DOCUMENT RESUME

ED 399 299 TM 025 743

AUTHOR	van der Linden, Wim J.
TITLE	Assembling Tests for the Measurement of Multiple Abilities.
INSTITUTION	Twente Univ., Enschede (Netherlands). Dept. of Education.
REPORT NO	RR-95-03
PUB DATE	Dec 95
NOTE	39p.
AVAILABLE FROM	Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P. O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Estimation (Mathematics); Foreign Countries; *Item Banks; *Measurement Techniques; *Test Construction
IDENTIFIERS	*Multidimensionality (Tests); *Variance (Statistical)

ABSTRACT

It is proposed that the assembly of tests for the measurement of multiple abilities be based on targets for the (asymptotic) variance functions of the estimators in each of the abilities. A linear programming model is presented that can be used to computerize the assembly process. Several cases of test assembly dealing with multidimensional abilities are distinguished, and versions of the model applicable to each of these cases are discussed. An empirical example of a test assembly program from a two-dimensional mathematics item pool concludes the paper. (Contains 2 tables, 2 figures, and 27 references.) (Author/SLD)

Assembling Tests for the Measurement of Multiple Abilities

Research Report 95-03

Wim J. van der Linden

department of EDUCATION

Division of Educational Measurement
and Data Analysis

University of Twente

Assembling Tests for the Measurement of Multiple Abilities

Wim J. van der Linden

Assembling tests for the measurement of multiple abilities, Wim J. van der Linden - Enschede: University of Twente, Faculty of Educational Science and Technology, December 1995. - 34 pages.

Abstract

It is proposed to base the assembly of tests for the measurement of multiple abilities on targets for the (asymptotic) variance functions of the estimators of each of the abilities. A linear programming model is presented which can be used to computerize the assembly process. Several cases of test assembly dealing with multidimensional abilities are distinguished, and versions of the model applicable to each of these cases are discussed. An empirical example of a test assembly problem from a two-dimensional mathematics item pool concludes the paper.

Assembling Tests for the Measurement of Multiple Abilities

A standard procedure for assembling tests from an item pool fitting a unidimensional IRT model was suggested by Birnbaum (1968). The central quantity in his suggestion is the test information function which is defined as Fisher's information about the unknown ability parameter θ in the responses to the test taken as a function over the range of possible values of the parameter. For a one-dimensional IRT model of choice, let $L(\theta)$ be the likelihood statistic associated with the responses to the test. The test information function is given by

$$I(\theta) = -E(\frac{\partial^2 \ln L}{\partial\theta^2}), \qquad -\infty<\theta<\infty. \tag{1}$$

Birnbaum's suggestion was to first design a target for the information function of the test and then select items in the test such that the sum of their information functions matches the target. The procedure capitalizes on the fact that local independence between item responses guarantees additivity of the item information functions. If $I_i(\theta)$ is the information function of item i (i=1,...,n), defined analogously to Equation 1 for the likelihood statistic associated with the response to this item, it holds that

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

where n is the number of items in the test.

If $\hat{\theta}$ is the maximum-likelihood estimator (MLE) of θ,

$$\mathrm{Var}(\hat{\theta}\,|\,\theta) \rightarrow 1/I(\theta) \quad \text{for} \quad n \rightarrow \infty \tag{3}$$

(e.g., Kendall & Stuart, 1976, chap. 18). Because of this reciprocity, setting a target for the information function is equivalent to setting one for the (asymptotic) variance function of the ability estimator.

However, in practice, in spite of the additivity of the item information functions, the problem of picking n items from a pool of a realistic size such that the sum of the functions matches the target best over the range of possible values of θ is not a trivial task. The prohibitively large number of possible combinations rules out optimal test assembly by hand. In fact, even for a high-speed computer explicit enumeration of all possible solutions and picking out the best is an unrealistic job. The problem gets more difficult still if the test has to meet various constraints on the selection of the items related to the distributions of, for example, item content, item format, or the values of certain item parameters. To implement Birnbaum's procedure, efficient algorithms which reduce the set of feasible solutions to a smaller set of candidate solutions and then select an optimal one are badly needed.

Application of Linear Programming

Formally, the problem of test assembly is an instance of the problem of constrained combinatorial optimization which, in its mathematical generality, has been studied in such fields as applied mathematics, decision theory, and operations research. It should therefore not come as a surprise that attempts to implement Birnbaum's procedure in a computer algorithm have been based on techniques of combinatorial optimization, in particular on techniques of (mixed) integer programming from the field of Linear Programming (LP). Though suggestions to resort to LP for solving test assembly problems were made earlier (Feuerman & Weiss, 1973; Votaw, 1952; Yen, 1983), the first LP model for a variation of Birnbaum's procedure was published in Theunissen (1985). Ever since, modeling various test assembly problems as an LP problem and finding algorithms and heuristics to solve the model for an optimal solution has been a fruitful field of research. Some references to relevant papers are: Adema (1990, 1992); Adema, Boekkooi-Timminga, and van der Linden (1991); Adema and van der Linden (1989); Amstrong and Jones (1992); Amstrong, Jones, and Wu (1992); Boekkooi-Timminga (1987, 1990); Timminga and Adema (1995); van der Linden (1994); van der Linden and Boekkooi-Timminga (1988, 1989); van der Linden and Luecht (in press). Important heuristic approaches to the same problems have been presented in Ackerman (1989), Luecht and Hirsch (1992), and Swanson and Stocking (1993).

Maximin Model

The model taken as a starting point for the problem of multidimensional test assembly is the maximin model for unidimensional assembly in van der Linden and Boekkooi-Timminga (1989). It is assumed that a test of n items has to measure an interval of possible θ values with uniform accuracy, and that the test assembler wants to control this behavior at the ability points θ_k, k=1,...,K. To attack the problem, decision variables x_i, i=1,...,I are defined, one for each item in the pool, which take the value 1 if the item is included into the test and the value 0 otherwise. The maximin model is as follows:

$$\text{maximize } y \tag{4}$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i - y \geq 0, \qquad k=1,...K, \tag{5}$$

$$\sum_{i=1}^{I} x_i = n \; , \tag{6}$$

$$x_i \in \{0,1\}, \qquad i=1,...,I, \tag{7}$$

$$y \geq 0 \; . \tag{8}$$

The model is based on the idea that a common lower bound y to each of the values of the test information function defined by the inequality in Equation 5 should be maximized, as is done by the objective function in Equation 4. At the same time, Equation 6 constrains the length of the test to size n. The last two equations define the ranges of values of the decision variables in the model.

The model can be generalized to a target for the test information function of any shape by providing the variable y in Equation 5 with coefficients r_k which govern the relative heights of the test information function at $\theta_1 ... \theta_K$ (van der Linden & Boekkooi-Timminga, 1989). For ease of exposition only the case of a uniform target will be considered in this paper. Also, a catalogue of additional linear constraints is available to model test specifications with respect to such categories as item content, format, response time available, the values of classical or IRT item parameters, interdependencies between test items, etc. (van de Linden & Boekkooi-Timminga, 1989). For an illustration of the use of some of the constraints, see the empirical example below. The model has been implemented as one of the options in the computer program CONTEST (Timminga & van der Linden, 1995) which contains a large choice of algorithms and heuristics to solve the model for an optimal combination of values for its decision variables. Quick heuristics to solve certain test assembly problems have been presented in Ackerman (1989) and Leucht and Hirsch (1992). If the model has a network flow structure, computation of an optimal solution simplifies dramatically (e.g., Amstrong, Jones & Wu, 1992).

It is the purpose of this paper to present models for the optimal assembly of tests measuring more than one ability. However, unlike the case of an IRT model with a single ability parameter, for a model with multiple ability parameters Fisher's information measure is no longer a scalar but a (non-diagonal) matrix. Also, the (asymptotic) variances of the MLEs of the ability parameters are not given by the reciprocals of the diagonal elements of the information matrix, but by the diagonal elements of the variance-covariance matrix, which is the inverse of it. Hence, the motivation to use a target directly for Fisher's information measure fails for the case addressed in this paper. To solve the problem, the use of targets for the variance functions in the model will be explored. Then a generalization of the maximin model and a heuristic for the assembly of tests in the presence of multiple ability parameters will be proposed. Next, various cases of multidimensional test assembly will be discussed, and it will be shown how the model applies to each of these cases. The cases include such examples as the assembly of a test measuring two intentional abilities, a test insensitive to a nuisance ability, or a test required to have a simple structure of underlying abilities. The paper concludes with an empirical example of the use of the model to a test assembly problem.

Multidimensional Test Assembly

The multidimensional IRT model considered in the paper is the logistic model discussed by McKinley and Reckase (1983), Reckase (1985; in press), and Samejima (1974). To simplify notation, only the case of two ability parameters,

(θ_1, θ_2), is considered. Let the response variables U_{ij} take the value 1 if the response of person j=1,...,N to item i=1,...,n is correct and the value 0 otherwise. The model is defined by the following logistic response function:

$$P_i(\theta_1,\theta_2) \equiv P(U_{ij}|a_{1i},a_{2i},d_i,\theta_1,\theta_2)$$
$$\equiv \frac{\exp(a_{1i}\theta_1 + a_{2i}\theta_2 + d_i)}{1 + \exp(a_{1i}\theta_1 + a_{2i}\theta_2 + d_i)}, \tag{9}$$

where (a_{1i},a_{2i}) are the discrimination parameters of item i along the abilities θ_1 and θ_2, respectively, and d_i can be interpreted as a composite parameter representing the difficulty of the item. In the remainder of this paper, it is assumed that these item parameters are known and that the model is used to estimate the abilities $(\theta_{1j},\theta_{2j})$ from a realization of the response variables $U_{ij}=u_{ij}$ for i=1,...,n and j=1,...,N.

Variance Functions

For the case of two ability parameters Fisher's information matrix is defined as:

$$I(\theta_1,\theta_2) \equiv -E\begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial\theta_1^2} & \dfrac{\partial^2 \ln L}{\partial\theta_1\partial\theta_2} \\ \dfrac{\partial^2 \ln L}{\partial\theta_1\partial\theta_2} & \dfrac{\partial^2 \ln L}{\partial\theta_2^2} \end{bmatrix}, \tag{10}$$

where L now is the likelihood statistic associated with the data under the model in Equation 9. Following the derivation in Ackerman (1994, Appendix) and using the notation $P_i \equiv P_i(\theta_1,\theta_2)$, the following result is obtained for the model in Equation 9:

$$I(\theta_1,\theta_2) = \begin{bmatrix} \sum_{i=1}^{n} a_{1i}^2 P_i Q_i & \sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i \\ \sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i & \sum_{i=1}^{n} a_{2i}^2 P_i Q_i \end{bmatrix}. \tag{11}$$

Standard techniques for matrix inversion yield the variance-covariance matrix of the MLE's of (θ_1,θ_2):

$$V(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2) = \begin{bmatrix} \dfrac{\sum_{i=1}^{n} a_{2i}^2 P_iQ_i}{|I(\theta_1,\theta_2)|} & \dfrac{-\sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i}{|I(\theta_1,\theta_2)|} \\ \dfrac{-\sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i}{|I(\theta_1,\theta_2)|} & \dfrac{\sum_{i=1}^{n} a_{1i}^2 P_iQ_i}{|I(\theta_1,\theta_2)|} \end{bmatrix}, \tag{12}$$

where

$$|I(\theta_1,\theta_2)| = (\sum_{i=1}^{n} a_{1i}^2 P_iQ_i)(\sum_{i=1}^{n} a_{2i}^2 P_iQ_i) - (\sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i)^2 \tag{13}$$

is the determinant of the matrix in Equation 11, which is assumed to be nonzero throughout this paper. The diagonal elements of the matrix in Equation 12 are the (asymptotic) variances of the MLEs of θ_1 and θ_2, respectively:

$$\begin{aligned} Var(\hat{\theta}_1|\theta_1,\theta_2) = {} & (\sum_{i=1}^{n} a_{2i}^2 P_iQ_i)[(\sum_{i=1}^{n} a_{1i}^2 P_iQ_i)(\sum_{i=1}^{n} a_{2i}^2 P_iQ_i) \\ & - (\sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i)^2]^{-1}, \end{aligned} \tag{14}$$

$$\begin{aligned} Var(\hat{\theta}_2|\theta_1,\theta_2) = {} & (\sum_{i=1}^{n} a_{1i}^2 P_iQ_i)[(\sum_{i=1}^{n} a_{1i}^2 P_iQ_i)(\sum_{i=1}^{n} a_{2i}^2 P_iQ_i) \\ & - (\sum_{i=1}^{n} a_{1i}a_{2i}P_iQ_i)^2]^{-1}. \end{aligned} \tag{15}$$

Observe that Equation 14 gives the (asymptotic) variance of $\hat{\theta}_1$ for the true ability of the examinee being a point in a two-dimensional space. Thus, the variance of $\hat{\theta}_1$ not only depends on the true value of θ_1 but also on the value of θ_2. The same holds for the (asymptotic) variance of $\hat{\theta}_2$ in Equation 15. Also, observe that the two variances differ only by the factors a_{2i}^2 and a_{1i}^2 in the two numerators.

If the variances in Equations 14-15 are taken as functions over the complete two-dimensional ability space, two variance functions are defined, one for $\hat{\theta}_1$ and the other for $\hat{\theta}_2$. Figure 1 shows the plots of three pairs of variance functions each for a different

[Figure 1 about here]

test. The first test consists of nine items which have larger values for the first than for the second discrimination parameter: a_1=(2.0, 2.0, 2.6, 1.2, 1.5, 1.7, 1.2, 0.8, 0.9); a_2=(0.1, 1.1, 1.7, 2.4, 2.0, 3.0, 1.9, 2.1, 1.8); and d_i=0.0 for all items. In the second test, the values for the two discrimination parameters are equal on average for the six items in the test: a_1=(1.8, 2.6, 1.7, 1.8, 2.2, 2.0), a_2=(2.0, 1.8, 1.9, 1.7, 1.8, 1.7); and d_i=-2.0 for al items. The six items in the third test have values for the first discrimination parameter exactly twice the ones for the second parameter, the only exception being Item 6 for which the values slightly deviate from this proportion: a_1=(2.0, 2.0, 2.6, 2.4, 2.0, 3.0); a_2=(1.0, 1.0, 1.3, 1.2, 1.0, 1.7); and d_i=0.0 for all items. The result of this case of "weak identifiability" is a variance function for $\hat{\theta}_1$ which is low only locally along a line in the ability plane and a function for $\hat{\theta}_2$ which never takes on any small value. (Note that for readability in all three figures the surfaces are cut at a height of 100.)

Targets for Variance Functions

It is proposed to define targets for the two variance functions to guide the multidimensional test assembly process. Graphically, the proposals means that tests are assembled such that the plots of their variance functions meet previously defined forms. For example, if ability θ_1 is considered to be more important than θ_2, a target for $\mathrm{Var}(\hat{\theta}_1|\theta_1,\theta_2)$ uniformly lower than the one for $\mathrm{Var}(\hat{\theta}_2|\theta_1,\theta_2)$ over the ability area of interest makes sense. The proposal implies that the covariance function in Equation 12 can be ignored. This implication is in agreement with the fact that test assembler are typically not interested in covariances between ability estimators.

Computational Complications

Test assembly with simultaneous targets for two distinct functions is an example of a multi-objective decision problem. Standard approaches to decision problems with two objectives are, for example, to combine the two objectives into one objective function or to focus on one as the objective function and represent the other by a constraint with an optimally chosen bound. More important, however, is the fact that the two expressions in Equations 14-15 are nonlinear. A realistic objective function based on the difference between the two expressions and their targets will also be nonlinear. Due to this complication, algorithms allowing for optimal multidimensional test assembly which operate in polynomial time are not available. Hence, unless the problem is trivially small, the use of a heuristic which yield good but not necessarily best solutions seems the only possibility left.

Multidimensional Maximin Model

Further analysis of the variance functions in Equations 14-15 reveals that, though nonlinear, they consists of sums each of which is additive in the items. The role of these sums becomes more obvious if decision variables are added to Equation 14, and this variance function of $\hat{\theta}_1$ is written as:

$$\mathrm{Var}(\hat{\theta}_1 | \theta_1,\theta_2) = [\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i - (\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i)^2 / (\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i)]^{-1}. \qquad (16)$$

It is now immediately clear that, for a fixed value of (θ_1,θ_2), the function in Equation 16 decreases in value if the values of the decisions variables x_i, $i=1,...,I$, are chosen such that

$$\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i \qquad \text{increases;} \qquad (17)$$

$$\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i \qquad \text{increases;} \qquad (18)$$

$$\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i \qquad \text{decreases.} \qquad (19)$$

However, note that for a fixed set of item parameter values, the expression in Equation 19 can not decrease independently of the expressions Equations 17-18. In fact, a tradeoff exists between these two sets of expressions because any choice of parameter values which decreases the last expression also decreases the first two. The optimum value of Equation 16 thus depends on the relative rates of change of the three expressions. This fact suggests an approach in which the expression in Equation 19 is minimized for a systematically varying series of lower bounds on the expressions in Equations 17-18.

Consider the following variant of the maximin model in Equations 4-7 in which, for a selection of ability points $(\theta_{1p},\theta_{2q})$, p=1,...,P, q=1,...,Q, minimization of the expression in Equation 19 is taken as the objective function and the expressions in Equations 18-19 are constrained by lower bounds:

$$\text{minimize } y \tag{20}$$

subject to

$$\sum_{i=1}^{I} a_{1i}a_{2i}P_i(\theta_{1p},\theta_{2q})Q_i(\theta_{1p},\theta_{2q})x_i - y \geq 0, \qquad p=1,...,P,\ q=1,...,Q, \tag{21}$$

$$\sum_{i=1}^{I} a_{1i}^2 P_i(\theta_{1p},\theta_{2q})Q_i(\theta_{1p},\theta_{2q})x_i \geq c_1, \qquad p=1,...,P,\ q=1,...,Q, \tag{22}$$

$$\sum_{i=1}^{I} a_{2i}^2 P_i(\theta_{1p},\theta_{2q})Q_i(\theta_{1p},\theta_{2q})x_i \geq c_2, \qquad p=1,...,P,\ q=1,...,Q, \tag{23}$$

$$\sum_{i=1}^{I} x_i = n, \tag{24}$$

$$x \in \{0,1\}, \qquad i=1,...,I, \tag{25}$$

$$y \geq 0. \tag{26}$$

The basic idea is now to run this model systematically varying the values of c_1 and c_2 until optimal variance functions are found.

Choices of Values for c_1 and c_2

First, note that the following inequalities hold:

$$0 \leq \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i \leq .25 \sum_{\max} a_{1i}^2 , \tag{27}$$

$$0 \leq \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i \leq .25 \sum_{\max} a_{2i}^2 , \tag{28}$$

where the right-hand sums are taken over the n items with the largest values for a_{1i} and a_{2i} in the item pool, respectively. Thus, the right-hand sides can be used as upper bounds for c_1 and c_2. Observe, however, that items with high values for a_{1i} are not necessarily the ones with high values for a_{2i} and vice versa; therefore these bounds will seldom be reached in practice.

Second, if c_1 and/or c_2 are set high, overconstraining may occur and no feasible solution found. If infeasibility is found for certain values of c_1 and c_2, no larger values have to be tried since these will also yield infeasibility.

Third, for brevity, let

$$u \equiv \sum_{i=1}^{I} a_{1i}^2 P_i Q_i , \tag{29}$$

$$v \equiv \sum_{i=1}^{I} a_{2i}^2 P_i Q_i , \tag{30}$$

$$w \equiv \sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i . \tag{31}$$

Suppose no dependences existed between u,v, and w. The following partial derivatives then show the impact of u and v on the variance function of $\hat{\theta}_1$:

$$\frac{\partial \mathrm{Var}(\hat{\theta}_1 | \theta_1, \theta_2)}{\partial u} = \frac{-v^2}{(uv - w^2)^2} \tag{32}$$

and

$$\frac{\partial \mathrm{Var}(\hat{\theta}_1 | \theta_1, \theta_2)}{\partial v} = \frac{-w^2}{(uv - w^2)^2} . \tag{33}$$

As u,v,w ≥ 0, the derivatives are negative for all possible values of u, v, and w (provided uv≠w). Consequently, as already assumed in Equations 17-18, for a fixed value of (θ_1,θ_2), $\mathrm{Var}(\hat{\theta}_1 | \theta_1,\theta_2)$ is minimal for u and v maximal. However, in the model, w is minimized, and the derivatives in Equations 32-33 show that if an optimal solution is approached, the marginal contribution of v to $\mathrm{Var}(\hat{\theta}_1 | \theta_1,\theta_2)$ is likely to be smaller than the one of u. If w approaches zero, the contribution of v even becomes negligible. By symmetry, the reverse conclusion holds for the contributions of u and v to the variance function of $\hat{\theta}_2$. These relations seem to suggest that a larger value of c_1 relative to c_2 favors minimization of $\mathrm{Var}(\hat{\theta}_1 | \theta_1,\theta_2)$ whereas a smaller value of c_1 favors minimization of $\mathrm{Var}(\hat{\theta}_2 | \theta_1,\theta_2)$. However, the actual problem is one of combinatorial optimization over a finite pool of possible values for the item parameters. Also, as already noted these values create dependencies between the expressions in Equations 17 through 19. It is therefore recommended to always check this suggestion for the actual item pool in use. For an empirical example, see the analyses further on in this paper.

A Heuristic

The following heuristic can be used to find a (nearly) optimal solution to the test assembly problem:

1. Choose a grid of values for $(\theta_{1p},\theta_{2q})$ which covers the ability area of interest. As the variance functions are well-behaved smooth functions, a 3x3 or 4x4 grid will generally do. There is no need to space the points evenly or to have the same numbers of point along both dimensions.
2. Choose a series of values for (c_1,c_2) covering the range of possible values below the upper bounds in Equations 27-28, taking into account the distribution of the values of the item parameters in the pool as well as the goal of the test (see the following section) ;
3. Solve the model in Equations 20-26 using, for example, one of the algorithms or heuristics in CONTEST (Timminga & van der Linden, 1995);
4. Calculate the two variance functions for each solution in the previous step;
5. Based on an inspection of the results, repeat steps 3-4 for a finer grid of values for (c_1,c_2) in the neighborhood of the value for which the best variance functions were obtained;
6. Repeat steps 5 until favorable variance functions are obtained.

Based on experiences with a previous test assembly problem for a given item pool, the first selection of values for (c_1,c_2) can be made more effective. For example, if infeasibility was met for certain values of (c_1,c_2) in a previous run, it makes no sense to use larger values for (c_1,c_2) in any later test assembly problem for the same item pool whatever other constraints are added to the model. An implementation of the heuristic for the case of two flat variance functions will be given in the empirical example below.

Different Cases of Multidimensional Test Assembly

Five different cases of test assembly are considered in which multidimensionality of the item pool plays a role. For each case a different use of the multidimensional model in Equations 20-26 is proposed, with the exception of one case which leads to the use of a modified version of the unidimensional model in Equations 4-8. The main criteria used to classify the five cases are: (1) whether the abilities are intentional or should be viewed as "nuisance abilities"; and (2) whether or not the abilities underlying the test should display a "simple structure".

1. Two intentional abilities. In this case, the test items are designed to measure two abilities, and it is intended to report scores on both abilities for each examinee. Thus, for each possible ability point (θ_1,θ_2) the test should produce variances of $\hat{\theta}_1$ and $\hat{\theta}_2$ which meet realistic targets.

The model to be used in this case is the multidimensional maximin model in Equations 20-26 possibly provided with additional (linear) constraints to deal with remaining test specifications. As already suggested, the relative sizes of the values of c_1 and c_2 can be used to control for the importance of the two variance functions.

2. One intentional and one nuisance ability. The test items in the pool are designed to measure one intentional ability but happen to be sensitive to another ability. When scoring the test, the nuisance ability is ignored and only a score for the intentional ability is reported. An obvious example is the case of a nuisance ability which creates "differential item functioning" because a focal and a reference group have different distributions on it. Removing the effect of the nuisance ability by fitting a two-dimensional IRT model and scoring only for the intentional ability is believed to be a strategy yielding more informative ability estimates than simply

removing all items sensitive to the nuisance ability from the test.

The best approach in this case seems to ignore the variance function for the estimator of the nuisance ability and set a target for the intentional ability only. If θ_2 is the nuisance ability, this approach is implemented if the same model as in the previous case is applied, but now with c_2 small relative to c_1. Again, additional linear constraints can be added to the model to deal with other test specifications.

3. <u>One composite ability</u>. This case arises if both abilities are intentional but an estimates of the linear combination $\beta_1\theta_1 + \beta_2\theta_2$, with β_1, $\beta_2 > 0$, are to be reported. A practical motivation for this case might be that the construct measured by the test is truly two dimensional but that test consumers want a single score equally reflecting both abilities. The variance function of the estimator of the linear composite is equal to

$$\begin{aligned} \mathrm{var}(\beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2 | \theta_1,\theta_2) = \beta_1^2\mathrm{Var}(\hat{\theta}_1 | \theta_1,\theta_2) + \beta_2^2\mathrm{Var}(\hat{\theta}_2 | \theta_1,\theta_2) \\ + 2\beta_1\beta_2\mathrm{Covar}(\hat{\theta}_1,\hat{\theta}_2 | \theta_1,\theta_2). \end{aligned} \tag{34}$$

(Ackerman, 1994, Eqs. 15-16). Though this function is also an expression consisting of the sums of the elements in the information matrix in Equation 11, analysis of the expression shows that it misses the monotonicity which could lead to a conclusion as in Equations 17-19. The best solution in this case, therefore, is to rotate the ability space such that in the reparameterized model the composite corresponds to the first ability dimension. Henceforth, the case is identical to the previous one.

4. <u>Simple ability structure: Case I</u>. The item pool is again assumed to measure two intentional abilities but the test has to be assembled such that one subtest is maximally informative on θ_2 and another subtest on θ_2. This case may arise if for diagnostic purposes test performances have to be reported at item level

and it is thus necessary to know which items measure θ_1 best and which items θ_2.

Let n_1 be the number of items required to be informative on θ_1 and n_2 the number of items informative on θ_2. An obvious approach in this case seems to apply the multidimensional model in Equations 20-26 first to assembly n_1 items under the condition of $c_1>c_2$ and a second time to assemble n_2 items under the reverse condition of $c_1<c_2$ removing the items already selected from the pool. However, a clear disadvantage of a sequential approach is that some of the items fitting the constraints of the second subtest better have already gone into the former. Also, it is not possible to directly constrain item selection with respect to item content, format, etc., at the level of the complete test.

A more favorable solution, therefore, is to select the two subtest simultaneously. This choice leads to an adaptation of the multidimensional model in Equations 20-26. New decision variables x_{is} are introduced which take the value of 1 if item i is assigned to subtest s and the value 0 otherwise (s=1,2). The adapted model is:

$$\text{minimize } y \qquad (35)$$

subject to

$$\sum_{s=1}^{2}\sum_{i=1}^{I} a_{1i}a_{2i}P_i(\theta_{1p},\theta_{2q})Q_i(\theta_{1p},\theta_{2q})x_{is} - y \geq 0, p=1,..,P,\ q=1,...,Q, \quad (36)$$

$$\sum_{s=1}^{2}\sum_{i=1}^{I} a_{1i}^2P_i(\theta_{1p},\theta_{2q})Q_i(\theta_{1p},\theta_{2q})x_{is} \geq c_1, \qquad p=1,...,P,\ q=1,...,Q, \quad (37)$$

$$\sum_{s=1}^{2} \sum_{i=1}^{I} a_{2i}^{2} P_i(\theta_{1p},\theta_{2q}) Q_i(\theta_{1p},\theta_{2q}) x_{is} \geq c_2, \qquad p=1,\ldots,P,\ q=1,\ldots,Q, \quad (38)$$

$$\sum_{i=1}^{I} x_{is} = n_s, \qquad s=1,2, \quad (39)$$

$$\sum_{s=1}^{2} x_{is} \leq 1, \qquad i=1,\ldots,I, \quad (40)$$

$$x \in \{0,1\}, \qquad i=1,\ldots,I; \quad (41)$$

$$y \geq 0. \qquad (42)$$

New constraints in the model are the ones in Equation 39 which define the lengths of the two subtests as well as those in Equation 40 which prevent the items from being assigned to both subtests. The model can be solved using the algorithms and heuristics referred to earlier. However, the doubling of the number of decision variables generally has an effect on the speed of the algorithms and heuristics comparable to the one of doubling the size of the item pool, and, as a consequence, some of the heuristics slow down considerably.

5. Simple ability structure: Case II. For completeness' sake, the case of two subpools of items each fitting a unidimensional IRT but with the complete pool fitting only a two-dimensional model is mentioned. The practical motivation for assembling a test with this simple structure for its ability space is the same as the one in the previous case.

Again, a simple solution would seem to assemble the two subtest sequentially but the same objections to sequential assembly as before apply. A model for simultaneous assembly can be obtained by adapting the model in Equations 4-8 analogously to Equations 35-42.

Empirical Example

Data from an ACT Assessment Program Mathematics Item Pool were used to assemble a test. The pool consisted of 176 items to which a two-dimensional version of the model in Equation 9 showed an acceptable fit. The items in the pool were classified according to content (PG: Plane Geometry; PA: Pre-Algebra; EA: Elementary Algebra; CG: Coordinate Geometry; TG: Trigonometry; IA; Intermediate Algebra) as well as skill (BS: Basic Skill; AP: Application; AN: Analysis). It was attempted to assembly two tests with flat variance functions for both ability estimators over the complete grid of points defined by $\theta_1, \theta_2 = -2, -1, 0, 1, 2$, where measurement of both ability variables was assumed to be intentional and equally important. One test was assembled using the basic model in Equations 20-26 (Model I). The other test was assembled adding the following set of constraints to the previous model to simulate the presence of content and skill specifications in the assembly program (Model II):

$$\sum_{i \in V_{PG}} x_i \geq 5, \tag{43}$$

$$\sum_{i \in V_{PA}} x_i \geq 5, \tag{44}$$

$$\sum_{i \in V_{EA}} x_i \geq 5, \tag{45}$$

$$\sum_{i \in V_{CG}} x_i \geq 5, \tag{46}$$

$$\sum_{i \in V_{TG}} x_i \geq 5, \tag{47}$$

$$\sum_{i \in V_{IA}} x_i \geq 5, \tag{48}$$

$$\sum_{i \in V_{BS}} x_i \geq 15, \tag{49}$$

$$\sum_{i \in V_{AP}} x_i \geq 15, \tag{50}$$

$$\sum_{i \in V_{AN}} x_i \geq 5, \tag{51}$$

where, for example, V_{PG} is the set indices of the items with content classification Plane Geometry. For both models test length was set at n=50. The two models were solved using the First Acceptable Integer Solution algorithm as implemented in the CONTEST program (Timminga & van der Linden, 1995, sect. 6.6; see also Adema, Boekkooi-Timminga & van der Linden, 1991). This algorithm was used to find the first integer solution with a value for the objective function within 5% from the optimal value for the fully relaxed model. Since the two variance functions were assumed to be equally important, the values of c_1 and c_2 in Equations 22-23 were set equal to each other.

For both models values of c_1=c_2 larger than or equal to 1.4 led to overconstraining and no feasible solutions was found. Therefore, the two models were run for c_1=c_2=0.0(0.1)1.3. The results are summarized in Table 1. As the variance functions had to be both low and flat, the mean value (μ) plus one standard deviation (σ) of the values of the two variance functions over de 25 points of the grid of (θ_1,θ_2)

[Table 1 about here]

values was used as a summary measure. For both models the solutions for the smallest values of $c_1=c_2$ not only yielded the same value for this measure but also contained the same set of items. The best solution for Model I was obtained for $c_1=_2=1.0$ Plots of the variance functions $\mathrm{Var}(\hat{\theta}_1|\theta_1,\theta_2)$ and $\mathrm{Var}(\hat{\theta}_2|\theta_1,\theta_2)$ associated with the items in this solution are given in Figure 2. Both functions show a flat surface over the ability space considered, albeit the one for $\hat{\theta}_1$ has a tendency to slightly increase for θ_1 approaching the value of 2.0, whereas the one for $\hat{\theta}_2$ goes up for θ_2 approaching -2.0. For Model II, the best solution was obtained for $c_1=c_2=0.9$. Both the numerical results in Table 1 and the plots of the two variance functions for the solution obtained for this value show that adding the additional constraints in Equations 43 through 51 to the model hardly deteriorates the results.

[Figure 2 about here]

To assess the numerical effects of setting c_1 lower or higher than the value of c_2, solutions for Model I were computed over the full range of possible values for c_2 both for $c_1=0.2$ and $c_1=1.2$. Observe that these two value for c_1 are near the extremes of the range of values in Table 1 for which feasible solutions were obtained. The results are presented in Table 2. The general conclusion from this table is that the lower value for c_1 favors minimization of the variance function for $\hat{\theta}_2$ both in terms of its average value and spread, whereas the higher value of c_1 favors minimization of the function for $\hat{\theta}_1$. These results are as predicted.

[Table 2 about here]

However, surprisingly, none solutions was better than the one in Table 1 for $c_1=c_2=1.0$.

The software in the examples was run on a PC with a 486 processor (66 Mhz). None of the runs took more than 2 seconds of computing time to reach a solution.

Discussion

The choice to base the assembly of tests measuring multiple abilities on the variance functions associated with the ability estimators seems obvious. However, as indicated earlier, the choice involves a multi-objective decision problem with nonlinear objective functions. The current paper offers a model along with a heuristic scheme to solve the problem. Implementations of the heuristic for other targets than the one for the case of two intentional abilities in the empirical example above still have to be examined. It is not unlikely that practical experience with the heuristic will reveal that, for some of the cases discussed above, certain patterns of item parameter values guarantee optimal variance functions. If so, this knowledge, in turn, could be used to further improve the focus of the heuristic.

References

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

Ackerman, T.A. (1994). Creating a test information profile for a two-dimensional latent space. Applied Psychological Measurement, 18, 257-275.

Adema, J.J. (1990). The construction of customized two-staged tests. Journal of Educational Measurement, 27, 241-253.

Adema, J.J. (1992). Methods and models for the construction of weakly parallel tests. Applied Psychological Measurement, 16, 53-63.

Adema, J.J., Boekkooi-Timminga, E., & van der Linden, W.J. (1991). Achievement test construction using 0-1 linear programming. European Journal of Operations Research, 55, 103-111.

Adema, J.J. & van der Linden, W.J. (1989). Algorithms for computerized test construction using classical item parameters. Journal of Educational Statistics, 14, 279-290.

Amstrong, R.D. and Jones, D.H. (1992). Polynomial algorithms for item matching. Applied Psychological Measurement, 16, 365-373.

Amstrong, R.D., Jones, D.H., & Wu, I-L. (1992). An automated test development of parallel tests. Psychometrika, 57, 271-288.

Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick), Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1*, 1101-112.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, 15, 129-145.

Feuerman, F. & Weiss, H. (1973). A mathematical programming model for test construction and scoring. *Management Science, 19*, 961-966.

Kendall, M.G., & Stuart, A. (1976). *The advanced theory of statistics* (Volume 2; 4th edition). London: Griffin & Co.

Luecht, R.M. & Hirsch, T.M. (1992). Computerized test construction using average growth approximation of target information functions. *Applied Psychological Measurement, 16*, 41-52.

McKinley, R.L., & Reckase, M.N. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research report ONR 83-2). Iowa City, IA: American College Testing.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M.D. (in press). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. New York City, NY: Springer-Verlag.

Samejima, F. (1974). Normal ogive model for the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

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.

Timminga, E. & Adema, J.J. (1995). Test construction from item banks (pp. 111-127). In G. H. Fischer & I.W. Molenaar (Eds.), The Rasch model: Foundations, recent developments, and applications. New York: Springer-Verlag.

van der Linden, W.J. (1994). Optimum design in item response theory: Applications to test assembly and item calibration. In G.H. Fischer & D. Laming (Eds,), Contributions to mathematical psychology, psychometrics, and methodology (pp. 308-318). New York: Springer-Verlag.

van der Linden, W.J. & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. Applied Psychological Measurement, 12, 201-209.

van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 53, 237-247.

van der Linden, W.J. & Luecht, R.M. (in press). An optimization model for test assembly to match observed-score distributions. In G. Engelhard & M.. Wilson (Ed.), Objective measurement: Theory into practice (Vol.3). Norwood, New Jersey: Ablex Publishing Company.

Votaw, D.F. (1952). Methods of solving some personnel classification problems. Psychometrika, 17, 255-266.

Yen, W.M. (1983). Use of the three-parameter model in the development of standardized achievement tests. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Table 1

Values of μ and σ for selected values of $c_1=c_2$ for Model I and Model II

$c_1=c_2$	Model I			Model II		
	μ	σ	$\mu+\sigma$	μ	σ	$\mu+\sigma$
0.0	1.551	0.833	2.384	1.586	0.922	2.508
0.1	1.551	0.833	2.384	1.586	0.922	2.508
0.2	1.551	0.833	2.384	1.586	0.922	2.508
0.3	1.551	0.833	2.384	1.586	0.922	2.508
0.4	1.551	0.833	2.384	1.562	0.833	2.395
0.5	1.386	0.505	1.891	1.387	0.508	1.895
0.6	1.311	0.373	1.684	1.335	0.384	1.719
0.7	1.286	0.369	1.655	1.277	0.363	1.640
0.8	1.180	0.313	1.493	1.189	0.322	1.511
0.9	1.057	0.295	1.352	1.085	0.294	1.379
1.0	1.037	0.281	1.318	1.104	0.294	1.398
1.1	1.169	0.320	1.489	1.231	0.325	1.556
1.2	1.500	0.437	1.937	1.479	0.410	1.889
1.3	1.826	0.486	2.313	1.907	0.509	2.416
1.4		Inf			Inf	

Note. "Inf" means no feasible solution available.

Table 2

Values of μ_1 and μ_2 for c_1=0.2 and c_1=1.2 (Model I)

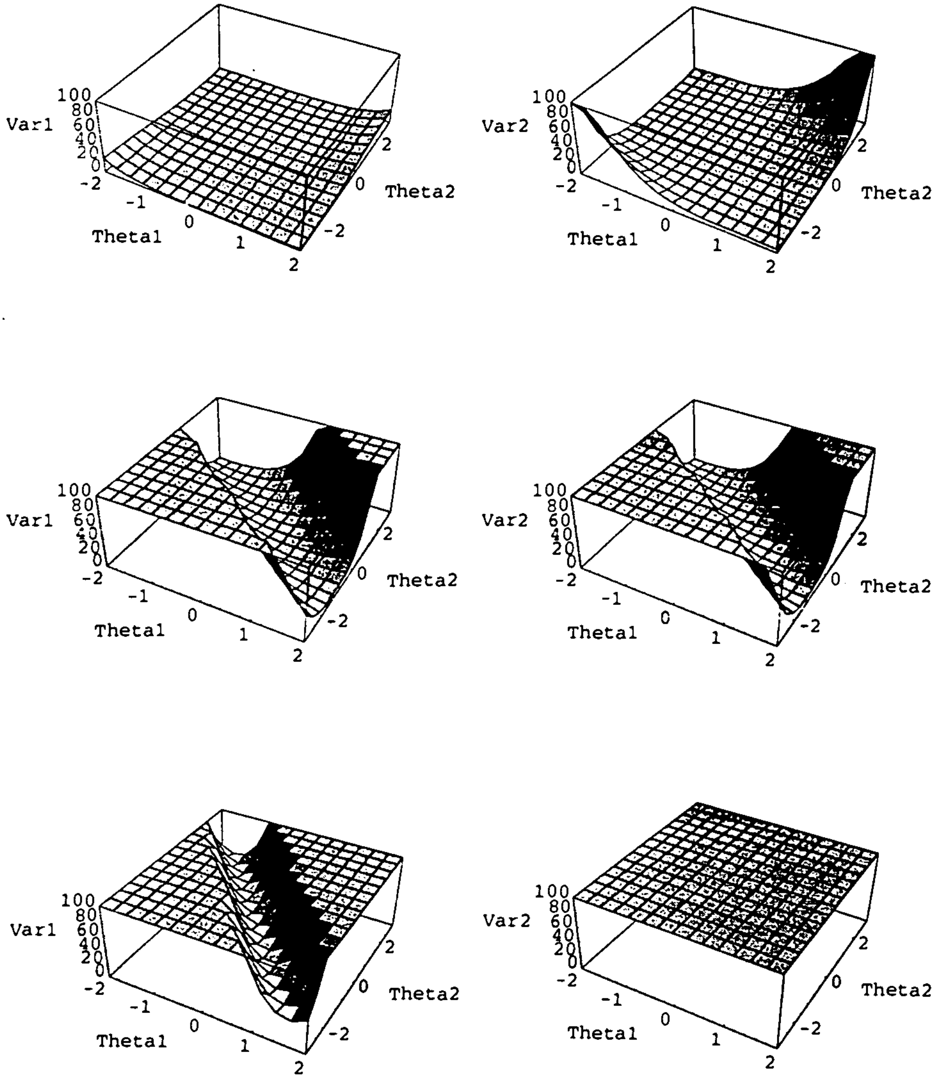
c_2	c_1=0.2				c_1=1.2			
	μ_1	σ_1	μ_2	σ_2	μ_1	σ_1	μ_2	σ_2
0.0	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.1	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.2	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.3	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.4	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.5	2.029	0.893	1.073	0.038	1.296	0.265	3.472	0.516
0.6	2.029	0.893	1.073	0.038	1.235	0.230	3.085	0.373
0.7	2.029	0.893	1.073	0.038	0.974	0.102	2.069	0.123
0.8	2.029	0.893	1.073	0.038	1.010	0.110	1.698	0.081
0.9	2.412	1.624	1.014	0.042	0.923	0.096	1.456	0.063
1.0	2.663	2.330	0.949	0.042	0.981	0.107	1.441	0.066
1.1	7.123	2.449	1.426	0.062	1.212	0.166	1.441	0.091
1.2	12.819	1.650	2.036	0.123	1.428	0.232	1.572	0.141
1.3	11.606	2.461	1.951	0.140	1.564	0.276	1.533	0.146
1.4	8.520	1.769	1.661	0.088	2.230	0.373	1.889	0.243
1.5	7.193	1.254	1.643	0.107		Inf		
1.6	4.554	1.047	1.470	0.121		Inf		
1.7		Inf				Inf		

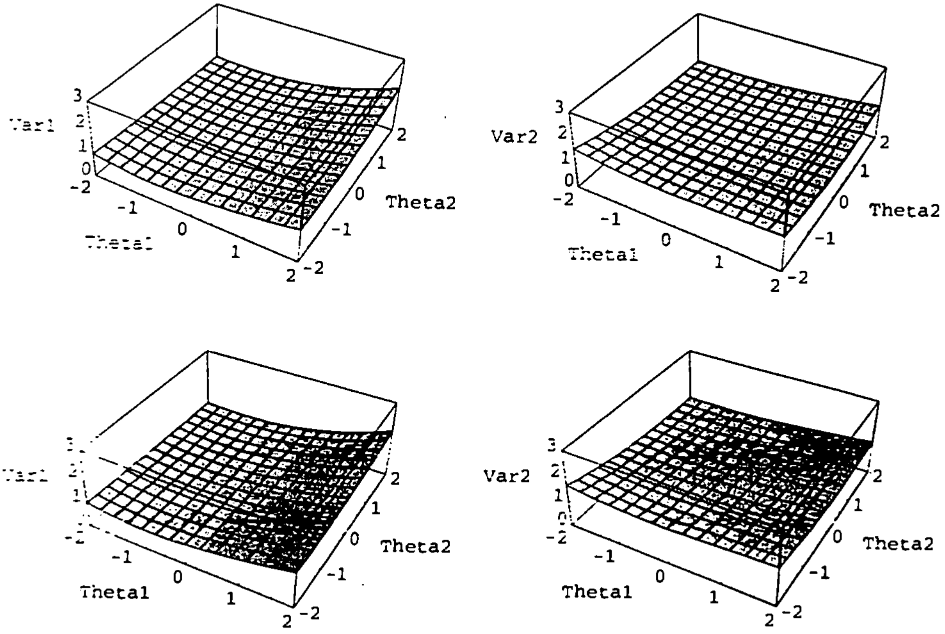
Note. "Inf" means no feasible solution present.

Figure Captions

Figure 1. Three examples of variance functions for $\hat{\theta}_1$ and $\hat{\theta}_2$ (for the values of the item parameters, see the text).

Figure 2. Variance functions for the tests assembled under Model I (upper panel) and Model II (lower panel).


100
80
60
40
20
0
Var1
-2
-1
0
1
2
Theta1
2
0
-2
Theta2
Var2
Var1
Var2
Var1
Var2


Var1
3
2
1
0
-2
-1
0
1
2
Theta1
2
1
0
-1
-2
Theta2
Var2
3
2
1
0
-2
-1
0
1
2
Theta1
2
1
0
-1
-2
Theta2
Var1
Theta1
Theta2
Var2
Theta1
Theta2

Author Note

This paper was prepared for T.A. Ackerman (Ed.), Multidimensional IRT, Applied Psychological Measurement [Special Issue]. The author is indebted to Wim M.M. Tielen for his computational support and to Terry A. Ackerman for the data set used in the empirical example.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis.
University of Twente, Enschede,
The Netherlands.

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score pattems: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score pattems on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede,
The Netherlands.

department of

EDUCATION

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands