Psychological Test and Assessment Modeling, Volume 55, 2013 (1), 105-123

The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality

Nicki-Nils Seitz[1] *& Andreas Frey*[2]

Abstract

It is examined whether the unidimensional Sequential Probability Ratio Test (SPRT) can be productively combined with multidimensional adaptive testing (MAT). With a simulation study, it is investigated whether this combination results in more accurate simultaneous classifications on two or three dimensions compared to several instances of unidimensional adaptive testing (UCAT) in combination with SPRT. The number of cut scores, and the correlation between the dimensions measured were varied. The average test length was mainly influenced by the number of cut scores (one, four) and the adaptive algorithm (MAT, UCAT). With MAT, a lower average test length was achieved in comparison to the UCAT. It is concluded that MAT will result in a higher percentage of correct classifications than UCAT when more than two dimensions are measured.

Key words: classification, computerized adaptive testing, item response theory, multidimensional adaptive testing, sequential probability ratio test

---

[1] *Correspondence concerning this article should be addressed to:* Nicki-Nils Seitz, Institute of Educational Science, Department of Research Methods in Education, Friedrich-Schiller-University Jena, Am Planetarium 4, 07737 Jena, Germany; email: nicki-nils.seitz@uni-jena.de

[2] Institute of Educational Science, Department of Research Methods in Education, Friedrich-Schiller-University Jena, Germany

Multidimensional adaptive testing (MAT) is a special approach to the assessment of two or more latent abilities in which the selection of the test items presented to the examinee is based on the responses given by the examinee to previously administered items (e.g., Frey & Seitz, 2009). The main advantage of MAT is its capacity to substantially increase measurement efficiency compared to sequential testing or unidimensional computerized adaptive testing (UCAT). Most of the studies on MAT are focusing its application for assessing individual abilities located on continuous scales. Currently, only very little is known about the capabilities of MAT regarding the classification of test takers to one of several ability categories (e.g., pass vs. fail). To fill in this gap, the present paper focuses on the combination of MAT with the sequential probability ratio test (SPRT; e.g., Kingsbury & Weiss, 1983; Reckase, 1983). The SPRT is a classification method that already has been used successfully in combination with UCAT (e.g., Eggen, 1999; Eggen & Straetmans, 2000; Spray & Reckase, 1996; Thompson, 2007b).

Regarding MAT, Spray, Abdel-fattah, Huang, and Lau (1997) made an attempt to modify the SPRT in order to use it with MAT based on items with within-item multidimensionality. Items with within-item multidimensionality are allowed to measure more than one dimension simultaneously (Wang, Wilson, & Adams, 1997). Dealing with within-item multidimensionality, the multidimensional item response theory (IRT) model used with MAT is a compensatory model (e.g., Reckase, 2009). With such an IRT-model, the linear combination of the abilities measured leads to a curvilinear function. Therefore, the test statistic of the SPRT, which is a likelihood ratio test, cannot be updated by two unique values required by the SPRT. For details, see Spray et al. (1997). Considering multidimensional pass-fail tests, Spray and colleagues did not find a satisfactory solution for implementing a multidimensional SPRT into such a MAT.

Nevertheless, from a practical point of view, tests entailing items measuring exactly one dimension each (between-item multidimensionality) are much more common than tests based on an item pool with within-item multidimensionality. Hence, the present paper focusses on the combination of MAT and SPRT for items with between-item multidimensionality. Note that when the MAT approach of Segall (1996) is used for items with between-item multidimensionality, information from items which measure one dimension is used as information about the person's score on other dimensions. This is done by incorporating assumption about the multivariate ability distribution in terms of correlations between the measured dimensions. Several studies showed that using this information results in substantial increase in measurement efficiency compared to using several unidimensional adaptive tests (e.g., Frey & Seitz, 2010; Luecht, 1996). Moreover, Wang and Chen (2004) showed that the measurement efficiency of MAT increases with increasing correlations and also with an increasing number of dimensions. Thus, in tests measuring several abilities, assumed to be moderately or highly correlated, MAT will generally outperform UCAT when the derivation of ability scores on continuous scales is the aim.

Nevertheless, not much is known about potential benefits of MAT compared to UCAT when classifications should be made. The results derived from the studies focusing on the estimation of abilities on continuous scales cannot directly be transferred to using MAT for classification purposes, because the SPRT is not based on the provisional abil-

ity estimates which are estimated in the process of adaptive testing. Thus, a potential advantage of MAT compared to UCAT can only be caused by using the correlation between the measured dimensions for item selection. Therefore, it is unclear whether and, if so, how strongly the number of dimensions measured and the magnitude of the correlations influences the test performance of the SPRT in terms of the classification efficiency.

Hence, the purpose of this article is threefold:

1. To demonstrate how the SPRT approach can be used in combination with MAT with between-item multidimensionality.
2. To investigate potential gains in the classification efficiency of SPRT in combination with MAT compared to several unidimensional adaptive tests, each with unidimensional SPRTs.
3. To investigate potential gains in the classification efficiency of SPRT in combination with MAT with an increasing number of measured dimensions and an increasing magnitude in the correlation between dimensions.

The article is organized as follows: Firstly, an overview of the main ideas of the MAT approach of Segall (1996) is given. Secondly, the original SPRT is introduced, the extended SPRT approach of Armitage (1950) is presented, and the combination of the latter with MAT is formally described. Thirdly, the methods of the present simulation study are presented. Then, the results are reported and finally discussed.

Multidimensional adaptive testing

For MAT, Segall (1996) proposed a multidimensional Bayesian ability estimation method and a multidimensional Bayesian item selection method. In general, Bayesian methods consider the underlying parameters, for example, in adaptive testing the ability vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ for P dimensions, as random variables with a known distribution, that is, the prior distribution. Thereby, previous knowledge or assumptions about the ability distribution can be considered (Bernardo & Smith, 1994). The characteristics of the prior distribution are often based on knowledge stemming from empirical studies carried out with the same instruments in the past. Segall (1996) proposed to assume that the abilities follow a multivariate normal distribution, $\boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Phi})$, with the mean vector $\boldsymbol{\mu}$, the variance-covariance matrix $\boldsymbol{\Phi}$, and the probability density function $f(\boldsymbol{\theta}) = (2\pi)^{-P/2} \, |\boldsymbol{\Phi}|^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Phi}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right]$ as prior distribution.

Based on the Bayes Theorem, the ability vector $\boldsymbol{\theta}$ is estimated using the posterior density function $f(\boldsymbol{\theta}|\mathbf{u}) = L(\mathbf{u}|\boldsymbol{\theta})\frac{f(\boldsymbol{\theta})}{f(\mathbf{u})}$ containing information from the responses $\mathbf{u} = (u_1, \ldots, u_t)$ given on *t* items. Here, $f(\mathbf{u})$ is the marginal density function of $\mathbf{u}$, and $L(\mathbf{u}|\boldsymbol{\theta}) = \prod_{i=1}^{t} P_i(\boldsymbol{\theta})^{u_i}\left(1 - P_i(\boldsymbol{\theta})\right)^{1-u_i}$ is the likelihood of the response pattern with $P_i(\boldsymbol{\theta})$, the probability of a correct response to an item *i* based on a multidimensional item response theory model (MIRT model).

In general, MIRT models define the interaction between the test taker's abilities and the item characteristics expressed in the person's response to an item. Frequently used MIRT models assume that the items are dichotomous, that is, they have two score categories (e.g., correct-incorrect answer; Reckase, 2009). The probability of a correct answer $u_i = 1$ to a dichotomous item *i* conditional on the underlying ability vector $\boldsymbol{\theta}$ is frequently expressed by the multidimensional three-parameter logistic model (M3PL model). For the M3PL, Segall (1996) introduced:

$$P_i(\boldsymbol{\theta}) = P_i(u_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp\left[-\mathbf{a}_i'(\boldsymbol{\theta} - b_i \cdot \mathbf{1})\right]}. \quad (1)$$

The M3PL model in *Equation 1* includes three types of item parameter, the discrimination vector $\mathbf{a}_i' = (a_{i1} \ldots a_{iP})$, the item difficulty b_i, and the pseudo guessing parameter c_i. The **1** represents a $P \times 1$ vector filled with 1s. Thus, the same item difficulty is used for all dimensions measured. The multidimensional two-parameter logistic model (M2PL model) does not account for a pseudo guessing parameter ($c_i = 0$). A non-zero entry in the discrimination vector $\mathbf{a}_i'$ indicates which dimension is measured with the item (i.e., the *item loading*).

As the Bayesian point estimate of $\boldsymbol{\theta}$, either the mean or the mode of the posterior distribution are commonly used. Because of its easier calculations, Segall (1996) proposed to use the Bayesian modal estimator, that is, the maximum a-posteriori (MAP), which is calculated from the maximum of the log posterior density function $\ln f(\boldsymbol{\theta}|\mathbf{u})$. Since no closed form solution is given for obtaining the maximum of the log posterior density function, numerical estimation methods like the Newton-Raphson method have to be used instead (see Segall, 1996).

Finding the maximum of the posterior density function also depends on the variance-covariance matrix $\boldsymbol{\Phi}$ of the prior density function. In the variance-covariance matrix, the non-diagonal elements can differ from zero, that is, the covariances, which can be transformed into correlations. Using non-zero covariances or correlations for item selection and ability estimation generally increases measurement efficiency (e.g., Segall, 1996; Wang & Chen, 2004). Nevertheless, Reckase (2009) pointed out that usage of the Bayesian MAT approach as compared with the Maximum Likelihood MAT approach may

result in poorer estimates when the correlations between the measured dimensions are high and the abilities do not fit to the prior density function. Similar findings were reported by Diao (2009). Diao found that the ability estimates were pulled towards the mean of the prior distribution. However, these effects are only of a relevant magnitude for short tests. For longer tests, the information derived from the responses given becomes much more important and regressions to the centroid of the prior distribution are minimized.

The item selection method of Segall (1996) is based on a maximum information criterion which chooses the item from the item pool with the maximum of the determinant

$$\left|\mathbf{W}\right| = \left|\mathbf{I}\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}\right) + \mathbf{I}\left(\boldsymbol{\theta},u_{i^*}\right) + \boldsymbol{\Phi}^{-1}\right| . \tag{2}$$

Here, $\mathbf{I}\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}\right)$ represents the information matrix of the items already administered, $\mathbf{I}\left(\boldsymbol{\theta},u_{i^*}\right)$ is the information matrix of the candidate item *i**. Segall (1996) showed that when the item *i** provides the maximum value of *Equation 2*, this leads to the largest decrement in the volume of the Bayesian credibility ellipsoid. Thus, administering this item provides the largest possible increase in the precision of the multidimensional ability estimate. The optimizing criterion *Equation 2* is sometimes also referred to as a D-optimality criterion (i.e., D for determinant; Atkinson & Donev, 1992).

Sequential probability ratio test

The SPRT (Wald, 1945, 1947) is based on the principle of hypotheses testing. Reckase (1983) modified the SPRT approach of Wald in order to classify individuals to one of two categories which is also referred to, as mastery testing. For mastery testing, two hypotheses are tested against each other. Each hypothesis postulates the test taker's membership in one of two mutually exclusive ordinal categories (e.g., pass or fail). Based on a given set of responses, it is tested whether the test taker can be classified into one of these two categories with an acceptable probability of misclassification (i.e., α - and β -error), or whether more responses are needed to reach an acceptable error rate.

The categories are derived by setting cut scores on an underlying continuous ability scale. Besides the pass-fail situation with only one cut score, more than two categories, separated by more than two cut scores, can also be used (e.g., Spray, 1993). For each category, a hypothesis which postulates the test taker's membership in this category is formulated. In general, each hypothesis can be tested against each of the remaining hypotheses. With an increasing number of cut scores and categories, the number of possible pairs of hypotheses increases. Since a single hypothesis test only compares two hypotheses with each other, using several cut scores requires testing a series of hypotheses (Spray, 1993). Therefore, an overall test decision is needed. If an overall test decision is only based on the decision of one hypothesis test, contradictory test decisions from the other hypothesis tests may be disregarded (Wetherill & Glazebrook, 1986). To overcome this problem, all hypothesis tests need a clear decision (Sobel & Wald, 1949). For this

purpose, the SPRT approaches either of Sobel and Wald (1949) or of Armitage (1950) can be applied.

However, when more than three hypotheses are considered, the Sobel-Wald approach may also not be able to lead to a clear decision (Ghosh, 1970). Armitage (1950) described a solution to this problem. In general, the Armitage approach is applicable for any number of hypotheses. The hypotheses need to be exclusive and ordered. In contrast to the Sobel-Wald approach, with the Armitage approach, every hypothesis is tested against the remaining hypotheses (Govindarajulu, 1981). Additionally, the overall test only stops if all hypothesis tests stop simultaneously. Since the Armitage approach can be used for any number of cut scores, in the following section(s), only this more general approach is further elaborated upon.

The SPRT approach of Armitage

Before classifying with the SPRT, K cut scores $\theta_k \left(k = 1,\ldots,K\right)$ have to be located on a continuous ability scale, resulting in $K+1$ categories. Furthermore, indifference regions (IRs) defined by an lower bound $\theta_{\mathrm{L},k}$ and an upper bound $\theta_{\mathrm{U},k}$ are set around each cut score. The IR defines the region where no decision can be made, owing to measurement fallibility (Eggen & Straetmans, 2000). These bounds are chosen based on experience, that is, based on explorative studies (Thompson, 2007b). Often, symmetric bounds are chosen for all cut scores with a fixed distance $\delta > 0$ representing the width of the IR, so that $\theta_{\mathrm{L},k} = \theta_k - \delta$ and $\theta_{\mathrm{U},k} = \theta_k + \delta$.

With $K+1$ categories, $K+1$ hypotheses are considered. These lead to an overall of $\binom{K+1}{2} = \frac{1}{2} \cdot K \cdot (K+1)$ different pairs of hypotheses. Each pair of hypotheses consists of the hypotheses $H_{\mathrm{m}}: \theta \le \theta_{\mathrm{L},m}$ versus $H_n: \theta \ge \theta_{\mathrm{U},n-1},\ m < n \in \{1,\ldots,K+1\}$. The $\frac{1}{2} \cdot K \cdot (K+1)$ different pairs of hypotheses can be expressed in K independent test statistics, that is, a likelihood ratio (Armitage, 1950). If all K tests concerning hypothesis H_m accept hypothesis H_m, the overall test accepts hypothesis H_m (Ghosh, 1970).

In order to further clarify the procedure, let us consider an example: Setting $K = 3$ cut scores leads to four categories and, therefore, to four hypotheses ($H_1: \theta \le \theta_1$, $H_2: \theta_1 < \theta \le \theta_2$, $H_3: \theta_2 < \theta \le \theta_3$, and $H_4: \theta_3 < \theta$). Each hypothesis postulates that the test taker belongs to one of the four categories; hypothesis H_1, for example, postulates that the test taker belongs to category one. Overall, $\binom{3+1}{2} = \frac{1}{2} \cdot 3 \cdot (3+1) = 6$ different statistical tests need to be carried out. In order to decide which category the test taker belongs to, the decisions from only three tests are needed. The overall test accepts cate-

gory two, for example, if all three tests accept H_2 ($H_1: \theta \le \theta_{\mathrm{L},1}$ versus $H_2: \theta \ge \theta_{\mathrm{U},1}$, $H_2: \theta \le \theta_{\mathrm{L},2}$ versus $H_3: \theta \ge \theta_{\mathrm{U},2}$, $H_2: \theta \le \theta_{\mathrm{L},2}$ versus $H_4: \theta \ge \theta_{\mathrm{U},3}$).

The test statistic used by the SPRT is a likelihood ratio criterion. The likelihood ratio (LR) for the test between the two hypotheses $H_m: \theta \le \theta_{\mathrm{L},m}$ versus $H_n: \theta \ge \theta_{\mathrm{U},n-1}$, $m < n \in \{1,\ldots,K+1\}$ is calculated by

$$\mathrm{LR}_{m,n} = \frac{L\left(\theta_{\mathrm{U},n-1} \mid \mathbf{u}\right)}{L\left(\theta_{\mathrm{L},m} \mid \mathbf{u}\right)} = \frac{\prod_{i=1}^{t} P_i\left(\theta_{\mathrm{U},n-1}\right)^{u_i}\left(1-P_i\left(\theta_{\mathrm{U},n-1}\right)\right)^{1-u_i}}{\prod_{i=1}^{t} P_i\left(\theta_{\mathrm{L},m}\right)^{u_i}\left(1-P_i\left(\theta_{\mathrm{L},m}\right)\right)^{1-u_i}}, \tag{3}$$

with the likelihood of the upper bound $L\left(\theta_{\mathrm{U},n-1} \mid \mathbf{u}\right)$ of the cut score being θ_{n-1} and the likelihood of the lower bound $L\left(\theta_{\mathrm{L},m} \mid \mathbf{u}\right)$ of the cut score being θ_m, given the response pattern $\mathbf{u} = \left(u_1,\ldots,u_t\right)$ to *t* administered items. This presented unidimensional SPRT uses unidimensional item response theory models, like the unidimensional one-parameter logistic model (1PL model) $P_i(\theta) = \left[1+\exp\left(-\left(\theta - b_i\right)\right)\right]^{-1}$ (e.g., Hambleton & Swaminathan, 1984) based on the IR-bounds $\theta = \theta_{\mathrm{U},n-1}$ or $\theta = \theta_{\mathrm{L},m}$.

To come to a decision, the $\mathrm{LR}_{m,n}$ is compared to the values A and B. These values are approximated with the nominal error α and β, for example, with $\alpha = \beta = .05$ indicating a 5% error level, as $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ (Wald, 1947). After administering an item, a decision can be made about whether the test taker can be classified with acceptable error rates or whether the test needs to be continued: If $\mathrm{LR}_{m,n} > A$, the test classifies the test taker into the category above the cut score θ_{n-1}, or if $\mathrm{LR}_{m,n} < B$, the test classifies the test taker into the category below the cut score θ_m, but if $B \le \mathrm{LR}_{m,n} \le \mathrm{A}$, no decision can be made and another item is administered.

Using SPRT for MAT

In multidimensional adaptive testing based on items with between-item multidimensionality, each item loads on exactly one dimension. For example, in a test measuring three dimensions and allowing only 0s and 1s in the discrimination vector, $\mathbf{a}_i' = (0\,0\,1)$ represents an item *i* loading on the third dimension only. Considering the M2PL based on *Equation 1* (discarding the c_i-parameter) leads in the exponent to $\mathbf{a}_i'\left(\theta - b_i \cdot \mathbf{1}\right) = (0\ 0\ 1)\left(\theta_1\ \theta_2\ \theta_3\right)' - (0\ 0\ 1)\left(b_i\ b_i\ b_i\right)' = \theta_3 - b_i$. Thus, and as described by Wang and Chen (2004), in the case of between-item multidimensionality, the M2PL is equivalent to using a unidimensional 2PL for each of the measured dimensions.

Based on this consideration, the SPRT can be implemented in MAT with between-item multidimensionality as follows: In general, the number and the location of the cut scores can be chosen differently for each dimension. To simplify matters, in this case, the same number and location of cut scores is set for each dimension. For *K* cut scores on each of the *P* dimensions, a cut score is identified by $\theta_k^{(p)}, \left(k = 1,\ldots,K; p = 1, \ldots, P\right)$.

Then, for a fixed width of the IR $\left(\delta > 0\right)$, the upper bounds are identified by $\theta_{\mathrm{U},k}^{(p)} = \theta_k^{(p)} + \delta$ and the lower bounds by $\theta_{\mathrm{L},k}^{(p)} = \theta_k^{(p)} - \delta$. The multidimensional LR of dimension *p* is calculated with the ability vector for the upper bound $\boldsymbol{\theta}_{\mathrm{U},k}^{(p)}$ and the ability vector for the lower bound $\boldsymbol{\theta}_{\mathrm{L},k}^{(p)}$ with an entry on the position *p* of $\theta_{\mathrm{U},k}^{(p)}$ and $\theta_{\mathrm{L},k}^{(p)}$, respectively. Since no decision is needed on the other *P*-1 dimensions, the *P*-1 entries in the vectors $\boldsymbol{\theta}_{\mathrm{U},k}^{(p)}$ and $\boldsymbol{\theta}_{\mathrm{L},k}^{(p)}$ can be set to the provisional ability estimates $\hat{\theta}_1,\ldots,\hat{\theta}_{p-1},\hat{\theta}_{p+1},\ldots,\hat{\theta}_P$.

Following *Equation 3* and the Armitage approach for testing two hypotheses $\mathrm{H}_m^{(p)}: \boldsymbol{\theta}^{(p)} \le \boldsymbol{\theta}_{\mathrm{L},m}^{(p)}$ versus $\mathrm{H}_n^{(p)}: \boldsymbol{\theta}^{(p)} \ge \boldsymbol{\theta}_{U,n-1}^{(p)}$, $m < n \in \left\{1,\ldots,K+1\right\}$ against each other, the likelihood ratio for the cut scores on dimension *p* is calculated as

$$\mathrm{LR}_{m,n}^{(p)} = \frac{L\left(\boldsymbol{\theta}_{\mathrm{U},n-1}^{(p)} \mid \mathbf{u}\right)}{L\left(\boldsymbol{\theta}_{\mathrm{L},m}^{(p)} \mid \mathbf{u}\right)} = \frac{\prod_{i=1}^{\mathrm{t}} P_i\left(\boldsymbol{\theta}_{\mathrm{U},n-1}^{(p)}\right)^{u_i}\left(1 - P_i\left(\boldsymbol{\theta}_{\mathrm{U},n-1}^{(p)}\right)\right)^{1-u_i}}{\prod_{i=1}^{\mathrm{t}} P_i\left(\boldsymbol{\theta}_{\mathrm{L},m}^{(p)}\right)^{u_i}\left(1 - P_i\left(\boldsymbol{\theta}_{\mathrm{L},m}^{(p)}\right)\right)^{1-u_i}} \tag{4}$$

Since MAT with between-item multidimensionality only considers items measuring one dimension and the MIRT model is reduced to a unidimensional IRT-model, *Equation 4* simplifies to

$$\mathrm{LR}_{m,n}^{(p)} = \frac{L\left(\theta_{\mathrm{U},n-1}^{(p)} \mid \mathbf{u}_p\right)}{L\left(\theta_{L,m}^{(p)} \mid \mathbf{u}_p\right)} = \frac{\prod_{i=1}^{t_p} P_i\left(\theta_{\mathrm{U},n-1}^{(p)}\right)^{u_{i,p}}\left(1 - P_i\left(\theta_{\mathrm{U},n-1}^{(p)}\right)\right)^{1-u_{i,p}}}{\prod_{i=1}^{t_p} P_i\left(\theta_{\mathrm{L},m}^{(p)}\right)^{u_{i,p}}\left(1 - P_i\left(\theta_{\mathrm{L},m}^{(p)}\right)\right)^{1-u_{i,p}}} \tag{5}$$

Here, $\mathbf{u}_p = \left(u_{1,p}, \ldots, u_{t_p,p}\right)$ represents the responses to the t_p administered items loading on dimension *p*. Note that in *Equation 4* and *Equation 5* the provisional ability estimate is not included.

Method

Simulation study

To compare the performance of the two test algorithms, a MAT with multiple-unidimensional SPRTs and several UCATs, each with a unidimensional SPRT, a simula-

tion study was conducted. For this purpose, the combination of the SPRT approach of Armitage (1950) and the MAT approach of Segall (1996) was examined using SAS 9.3.

Design

Four independent variables were varied: (1) the adaptive algorithm, (2) the number of dimensions, (3) the correlation between the dimensions and (4) the number of cut scores.

For the comparison between several unidimensional and multidimensional tests, UCAT and MAT were considered as the *adaptive algorithm*. For both test algorithms, the items were assumed to load on exactly one dimension each, thus resulting in between-item multidimensionality. Since in UCAT items can only load on one dimension, this allows for unequivocal comparisons between MAT and UCAT, which would not be possible if within-item multidimensionality would have been used. Since the impact of the *number of dimensions* on the differences between MAT and UCAT with SPRT regarding their performance is not known so far, tests comprising two dimensions (2D) or three dimensions (3D) were run. For UCAT, three independent unidimensional adaptive tests, each with a unidimensional SPRT, were simulated. Then, two or three unidimensional adaptive tests were combined for tests measuring two or three dimensions, respectively. The MAT condition was based on the M2PL model and considered two or three dimensions. For MAT, the correlations between the dimensions were either assumed to be $\rho = .00$ or $\rho = .85$. To recreate tests that classify persons on power achievement levels, the tests had either one cut score ($\theta_1^{(p)} = 0.000$), or four cut scores ($\theta_1^{(p)} = -1.035$, $\theta_2^{(p)} = -0.115$, $\theta_3^{(p)} = 0.805$, and $\theta_4^{(p)} = 1.725$) on each dimension with $p \in \{1,2,3\}$. The levels of the correlations, as well as the number and the values of the cut scores, were chosen to represent a broad range of test situations that can be found in empirical studies. For the width of the IRs, δ was set at 0.3, and the nominal errors of the SPRTs were fixed at $\alpha = \beta = .05$. These values are frequently used for SPRT.

Data generation

Three *person parameter* populations were generated, each assumed to be normally distributed. The first person parameter population was generated for the UCAT condition. A total of 13,834 standard normal distributed parameters were generated for each of the three dimensions. The second and third person parameter populations were generated for the MAT condition. They were either assumed to be two-dimensional (2D) or three-dimensional (3D). For the MAT condition, the person parameters were also assumed to be multivariate normally distributed with a correlation of $\rho = .85$ between the dimensions. Using the Cholesky decomposition, two or three of the standard normal distributed parameter vectors from the UCAT condition were transformed into 2D or 3D, respectively, multivariate normally distributed parameters considering a correlation of $\rho = .85$.

The *item pool* consisted of 300 dichotomous items for each dimension, with each item measuring one dimension only. The difficulties of the items were assumed to be standard normally distributed, $b_i \sim \mathrm{N}(0,1)$. Thus, sufficient items were available where most simulees were located on the ability scale, that is, near the mean. The *responses* of the simulees to the dichotomous items were either based the 1PL model (UCAT) or a M2PL model (MAT).

Procedure

The person parameters, the item parameters and the responses were used to simulate tests under the conditions specified by the research design. The first item was selected randomly from the item pool. Beginning with the second item, the next item was chosen in order to maximize *Equation 2*. Abilities were estimated with the MAP estimator. For each test condition, ten replications were simulated. Overall, the results were based on a total of 138,240 tests. The test was terminated when either the simulees could be classified on the one dimension (UCAT) or on all the dimensions (MAT) being measured, or a total of 30 items for each dimension had been administered. This led to a maximum test length of 60 items (2D) or 90 items (3D). If one or more dimensions of an examinee were located near a cut score, the maximum test length of 30 items for one dimension was reached before a classification could be made. In order to classify a simulee in such a case, the best possible decision had to be made based on the information at hand. If 30 items for a dimension were administered, the simulee was classified according to the decision: if $\mathrm{LR}_{m,n}^{(p)} > 1$ the test accepts $\mathrm{H}_n^{(p)}$, or if $\mathrm{LR}_{m,n}^{(p)} \leq 1$ the test accepts $\mathrm{H}_m^{(p)}$ (Spray & Reckase, 1997). Spray and Rekase refer to such a decision as a *truncated SPRT*.

Dependent variables

In order to evaluate the goodness of classification with variable test length, the average test length (*ATL*) and the percentage of correct classifications (*PCC*) were calculated as dependent variables.

The *ATL* was calculated as the mean number of items t_j needed for classification, averaged across *n* simulees and *R* replications , $ATL = \frac{1}{R \cdot n} \Sigma_{r=1}^{R} \Sigma_{j=1}^{n} t_j$.

The goodness of classification in terms of the *PCC* was calculated as the average percentage of correct classifications. A correct classification for a simulee *j* on all dimensions *P* was indicated with $m_j = 1$. The average percentage of correct classifications on all measured dimensions was calculated using $PCC = 100 \cdot \frac{1}{R \cdot n} \Sigma_{r=1}^{R} \Sigma_{j=1}^{n} m_j$.

Results

The items in the item pool covered item difficulties from Min = -3.14 to Max = 4.00 with $M = 0.04$ and $SD = 1.01$. The person parameters used in the MAT conditions covered the range from -3.77 to 3.70, and in the UCAT conditions from -3.77 to 3.98.

Average test length

Although, for each dimension, a minimum of items was not required, for each measured dimension, at least six items were administered in all tests. In the one cut score conditions, 25% (2D) or 15% (3D) of the simulees reached the maximum test length of 60 (2D) or 90 (3D) items, whereas in the four cut score conditions, 88% (2D) or 84% (3D) reached the maximum test length. The mean *ATL* is shown in Table 1.

In general, the *ATL* was influenced by the number of cut scores. This goes along with earlier findings for UCAT (e.g., Spray, 1993). The *ATL* increases if the number of cut scores increases due to the increasing number of examinees located near a cut score. For such examinees, more information is needed for a clear classification. For the one cut score condition, the mean *ATL* was about 46.71 items for a 2D-test and 70.03 items on average for a 3D-test. For the 2D-tests and the 3D-tests, this was approximately 78% of the maximum test length of 60 items or 90 items, respectively. In the four cut score condition, per dimensional about six items more were needed for a 2D-test and a 3D-test. This led to a mean *ATL* of 58.61 items for the 2D-tests and 87.89 items for the 3D-tests, corresponding to 98% of the maximum test length of 60 items or 90 items, respectively. Since the percentage of the maximum test length was comparable in the 2D-tests and the

Table 1:
Mean and Standard Deviation of Average Test Length by Number of Cut Scores, Algorithm, and Correlation

			2D		3D	
Number of cut scores	Algorithm	ρ	*M*	*SE*	*M*	*SE*
one[1]	UCAT	n.a.	47.20	0.06	70.87	0.07
	MAT	.00	46.44	0.06	69.51	0.08
	MAT	.85	46.47	0.07	69.62	0.10
four[2]	UCAT	n.a.	58.68	0.02	88.00	0.03
	MAT	.00	58.60	0.02	87.89	0.02
	MAT	.85	58.54	0.02	87.76	0.04

Note. $N = 10$, [1] cut score at $\theta_1^{(p)} = 0.000$, [2] cut scores at $\theta_1^{(p)} = -1.035$, $\theta_2^{(p)} = -0.115$, $\theta_3^{(p)} = 0.805$, $\theta_4^{(p)} = 1.725$, ρ = correlation between the dimensions, 2D = tests measuring two dimensions, 3D = tests measuring three dimensions.

3D-tests, the number of measured dimensions does not have a substantial influence on the *ATL*. Note that the results of each dimension were not reported in order to avoid redundancy of results. For a 2D-test or a 3D-test, the number of items for one dimension was approximately 50% or 33% of the *ATL*, respectively.

When considering the one cut score conditions, *ATL* showed to be significantly higher for UCAT compared to the two MAT conditions with either $\rho = .00$ or $\rho = .85$. Note that in this section, we speak of *significant* difference if the difference between two outcomes is larger than $1.96 \cdot SE$. In the four cut score conditions, only in the MAT condition with $\rho = .85$ a significantly lower *ATL* than for UCAT is observed. Comparing the two MAT conditions with each other, the magnitude of the correlation was not found to cause substantial differences in the *ATL*.

The results of the 2D-MATs with $\rho = .85$ are also shown in *Figure 1*. Since the contour plots for the MAT condition with no correlation or the UCAT conditions were similar to the contour plots in *Figure 1*, only contour plots for the 2D-MAT condition with $\rho = .85$ are shown. *Figure 1* shows the observed *ATL* and *PCC* distribution as a function of the (true) person parameters as contour plots. Since the person parameter distribution was assumed to be multivariate normal, the plot appears to be an ellipsoid.

The contour plots for the *ATL* are in the first row. Dark areas represent high *ATL*s whereas light areas represent low *ATL*s. The first contour plot represents MAT with one cut score. Examinees located near the cut score on either one or both dimensions showed the highest *ATL* whereas examinees located far away from the cut score showed *ATL*s lower than about 20 items. For the four cut score condition, the dark area has expanded since the cut scores cover the range from -1.035 to 1.725. An *ATL* between 20 and 30 items was also needed for examinees located far away from the cut scores.

Percentage of correct classification

When considering several dimensions and several cut scores with the approach of Armitage, not only one SPRT but several SPRTs are needed, that is, multiple testing. However, multiple testing leads to a lower overall goodness of classification than is derived with only one test. Assuming independent tests, the expected goodness of classification considering P dimensions and K cut scores is calculated by $(1-\alpha)^{P \bullet K}$ (e.g., van Belle, 2002). Note that the desired confidence in the interval can be reached through classical alpha correction (e.g., Bonferroni) with α^* calculated from, for example, $0.95 = (1-\alpha^*)^{P \bullet K}$ when a confidence of 95% is needed. The resulting α^* is generally smaller than α.

Since we used the nominal error rate of $\alpha = .05$, instead of alpha correction, for different conditions in the simulation study, different *PCC*s were expected. For the conditions including only one cut score, *PCC*s of 90.25% (2D-tests) or *PCC*s of 85.74% (3D-tests), and for the conditions with four cut scores, *PCC*s of 66.34% (2D-tests) or *PCC*s of 54.04% (3D-tests) were expected.

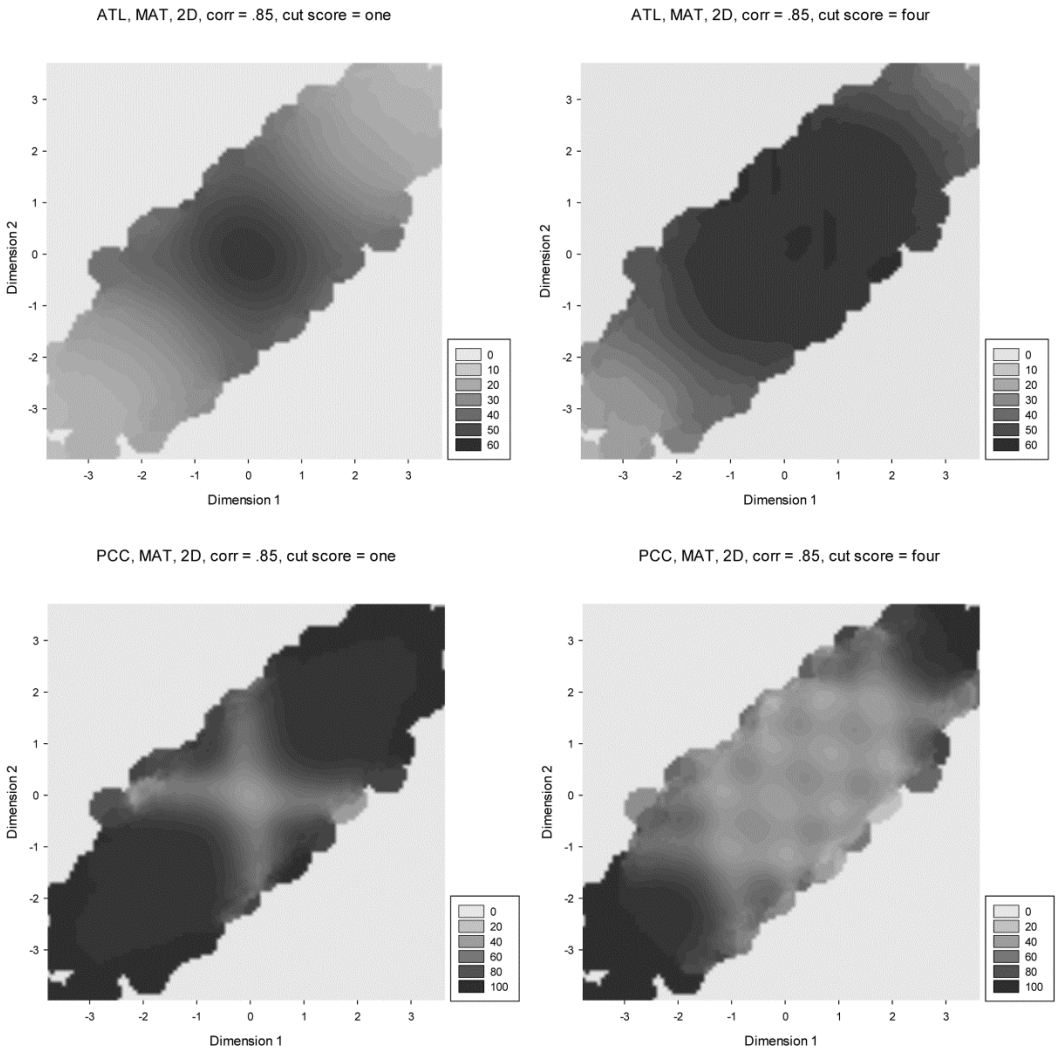

Figure 1:
Contour plots of *ATL* and *PCC* for the 2D-MATs ($\rho = .85$) with either one cut score ($\theta_1^{(p)} = 0.000$), or four cut scores ($\theta_1^{(p)} = -1.035$, $\theta_2^{(p)} = -0.115$, $\theta_3^{(p)} = 0.805$, $\theta_4^{(p)} = 1.725$) as a function of the true abilities (Dimension 1, Dimension 2).

Table 2:
Mean and Standard Error of Percentage of Correct Classification by Number of Cut Scores, Algorithm, and Correlation

			2D		3D	
Number of cut scores	Algorithm	ρ	*M*	*SE*	*M*	*SE*
one[1]	UCAT	n.a.	78.95	0.31	70.38	0.37
	MAT	.00	79.41	0.31	74.07	0.41
	MAT	.85	79.24	0.29	74.05	0.28
four[2]	UCAT	n.a.	52.24	0.33	41.48	0.15
	MAT	.00	49.80	0.28	44.79	0.55
	MAT	.85	49.80	0.43	45.03	0.24

Note. $N = 10$, [1] cut score at $\theta_1^{(p)} = 0.000$, [2] cut scores at $\theta_1^{(p)} = -1.035$, $\theta_2^{(p)} = -0.115$, $\theta_3^{(p)} = 0.805$, $\theta_4^{(p)} = 1.725$, ρ = correlation between the dimensions, 2D = tests measuring two dimensions, 3D = tests measuring three dimensions.

The mean *PCC* is shown in Table 2. In all test conditions, the *PCC* was significantly lower than the expected *PCC*. This might be due to the multivariate normal distribution that located most examinees near a cut score. This often led to a truncated SPRT and, therefore, to a higher probability of incorrect classifications. Note that in a previous simulation study, we assumed uniformly distributed person parameters instead of normally distributed person parameters. In such a case, the observed *PCC* resembled the expected *PCC* more closely. The results are not presented here because in empirical studies person parameters are mostly not assumed to be uniformly distributed.

For the one cut score conditions, the *PCC*s of the 2D-tests were in the same order of magnitude. However, when considering four cut scores, for MAT a substantially lower *PCC*s is observed compared to UCAT. But in the 3D-tests which considered either one or four cut scores, MAT always showed significantly higher *PCC*s than UCAT. For both MAT conditions, $\rho = .00$ and $\rho = .85$, the *PCC*s were in the same order of magnitude.

The contour plots for the *PCC*s are in the second row of *Figure 1*. Dark areas represent high *PCC*s whereas light areas represent low *PCC*s. In both contour plots, considering one or four cut scores, the low *PCC*s mirror the locations of the cut scores. Examinees located near a cut score show very low *PCC*s. This can be clearly seen in the one cut score condition. Here, most examinees are not located near the cut scores and, therefore, correctly classified up to 100%. However, when looking at the four cut score condition, the magnitudes of the person parameter distribution are located near the cut scores. This leads to a low *PCC* (< 40%) for most examinees.

Discussion

The purpose of the study was to evaluate the combination of multiple-unidimensional SPRTs and MAT in comparison with using a unidimensional SPRT in conjunction with several unidimensional adaptive tests. The study revealed small but interrelated effects on the *ATL* and *PCC*. Two main effects and one interaction effect were found.

The first main effect was that of the number of cut scores on the *ATL* and *PCC*. The more cut scores the test considered, the higher the *ATL* was, and the lower the *PCC* was. This can be explained by the fact that with an increasing number of cut scores, the number of test persons with true abilities located near a cut score increased. For these test persons, numerous items are needed for a clear test decision. This leads to long tests. Moreover, the more cut scores are considered, the lower the expected *PCC* is because of the increased probability of misclassifications with truncated SPRTs. This effect of the number of cut scores on *ATL* and *PCC* is inline with the findings of Spray (1993) for UCAT.

The second main effect was that of the used adaptive algorithm on the *ATL*. Overall, using MAT resulted in a lower *ATL* than using multiple unidimensional adaptive tests. Therefore, it can be concluded that classifying with MAT on several dimensions simultaneously leads to a decrease in test length compared to multiple unidimensional adaptive tests even though the differences are relatively small. Further reasearch should evaluate the practical significance.

The third effect was an interaction between the number of dimensions and the adaptive algorithm on *PCC*. The results were not as clear as for the main effects. In all the 3D-tests, MAT outperformed UCAT in terms of the *PCC* for the one cut score condition and the four cut score condition. Whereas, UCAT outperformed MAT in the 2D-tests with four cut scores. But in the one cut score condition, MAT and UCAT showed a comparable *PCC*. Overall, it is assumed that higher *PCC*s are obtained when dealing with more than two dimensions with MAT. Overall, MAT using an SPRT was superior to several unidimensional adaptive tests for most test conditions.

Thus, dealing with multivariate normal distributed abilities, the results indicate that MAT is advantageous in terms of a smaller *ATL* and a higher *PCC* compared to several UCATs. Although even for MAT, the *PCCs* were not as high as the expected *PCC*s, MAT might be a better choice instead of UCAT when individuals should be classified on more than two dimensions simultaneously.

In general, it can be stated that the number of items needed and the probability of a correct classification for an individual depends mainly on the location of the examinee relative to the locations of the cut scores. This was also shown in *Figure 1*. When the number of cut scores is increased, a correct classification achieved with a classification method like the SPRT goes along with high costs in terms of the number of items that need to be presented. However, for short tests with numerous cut scores, the probability of a high overall *PCC* is very low, even when MAT is used.

Nevertheless, the study has some limitations. Firstly, as described above, the proposed approach is a multiple-unidimensional SPRT. For every dimension, one unidimensional

SPRT was conducted with items each loading on one dimension only. Thereby, information from the items loading on other dimensions was ignored or was only indirectly used by the item selection for two or three dimensions simultaneously. However, since the likelihood ratio criterion cannot consider any information from the prior distribution, in MAT with between-item multidimensionality the SPRT will always be unidimensional. Calculating the likelihood, for example, one chooses the ability vector which consists of one bound of an IR for one dimension and the provisional ability estimation of the other dimensions. Since a likelihood function is calculated as the product of the probabilities of correct responses to the administered items, the probabilities based on the provisional ability estimations would all be the same for both likelihoods used in the likelihood criterion and, for that reason, could be disregarded. The challenge for further investigations is to find SPRT approaches for MAT with between-item multidimensionality which include information from all the items.

Secondly, for examinees located near a cut score, the number of items needed for a classification with acceptable probability of misclassification was substantially higher than for examinees located far away from a cut score. With test conditions comparable to this simulation study, Thompson (2009) found for SPRT striving for pass-fail decisions on one dimension, allowing a maximum test length of 200 items, an *ATL* of 49.11 items was needed to classify all examinees while testing. However, from a practitioner's point of view, the tests should be as short as possible in order to reducing test load. Hence, a maximum test length is needed. However, this leads to truncated SPRTs. When the maximum test length is reached, persons who cannot be classified with acceptable error rate and have thus be placed in the most likely category. Therefore, in the present study a maximum test length of 30 items for each measured dimension was chosen. This test length might be reasonable for many operational tests and is also sufficiently large to not lead to a high proportion of truncated tests. However, truncating SPRT will hardly ever be necessary in realistic test settings. Hence, different methods for truncated SPRTs should be developed and compared with each other. With UCAT, for example, the use of stochastic curtailment has been discussed as an alternative approach for ending a classification test (e.g., Finkelman, 2008; Wouden & Eggen, 2009). With stochastic curtailment, the test stops if additional items might not lead to improved classification with the SPRT.

Thirdly, we used estimation-based item selection. For the unidimensional SPRT, the question about which item selection method is best for the SPRT – estimation-based or cut score-based – has been intensively discussed. Cut score-based item selection choses items which have the highest information at the cut score and lead to better discrimination between the lower bound $\theta_{L,k}^{(p)}$ and the upper bound $\theta_{U,k}^{(p)}$ (e.g., Thompson, 2009). Dealing with several cut scores, the item with the highest information with respect to the nearest cut score is selected (e.g., Eggen, 1999). Eggen (1999) compared choosing the items according to maximum information at either the provisional ability estimate or the nearest cut score. Dealing with relatively short tests ($\leq$ 25 items, $\alpha = \beta = .05$), Eggen reported a difference of less than one item between the *ATL* for both item selection methods in favor of the estimation-based item selection. Thompson (2009) reported

results preferring cut score-based item selection dealing with pass-fail SPRTs and variable test length with a maximum test length of up to 200 items. However, the focus of the present study was on using the SPRT approach with the original approach of MAT (Segall, 1996), without modifying the original MAT approach, especially the item selection method. Therefore, we used an estimation-based item selection, that is, the item with the highest information with respect to the provisional ability estimation was administered. But as we used this item selection procedure for all test conditions, we had expected to find the same effects using cut score-based item selection, probably with shorter tests. Moreover, there are currently no studies that discuss item selection-based on cut scores in the multidimensional ability space. With the transition from unidimensional to multidimensional classification, for the choice of the nearest cut score, several criteria are available. With several cut scores for each dimension, the nearest cut score could be determined with a measure of distance, for example, the Euclidean distance or the Mahalanobis distance. Future investigations may examine other item selection methods for using the SPRT with MAT in order to identify possibilities for further improving its performance.

An interesting aspect of MAT as a classification instrument and an area of possible future research is its capability to use conjunctive or compensatory methods. This opens up the possibility of focusing on single dimensions or of gauging the performance of the test taker from a multidimensional viewpoint. Above all, tests combining the two classification methods could be developed. A test might, for example, measure four dimensions on continuous scales but classify test takers on only two overall dimensions based on two of the four dimensions. With such MAT classification methods, in combination with the item characteristics (e.g., between-item multidimensionality versus within-item multidimensionality) and the underlying MIRT model, classification with MAT is much more flexible than classification with UCAT. Taking also the findings of Spray et al. (1997) into account, this study does not complete the investigation of the use of the SPRT with MAT, but can be regarded as a further step towards enhancing the capability of classification with MAT.

Acknowledgement

This research was supported by grant FR 2552/2-3 from the German Research Foundation (DFG) in the Priority Programme “Models of Competencies for the Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes” (SPP 1293).

References

Armitage, P. (1950). Sequential Analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, B, 12*, 137-144.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum Experimental Design*. Oxford: Calerndon Press.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

Diao, Q. (2009). *Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing.* Unpublished doctoral dissertation. East Lansing: Michigan State University.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*, 713-734.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*, 442-463.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35,* 89-94. doi:10.1016/j.stueduc.2009.10.007

Frey, A., & Seitz, N. N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz [Multidimensional adaptive competence diagnostics. Results for measurement efficiency]. *Zeitschrift für Pädagogik, Beiheft 56*, 40-51.

Ghosh, B. K. (1970). *Sequential tests of statistical hypotheses.* Reading, MA: Addison-Wesley.

Govindarajulu, Z. (1981). *The sequential statistical analysis of hypothesis testing, point and interval estimation, and decision theory.* Columbus, OH: American Science Press, Inc.

Hambleton, R. K. & Swaminathan, H. (1984). *Item response theory: Principles and Applications*. Boston: Kluwer Nijhoff Publishing.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison if IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Reckase, M. D. (2009). *Multidimensional item response theory.* Dortrecht: Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Science, 20*, 502-522.

Spray, J. (1993). *Multiple-category classification using a sequential probability ratio test* (Research report 93-7). Iowa City IA: ACT, Inc.

Spray, J., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405-414.

Spray, J. A., Abdel-fattah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations of a computerized test when the item pool and latent space are multidimensional* (Research report 97-5). Iowa City IA: ACT, Inc.

Thompson, N. A. (2007a). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*, 1-13. Retrieved from http://pareonline.net/Home.htm

Thompson, N. A. (2007b). *A comparison of two methods of polytomous computerized classification testing for multiple cut scores*. Unpublished doctoral dissertation. Minnesota: University of Minnesota.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793. doi:10.1177/0013164408324460

van Belle, G. (2002). *Statistical rules of thumb*. New York, NY: John Wiley & Sons, Inc.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer* (2nd Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics, 16,*117-186.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316.

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139-155). Norwood, NJ: Ablex.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wetherill, G. B., & Glazebrook, K. D. (1986). *Sequential Methods in Statistics*. London: Chapman and Hall Ltd.

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In: D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.