Applied Psychological Measurement
2020, Vol. 44(1) 17–32
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621619827586
journals.sagepub.com/home/apm
SAGE

Multidimensional Test Assembly Using Mixed-Integer Linear Programming: An Application of Kullback–Leibler Information

Dries Debeer[1], Peter W. van Rijn[2] and Usama S. Ali[3,4]

Abstract

Many educational testing programs require different test forms with minimal or no item overlap. At the same time, the test forms should be parallel in terms of their statistical and content-related properties. A well-established method to assemble parallel test forms is to apply combinatorial optimization using mixed-integer linear programming (MILP). Using this approach, in the unidimensional case, Fisher information (FI) is commonly used as the statistical target to obtain parallelism. In the multidimensional case, however, FI is a multidimensional matrix, which complicates its use as a statistical target. Previous research addressing this problem focused on item selection criteria for multidimensional computerized adaptive testing (MCAT). Yet these selection criteria are not directly transferable to the assembly of linear parallel test forms. To bridge this gap the authors derive different statistical targets, based on either FI or the Kullback–Leibler (KL) divergence, that can be applied in MILP models to assemble multidimensional parallel test forms. Using simulated item pools and an item pool based on empirical items, the proposed statistical targets are compared and evaluated. Promising results with respect to the KL-based statistical targets are presented and discussed.

Keywords

automated test assembly, MIRT, linear programming, Kullback–Leibler information

In educational measurement there are many cases where different sets of items (i.e., test forms or item pools) are required to be parallel. That is, test forms[1] are required to be equivalent in terms of both their statistical and content-related properties. Without being exhaustive, the need for parallel test forms can arise in the following situations. High-stakes educational testing programs may require multiple parallel test forms with minimal or no item overlap for fair, secure,

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

Corresponding Author:
Dries Debeer, Department of Psychology, Psychological Methods, Evaluation and Statistics, University of Zurich,
Binzmuehlestrasse 14, Box 27, 8050 Zurich, Switzerland.
Email: dries.debeer@uzh.ch

and valid test scores. Computerized adaptive testing (CAT) administrations may need multiple parallel item pools, which can be applied at different time points. Multistage testing may require the subtests at each stage to be parallel with an ideal, preconceived subtest to meet measurement precision and content-related requirements. Large-scale low-stakes assessments that apply a balanced incomplete block design often require item blocks that (a) are nonoverlapping, (b) have a prespecified expected overall response time, and (c) meet a specified precision across a predefined ability range. Yet, regardless the situation, constructing these parallel test forms is not straightforward.

Assembling multiple parallel test forms out of a calibrated item pool has proven to be a complicated combinatorial optimization problem. Rather than by hand, test forms are comonly assembled automatically. Different automated test assembly (ATA) methods are available; some are based on sampling-and-classification methods such as the Cell Only and the Cell and Cube methods (Chen, Chang, & Wu, 2012), and others rely on constrained combinatorial optimization techniques (e.g., Finkelman, Kim, & 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 & van der Linden, 2011; Theunissen, 1985; van der Linden, 2005). This article focuses on the MILP approach for ATA problems, and specifically on the Minimax approach for the assembly of parallel test forms (Boekkooi-Timminga, 1990; van der Linden, 2005) within item response theory (IRT) framework.

The literature on parallel test assembly is currently limited to unidimensional IRT (UIRT). Yet multidimensional test administrations may also require parallel test forms. Therefore, the main aim of this article is to extend the Minimax approach to multidimensional IRT (MIRT; Reckase, 2009). To bridge this gap, the authors propose new statistical targets and evaluate their performance in both unidimensional and multidimensional cases. Although the article focuses on assembling multiple test forms with parallel expected score distributions and parallel measurement error, the proposed methods can also consider different targets (e.g., testing time and precision) or even assemble one optimal test form.

The remainder of this article is organized as follows. First, the authors introduce the item response and Fisher information (FI) functions in MIRT. Then, they discuss the MILP model approach for ATA, with a focus on Minimax models (van der Linden, 2005). Subsequently, they explain the challenge that arises with ATA of parallel test forms in the multidimensional case and propose six approaches that mitigate this issue. They first translate the two approaches proposed by Veldkamp (2002) to the case of parallel test assembly. Then, they introduce the Kullback–Leibler (KL) divergence and subsequently propose four KL-based approaches. These sections can be found in Online Appendices A and B. In the "Method" and "Results" sections, the six approaches are evaluated using simulated one- and two-dimensional item pools, as well as a three-dimensional bifactor model case that is based on data from an operational language aptitude test. The article ends with a discussion of the results.

MIRT

Due to its diagnostic feature and potential for exploiting correlated subskills, MIRT has gained more attention recently. Many certification and admission boards want to enhance their high stakes tests with multidimensional scale scores that can be used as a diagnostic service to inform candidates (Mulder & van der Linden, 2009b), but also for decision making. MIRT reflects testing for diagnosis by viewing ability as a p-dimensional vector, while still allowing an overall score (e.g., Haberman & Sinharay, 2010).

According to the multidimensional extension of the two-parameter logistic model (2PLM; Birnbaum, 1968), the probability of a correct response on item i given ability $\boldsymbol{\theta}$ is

$$P_i(\boldsymbol{\theta}) = \Pr(X_i = 1|\boldsymbol{\theta}) = \frac{1}{1 + \exp[-(\mathbf{a}_i^T\boldsymbol{\theta} - b_i)]}, \quad (1)$$

where X_i is the response on item i, and the ability $\boldsymbol{\theta}$ is a p-dimensional latent variable. $\mathbf{a}_i^T$ is a vector of slope or discrimination parameters corresponding to the p ability values, and b_i is the intercept or location for item i. For simplicity, in the remainder of this article all items are assumed to follow the item response function in Equation 1. Extensions to other MIRT models are straightforward.

The test characteristic curve (TCC) for a test with n items describes the expected test score for a given ability $\boldsymbol{\theta}$. Under the local independence assumption, the TCC, denoted $T(\boldsymbol{\theta})$, is the sum of the $P_i(\boldsymbol{\theta})$ of all n items in a test:

$$T(\boldsymbol{\theta}) = \sum_{i=1}^{n} P_i(\boldsymbol{\theta}). \quad (2)$$

In addition, under local independence, the likelihood of a response pattern $\boldsymbol{X}$ on a test simplifies to

$$f(X; \boldsymbol{\theta}) = L(\boldsymbol{\theta}, X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} [P_i(\boldsymbol{\theta})]^{X_i}[1 - P_i(\boldsymbol{\theta})]^{1-X_i}. \quad (3)$$

Fisher Information

In maximum likelihood theory, the variance of the score function is referred to as FI. It is a measure of the information that a random variable carries about an unknown parameter. In MIRT, with X as the (vector-valued) random variable and assuming item parameters are known, the FI that concerns the latent ability parameter vector $\boldsymbol{\theta}$ is defined by

$$I(\boldsymbol{\theta}) = E\left(\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X; \boldsymbol{\theta})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X; \boldsymbol{\theta})\right]^{\mathrm{T}} |\boldsymbol{\theta}\right). \quad (4)$$

The inverse of this $p \times p$ FI matrix is the asymptotic variance–covariance matrix of the maximum likelihood estimate (MLE) of the ability parameter $\hat{\boldsymbol{\theta}}$: $V(\hat{\boldsymbol{\theta}}) = I(\hat{\boldsymbol{\theta}})^{-1}$. The FI matrix for a single item i, denoted as $I_i(\hat{\boldsymbol{\theta}})$, is usually referred to as the item information function (IIF):

$$I_i(\boldsymbol{\theta}) = E\left(\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i; \boldsymbol{\theta})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i; \boldsymbol{\theta})\right]^{\mathrm{T}} |\boldsymbol{\theta}\right), \quad (5)$$

where $f(X_i; \boldsymbol{\theta}) = [P_i(\boldsymbol{\theta})]^{X_i}[1 - P_i(\boldsymbol{\theta})]^{1-X_i}$. The IIF is typically interpreted as the information that an item i carries about the ability $\boldsymbol{\theta}$. For a given n-item test, the FI is commonly referred to as the test information function (TIF). Under local independence, the TIF and thus the asymptotic variance–covariance matrix of the MLE simplify to

$$I(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_i(\boldsymbol{\theta}) \quad \text{and} \quad V\left(\hat{\boldsymbol{\theta}}\right) = \left[\sum_{i=1}^{n} I_i\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}. \quad (6)$$

Test Assembly via MILP

Translating the ATA problem to a MILP model is a common technique (Diao & van der Linden, 2011; Theunissen, 1985; van der Linden, 2005). Because MILP and linear programming models are widely used in various industries, efficient algorithms have been developed and various software packages (usually referred to as ''solvers'') are available. For an overview of both commercial and open source solvers that can be used for ATA problems, see Donoghue (2015).

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 into a specific test form. In general, a MILP model includes two main parts (Ali & van Rijn, 2016; Diao & van der Linden, 2011): (a) the objective function,

$$\min \mathbf{c}^T\mathbf{x}, \tag{7}$$

and (b) a set of inequality constraints,

$$\mathbf{A}\mathbf{x} \leq \mathbf{d}, \tag{8}$$

where $\mathbf{x} = (x_1, \ldots, x_E)^T$ is a vector of length E that needs to be solved for, and $\mathbf{c} = (c_1, \ldots, c_E)^T$ is a numeric vector of known coefficients for the objective function. $\mathbf{A}$ is an $O\times E$ matrix with known coefficients, one row for each of the left-hand side values of the O linear inequality constraints. The columns of $\mathbf{A}$ correspond to the elements of $\mathbf{x}$. Finally, $\mathbf{d} = (d_1, \ldots, d_{\mathrm{O}})^T$ is a vector with the corresponding right-hand side 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 with inequality constraints and an objective function is generally straightforward.

Minimax Models

Although there are other MILP approaches available for the ATA of parallel test forms, the authors only focus on the Minimax approach (Boekkooi-Timminga, 1990; van der Linden, 2005). The Minimax approach is one of the most commonly used MILP approaches and it works as follows. First, the desired reference or target test form, which can be based on either an existing or an ideal test form, is decided. Then, after choosing the statistical target (i.e., the statistic that is used to attain parallelism), the Minimax model minimizes the maximum distance between the statistical targets of each assembled test form and the statistical target of the reference test form (Chen et al., 2012).

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 $\mathbf{x}$ that indicate if item j is selected for test form f. If some statistical target Q is chosen, a Minimax model minimizes the maximum absolute or relative distance between the Q of the assembled test forms and the Q of the reference test form (also referred to as the target Q) at H prespecified ability values (or $\boldsymbol{\theta}$-points). For instance, when the maximum relative distance is minimized, the following constraints are used:

$$\begin{aligned}&\sum_{j=1}^{J}\frac{Q_j(\boldsymbol{\theta}_h)}{Q^{(t)}(\boldsymbol{\theta}_h)}x_{jf}-y_hz\leq 1,\\&\sum_{j=1}^{J}\frac{Q_j(\boldsymbol{\theta}_h)}{Q^{(t)}(\boldsymbol{\theta}_h)}x_{jf}+y_hz\geq 1,\quad \text{for}\quad h=1,2,\ldots,H;\ \text{and}\quad f=1,2,\ldots,F.\end{aligned}\tag{9}$$

where $Q_j(\boldsymbol{\theta}_h)$ is the item-level statistic Q of item j at $\boldsymbol{\theta}$-point h and $Q^{(t)}(\boldsymbol{\theta}_h)$ is the test-level statistic of the reference test form at $\boldsymbol{\theta}$-point h. The y_h are user-defined (positive) weights that express the (relative) importance of the difference between assembled and reference test forms at the different $\boldsymbol{\theta}$-points h, and z is an additional positive real-valued variable (i.e., $z\geq 0$) in x that should be minimized: min(z).

When all $y_h=1$, z corresponds to the maximal relative difference between the Q of a new test form and the target Q, at one of the prespecified $\boldsymbol{\theta}$ points. When parallelism with respect to Q is considered to be more important at a specific θ-range, the weights y_h can be adjusted accordingly. As a result, in Equation 9, the differences between the Q of the assembled test(s) and the target Q will be weighed more heavily at the theta-points with higher y_h values. The constraints in Equation 9 translate to $2\times H\times F$ rows in **A** (cf. Equation 8). For instance, for the first $\boldsymbol{\theta}$-point ($h=1$) and for the first test form ($f=1$), the left-hand side of the upper equation in Equation 9 results in the following row in **A**: $(\frac{Q_1(\boldsymbol{\theta}_1)}{Q^{(t)}(\boldsymbol{\theta}_1)},\frac{Q_2(\boldsymbol{\theta}_1)}{Q^{(t)}(\boldsymbol{\theta}_1)},\ldots,0,0,y_1)$. The $E=J\times F+1$ columns of A correspond to the $J\times F$ binary variables x_{jf} and the continuous variable z in x. Thus, $\mathbf{x}=(x_{11},x_{21},\ldots,x_{JF},z)^T$. The right-hand values in Equation 9 translate to $2\times H\times F$ ones in **d**. Furthermore, $\mathbf{c}=(0,0,\ldots,1)^T$ is a vector of length $J\times F+1$ containing $J\times F$ zeros and one 1, corresponding with respectively the x_{jf} and z in x, so that Equation 7 becomes min(z). Constraints related to test length, item content, item format, and so on can be added in a similar way. For instance, if item overlap is not allowed among 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**.

Statistical Targets in Minimax Models

As can be seen in Equation 9, there is an important restriction for a statistical target Q. To be applicable in Minimax models—and in ATA using MILP in general—Q at the test level has to be a linear combination of the Q_i at the item level. Both the TCC and the TIF satisfy this condition, while the variance of the MLE $V(\hat{\boldsymbol{\theta}})$ does not (see Equations 2 and 6). When the TIF is chosen as the statistical target (i.e., parallel precision) in the unidimensional case, $Q_j(\boldsymbol{\theta}_h)$ and $Q^{(t)}(\boldsymbol{\theta}_h)$ in Equation 9 are replaced by $I_j(\boldsymbol{\theta}_h)$ and $I^{(t)}(\boldsymbol{\theta}_h)$, respectively. Similarly, with the TCC as statistical target (i.e., parallel expected test scores), $P_j(\boldsymbol{\theta}_h)$ and $P^{(t)}(\boldsymbol{\theta}_h)$ are used. Note that when using the TCC, it is more logical to minimize the maximal absolute, rather than the relative difference (Debeer, Ali, & van Rijn, 2017).

The Multidimensionality Issue

Regardless the dimensionality of the ability $\boldsymbol{\theta}$, $P_i(\boldsymbol{\theta})$ and $T(\boldsymbol{\theta})$ are always scalar-valued. Therefore, apart from the number and selection of H $\boldsymbol{\theta}$-points, the use of the TCC as statistical target (Equation 9) does not differ between the unidimensional and the multidimensional case.

In contrast, $I_i(\boldsymbol{\theta})$ and $I(\boldsymbol{\theta})$ change from scalars in UIRT to $p\times p$ matrices in MIRT. As a result, ATA methods based on these statistical targets do not simply translate from the unidimensional to the multidimensional case.

Previously proposed strategies to mitigate this multidimensionality issue mainly focus on MCAT, where the item selection procedure is based on sequentially selecting the next item. These MCAT item selection criteria either transform the test information matrix into a scalar using a function (e.g., D-optimality), or define a new (scalar) statistical target that is derived from other information measures, such as Shannon's Entropy (Wang & Chang, 2011), mutual information (Mulder & van der Linden, 2009a), or the KL divergence (Veldkamp & van der Linden, 2002). The item selection criteria from MCAT are, however, not directly transferable to the ATA of linear test forms. In addition, the few strategies that have been proposed specifically for the ATA of multidimensional linear tests (a) require interventions from the test assembler (van der Linden, 1996) or (b) focus on the construction of one optimal test, rather than multiple parallel test forms (van der Linden, 1996; Veldkamp, 2002). Therefore, our main aim is to contribute to the literature by presenting and evaluating six statistical targets that mitigate the multidimensionality issue, with a focus on the assembly of multiple parallel test forms. The authors first extend the approach proposed by Veldkamp (2002) to parallel test assembly. Then, they propose four new statistical targets based on KL divergence.

Scalar Aggregates of FI Matrix and Its Inverse

In MCAT, several strategies have been proposed to reduce the FI matrix (or the variance-covariance matrix) into a single scalar. Two of these methods are A-optimality and D-optimality. These strategies are, however, not directly applicable in MILP. To overcome this issue, Veldkamp (2002) proposed to use linear approximation of the A-optimality and D-optimality constraints. The work of Veldkamp (2002) was limited to the assembly of one optimal test. In this article the authors extended it to the assembly of multiple parallel test forms. For both D-optimality and A-optimality the resulting constraints are

$$\begin{aligned}&\sum_{j=1}^{J} k_{jh}x_{jf} - y_h z \leq 1,\\&\sum_{j=1}^{J} k_{jh}x_{jf} + y_h z \geq 1, \quad \text{for} \quad h=1,2,\ldots,H; \quad \text{and} \quad f=1,2,\ldots,F.\end{aligned} \tag{11}$$

The principle behind A-optimality and D-optimality, and the formulae and derivations of the linear approximations k_{jh} are found in Online Appendix A.

KL-Based Statistical Targets

The FI is not the only measure that can be used to quantify the information contained in a random variable or in a distribution. Alternative information measures have been proposed, such as the KL divergence, which expresses the divergence (i.e., nonsymmetric distance) between two probability distributions $f(Y)$ and $g(Y)$ over the same parameter space (Cover & Thomas, 2012; Lehmann & Casella, 1998; Wang, Chang, & Boughton, 2011). Although the KL divergence has been applied in CAT and MCAT, to our knowledge it has not been applied to the problem of ATA of multidimensional test forms.

Online Appendix B discusses how an KL Index (KLI) can be derived from the KL divergence and applied in CAT and MCAT. From this starting point, it is explained how KL-based statistical targets can be defined. Three KLIs that can be applied as statistical targets in ATA using MILP are proposed: the KLI^{MLE}, the KLI^{Local}, and the KLI^{Global}.

Method

In this section the authors describe the design and the technical details of three simulation studies. Simulation 1 focuses on the unidimensional case and compares the performance of the proposed KLIs with the performance of the FI and TCC as the statistical targets in Minimax models. Simulation 2 applies the KLIs, the linear approximations of D- and A-optimality, and the TCC to two-dimensional item pools. Finally, Simulation 3 is based on an operational language aptitude assessment and focuses on a three-dimensional bifactor model.

Simulation 1: Unidimensional Item Pool

In the first simulation, five 30-item test forms are assembled out of an item pool with $J = 450$ unidimensional 2PLM items. Because solving a MILP model is time consuming, only 20 replications were run using the following settings.

Item pool. In each replication an item pool was generated by sampling the item discriminations from a uniform distribution $a_i \sim U(.5, 2)$ and the item locations from a standard normal distribution $b_i \sim N(0, 1)$. In addition, each item was randomly assigned to one of five content types (c_1–c_5) and one of three item formats (v_1–v_3) while maintaining a balanced design with 30 items in each of the $5 \times 3 = 15$ cells.

Reference test form. Because a simulated rather than an empirical item pool was used, each reference test was proportional to the corresponding item pool. The target values for the statistical targets $Q(\boldsymbol{\theta}_h)$ at $\boldsymbol{\theta}_h$ were computed as follows:

$$Q^{(t)}(\boldsymbol{\theta}_h) = \frac{n}{J}\sum_{j=1}^{J} Q_i(\boldsymbol{\theta}_h). \tag{12}$$

Minimax models. In total, for each generated item pool, 2 sets of $\boldsymbol{\theta}$-points $\times 5$ statistical targets $\times 2$ approaches (single or combined) –2 (no combined approach with the TCC as statistical target) = 18 Minimax models were applied. The number of $\boldsymbol{\theta}$-points H was either three (-1.5, 0, and 1.5) or five (-2, -1, 0, 1, and 2) per dimension. The statistical targets were the TIF, the KLI^{MLE}, the KLI^{Local}, the KLI^{Global}, and the TCC. Finally, either a *single* or a *combined approach* was applied. In the single approach, the Minimax models included the constraints in Equation 9 with respect to a single statistical target.[2] In the combined approach, the constraints used in the single approach were combined with additional constraints pertaining to a second statistical target: the TCC. For a given statistical target, the combined approach was equivalent to the single approach but with the following additional constraints:

$$\begin{aligned}
&\sum_{j=1}^{J} P_j(\boldsymbol{\theta}_h)x_{jf} \leq T^{(t)}(\boldsymbol{\theta}_h) + y_h^T,\\
&\sum_{j=1}^{J} P_j(\boldsymbol{\theta}_h)x_{jf} \geq T^{(t)}(\boldsymbol{\theta}_h) - y_h^T, \quad \text{for} \quad h = 1, 2, \ldots, H, \text{ and } \quad f = 1, 2, \ldots, F,
\end{aligned} \tag{13}$$

so that the TCCs of the assembled test forms were constrained to be within an absolute distance of y_h^T from the target TCC at $\boldsymbol{\theta}$-point h (i.e., one of the combined approaches proposed by Debeer et al., 2017). In Equation 13 y_h^T was set to 0.25 at all $\boldsymbol{\theta}$-points. Hence, the difference between the TCCs of two assembled test at the $\boldsymbol{\theta}$-points could be 0.5 points maximum.[3]

Constraints with respect to item content/format were added to all the Minimax models:

$$\sum_{j \in V_w} x_{jf} = \frac{n}{J} \times n_w, \quad \text{for all} \quad w \in (c_1, c_2, c_3, c_4, c_5, v_1, v_2, v_3), \text{ and} f = 1, 2, \ldots, F, \tag{14}$$

where V_w is the subset of items in the item pool belonging to item format or item content w, and n_w is the cardinal number of V_w. As a result, the new test forms were constrained to contain a number of items from each item format and item content that was proportional to the number of items in the item pool belonging to that item format or item content. Finally, item overlap was not allowed by including the constraints in Equation 10.

Simulation 2: Two-Dimensional Item Pool

In the second simulation study, the authors focused on a scenario where five 30-item parallel test forms are assembled out of an item pool with $J = 450$ two-dimensional items. Twenty replications were run, using similar settings as in the unidimensional simulation study.

Item pool. In each replication the item pool consisted of three types of items, from each item type 150 items were generated. Type 1 items were generated by sampling the slope parameters for the first and second dimension from respectively a uniform distribution between 0 and 0.5 and a uniform distribution between 1.5 and 2: $a_{1i} \sim U(0, 0.5)$ and $a_{2i} \sim U(1.5, 2)$. Type 2 items were generated with respectively $a_{1i} \sim U(1.5, 2)$ and $a_{2i} \sim U(0, 0.5)$. Finally, for Type 3 respectively $a_{1i} \sim U(1, 1.5)$ and $a_{2i} \sim U(1, 1.5)$ were used. Note that the three types have roughly the same expected multidimensional discrimination (Reckase & McKinley, 1991). For all the items, the locations were randomly sampled from a standard normal distribution: $b_i \sim N(0, 1)$. In addition, each item was randomly assigned to one of five content types (c_1–c_5) and one of three item formats (v_1–v_3) while maintaining a balanced design with 10 items in each of the $3 \times 5 \times 3 = 45$ cells.

Reference test form and Minimax models. As in Simulation 1, the target values were obtained using Equation 12. The number of $\boldsymbol{\theta}$-points per dimension was either three ($H = 3^2$) or five ($H = 5^2$), with the ability values used in Simulation 1. In addition to the TCC, the KLI^{MLE}, the KLI^{Local}, and the KLI^{Global}, the linear approximations of the A- and D-optimality criteria and the trace of the TIF were the applied statistical targets. Both the single and the previously described combined approach were applied, except when the TCC was the statistical target. Hence, in total, for each generated item pool $2 \times 7 \times 2 - 2 = 26$ Minimax models were applied. In the application of the A-optimality approximation the weights of the diagonal $\mathbf{M}$ matrix (cf. Equation A.2 in Online Appendix A) were set to $m_1 = m_2 = 1$.

As in Simulation 1, item content and item format constraints (cf. Equation 14) as well as item-overlap constraints (cf. Equation 10) were added to each Minimax model. Note that there were no constraints related to the three item types.

Simulation 3: Three-Dimensional Bifactor Item Pool

The third simulation study was based on empirical data from an operational language aptitude assessment. Twenty item pools with $J = 750$ items were randomly created based on operational items following a bifactor structure with one common and two group dimensions. From each item pool five parallel test forms were assembled, each containing $n = 50$ items.

Item pool. Data from an operational language aptitude assessment consisting of 79 items were fit to several IRT models. A bifactor model with one common dimension and two orthogonal group dimensions provided the best relative fit. For each of the 20 item pools 750 new binary items were generated in two steps. First, 750 items were randomly drawn (with replacement) from the 73 binary items.[4] Second, for each drawn original item a new item was generated by sampling its parameters from normal distributions using the parameter estimates and standard errors of the original item. Online Appendix C lists the calibrated item parameters and standard errors of the 73 binary operational items.

Reference test form and Minimax models. As in Simulations 1 and 2, the target values were obtained using Equation 12. The number of theta-points per dimension was either three ($H = 3^3$) or five ($H = 5^3$). The TCC, the KLI^{MLE}, the KLI^{Local}, and the KLI^{Global}, and the trace of the TIF were the statistical targets within either the single or the previously described combined approach. Hence, for each generated item pool, $2 \times 5 \times 2 - 2 = 18$ Minimax models were applied to assemble the test forms. No content or item type constraints were included, only the item-overlap constraints (cf. Equation 10) were added to each Minimax model.

Evaluation Criteria

The parallelism of the constructed test forms in each simulation study was evaluated using different statistics. First, for each test form the root mean square deviation (RMSD) statistic was computed with respect to the target TCC (RMSD_T). For a assembled test form f,

$$\text{RMSD}_T = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[T^{(f)}(\theta_k) - T^{(t)}(\theta_k)\right]^2}, \tag{15}$$

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. Furthermore, K is the number of $\boldsymbol{\theta}$-points over which the RMSD_T is computed. The RMSD_T can be interpreted as the average difference between the expected test scores of the assembled test form and the reference test form.

Second, with respect to the variance of the MLE estimate (i.e., the error variance) in dimension p, the mean relative deviation (MRD) was computed. The MRD is the average relative deviation of the error variance in dimension p of a new test form, from the target error variance in that dimension, expressed as a proportion of the target error variance:

$$\text{MRD}_{V_p} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{V_p^{(f)}\left(\hat{\theta}_k\right) - V_p^{(t)}\left(\hat{\theta}_k\right)}{V_p^{(t)}\left(\hat{\theta}_k\right)}\right|. \tag{16}$$

Third, as a more locally oriented evaluation criterion with respect to the TCC, the maximum score gap (MSG) was computed for each Minimax model. The MSG is the maximal absolute difference between the conditional expected test scores of the assembled forms, within a specified ability range:

$$\text{MSG} = \max_{\theta}\left\{\max_{f}\left[T^{(f)}(\theta)\right] - \min_{f}\left[T^{(f)}(\theta)\right]\right\}. \tag{17}$$

Finally, the maximum relative error variance gap (MREVG) was computed (per dimension). That is, the maximal relative difference in conditional error variance in dimension p across the test forms, expressed as a proportion of the target error variance in dimension p:

$$\mathrm{MREVG}_p = \max_{\theta}\left\{\frac{\max_{f}\left[V_p^{(f)}(\hat{\theta})\right] - \min_{f}\left[V_p^{(f)}(\hat{\theta})\right]}{V_p^{(t)}(\theta)}\right\}. \tag{18}$$

In the unidimensional case, $K = 81$ equally spaced $\boldsymbol{\theta}$-points ranging from -2 to $+2$ were considered to compute the four statistics. In the two- and three-dimensional case, respectively, a square grid with $K = 21^2$ and a cubic grid with $K = 11^3$ equally spaced $\boldsymbol{\theta}$-points ranging from -2 to $+2$ on all dimensions were used.

In addition, as more global measures of performance, the RMSD_T and the MRD_{V_p} were averaged across the five test forms (denoted by $\overline{\mathrm{RMSD}_T}$ and $\overline{\mathrm{MRD}_{V_p}}$, respectively).

Technical Details

To compute the KLI^{MLE} at each $\boldsymbol{\theta}$-point h, 10 Legendre quadrature points was used per dimension. By setting $r = 1\%, 68\%, 47\%,$ and 32% of the MLE confidence region was covered in the uni-, two-, and three-dimensional case, respectively. The KLI^{Global} was computed using 10 Legendre quadrature points per dimension across a 0.01-wide ability region. Finally, for the KLI^{Global} 10 Gauss–Hermite quadrature points were used per dimension. In the two- and three-dimensional case, the ability dimensions were considered orthogonal. Solutions for each Minimax model were computed using GUROBI (Gurobi & Optimization, 2017) through R (R Development Core Team, 2017). For each Minimax model, the maximum computing time on a Fujitsu Lifebook S710 with an i5-560M 2.66 GHz processor was set to 20 min.

Results

In the three simulation studies, GUROBI found at least one feasible solution (satisfying all constraints) for all Minimax models, after 20 min of computing time. For each study, a graphical representation of the results can be found in Online Appendix D.

Simulation 1—Unidimensional Item Pool

As can be seen in Online Appendix D, the following was observed. First, as could be expected giving previous research (cf. Ali & van Rijn, 2016; Debeer et al., 2017), using the TCC as the statistical target results in a better performance with respect to the TCC, but a worse performance with respect to the error variance. Second, for all the other Minimax models the difference in performance with respect to the error variance is limited, but the combined approach with additional TCC constraints performs better with respect to TCC parallelism (cf. Ali & van Rijn, 2016; Debeer et al., 2017). Third, increasing the number of $\boldsymbol{\theta}$-points from three to five improves the parallelism both with respect to the error variance and the TCC. However, the effect is limited. Finally, and most important for this study, the performance of the three KLIs is very similar to the performance of the FI.

These results indicate that in the unidimensional case, the proposed KLIs are valuable replacements for the commonly used TIF as statistical targets in Minimax models.

Simulation 2—Two-Dimensional Item Pool

With respect to Simulation 2, the following was observed (cf. Online Appendix D). First, the combined approach with additional TCC constraints outperforms the single approach, both with respect to the error variance and the TCC evaluation criteria. Second, although increasing the number of $\boldsymbol{\theta}$-points per dimension from three to five has a beneficial effect, the performance with three $\boldsymbol{\theta}$-points per dimension is already at a satisfactory level. Third, the six FI- and KLI-based statistical targets all perform about equally well in the combined approach. In the single approach, the performance of the linear approximations and the KLI^{MLE} seems a little worse. Finally, and unexpectedly, the TCC as statistical target outperforms almost all other statistical targets in the single approach with respect to the error variance; furthermore, it performs only slightly worse than the combined approaches.

Overall, the results suggest that for two-dimensional item pools, the six proposed targets are applicable, preferably using a combined approach. When using the single approach there is a preference for the trace of the FI, the KLI^{Global} and the KLI^{Local}, and the TCC.

Simulation 3—Bifactor Item Pool

Overall, the results are similar to the results in the two-dimensional case (i.e., Simulation 2; see Online Appendix D). First, the combined approach with additional TCC constraints outperforms the single approach, but the difference in performance seems small. Second, although increasing the number of $\boldsymbol{\theta}$-points per dimension from three to five has a beneficial effect, the performance with three $\boldsymbol{\theta}$-points per dimension is already at a satisfactory level. Moreover, when using the combined approach, three $\boldsymbol{\theta}$-points per dimension (H = 27) seem to be enough. Third, the four KLI-based statistical targets perform about equally well in the combined approach. In the single approach, the performance of KLI^{MLE} seems a little worse. Finally, the Minimax model with only the TCC as the statistical target performs remarkably well. Even with respect to the error variance its performance competes with the best KLI-based statistical targets in the single and the combined approach.

The performance with respect to the measurement error on the third dimension is overall worse than those on the first and second dimensions. At the same time, only minor differences are present between the error variance performance on the first and second dimension. The worse performance on the third dimension, especially by the KLI^{Local} and the trace of the FI, is related to the lower IRT reliability of the third dimension (see Online Appendix C). Therefore, Dimension 3 can be considered noisy, and probably less relevant. In practice, one could decide to leave this dimension out of the computation of the different KLIs. However, because the goal of the simulation study was to investigate the feasibility of the proposed statistical targets in a three-dimensional situation, the authors chose to consider all estimated dimensions.

Conclusion

The results of the three simulation studies combined show that the proposed KLIs can be used as statistical targets for assembling test forms using Minimax models via MILP. In the unidimensional case the three proposed KLIs' performance is similar to the performance of FI. In the multidimensional cases, the parallelism achieved by using the KLIs was at an overall satisfactory level. The performance of the linear approximations of the A- and D-optimality criteria was similar, but these statistical targets are limited to the two-dimensional case.

Although some differences were found in the performance of the KLIs, when applied in the single approach, these differences mostly vanished when they were applied in the combined approach with the additional TCC constraints. Moreover, the combined approach generally outperformed the single approaches, with the exception of the single approach with the TCC as the statistical target, for which the performance in the multidimensional cases was remarkably good. Hence, in practice, including the TCC (e.g., in a combined approach) in the Minimax model is recommended, especially when number-correct scoring is to be used.

Discussion

Many testing scenarios require multiple test forms that are parallel with respect to their content and statistical properties. A popular and efficient way to automatically assemble parallel test forms is to write the assembly problem as a Minimax model and solve this model via MILP. In this article the authors focused on the ATA of multidimensional parallel test forms. The statistical targets that are commonly applied in UIRT are not directly applicable in MIRT. Therefore, they proposed six statistical targets that can mitigate the multidimensionality issue. Two of those are based on linear approximations of either the A- or D-optimality criterion (cf. Veldkamp, 2002), and are limited to the two-dimensional case. The other four statistical targets are based on the KL divergence. They can be interpreted as alternative information indexes, and they can be applied to both uni- and multi-dimensional cases.

Using simulated item pools, of which some were based on operational items, the performance of the proposed statistical targets was compared in a uni-, a two-, and a three-dimensional bifactor case. The results indicated that (a) the KLIs can replace the traditional TIF as statistical target in the unidimensional case, (b) the proposed statistical targets can be successfully applied in two- and three-dimensional cases (note that the linear approximations are limited to the two-dimensional case), (c) the combined approach with additional TCC constraints (Ali & van Rijn, 2016; Debeer et al., 2017) outperforms the single approach regardless of the item pool or the chosen statistical target, and (d) in contrast with the unidimensional case, using only the TCC as the statistical target also results in good parallelism with respect to the error variance in the multidimensional case.

In the remainder of this discussion, the practical implications related to the proposed statistical targets and their performance are discussed. Subsequently, the limitations of our study are critically discussed and suggestions for future extensions are provided.

Practical Implications

Given the findings in the simulation studies, the authors would advise to use one of the proposed KLIs in multidimensional situations combined with constraints with respect to the TCC when this is applicable. Depending on the situation, other combined approaches (cf. Debeer et al., 2017) may also be appropriate. When it is not possible or not advisable to include constraints with respect to the TCC (see below) in the Minimax model, the KLI^{Global} should be preferred, because, generally, it demonstrated the best performance in the single approach. The linear approximations of A- and D-optimality can also be applied, but only in the two-dimensional case. In the unidimensional case, there is no need to replace the FI as the statistical target in Minimax models. Nevertheless, the proposed KLIs seem valid alternatives.

The simulation study suggests that, when applying the combined approach with additional TCC constraints, the performance of the different KLIs is similar. Therefore, the choice for a specific KLI can be driven by factors other than the performance with respect to parallelism. First, the interpretation of the KLIs can be a deciding factor. Practitioners responsible for the assembly of new parallel test forms can decide on which KLI to use, based on which type of information index they find most appropriate. Should the used information measure quantify how well an item/test can discriminate ability values from (a) other abilities in a very small region around the ability value (cf. KLI^{Local}), (b) other abilities in a region based on the MLE confidence (cf. KLI^{MLE}), or (c) all abilities in the (assumed) population (cf. KLI^{Global})?

A second factor that may guide the choice for a specific KLI is computation speed. First, in the multidimensional case, and when appropriate, using the TCC is a straightforward and good performing alternative. Second, the trace of the FI, which is the (proportional) linear approximation of the KLI^{Local}, computes faster than the three other KLIs, because the latter require Gaussian quadrature integration. However, even for larger multidimensional item pools, the computation of the KLIs for a high number of $\boldsymbol{\theta}$-points should not take longer than 10 to 15 min. In addition, the computation of the KLIs can be done well before the actual ATA. Moreover, regardless the chosen statistical target, the assembly of the test forms (using MILP) will be the real computational burden.

In situations where there is ample time for the ATA, the authors would suggest to implement the assembly using different statistical targets. Then, the parallelism performance with respect to the relevant statistics (i.e., measurement error, expected test score) can be compared across the different solutions, and the best performing solution can be chosen.

In contrast with the undimensional case (Ali & van Rijn, 2016; Debeer et al., 2017), the performance of the TCC as the statistical target in Minimax models in the multidimensional case was very good. Therefore, it seems advisable to include constraints with respect to the TCC in the ATA Minimax model. However, this may not always be appropriate. First, when pattern scoring methods (e.g., maximum likelihood estimation, expected a priori estimation) are used, and when the expected testing time is the target, rather than the number of items, constraints with respect to the TCC may be undesirable. Such situations are not unlikely when assembling multi-stage assessments, where modules should have equal testing time but not necessarily equal number of items. In these cases, constraining the TCC could result in harder MILP models and suboptimal test assembly. Second, previous research indicated that in the unidimensional case the performance of using only the TCC decreases when the item pool consist of mixed format items (Debeer et al., 2017). A similar deterioration could also be present in multidimensional mixed format item pools.

Although the focus of this article was on the assembly of parallel test forms, the described approaches also apply to cases where only one test form is assembled to be as close as possible to some ideal or existing reference test form. For instance, when only one test form/item set needs to be assembled, rather than using optimal test assembly—which usually involves maximizing the precision—one can specify an ideal test form and and apply the Minimax approach to approximate this ideal form as accurate as possible. The advantage of this approach is that the assembled test satisfies the required properties, without depleting the item pool. Especially when the item pool will be used in the future, the Minimax approach should be preferred over the optimal test assembly.

Finally, when optimal test assembly is the preferred option, both the KLI^{Local} and the *KLI*^{Global} can be applied as the statistical targets. The KLI^{MLE} cannot be used because there is no reference test form that can be used to specify the confidence regions.

Limitations and Extensions

Simulation studies can never completely cover all the practically relevant situations, nor provide an exhaustive evaluation of methods. In that sense, the present simulation study is also limited. In addition, due to the computational demands of assembly problems, we were required to limit the number of conditions and replications to keep the overall computing time reasonable. Therefore, additional simulations are desirable, as they will be able to provide evidence that the proposed statistical targets are also feasible in bigger/smaller item pools, for different reference test forms (for instance, with a flat information surface), for other numbers of test forms, and for mixed format item pools.

In the operationalization of the different KLIs, there were several tuning parameters for which the authors attempted to use sensible values. First, to compute the KLI^{Local} they used a hypercube with a volume of 0.01^p around the ability point as the region for integration. Second, for the KLI^{MLE} the used confidence values $(1 - \alpha)$ decreased with the dimensionality of the item pool, because in MIRT the confidence regions can grow rapidly in low precision regions. Future research could investigate the impact of these tuning parameters on the test assembly performance. Ideally, optimal values would be found.

According to the rationale behind the KLI^{Global}, the extent to which an item/test can discriminate one ability value from all the other ability values in the population of interest is a relevant information measure. Since the actual distribution is generally unknown, the authors suggested to assume that the ability is normally distributed in the population. The normal distribution not only facilitates the computation of the KLI^{Global}, for many human characteristics it is also the commonly accepted population distribution. Nevertheless, in some situations intelligible assumptions about the distribution of ability can be made, for which the normal distribution (or a mixture of normals) is not a good approximation. It is not yet clear how the KLI^{Global} can be computed efficiently in these situations. Future research could address this issue. However, whether this issue would be relevant for ATA in practice seems questionable.

As an alternative for the trace of the FI, which the authors introduced as an proportional approximation of the KLI^{Local}, a reviewer proposed to separately use the FI matrix's diagonal elements as multiple statistical targets in a Minimax model. Like the trace of the FI, this approach would reduce to the traditional (scalar) FI in the unidimensional case. In the multidimensional case, however, this alternative approach would result in a multi-objective problem. Future research could compare the performance of this alternative approach with the performance of the approaches proposed in this article. In addition, it would be interesting to investigate whether and to what extent this multi-objective approach is sensitive to the multidimensionality issue.

In our simulation studies, the authors assumed that the ability dimensions were orthogonal. However, in certain MIRT models ability dimensions may be correlated (e.g., simple-structure models). Although it is not impossible to incorporate them into FI-based statistical targets, the KLI^{Global} can take these correlations more explicitly into account. Future research could investigate whether including or ignoring the correlation structure of ability has a differential impact on the assembly performance when the KLI^{Global} is used as the statistical target.

Also, the authors did not take into account the uncertainty in the item parameter estimates. This can impact the TCC-, FI-, and KL-based statistical targets and, in practice, lead to assembled test forms, whose true FI and TCCs are less parallel than expected. Future research could assess whether or not the Minimax approach is more robust against these uncertainties than optimal test assembly, and whether the approach of Veldkamp, Matteucci, and de Jong (2013) can be extended to the Minimax approach and the statistical targets proposed in this article.

Finally, in this article the authors focused on the unidimensional and multidimensional versions of the 2PLM. The current methods could be extended for application to other commonly used compensatory IRT models, but also to noncompensatory MIRT models. From a theoretical perspective, there are no obstacles to compute and apply the KLIs to other IRT models. Yet from a practical viewpoint, the R-code provided in the online appendix needs to be extended.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been partially funded by the Educational Testing Service.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The authors consistently discuss *test forms*, yet the described methods also apply to other types of *item sets*.
2. The absolute rather than the relative maximum distance to the target was minimized when the test characteristic curve (TCC) was used as statistical target in the single approach.
3. Because the constraints in Equation 13 are redundant when the absolute distance with respect to the target TCC is minimized, only the single approach was applied for the TCC as the statistical target.
4. The six polytomous items were not considered.

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. doi: 10.1177/0146621615613308

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinees ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational and Behavioral Statistics*, *15*, 129-145. doi:10.2307/1164766

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. doi:10.1177/0013164412443688

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. New York, NY: John Wiley. doi:10.1002/047174882x

Debeer, D., Ali, U. S., & van Rijn, P. W. (2017). Evaluating statistical targets for assembling parallel mixed-format test forms. *Journal of Educational Measurement*, *54*, 218-242. doi:10.1111/jedm.12142

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. doi:10.1177/0146621610392211

Donoghue, J. R. (2015). *Comparison of integer programming (IP) solvers for automated test assembly (ATA)* (ETS Research Report RR-15-05). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12051

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. doi: 10.1111/j.1745-3984.2009.00081.x

Gurobi Optimization, I. (2017). *Gurobi optimizer reference manual*. Available from http://www.gurobi.com

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209-227. doi:10.1007/s11336-010-9158-4

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (Vol. *31*). New York, NY: Springer Science & Business Media. doi:10.2307/1270597

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236. doi:10.1177/01466216980223003

Mulder, J., & van der Linden, W. J. (2009a). Multidimensional adaptive testing with Kullback–Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77-101). New York, NY: Springer. doi:10.1007/978-0-387-85461-8 4

Mulder, J., & van der Linden, W. J. (2009b). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273-296. doi:10.1007/s11336-008-9097-5

R Development Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. doi:10.1007/978-0-387-89976-3

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361-373. doi:10.1177/014662169101500407

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika*, *50*, 411-420. doi:10.1007/bf02296260

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, *20*, 373-388. doi:10.1177/014662169602000405

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer-Verlag. doi:10.1007/0-387-29054-0

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement*, *26*, 133-146. doi:10.1177/01421602026002002

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. doi:10.1177/0146621612469825

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575-588. doi:10.1007/bf02295132

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika*, *76*, 363-384. doi:10.1007/s11336-011-9215-7

Wang, C., Chang, H.-H., & Boughton, K. A. (2011). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13-39. doi:10.1007/s11336-010-9186-0