 2, \ldots, J, \tag{10}$$

results in $J$ extra rows in **A** and **d**.

When the TIF is used as the statistical target to attain parallelism in the Minimax model, the maximum absolute distance between the TIF of the new test forms and TIF of the reference test form (also referred to as target TIF) is minimized at $H$ prespecified θ-points. In this situation, the constraints in Equation 8 are replaced by

$$\begin{aligned} &\sum_{j=1}^{J} I_j(\theta_h)x_{jf} - u_h z \leq I^{(t)}(\theta_h), \\ &\sum_{j=1}^{J} I_j(\theta_h)x_{jf} + u_h z \geq I^{(t)}(\theta_h), \quad \text{for } h = 1, 2, \ldots, H, \\ &\quad \text{and for } f = 1, 2, \ldots, F, \end{aligned} \tag{11}$$

where $I_j(\theta_h)$ is the IIF of item $j$ at θ-point $h$ and $I^{(t)}(\theta_h)$ is the target TIF at θ-point $h$. When the weights $u_h$ are 1 for all θ-points $h$, $2z$ corresponds to the maximum information gap between two new test forms at one of the prespecified θ-points. Note that the value of such an information gap is not directly interpretable because the scale of the TIF is unknown. Information gaps, however, can be interpreted using the relative difference between the TIF values of the two test forms at the specific θ-point. Therefore, as an alternative, we propose to minimize the maximum relative distance between the TIF of the assembled test forms and the target TIF:

$$\begin{aligned} &\sum_{j=1}^{J} \tfrac{I_j(\theta_h)}{I^{(t)}(\theta_h)} x_{jf} - u_h z \leq 1, \\ &\sum_{j=1}^{J} \tfrac{I_j(\theta_h)}{I^{(t)}(\theta_h)} x_{jf} + u_h z \geq 1, \quad \text{for } h = 1, 2, \ldots, H, \\ &\quad \text{and for } f = 1, 2, \ldots, F. \end{aligned} \tag{12}$$

When all the weights $u_h$ are 1, $z$ is the maximal relative difference between the TIF of a new test form and the target TIF, at one of the prespecified θ-points. Because this relative difference is expressed as a proportion of the target TIF at the corresponding θ-point, the difference is directly interpretable. In addition, the maximal absolute difference in precision is automatically adjusted according to the desired precision at the prespecified θ-points.

**Minimax models with combined statistical targets.** In their study, Ali and van Rijn (2016) proposed combining the two statistical targets in a Minimax model by minimizing the maximum distance between the new test forms and a reference test form with respect to both the TIF and the TCC simultaneously. Hence, Equations 11 and 8 are combined, with weights $w_I$ and $w_T$ to specify the relative importance of the TIF and the TCC, respectively:

$$\begin{aligned}
&\sum_{j=1}^{J} I_j(\theta_h)x_{jf} - w_I z \leq I^{(t)}(\theta_h),\\
&\sum_{j=1}^{J} I_j(\theta_h)x_{jf} + w_I z \geq I^{(t)}(\theta_h),\\
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} - w_T z \leq T^{(t)}(\theta_h),\\
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} + w_T z \geq T^{(t)}(\theta_h), \quad \text{for } h = 1, 2, \ldots, H,\\
&\quad \text{and for } f = 1, 2, \ldots, F. \qquad (13)
\end{aligned}$$

This approach retains a single objective function (cf. Equations 9 and 6), which is convenient because many solvers can only deal with a single objective function. A disadvantage of this approach is that specifying the weights $w_I$ and $w_T$ is not straightforward because the TCC and the TIF do not have the same metric. Furthermore, the value of the objective function is not directly interpretable. For simplicity, the weights $u_h$ were omitted in Equation 13. If score gaps or information gaps are considered more problematic at certain $\theta$-points, the weights $u_h$ can be added to the constraints.

As an alternative way to combine the TIF and the TCC as statistical targets to obtain parallelism in a Minimax model, we propose minimizing the distances with respect to the statistical target that is considered most important, and specifying a maximum allowable difference (or absolute tolerance) to the other statistical target. For instance, when the TIF is considered most important, the constraints in Equation 11 or Equation 12 can be combined with the following constraints for the TCC:

$$\begin{aligned}
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} \leq T^{(t)}(\theta_h) + y_h^T,\\
&\sum_{j=1}^{J} P_j(\theta_h)x_{jf} \geq T^{(t)}(\theta_h) - y_h^T, \quad \text{for } h = 1, 2, \ldots, H,\\
&\quad \text{and for } f = 1, 2, \ldots, F, \qquad (14)
\end{aligned}$$

where $y_h^T$ is the absolute tolerance with respect to the TCC at $\theta$-point $h$. To set the absolute tolerance the "difference that matters criterion" (proposed by Dorans and

Feigenbaum (1994) in the context of equating) can be used. For instance, if the unit of the score scale in the test forms is 1, the absolute tolerance can be set to $y_h^T = 0.25$ so that the TCCs of the new test forms are constrained to lie within a distance of 0.25 of the target TCC at θ-point $h$. Hence, the expected test scores for the new test forms will differ at the most $2y_h^T = 0.5$ points at that θ-point.

Similarly, Equation 8 can be combined with constraints that specify a maximum allowable difference with respect to the target TIF. Although absolute tolerances can be specified with respect to the TIF, an alternative is to use a relative tolerance:

$$\begin{aligned}
&\sum_{j=1}^{J} I_j(\theta_h)x_{jf} \leq (1 + y_h^I) \times I^{(t)}(\theta_h), \\
&\sum_{j=1}^{J} I_j(\theta_h)x_{jf} \geq (1 - y_h^I) \times I^{(t)}(\theta_h), \quad \text{for } h = 1, 2, \ldots, H, \\
&\qquad \text{and for } f = 1, 2, \ldots, F,
\end{aligned} \tag{15}$$

where $y_h^I$ is the relative tolerance with respect to the TIF at θ-point $h$, defined as a proportion of the target TIF. To set the relative tolerance for the TIF in Equation 15, the difference that matters criterion can be used. For instance, if the relative tolerance is set to $y_h^I = \frac{41}{841}$, the maximum relative difference between the standard errors for the maximum likelihood ability estimate at θ-point $h$ in the new test forms will be at most 5%.

**Some remarks.** For an existing assessment, reference forms are available in the form of previously used test forms, and new test forms can be assembled to be parallel with this reference form. When there is no reference form, other MILP approaches are available. For instance, the TIF of the new test forms can be maximized (at prespecified θ-points), while the distances between the TIF and/or TCC of the new test forms are minimized (at prespecified θ-points) to attain parallelism. There are, however, some caveats for approaches that maximize the TIF. First, maximizing the TIF favors highly discriminating items, and can result in an unbalanced usage of the item pool. It can deplete the item pool, which can make the assembly of new parallel test forms in a later phase problematic. Second, the number of constraints needed to specify the distances between new test form with respect to the used statistical target increases exponentially with the number of test forms, which can have a substantial impact on the solving time (Ali & van Rijn, 2016). As an alternative, in these situations, an ideal test form can be fabricated, taking into account the goal of the assessment and the composition of the item pool. The ideal test form can then be used as the reference form in a Minimax model.

## Method

The aim of this study is to evaluate the performance of a Minimax model with one statistical target versus the combined Minimax models described above in attaining parallelism. These models are compared in item pools with different numbers of polytomous items. We also took into account the item response models used in the

Table 1
*Design of the Study*

| Item Types | IRT Model Combination | | |
|---|---|---|---|
| | 1PLM & PCM | 2PLM & GPCM | 3PLM & GPCM |
| Only Dichotomous | I | II | III |
| Dichotomous & 3-category Polytomous | IV | V | VI |
| Dichotomous & 3-, 4-, or 5-category Polytomous | VII | VIII | IX |

item pools. Based on mathematics items from the 1996 National Assessment of Educational Progress (NAEP; Allen, Jenkins, Kulick, & Zelenak, 1996), nine research item pools were constructed. For each item pool a reference test form with target TIF and TCC was created and five new forms were assembled.

## Item Pools

In correspondence with the nine conditions in the $3 \times 3$ design depicted in Table 1, we constructed nine research item pools. The first factor in the design was the composition of the item pool, with three levels: (1) only dichotomous items; (2) dichotomous items and polytomous items with three response categories; and (3) dichotomous items and polytomous items with three, four or five response categories. The second factor, also with three levels, related to parameters that were used in the item response models for the items: (1) only location parameters (i.e., 1PLM and PCM); (2) location and slope parameters (i.e., 2PLM and GPCM); and (3) location, slope and—only for the dichotomous items—guessing parameters (i.e., 3PLM and GPCM). Hence, the nine item pools differed with respect to their number and type of polytomous items, as well as with respect to the used item response models, going from less to more flexible IRT models. Throughout the article, the nine conditions will be referred to by their roman numbers (I to IX), as listed in Table 1.

Item parameters for the items in the nine pools were based on the calibrated items of the 1996 NAEP mathematics assessment (Allen et al., 1996). In the 1996 NAEP assessment a total of 521 items were calibrated for three grade levels (i.e., Grades 4, 8, and 12) using the 2PLM and the 3PLM for dichotomous items, and the GPCM for polytomous items. Each item belonged to one of the five mathematics content areas: algebra and functions (ALG); data analysis, statistics and probabilities (DAT); geometry and spatial sense (GEO); measurement (MEA); and number sense, properties and operations (NUM).

After inspection, we dropped nine items because they were outliers with respect to their maximum IIF value. A table that lists item parameter statistics of the remaining 512 items is found in the online Supporting Information. To increase the number of polytomous items, we added 40 extra items with three, four, and five response

Table 2
*Item Pool Composition by Item Type and Mathematics Content Area*

| Number of | Mathematics Content Area | | | | | |
|---|---|---|---|---|---|---|
| Response Categories *R* | ALG | DAT | GEO | MEA | NUM | Total |
| 2 | 73 | 57 | 83 | 61 | 158 | 432 |
| 3 | 14 | 7 (+2) | 9 (+1) | 10 | 17 | 60 |
| 4 | 3 (+3) | 6 (+2) | 2 (+6) | 3 (+3) | 3 (+9) | 40 |
| 5 | 1 (+3) | 3 (+1) | 1 (+3) | (+4) | 1 (+3) | 20 |
| Total | 97 | 78 | 105 | 81 | 191 | 552 |

*Note.* Each cell gives the number of items in the final item pool, according to content area and item type. The number of additional generated items in a cell is given between parentheses.

categories, resulting in a total of 552 items. The item parameters for these items were randomly sampled from uniform distributions with minimum and maximum values equal to the minimum and maximum values of the corresponding item parameters in the remaining original NAEP pool. Then, we randomly assigned the 40 generated items to one of the five math content areas. In Table 2, the 552 items are grouped according to item type and content area; the number of generated items in each cell is presented between parentheses.

To correspond with the nine conditions in Table 1, we constructed the nine different item pools as follows. In condition IX the total item pool was used and all the item parameters were considered, corresponding with 2PLM and 3PLM dichotomous items and GPCM polytomous items. The item pools in conditions VII and VIII also consisted of all the 552 items. However, to obtain an item pool without 3PLM items in condition VIII, the guessing parameters were ignored (i.e., set to 0), resulting in a pool with 2PLM dichotomous and GPCM polytomous items. Further, in condition VII only the location parameters of the items were considered (i.e., setting the slope parameters to 1 and the guessing parameters to 0), resulting in an item pool of 1PLM and PCM items. In conditions IV, V, and VI, the polytomous items with four and five response categories were dropped, resulting in item pools of 492 items. In condition VI all item parameters were used, in condition V the guessing parameters were ignored (i.e., only 2PLM and GPCM items), and in condition IV only the location parameters were considered (i.e., only 1PLM and PCM items). Finally, in conditions I, II, and III, only the 432 dichotomous items were retained. In condition III, all the item parameters were used (i.e., 2PLM and 3PLM items), in condition II the guessing parameters were ignored (i.e., only 2PLM items), and in condition I only the location parameters were used (i.e., only 1PLM items). Hence, a different item pool was used in each condition, and this should be taken into account in the comparison of results across conditions. Although the item pools were constructed artificially, during the ATAs each pool was treated as if its items were calibrated on the same ability scale,[1] using the correct IRT models. Moreover, in each condition the item pool was treated as if it was created especially for the corresponding (fictional) assessment, and therefore was representative of the goal of the assessment.

**Defining the Reference Test Form**

We based the reference test forms in each condition on the characteristics of the corresponding item pools. First we specified constraints with respect to item type and item content in each condition, then we defined the target TIFs and target TCCs.

**Constraints.** With respect to item type, let $V_r$ be the subset of all items in the pool that belong to item type $r$, where $r = 2$ for dichotomous items, $r = 3, 4$, or 5 for polytomous items with respectively three, four, or five response categories. Further, $n_r$ is the number of items from set $V_r$ that should be selected in each new test form. The item type constraints in each condition were translated to the Minimax models by

$$\sum_{j \in V_r} x_{jf} = n_r, \quad \text{for all } r = 2, \ldots, R, \quad \text{and for } f = 1, 2, \ldots, F, \tag{16}$$

where $R$ is the maximum number of response categories for items in the item pool. Depending on the condition, $R = 2$, 3, or 5. In conditions I, II, and III, $R = 2$ and the number of dichotomous items to be selected in the new test form was set to $n_2 = 28$. In conditions IV, V, and VI, $R = 3$ and the $n_r$ were set to $n_2 = 20$ and $n_3 = 3$. Finally, in conditions VII, VIII, and IX, $R = 5$ and the $n_r$ were set to $n_2 = 17$, $n_3 = 3$, $n_4 = 2$, and $n_5 = 1$.

With respect to the math content areas, let $V_{Co}$ be the subset of all items in the pool that belong content area $Co$, where $Co =$ ALG, DAT, GEO, MEA, or NUM. Let $n_{Co}$ be the number of items form from set $V_{Co}$ that should be selected in each new test form. The content constraints in each condition were translated to the Minimax models by

$$\sum_{j \in V_{Co}} x_{jf} = n_{Co}, \quad \text{for all } Co, \quad \text{and for } f = 1, 2, \ldots, F. \tag{17}$$

In conditions I, II, and III, the $n_{Co}$ were set to $n_{\text{ALG}} = 5$, $n_{\text{DAT}} = 4$, $n_{\text{GEO}} = 5$, $n_{\text{MEA}} = 4$, and $n_{\text{NUM}} = 10$. In conditions IV to IX, the $n_{Co}$ are set to $n_{\text{ALG}} = 4$, $n_{\text{DAT}} = 3$, $n_{\text{GEO}} = 4$, $n_{\text{MEA}} = 3$, and $n_{\text{NUM}} = 9$.

**Target TIF and TCC.** The TIF of the reference test (i.e., the target TIF), was based on the item pool TIF, taking the item type constraints into account. More specifically, in each condition, the target TIF ($I^{(t)}$) was constructed as

$$I^{(t)} = \sum_{r=2}^{R} \frac{n_r}{l_r} \times \sum_{i \in V_r} I_{ri}, \tag{18}$$

where $l_r$ is the number of items in $V_r$ (i.e., the number of items with $r$ response categories in the item pool) and $I_{ri}$ is the IIF of the $i$th item in $V_r$. The same rationale was used to create the target TCC ($T^{(t)}$):

$$T^{(t)} = \sum_{r=2}^{R} \frac{n_r}{l_r} \times \sum_{i \in V_r} P_{ri}, \tag{19}$$

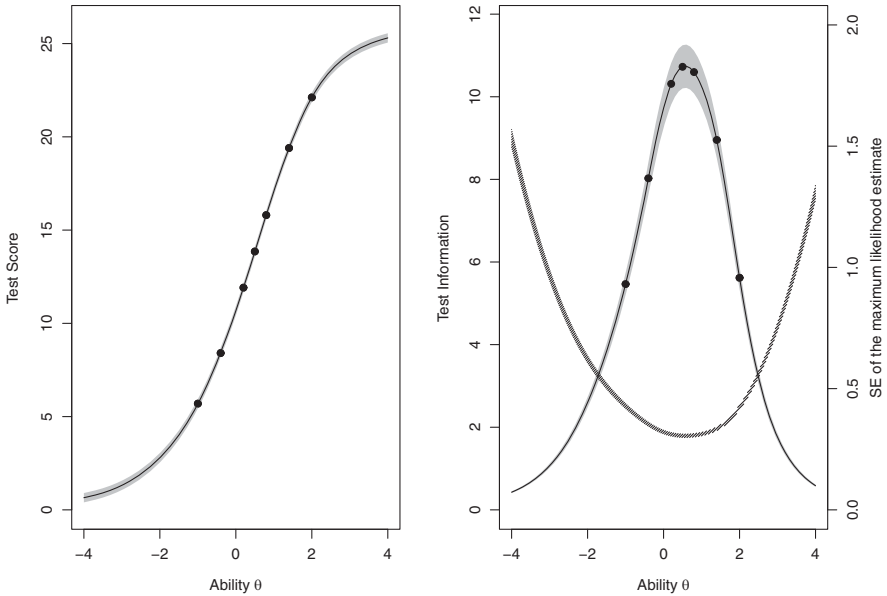


*Figure 2.* The target TCC and TIF in condition V. The black lines are the target TCC (left-hand panel) and the target TIF (right-hand panel). The black dots are the target values at the seven used θ-points. In the left-hand panel, the gray area represents the band around the target TCC in which the TCCs of the new test forms should lie, created by an absolute tolerance of $y^T = 0.25$. In the right-hand panel, the gray area represents the band around the target TIF in which the TIFs of the new forms should lie, created by a relative tolerance of $y^I = \frac{41}{841}$. Further, the shaded area in the right-hand panel is the corresponding band with respect to the measurement error.

with $P_{ri}$ representing the ICC of the $i$th item in $V_r$. In Equations 18 and 19 the $\frac{n_r}{l_r}$ can be seen as weights that bring the item pool TIF and TCC to the scale of the reference test form, taking into account the fact that the ratio of item types in the item pool is different from the ratio of item types in the reference test form.

For example in condition V, $R = 3$, $n_2 = 20$, $n_3 = 3$, $l_2 = 432$, and $l_3 = 60$. The $I_{2i}$ and $P_{2i}$ for the 432 dichotomous items were computed according to the 2PLM, and the $I_{3i}$ and $P_{3i}$ for the 60 polytomous items with three response categories were computed according to the GPCM. The resulting target TCC and TIF are displayed in Figure 2 as black solid lines, in the left-hand and right-hand panel, respectively.

### Five Minimax Models

Five Minimax models were applied in each condition. The first Minimax model minimized the maximum absolute distance between the reference and the new test forms with respect to the TIF at prespecified θ-points (cf. Equations 11 and 9). In the second Minimax model the maximum distance with respect to the TCC was minimized (cf. Equations 8 and 9). These two Minimax models only used a single statistical target. In the remainder of this article they will be referred to as *Only TIF* and *Only TCC*, respectively.

The third Minimax model minimized the maximum absolute distance with respect to both the TIF and the TCC simultaneously, as proposed by Ali and van Rijn (2016) (cf. Equations 13 and 9). In this study, the weights that specify the relative importance of the TIF and the TCC in Equation 13 were set to $w_I = w_T = 1$. The approach of Ali and van Rijn (2016) will be referred to as *TIF & TCC*. In the fourth Minimax model, the maximum relative distance with respect to the TIF was minimized, while setting an absolute tolerance for the distance with respect to the TCC (cf. Equations 12, 14, and 9). The absolute tolerance with respect to the TCC was set to $y_h^T = 0.25$ for all θ-points $h$. Using this absolute tolerance can be interpreted as creating a band of width $2y_h^T = 0.5$ (around the target TCC) in which the TCCs of the new forms should lie. The gray area in the left-hand panel of Figure 2 represents this band for condition V. Finally, the fifth Minimax model minimized the maximum absolute distance with respect to the TCC, while setting a relative tolerance for the distance with respect to the TIF (cf. Equations 8, 15, and 9). The relative tolerance was set to $y_h^I = \frac{41}{841}$ for all θ-points, creating a band (around the target TIF) in which the TIFs of the new test forms should lie. The gray area in the right-hand panel of Figure 2 represents the band around the target TIF in which the TIF of the new test forms should lie in condition V. In the same panel, the corresponding band with respect to the measurement error is represented by the shaded area. Because the last two Minimax models minimized the maximum distances with respect to one statistical target, while using an upper and a lower bound for the other statistical target, they will be referred to as *TIF & bounded TCC* and *TCC & bounded TIF*, respectively.

### Specifying the θ-Points

According to van der Linden and Adema (1998), three to five θ-points is sufficient when the target TIF (or TCC) is continuous and well-behaved. However, because it is expected that the TIF and TCC will not behave as well for mixed-format test forms, three θ-points may be insufficient. Therefore, for each of the five Minimax models, a version with $H = 3$ and another version with $H = 6$ θ-points was used to assemble the test forms. θ-points were chosen to cover the ability-range for which the item pools were most informative. The used θ-points were $-0.4$, 0.5, and 1.4 when $H = 3$ and $-1.0$, $-0.4$, 0.2, 0.8, 1.4, and 2 when $H = 6$. The target TIF and TCC values corresponding to these θ-points in condition V are displayed in Figure 2 as black dots.

Given the design with nine conditions, five Minimax models with two versions each, in total $3 \times 3 \times 5 \times 2 = 90$ Minimax models were formulated.

### Evaluation Criteria

To evaluate the parallelism of the constructed test forms in each condition, besides graphical presentations of the TIFs or TCCs, the root mean square deviation (RMSD) statistic was computed with respect to the target TCC ($\text{RMSD}_T$) for each assembled

test form. For assembled test form $f$,

$$\text{RMSD}_T = \sqrt{\frac{\sum_{k=1}^{K}\left[T^{(f)}(\theta_k) - T^{(t)}(\theta_k)\right]^2}{K}}, \tag{20}$$

where the superscript $(f)$ refers to the observed value for the assembled test form $f$, with $f = 1, 2, 3, 4$ or 5 in this study. Further, $K$ is the number of $\theta$-points over which the $\text{RMSD}_T$ is computed. The $\text{RMSD}_T$ can be interpreted as the average difference between the expected test scores of the assembled test form and the reference test form. Although a similar RMSD can be computed with respect to the TIF, its interpretation is less straightforward. Therefore, as an alternative criterion, we propose to compute the mean relative deviation (MRD) with respect to the TIF:

$$\text{MRD}_I = \frac{\sum_{k=1}^{K}\left|\frac{I^{(f)}(\theta_k) - I^{(t)}(\theta_k)}{I^{(t)}(\theta_k)}\right|}{K}. \tag{21}$$

The $\text{MRD}_I$ can be interpreted as the average proportion of the target TIF that a new test form differs from the target TIF.

In the present study, $K = 61$ ability points ranging from $-3$ to $+3$ in increments of 0.1 units were considered to compute both the $\text{RMSD}_T$ and the $\text{MRD}_I$. Further, the $\text{RMSD}_T$ and the $\text{MRD}_I$ were averaged across the five test forms resulting from the Minimax model (denoted by $\overline{\text{RMSD}_T}$ and $\overline{\text{MRD}_I}$, respectively) as more global measures of performance.

Besides the average performance, test developers are also concerned with the local performance with respect to parallelism. Therefore, as more locally oriented evaluation criteria, across the assembled test forms for each Minimax model we introduce the maximum score gap (MSG) and the maximum relative information gap (MRIG):

$$\text{MSG} = \max_{\theta}\left\{\max_{f}\{T^{(f)}(\theta)\} - \min_{f}\{T^{(f)}(\theta)\}\right\}, \tag{22}$$

and

$$\text{MRIG} = \max_{\theta}\left\{\frac{\max_f\{I^{(f)}(\theta)\} - \min_f\{I^{(f)}(\theta)\}}{I^{(t)}(\theta)}\right\}. \tag{23}$$

Both criteria are relevant for practice and directly interpretable. The maximum score gap is the maximal absolute difference between the conditional expected test scores of the new forms, within a specified ability range. The maximum relative information gap is the maximal relative difference in conditional precision across the test forms, expressed as a proportion of the target TIF.

We computed the maximum score gap and the maximum relative information gap for the ability values ranging from $-1.0$ to 2.0 using increments of 0.1 units. This range corresponds to the range covered by the $\theta$-points. To evaluate the performance of the assembly, the difference that matters criterion can be applied to the

different criteria. A $\overline{\text{MRD}_I}$ and a maximum relative information gap below $10\%^2$, and a $\overline{\text{RMSD}_T}$ and a maximum score gap below .5 score points were considered satisfactory.

### Expectations

We expect that the Minimax models with only the TIF or only the TCC as statistical target will have a good overall performance for an item pool with only 1PLM items (i.e., condition I). In item pools with 2PLM and 3PLM items (i.e., condition II and III) we expect satisfactory performance with respect to the TIF but not with respect to the TCC for the *Only TIF* Minimax model, while the opposite is expected for the *Only TCC* Minimax model (Ali and van Rijn, 2016; Chen et al., 2012). Further, a similar pattern is expected for the mixed-format item pools, regardless the used item response models (i.e., in condition IV through IX). Finally, we expect the combined Minimax models to mitigate these problems, and deliver test forms that are parallel both with respect to the TIF and TCC.

### Technical Details and Comparability

Solutions for each Minimax model were computed using GUROBI (Gurobi Optimization, 2016) and lp_solve (Berkelaar et al., 2004) through R (R Development Core Team, 2015). All computations were done on a Dell Optiplex 7010 desktop with an i7-3770 3.40GHz processor. Both programs solve MILP models using a branch-and-bound algorithm (Land & Doig, 1960). However, lp_solve was unable to find solutions for multiple Minimax models within reasonable computing times. More specifically, when $H = 6$, or when the *TIF & TCC* Minimax model was used in conditions with polytomous items included in the item pools, no solutions were found. Therefore, only the analyses and results using GUROBI are discussed.

In practice, the time required to find the optimal solution of a MILP model can become extremely long. One strategy to reduce solving times is to specify a tolerance for the objective value (i.e., $z$ in Equation 9). The first feasible solution for which $z$ meets the tolerance is then considered a good (enough) solution, despite not being optimal. A second strategy is to set a time limit. Because our goal was to compare the performance of the Minimax models, and because the interpretation of the objective values differs across the five Minimax models, we chose to specify a time limit of 30 minutes for all Minimax models. In all 90 assembly problems, feasible solutions were found within one minute, indicating that 30 minutes was a reasonable time limit.

Comparing the performance of the five Minimax models is not straightforward. Ideally, without any restriction in computing time, one would have look at all the feasible solutions within each condition, and check the relation between the objective value and the parallelism of the assembled test forms, for each Minimax model. An efficient Minimax model would result in an increasing parallelism when the objective value decreases. Such a strategy, however, is unrealistic. The rationale behind the current comparison is as follows: if a Minimax model efficiently improves the parallelism of the test forms by minimizing the objective value, a good solution

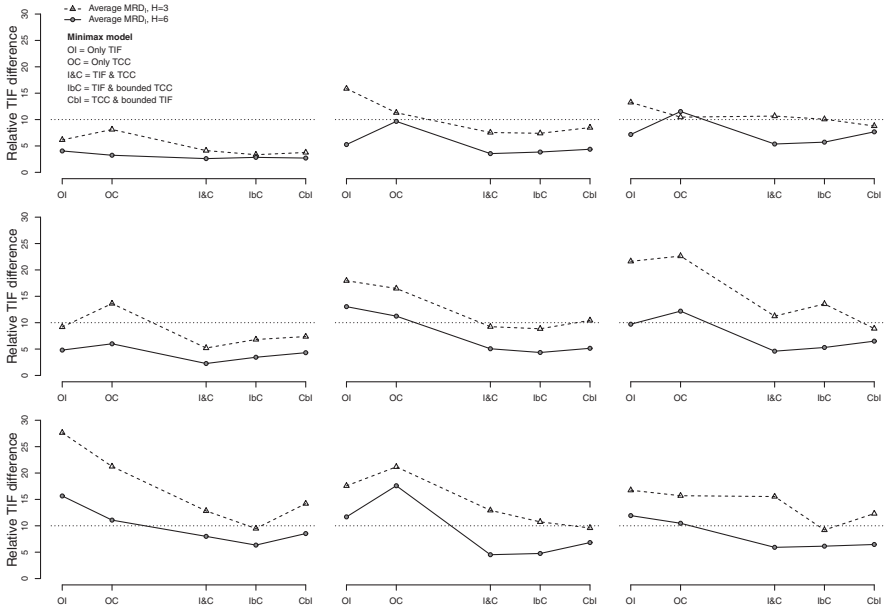


*Figure 3.* The average mean relative difference with respect to the TIF ($\overline{\text{MRD}_I}$) across the five Minimax models with three and six θ-points in each condition. The *y*-axis is the percentage relative TIF difference. The horizontal dotted line marks a relative difference in TIF of 10%, which corresponds to about a 5% difference in measurement error.

should be found after 30 minutes. Likewise, a bad solution after 30 minutes can be seen as an indication of inefficiency.

## Results

In each of the 90 assembly problems, five new test forms were assembled, with all content constraints satisfied. A graphical presentation of the TIFs and TCCs of the assembled test forms, across all Minimax models, in each condition, can be found online.[3] Two tables with the $\overline{\text{MRD}_I}$ and the $\overline{\text{RMSD}_T}$ for the five Minimax models in the nine conditions for, respectively, the version with $H = 3$ and $H = 6$ are found in the online Supporting Information. The panels in Figure 3 present the $\overline{\text{MRD}_I}$ across the five Minimax models with three and six θ-points in each of the nine conditions. Similar figures are made for the maximum relative information gap (Figure 4), the $\overline{\text{RMSD}_T}$ (Figure 5), and the maximum score gap (Figure 6). In each of the Figures 3 to 6, the nine panels correspond to the nine conditions, with condition I in the top left corner, condition III top right corner, and condition IX in the bottom right corner. In each panel, the five Minimax models are located on the *x*-axis.

### Parallelism With Respect to the TIF

Figures 3 and 4 show that in condition I all Minimax models performed equally well, and three θ-points seems to be enough to attain parallelism with respect to the

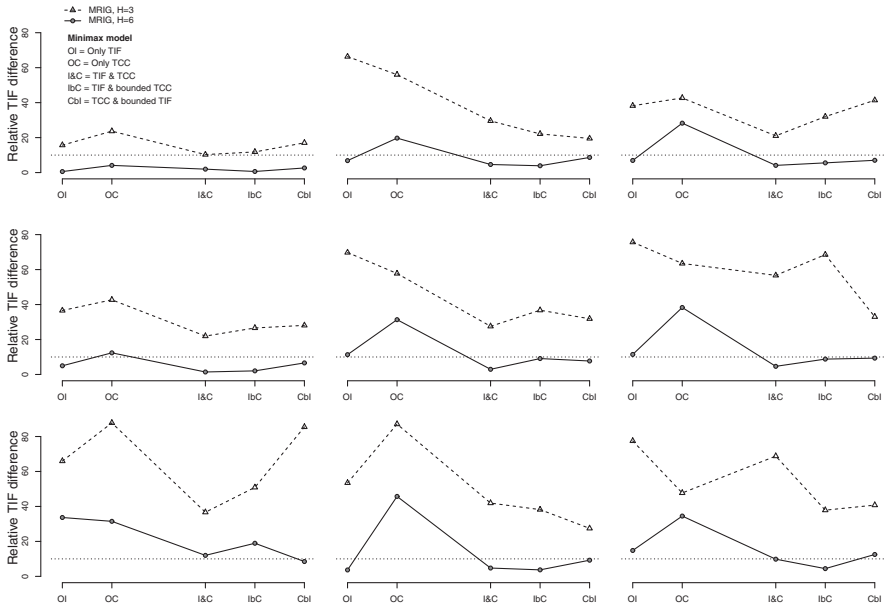


*Figure 4.* The maximum relative information gap (MRIG) across the five Minimax models with three and six θ-points in each condition. The *y*-axis is the percentage relative TIF difference. The horizontal dotted line marks a relative difference in TIF of 10%, which corresponds to about a 5% difference in measurement error.

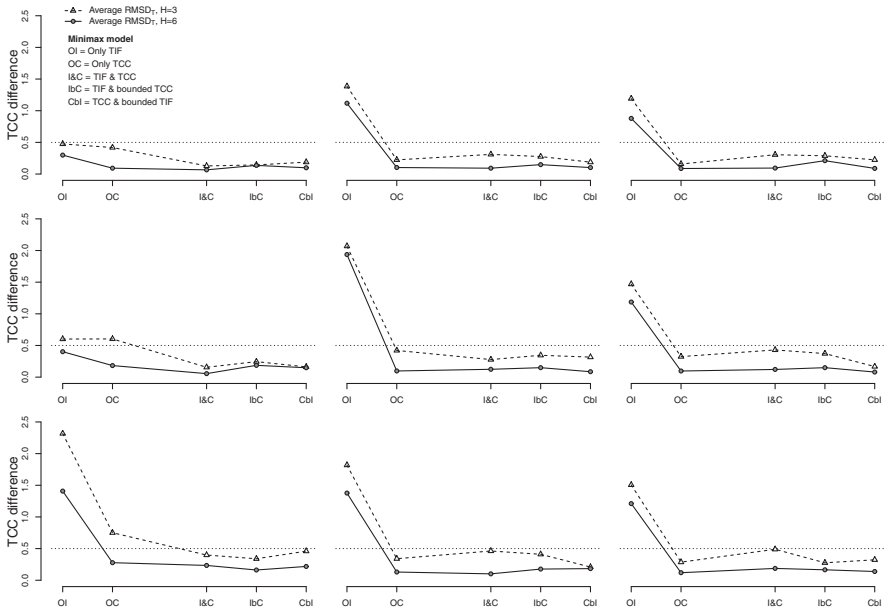


*Figure 5.* The $\overline{\text{RMSD}}_T$ across the five Minimax models with three and six θ-points in each condition. The *y*-axis is the TCC difference. The horizontal dotted line marks the difference of 0.5 score points.

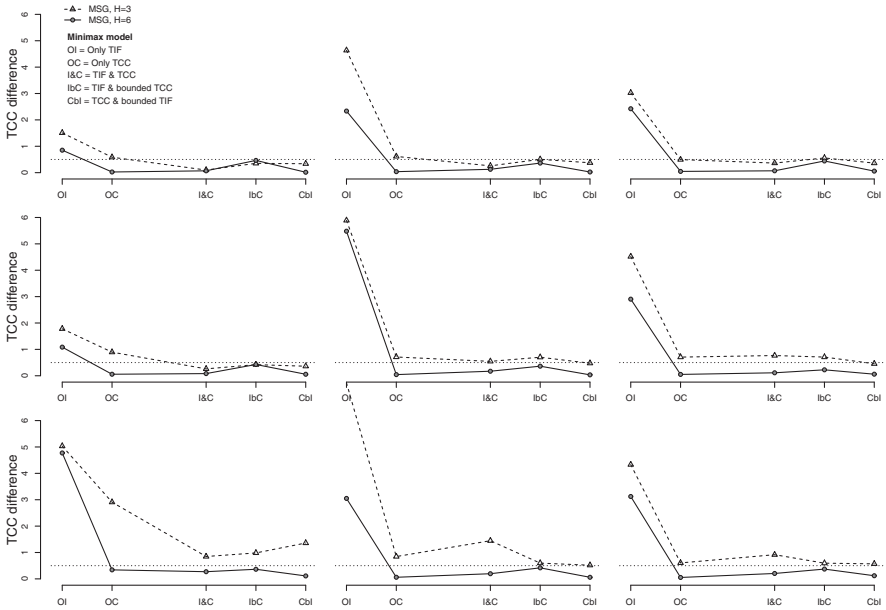


*Figure 6.* The maximum score gap (MSG) across the five Minimax models with three and six θ-points in each condition. The *y*-axis is the TCC difference. The horizontal dotted line marks the difference of 0.5 score points.

TIF, in item pools with only 1PLM items. As expected, the performance of the *Only TCC* Minimax model was worse in condition II and especially in condition III. The performance of the combined models was at the same level of the *Only TIF* Minimax model. Note that in condition II, and even more so in condition III, six rather than three θ-points were needed to keep the maximum information gap below 10%.

In conditions IV to XI, the performance of the *Only TCC* Minimax model was bad, especially with respect to the maximum relative information gap (see Figure 4). Interestingly, the *Only TIF* Minimax model also displayed bad performance in some of these conditions, even when $H = 6$. The graphical presentations reveal that, although the TIFs of the new test forms (and the target TIF) were almost equal at the θ-points, large information gaps arose in between the θ-points. As an example, the left-hand panel of Figure 7 gives the TIFs in condition VII using the *Only TIF* Minimax model with $H = 3$. Considerable information gaps are observed. The performance of the combined Minimax models was clearly better. However, in some conditions, six θ-points were needed to keep the maximum relative information gap below or around 10%.

### Parallelism With Respect to the TCC

Figures 5 and 6 show that, as expected, in condition I the performance of the *Only TIF* Minimax model was good, although the maximum score gap was higher than 0.5 even when $H = 6$. In all other conditions the performance of the *Only TIF* Minimax model was far from satisfactory, with condition IV as the only exception. But even

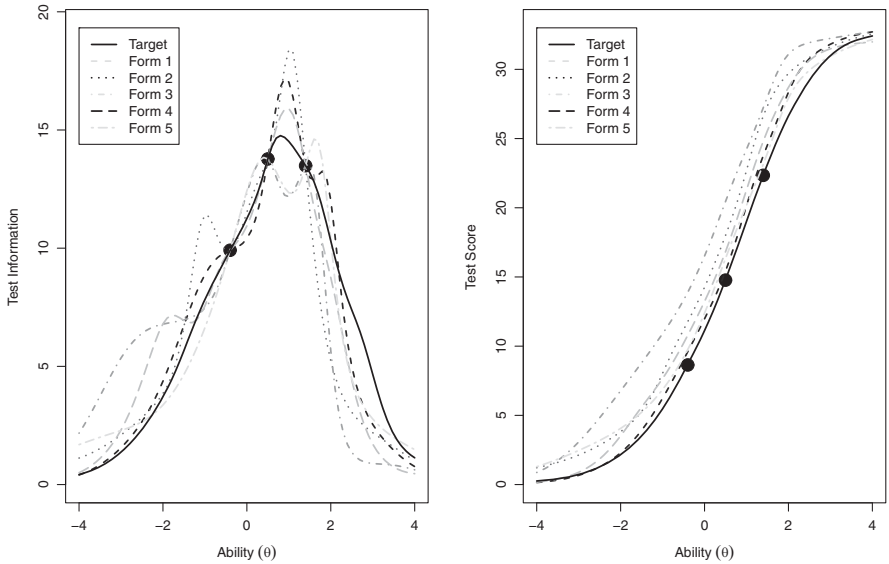


*Figure 7.* The TIF (left-hand panel) and TCC (right-hand panel) of the new test forms assembled using the *Only TIF* Minimax model with $H = 3$ θ-points, in condition VII. The black curves are the target TIF and target TCC, and the black dots are the target TIF and target TCC values corresponding to the three θ-points, in the left-hand and right-hand panel, respectively.

in that condition, the performance of the *Only TIF* Minimax model was worse than that of the other Minimax models. The combined Minimax models performed just as well as the *Only TCC* Minimax model in all conditions. For the four Minimax models that use the TCC in the statistical target, the parallelism with respect to the TCC was satisfactory in all conditions when $H = 6$.

### Overall Performance

Taken together, the results show that the Minimax models with a single target could only compete with the combined Minimax models in condition I. When the item pools were more diverse (i.e., also contained polytomous items or used more flexible item response models) the *Only TCC* Minimax model produced considerable information gaps while the *Only TIF* Minimax model produced large score gaps. Moreover, in the situations with more polytomous items (with more response categories), information gaps were present even when the *Only TIF* Minimax model was applied. In contrast, combining the TIF and TCC in a Minimax model resulted in test forms that were parallel with respect to both TIF and TCC in all conditions. Finally, when using the combined approaches in conditions with only 1PLM or 2PLM items, three θ-points was enough. However, in the other conditions, six θ-points were necessary to produce satisfactory results.

## Discussion

In linear test assembly, the goal is usually to construct test forms that satisfy a series of content specifications while simultaneously trying to achieve statistical parallelism with respect to a reference test form. Often this parallelism is pursued by minimizing the maximum distance with respect to a specific statistical target of the reference test form, using what is called a Minimax MILP model (van der Linden, 2005). Commonly only the TIF (Swanson & Stocking, 1993; van der Linden & Adema, 1998; Veldkamp, Matteucci, & de Jong, 2013), or on some occasions only the TCC (Armstrong, Jones, & Kunce, 1998; Armstrong et al., 2005) are used as statistical targets. However, both statistical targets can be combined in one Minimax model (Ali & van Rijn, 2016). In the present study, we established three methods within the combined approach: (a) minimizing the distance with respect to both the TIF and TCC simultaneously; (b) minimizing the distance with respect to the TIF, while constraining the distance to the TCC; and (c) vice versa. Further, we proposed to use the relative distance for constraints related to the TIF, which results in meaningful tolerances and interpretable objective function values.

Using the newly proposed criteria, the performance of the two single and three combined Minimax models was evaluated in mixed-format testing situations using realistic item pools. The results showed that Minimax models that use a single statistical target can result in considerable score and information gaps. The combined approach, however, effectively reduced the score and information gaps and performed better in achieving parallelism across test forms, especially for mixed-format tests. The results were similar across the three combined approaches.

It is common for Minimax models to include just a few well chosen ability values (i.e., $\theta$-points) at which the maximal distance to the target TIF is minimized. For test forms with only dichotomous items, TIFs usually have a well behaved continuous form, and practical experience shows that three to five $\theta$-points are sufficient (van der Linden & Adema, 1998). This observation was confirmed by this study. In mixed-format tests, however, the TIF can have a less well behaved shape, with multiple peaks. Therefore, information gaps between the evaluation points can arise even when the local minimization at the $\theta$-points is optimal (cf. condition VII), indicating that for these cases the *Only TIF* Minimax model is not efficient. Although adding more $\theta$-points seems a logical solution, there are, in our experience, some important caveats. First, adding more $\theta$-points does not automatically imply better parallelism throughout the whole ability range. Second, in almost all conditions with polytomous items in the item pool, the *TIF & bounded TCC* Minimax model with $H = 3$ outperformed the *Only TIF* Minimax model with $H = 6$ (see Figures 3–6). This result indicates that, when the maximum distance to the target TIF is minimized, it is better to add constraints with respect to the TCC to increase the efficiency of the Minimax model than to add an equal number of constraints with respect to the TIF (i.e., by adding extra $\theta$-points). This holds for obtaining better parallelism with respect to both the TCC and the TIF.

Within the combined approach, the proposed methods that specify tolerances with respect to one statistical target (methods b and c) have an advantage compared to Ali

and van Rijn (2016)'s method a. Because the TIF and TCC have a different metric, specifying the simultaneous minimization with respect to the TIF and TCC in the method of Ali and van Rijn (2016) is not straightforward. In contrast, absolute tolerance for the TCC and the relative tolerance for the TIF in the two newly proposed methods b and c can be set independently, and in a meaningful and interpretable way. A disadvantage of using the tolerances is that infeasibility problems may arise when the tolerances are too strict for the item pool. Different methods have been proposed to detect the causes and deal with infeasibility in ATA (Huitzing, Veldkamp, & Verschoor, 2005). However, when reasonable values are chosen for the tolerances, and the quality and size of the item pool is sufficient, problems of infeasibility should be the exception.

**Implications for Practice**

Due to the nature of assembly problems, generalizing the current findings to other test situations and item pools is not straightforward. The combinatorial problems encountered in the assembly of parallel test forms are nonlinear, and highly dependent on the item pool and the applied constraints. Nevertheless, given the results in our study, we recommend using one of the combined approaches, and six well chosen $\theta$-points rather than three, especially when dealing with mixed-format tests. Further, the choice between one of the combined Minimax models and decisions about the size of the tolerance, the number and location of the $\theta$-points, etc., are not straightforward. We propose that test-designers should thoughtfully decide on these choices, taking into account the purpose of the test as well as the assembly conditions such as item pool size, number of items in the test forms, and so on. For instance, in some situations it might be reasonable to have different $\theta$-points specific to each of the TIF and the TCC.

The combined approach has already been successfully applied to a reasoning test with 2PLM items (Weeks, 2015). High-stakes testing programs that require multiple parallel, mixed-format test forms could benefit from the combined approach. In addition, the methodology could also be applied to low-stakes assessments such as the Program for International Student Assessment (PISA), where for each domain items are divided in different blocks that should be parallel with respect to testing time, difficulty, and precision.

**Limitations and Extensions**

The assembly problems in this study proved to be complicated. Although the commercial solver GUROBI was able to quickly find good solutions for the Minimax models, the free solver lp_solve was unable to find solutions within a reasonable time frame for several models. However, other free solvers, such as GLPK, may also be able to deal with these complex problems.

This study focused on situations where number-correct sum scores are used, and is limited to unidimensional IRT models. In addition to number-correct scoring, formula scoring, or pattern scoring methods such as maximum likelihood estimation (MLE) and expected a priori (EAP) are also widely used. As the choice of a certain IRT ability estimator can have significant practical implications (Kolen & Tong,

2010), future research will have to show whether our results can be generalized to other scoring rules and ability estimators. Further, because not all item pools consist of unidimensional items, it would be interesting to investigate whether using a combined target is feasible for multidimensional items.

Finally, this study is limited to Minimax models. There are, however, other models available within the MILP approach for ATA. For example a MILP model can be formulated where the TIFs of the new test forms are maximized at prespecified $\theta$-points while simultaneously the maximum distance between the TIFs of the new test forms is minimized. An advantage of this approach is that it does not require specifying a target TIF (or target TCC). A combined approach for these models, where constraints are specified both with respect to the TIF and to the TCC would be an interesting extension of the current study.

## Acknowledgments

This work has been partially funded by the Educational Testing Service.

## Notes

[1]In NAEP, the item parameters were obtained through three separate calibrations, one for each grade level.

[2]A relative difference in TIF of 10% corresponds to a difference in measurement error of about 5%.

[3]https://statistiek-4.shinyapps.io/ATA_mixed-format-test-forms/

## References

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*, 163–179.

Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1996). *Technical report of the NAEP 1996 state assessment program in mathematics*. Washington, DC: National Center for Education Statistics.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Armstrong, R. D., Belov, D., & 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.

Berkelaar, M., Eikland, K., & Notebaert, P. (2004–2016). lp_solve: Open source (Mixed-Integer) Linear Programming system. Available at http://gts.sourceforge.net/

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). 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.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*, 398–409.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*, 295–311.

Donoghue, J. R. (2015). *Comparison of integer programming (IP) solvers for automated test assembly (ATA)* (Research Report 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 *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10, pp. 91–122). Princeton, NJ: Educational Testing Service.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*, 137–154.

FICO. (2016). Xpress-Optimizer Reference manual. Retrieved from http://www.fico.com/en/products/fico-xpress-optimization-suite

Finkelman, M., Kim, W., & Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement*, *46*, 273–292.

GNU linear programming kit. (2016). Retrieved from https://www.gnu.org/software/glpk/

Gurobi Optimization. (2016). Gurobi optimizer reference manual. Retrieved from http://www.gurobi.com

Huitzing, H. A., Veldkamp, B. P., & Verschoor, A. J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement*, *42*, 223–243.

IBM. (2016). IBM ILOG CPLEX Optimization Studio [Computer software]. Retrieved from http://www-01.ibm.com/software/commerce/optimization/cplex-cp-optimizer/

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14.

Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, 497–520.

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

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Pshychological Measurement*, *22* , 224–236.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

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

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*, 351–363.

R Development Core Team. (2015). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria: Author. Retrieved from. http://www.R-project.org

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Available at https://eric.ed.gov/?id=ED419814

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*.

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. (2005). *Linear models for optimal test design*. New York, NY: Springer.
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. (1998). Observed-score equation as a test assembly problem. *Psychometrika*, *63*, 401–418.
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.
Weeks, J. (2015, April). *Optimizing the assembly of parallel test forms via linear programming*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

## Authors

DRIES DEBEER is Postdoctoral Assistant at the University of Zurich, Department of Psychology, Research Unit Psychological Methods, Evaluation and Statistics, Binzmuehlestrasse 14 (PB 27), 8050 Zurich, Switzerland; dries.debeer@uzh.ch. His primary research interests include psychometrics, item response modeling, and statistical methods in psychology.
USAMA S. ALI is Psychometrician at Educational Testing Service, 666 Rosedale Rd, Princeton, NJ 08541, and Assistant Professor, Department of Educational Psychology, South Valley University, Qena, Egypt; uali@ets.org. His primary research interests include educational measurement and psychometrics.
PETER W. VAN RIJN is Senior Research Scientist at ETS Global, Strawinskylaan 929, 1077XX Amsterdam, the Netherlands; pvanrijn@etsglobal.org. His research focuses on educational measurement and psychometrics.

## Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.

**Table A1**: Statistics of Item Parameters of the 512 NAEP Items and 40 Generated Items by Item Type.
**Table A2**: Mean, Minimum and Maximum Values for the $\overline{\text{MRD}_I}$ and $\overline{\text{RMSD}_T}$ across the Five Minimax Models with $H = 6$, in Conditions I to IX.
**Table A3**: Mean, Minimum and Maximum Values for the $\overline{\text{MRD}_I}$ and $\overline{\text{RMSD}_T}$ across the Five Minimax Models with $H = 3$, in Conditions I to IX.