

Chapter 14

Multidimensional Computerized Adaptive Testing for Classifying Examinees



Maaike M. van Groen, Theo J. H. M. Eggen and Bernard P. Veldkamp

Abstract Multidimensional computerized classification testing can be used when classification decisions are required for constructs that have a multidimensional structure. Here, two methods for making those decisions are included for two types of multidimensionality. In the case of between-item multidimensionality, each item is intended to measure just one dimension. In the case of within-item multidimensionality, items are intended to measure multiple or all dimensions. Wald's (1947) sequential probability ratio test and Kingsbury and Weiss (1979) confidence interval method can be applied to multidimensional classification testing. Three methods are included for selecting the items: random item selection, maximization at the current ability estimate, and the weighting method. The last method maximizes information based on a combination of the cutoff points weighted by their distance to the ability estimate. Two examples illustrate the use of the classification and item selection methods.

14.1 Introduction

Computerized classification tests, like computerized adaptive tests, adapt some test elements to the student. Test length and/or item selection is tailored during test administration. The goal of a computerized classification test is to classify the examinee into one of two levels (e.g., master/nonmaster) or into one of multiple levels (e.g., basic/proficient/advanced). Classification methods stop testing when enough confi-

M. M. van Groen (✉) · T. J. H. M. Eggen
Cito, Amsterdamseweg 13, 6814, CM, Arnhem, The Netherlands
e-mail: maaike.vangroen@cito.nl

T. J. H. M. Eggen
e-mail: theo.eggen@cito.nl

T. J. H. M. Eggen · B. P. Veldkamp
Department of Research Methodology, Measurement, and Data Analysis,
University of Twente, P.O. Box 217, 7500, AE Enschede, The Netherlands
e-mail: b.p.veldkamp@utwente.nl

© The Author(s) 2019
B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_14

dence has been established to make the decision. An item selection method selects items for the examinee on the fly.

Computerized classification testing was developed several decades ago for tests intended to measure one construct or dimension. Unidimensional item response theory (UIRT) is used to model the student's performance, to select the items, to decide whether testing can be stopped, and to make the classification decisions. More recently, classification methods were developed for tests measuring multiple constructs or dimensions [(i.e., multidimensionality; Seitz and Frey (2013a, b), Spray et al. (1997), Van Groen et al. (2014b, c, 2016))] using multidimensional item response theory (MIRT). Classification decisions can be made for the entire test or for one or more dimensions or subsets of items.

Two types of multidimensionality Wang and Chen (2004) are discussed in the next section. In the case of between-item multidimensionality, each item is intended to measure just one ability whereas the test as a whole measures multiple abilities. For example, a test contains items for mathematics and language, and each item measures either mathematics or language ability. In case of within-item multidimensionality, items are intended to measure multiple abilities. Here, items measure both mathematics and language. The relative contribution of language and mathematics must vary over items. If not, no distinction can be made between the dimensions and the test should be modeled using UIRT. In what follows, adaptations of two unidimensional classification methods to multidimensional classification testing are discussed. These methods are based on Wald's sequential probability ratio test SPRT; (1947/1973) and Kingsbury and Weiss' confidence interval method (1979). Subsequently, three item selection methods are discussed. The efficiency and accuracy of the classification and item selection methods are investigated in the simulations section. Limitations and recommendations for further research are discussed in the final section.

14.2 Multidimensional Item Response Theory

Multidimensional computerized classification testing requires a statistical framework to model the student's performance, to obtain item parameters, to select the items, to make the classification decisions, and to decide whether testing can be stopped before reaching the maximum test length. MIRT (Reckase 2009) provides such a framework. In a calibrated item bank, model fit is established, item parameter estimates are available, and items with undesired characteristics are removed (Van Groen et al. 2014a). During testing, it is assumed that the item parameters have been estimated with enough precision to consider them known (Veldkamp and Van der Linden 2002).

A vector θ of p person abilities is used in MIRT to describe the skills and knowledge required for answering an item (Reckase 2009). The dichotomous two-parameter logistic model is used here. The probability of a correct answer, $x_i = 1$, to item i is given by Reckase (2009)

$$P_i(\boldsymbol{\theta}) = P_i(x_i = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}) = \frac{\exp(\mathbf{a}_i' \boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i' \boldsymbol{\theta} + d_i)}, \quad (14.1)$$

where \mathbf{a}_i is the vector of discrimination parameters and d_i is the easiness of the item.

Ability estimates can be used during testing by the item selection and classification methods in computerized classification testing. Ability is estimated using the likelihood function. The likelihood of a vector of observed responses $\mathbf{x}_j = (x_{1j}, \dots, x_{kj})$ to items $i = 1, \dots, k$ for an examinee j with ability $\boldsymbol{\theta}_j$ equals the product of the probabilities of the responses to the administered items (Segall 1996):

$$L(\boldsymbol{\theta}_j | \mathbf{x}_j) = \prod_{i=1}^k P_i(\boldsymbol{\theta}_j)^{x_{ij}} Q_i(\boldsymbol{\theta}_j)^{1-x_{ij}}, \quad (14.2)$$

where $Q_i(\boldsymbol{\theta}_j) = 1 - P_i(\boldsymbol{\theta}_j)$. This likelihood can be used due to the local independence assumption, and it uses the fixed item parameters from the item bank.

The vector of values, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$, that maximizes Eq. 14.2 is used as the ability estimate $\boldsymbol{\theta}_j$ (Segall 1996). Because the equations used to find the maximum likelihood (ML) estimates have no closed-form solution, an iterative search procedure, such as Newton-Raphson, is used. Weighted maximum likelihood (Tam 1992; Van Groen et al. 2016; Warm 1989) estimation reduces the bias in the ML estimates. Alternatively, Bayesian ability estimation approaches are also available (Segall 1996).

The two types of multidimensionality can be distinguished by the structure of their item parameters. If more than one discrimination parameter is nonzero for one or more items, within-item multidimensionality is present (Wang and Chen 2004), and items are intended to measure multiple (or all) abilities. Using a within-item multidimensional model, complex domains can be modeled while taking into account several abilities simultaneously (Hartig and Höhler 2008). Different combinations of abilities can be represented for different items (Hartig and Höhler 2008). If just one discrimination parameter is nonzero per item in the test, the test is considered to have between-item multidimensionality (Wang and Chen 2004). Items are then intended to measure just one ability, and the test contains several unidimensional subscales (Hartig and Höhler 2008).

14.3 Classification Methods

Multidimensional classification testing requires a method that decides whether enough evidence is available to make a classification decision (i.e., testing is stopped and a decision is made). The decision can be a classification into one of two levels (e.g., master/nonmaster) or into one of multiple levels (e.g., basic/proficient/advanced). Two well-known unidimensional methods, the SPRT (Eggen 1999; Reckase 1983; Spray 1993) and the confidence interval method (Kingsbury and

Weiss 1979), were adapted to multidimensional tests. Both methods stop testing when a prespecified amount of confidence has been established to make the classification decision, but the way the decision is made differs. The methods can make a decision based on the entire test and all dimensions, or based on parts of the tests or dimensions. Different implementations of the methods are required for between-item and within-item multidimensionality.

14.3.1 The SPRT for Between-Item Multidimensionality

The SPRT (Wald 1947/1973) was applied to unidimensional classification testing by Eggen (1999), Reckase (1983), and Spray (1993). Seitz and Frey (2013a) applied the SPRT to between-item multidimensionality by making a classification decision per dimension. Van Groen et al. (2014b) extended the SPRT to make classification decisions on the entire between-item multidimensional test.

14.3.1.1 The SPRT for Making a Decision Per Dimension

When making a decision per dimension, a cutoff point, θ_{cl} , is set for each dimension l , $l = 1, \dots, p$ (Seitz and Frey 2013a). The cutoff point is used by the SPRT to make the classification decision. Indifference regions are set around the cutoff points and account for the measurement error in decisions for examinees with an ability close to the cutoff point (Eggen 1999). The SPRT compares two hypotheses for each dimension (Seitz and Frey 2013a):

$$H_{0l} : \theta_{jl} < \theta_{cl} - \delta, \quad (14.3)$$

$$H_{al} : \theta_{jl} > \theta_{cl} + \delta, \quad (14.4)$$

where δ is the distance between the cutoff point and the end of the indifference region.

The likelihood ratio between the hypotheses for dimension l after k items are administered is calculated for the SPRT (Seitz and Frey 2013a) after each item administration by

$$LR(\theta_{cl} + \delta; \theta_{cl} - \delta) = \frac{L(\theta_{cl} + \delta; \mathbf{x}_{jl})}{L(\theta_{cl} - \delta; \mathbf{x}_{jl})}, \quad l = 1, \dots, p, \quad (14.5)$$

in which $L(\theta_{cl} + \delta; \mathbf{x}_{jl})$ and $L(\theta_{cl} - \delta; \mathbf{x}_{jl})$ are calculated using Eq. 14.2 with those items included that load on dimension l .

Decision rules are then applied to the likelihood ratio to decide whether to continue testing or to make a classification decision (Seitz and Frey 2013a; Van Groen et al. 2014b):

$$\begin{array}{ll}
\text{administer another item if} & \beta/(1 - \alpha) < \text{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) < (1 - \beta)/\alpha; \\
\text{ability below } \theta_{cl} \text{ if} & \text{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \leq \beta/(1 - \alpha); \\
\text{ability above } \theta_{cl} \text{ if} & \text{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \geq (1 - \beta)/\alpha,
\end{array} \tag{14.6}$$

where α and β specify the acceptable classification error rates (Spray et al. 1997). Previous research has shown that the size of α , β and sometimes also δ has a limited effect on classification accuracy (Van Groen et al. 2014a, c). When classifications need to be made into one of multiple levels, the decision rules are applied sequentially for each of the cutoffs until a decision can be made (Eggen and Straetmans 2000; Van Groen et al. 2014b).

14.3.1.2 The SPRT for Making a Decision on All Dimensions

In many testing situations, a pass/fail decision is required on the entire test in addition to classifications on dimensions of the test. The likelihood then includes all dimensions and items (Van Groen et al. 2014b):

$$\text{LR}(\theta_c + \delta; \theta_c - \delta) = \frac{L(\theta_c + \delta; \mathbf{x}_j)}{L(\theta_c - \delta; \mathbf{x}_j)}, \tag{14.7}$$

where θ_c and δ include all dimensions. The decision rules for the entire test then become (Van Groen et al. 2014b)

$$\begin{array}{ll}
\text{administer another item if} & \beta/(1 - \alpha) < \text{LR}(\theta_c + \delta; \theta_c - \delta) < (1 - \beta)/\alpha; \\
\text{ability below } \theta_c \text{ if} & \text{LR}(\theta_c + \delta; \theta_c - \delta) \leq \beta/(1 - \alpha); \\
\text{ability above } \theta_c \text{ if} & \text{LR}(\theta_c + \delta; \theta_c - \delta) \geq (1 - \beta)/\alpha.
\end{array} \tag{14.8}$$

The SPRT can also make decisions based on a subset of the dimensions or a subset of the items (Van Groen et al. 2014b). This implies that only the selected dimensions and items are included in the likelihoods. This makes it possible to include an item in several decisions.

14.3.2 The Confidence Interval Method for Between-Item Multidimensionality

An alternative unidimensional classification method, developed by Kingsbury and Weiss (1979), uses the confidence interval surrounding the ability estimate. The method stops testing as soon as the cutoff point is outside the confidence interval.

Seitz and Frey (2013b) applied this method to tests with between-item multidimensionality to make a classification decision per dimension. Van Groen et al. (2014c) extended the method to make decisions based on a some or all dimensions within the test.

14.3.2.1 The Confidence Interval Method for Making a Decision Per Dimension

Seitz and Frey (2013b) used the fact that the likelihood function of a between-item multidimensional test reduces to a unidimensional function for each dimension. This unidimensional likelihood function is used to make the classification decisions. The method uses the following decision rules (Van Groen et al. 2014c):

$$\begin{aligned}
 &\text{administer another item if } \hat{\theta}_{jl} - \gamma \cdot \text{se}(\hat{\theta}_{jl}) < \theta_{cl} < \hat{\theta}_{jl} + \gamma \cdot \text{se}(\hat{\theta}_{jl}); \\
 &\text{ability below } \theta_{cl} \text{ if } \hat{\theta}_{jl} + \gamma \cdot \text{se}(\hat{\theta}_{jl}) < \theta_{cl}; \quad (14.9) \\
 &\text{ability above } \theta_{cl} \text{ if } \hat{\theta}_{jl} - \gamma \cdot \text{se}(\hat{\theta}_{jl}) > \theta_{cl},
 \end{aligned}$$

where $\hat{\theta}_{jl}$ is the estimate examinee’s current ability for dimension l , γ is a constant related to the required accuracy (Eggen and Straetmans 2000), and $\text{se}(\hat{\theta}_{jl})$ is the estimate’s standard error (Hambleton et al. 1991):

$$\text{se}(\hat{\theta}_{jl}) = \frac{1}{\sqrt{I(\hat{\theta}_{jl})}}, \quad (14.10)$$

where $I(\hat{\theta}_{jl})$ is the Fisher information available to estimate θ_{jl} (Mulder and Van der Linden 2009). Fisher information is given by Tam (1992)

$$I(\hat{\theta}_{jl}) = \sum_{i=1}^k a_{il}^2 P_i(\hat{\theta}_{jl}) Q_i(\hat{\theta}_{jl}). \quad (14.11)$$

14.3.2.2 The Confidence Interval for Making a Decision on All Dimensions

Seitz and Frey’s (2013b) application of the confidence interval method can be used to make decisions per dimension, but not to make decisions on the entire test. An approach developed by Van der Linden (1999) can be used as a starting point for applying the confidence interval method to between-item multidimensional classification testing (Van Groen et al. 2014c).

Van der Linden (1999) considered the parameter of interest to be a linear combination of the abilities, $\lambda'\theta$, where λ is a combination of p nonnegative weights. In contrast to Van Groen et al. (2014c), we set λ equal to the proportion of items

intended to measure the dimension at maximum test length. The composite ability is given by Van der Linden (1999)

$$\zeta = \sum_{l=1}^p \theta_l \lambda_l. \quad (14.12)$$

The confidence interval method requires the standard error of the ability estimates (Yao 2012):

$$\text{se}(\zeta) = \sqrt{V(\zeta)}, \quad (14.13)$$

with variance

$$V(\zeta) = \boldsymbol{\lambda}' V(\boldsymbol{\theta}) \boldsymbol{\lambda}, \quad (14.14)$$

and $V(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})^{-1}$. The nondiagonal elements of matrix \mathbf{I} are zero, and the diagonal elements can be calculated using Eq. 14.11.

The decision rules are (Van Groen et al. 2014c)

$$\begin{aligned} \text{administer another item if} & \quad \hat{\zeta}_j - \gamma \cdot \text{se}(\hat{\zeta}_j) < \zeta_c < \hat{\zeta}_j + \gamma \cdot \text{se}(\hat{\zeta}_j); \\ \text{ability below } \zeta_c \text{ if} & \quad \hat{\zeta}_j + \gamma \cdot \text{se}(\hat{\zeta}_j) < \zeta_c; \\ \text{ability above } \zeta_c \text{ if} & \quad \hat{\zeta}_j - \gamma \cdot \text{se}(\hat{\zeta}_j) > \zeta_c, \end{aligned} \quad (14.15)$$

where ζ_c denotes the cutoff point and $\hat{\zeta}_j$ is the estimated composite for the student (Van Groen et al. 2014c). The cutoff point is determined with Eq. 14.12 using the cutoff points for the dimensions. If a decision on several, but not all, dimensions simultaneously is required, only those dimensions should be included in the equations.

14.3.3 The SPRT for Within-Item Multidimensionality

The SPRT can also be used for tests with within-item multidimensionality. This does require the computation of the so-called reference composite (Van Groen et al. 2016). The reference composite reduces the multidimensional space to an unidimensional line (Reckase 2009; Wang 1985, 1986). This line is then used to make the classification decision (Van Groen et al. 2016).

The reference composite describes the characteristics of the matrix of discrimination parameters for the items in the item bank (Reckase 2009) or test. The direction of the line is given by the eigenvector of the \mathbf{aa}' matrix that corresponds to the largest eigenvalue of this matrix (Reckase 2009). The elements of the eigenvector determine the direction cosines, $\alpha_{\xi l}$, for the angle between the reference composite and the dimension axes.

θ -points can be projected onto the reference composite (Reckase 2009). A higher value on the reference composite, ξ_j , denotes a more proficient student than does a lower value (Van Groen et al. 2014c). Proficiency on the reference composite can be calculated using an additional line through the θ_j -point and the origin. The length of this line is given by Reckase (2009)

$$L_j = \sqrt{\sum_{l=1}^p \hat{\theta}_{jl}^2}, \tag{14.16}$$

and the direction cosines, α_{jl} , for this line and dimension axis, l , are calculated using (Reckase 2009)

$$\cos \alpha_{jl} = \frac{\hat{\theta}_{jl}}{L_j}, \quad l = 1, \dots, p. \tag{14.17}$$

The angle, $\alpha_{j\xi} = \alpha_{jl} - \alpha_{\xi l}$, between the reference composite and the student’s line is used to calculate the estimated proficiency, $\hat{\xi}_j$, on the reference composite:

$$\hat{\xi}_j = L_j \cos \alpha_{j\xi}. \tag{14.18}$$

The reference composite can now be used to make classification decisions with the SPRT (Van Groen et al. 2016). The cutoff point for the SPRT, ξ_c , and δ^ξ are specified on the reference composite. The boundaries of the indifference region need to be transformed to their θ -points using

$$\theta_{\xi_{c+\delta}} = \cos \alpha_\xi \times (\xi_c + \delta^\xi); \tag{14.19}$$

$$\theta_{\xi_{c-\delta}} = \cos \alpha_\xi \times (\xi_c - \delta^\xi), \tag{14.20}$$

where α_ξ includes the angles between the reference composite and all dimension axes. The likelihood ratio for the SPRT becomes (Van Groen et al. 2016)

$$LR(\theta_{\xi_{c+\delta}}; \theta_{\xi_{c-\delta}}) = \frac{L(\theta_{\xi_{c+\delta}}; \mathbf{x}_j)}{L(\theta_{\xi_{c-\delta}}; \mathbf{x}_j)}, \tag{14.21}$$

which can be used to make multidimensional classification decisions with the following decision rules (Van Groen et al. 2016):

- administer another item if $\beta/(1 - \alpha) < LR(\theta_{\xi_{c+\delta}}; \theta_{\xi_{c-\delta}}) < (1 - \beta)/\alpha$;
- ability below ξ_c if $LR(\theta_{\xi_{c+\delta}}; \theta_{\xi_{c-\delta}}) \leq \beta/(1 - \alpha)$;
- ability above ξ_c if $LR(\theta_{\xi_{c+\delta}}; \theta_{\xi_{c-\delta}}) \geq (1 - \beta)/\alpha$.

Decisions can be made using the reference composite for different subsets of items and for more than two decision levels (Van Groen et al. 2016).

14.3.4 The Confidence Interval Method for Within-Item Multidimensionality

The confidence interval method (Kingsbury and Weiss 1979) can also be used for tests with within-item multidimensionality. Again, the reference composite is used to make a classification decision (Van Groen et al. 2014c).

To determine whether testing can be stopped after an item is administered, the examinee's ability is estimated after each item. This estimate is projected onto the reference composite. The proficiency on the reference composite is then transformed to the corresponding point in the multidimensional space (Van Groen et al. 2014c):

$$\boldsymbol{\theta}_{\hat{\xi}_j} = \cos \alpha_{\xi} \times \hat{\xi}_j. \quad (14.23)$$

The reference composite is considered to be a combination of abilities. The weights between the abilities, λ_{ξ} , are based on the angles between the reference composite and the dimension axes; $\lambda_{\xi l} = 1/\alpha_{\xi l}$ (Van Groen et al. 2014c).

The standard error for the confidence interval method is given by Yao (2012)

$$\text{se}(\xi) = \sqrt{V(\xi)}, \quad (14.24)$$

with

$$V(\xi) = \boldsymbol{\lambda}'_{\xi} V(\boldsymbol{\theta}_{\xi}) \boldsymbol{\lambda}_{\xi}. \quad (14.25)$$

The variance at $\boldsymbol{\theta}_{\hat{\xi}_j}$ is approximated by the inverse of the information matrix at $\boldsymbol{\theta}_{\hat{\xi}_j}$. The decision rules are (Van Groen et al. 2014c)

$$\begin{aligned} \text{administer another item if } & \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c < \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j); \\ \text{ability below } \xi_c \text{ if } & \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c; \\ \text{ability above } \xi_c \text{ if } & \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) > \xi_c, \end{aligned} \quad (14.26)$$

where ξ_c denotes the cutoff point on the reference composite. The cutoff point is determined based on the cutoff points for each dimension. Again, decisions can be made using the reference composite for different subsets of items and for more than two decision levels (Van Groen et al. 2016).

14.4 Item Selection Methods

Computerized classification testing requires a method to select the items during test administration. Most methods select an optimal item given some statistical criterion. Limited knowledge is available for selecting items in multidimensional classification testing (Seitz and Frey 2013a; Van Groen et al. 2014b,c, 2016).

Van Groen et al. (2014c) based their item selection on knowledge about item selection methods from unidimensional classification testing (Eggen 1999; Spray and Reckase 1994; Van Groen et al. 2014a, c) and multidimensional computerized adaptive testing for ability estimation (Reckase 2009; Segall 1996; Yao 2012).

Following Van Groen et al. (2014c), three different types of item selection methods are included here for between-item and within-item multidimensionality. In unidimensional classification testing, items are often selected that maximize information at the current ability estimate or at the cutoff point (Eggen 1999; Spray 1993; Thompson 2009). Maximization at the cutoff point becomes more complicated when multiple cutoff points are included. One type of the methods available for selection with multiple cutoff points, that is weighting methods, is also used here (Van Groen et al. 2014a). The third item selection method selects items at random. This provides a benchmark that can be used to compare the efficiency and accuracy of the more complex methods. The method that selects items based on the ability estimate and the weighting method have different implementations for between- and within-item multidimensionality.

14.4.1 Item Selection Methods for Between-Item Multidimensionality

Unidimensional item selection methods can be used for tests with between-item multidimensionality (Van Groen et al. 2014b), but items must be selected per dimension. Here this is accomplished using the Kingsbury and Zara (1989) approach. This approach selects the items from the dimension for which the difference between the desired and achieved percentage is the largest (Van Groen et al. 2014c) and can be applied to the three described item selection methods. The unidimensional item selection method that maximizes information at the ability estimate is described first. The unidimensional weighting method is described thereafter.

14.4.1.1 Item Selection Using the Ability Estimate

In unidimensional classification testing, Fisher information is often maximized at the current ability estimate (Eggen 1999; Spray 1993). The aim of maximizing information at the current ability estimate is to reduce the estimate's standard error. In between-item multidimensional testing, this boils down to (Van Groen et al. 2014b)

$$\max I_i(\hat{\theta}_l), \quad \text{for } i \in V_{al}, \quad (14.27)$$

where V_{al} is the set of items available for selection for dimension l .

14.4.1.2 Item Selection Using the Weighting Method

Information is often maximized at the cutoff point in unidimensional classification testing (Eggen 1999; Spray 1993). In between-item multidimensional tests, items can be selected that maximize the information at the cutoff point using

$$\max I_i(\theta_{cl}), \quad \text{for } i \in V_{al} \quad (14.28)$$

as a consequence of the selection per dimension in such tests.

This method can be used if just one cutoff point is specified per dimension. Several methods have been developed for item selection with multiple cutoff points for unidimensional classification testing (Eggen and Straetmans 2000; Van Groen et al. 2014a; Wouda and Eggen 2009). The weighting method combines the objective functions per cutoff point into one weighted objective function. The weight for the cutoff points depends on the distance of the cutoff point to the current ability estimate (Van Groen et al. 2014a). As a result, the test is adapted to the individual student's ability. The item is selected for dimension l that fulfills (Van Groen et al. 2014b)

$$\max \sum_{c=1}^C \frac{1}{|\hat{\theta}_{jl} - \theta_{cl}|} I_i(\theta_{cl}), \quad \text{for } i \in V_{al}. \quad (14.29)$$

14.4.2 Item Selection Methods for Within-Item Multidimensionality

In contrast to between-item multidimensional tests, selecting items for within-item multidimensional tests has to take the multidimensional structure of the test into account. This implies that multidimensional item selection methods have to be used. However, these methods were developed for multidimensional computerized adaptive testing to obtain an efficient and precise ability estimate. Nevertheless, Segall's (1996) method, which selects the items at the current ability estimate, can be used for classification testing. This method was adapted by Van Groen et al. (2016) for selection at a weighted combination of the cutoff points.

14.4.2.1 Item Selection Using the Ability Estimate

Segall (1996) developed an item selection method focused on estimating ability as precisely as possible. Precision is reflected by the size of the confidence ellipsoid surrounding the estimate. The item is selected that results in the largest decrement of the volume of the confidence ellipsoid. The size of the ellipsoid can be approximated by the inverse of the information matrix, so the item is selected that maximizes (Segall 1996)

$$\max \det \left(\sum_{i=1}^k I(\hat{\theta}_j, x_{ij}) + I(\hat{\theta}_j, x_{k+1,j}) \right), \quad \text{for } k+1 \in V_{k+1}, \quad (14.30)$$

where $\sum_{i=1}^k I(\hat{\theta}_j, x_{ij})$ denotes the information that has been collected thus far and $I(\hat{\theta}_j, x_{k+1,j})$ is the information that a potential item provides. Because the item is selected that results in the largest determinant of the information matrix, the volume of the confidence ellipsoid is minimized (Reckase 2009).

14.4.2.2 Item Selection Using the Weighting Method

The weighting method for within-item multidimensional classification testing is based on maximization at a weighted combination of the cutoff points (Van Groen et al. 2016). The weights for the dimensions depend on the distance between the cutoff points and the student's proficiency. The weighting method then selects the item using (Van Groen et al. 2016)

$$\max \sum_{c=1}^C \frac{1}{|\xi_{\hat{\theta}_j} - \xi_c|} I_i(\theta_{\xi_c}), \quad \text{for } i \in V_{at}. \quad (14.31)$$

14.5 Examples

The accuracy and efficiency of the classification and item selection methods are illustrated using two example simulation studies. The first example deals with a test in which between-item multidimensionality is present. The second example shows a test with a within-item multidimensional structure. The examples demonstrate the differences in accuracy and efficiency in specific situations with specific settings for the classification methods.

14.5.1 Example 1: Between-Item Multidimensionality

The End of Primary School Test (Cito 2012) in the Netherlands can be calibrated with a between-item multidimensional model. The test provides a recommendation on the most suitable level of secondary education to pupils. As a test result, a scale score is given that relates to an advise on the most appropriate level of secondary education (Van Boxtel et al. 2011). The test is typically modeled using six unidimensional scales; four scales for Language (100 items), one scale for Mathematics (60 items), and one scale for Study Skills (40 items).

14.5.1.1 Simulation Design for Example 1

The example is based on the item parameters, ability distribution, and correlation structure from a NOHARM 4 (Fraser and McDonald 2012) calibration with 147,099 students. A dataset of 10,000 students was generated with abilities drawn from the observed correlation structure. The same dataset was used for all simulations. The generated abilities were used to determine the true classification for each student. The minimum test length was set at 20 items, with a maximum of 200 items. Three cutoffs were specified based on the 2012 ability and score distributions. Simulations were run with the SPRT and the confidence interval method. For the SPRT, the values for α and β were set equal and varied between 0.1 and 0.2, and δ was set to 0.4. Other values for δ resulted in similar test lengths and accuracy. For the confidence interval method, γ varied between values corresponding to 90–97.5% confidence intervals for each of the individual cutoff points. Items were selected using random selection, maximization at the ability estimate, and the weighting method. The content structure of the item pool was respected in the item selection using the Kingsbury and Zara approach (1989).

The choices made here were different from those made for actual test administration for the purpose of simplicity and comparability with example 2. The number of cutoffs was limited to three in example 1. The actual test provides eight different advises and also reports at the subject level. The choices made here imply that the presented simulation results cannot be translated to actual test administration.

14.5.1.2 Simulation Results, Example 1

The results of the simulations for example 1 are presented in Figs. 14.1 and 14.2. The former indicates the average test length (ATL) over all students per condition. The latter denotes the proportion of correct decisions (PCD) as an indication of the accuracy of the classification decisions. The PCD compares the true classification and the decision made by the classification methods.

The plot for average test length for the SPRT shows that higher values for α and β result in a shorter test. The plot for the confidence interval method shows that when the confidence interval is increased, the test length also increases. The SPRT results in longer tests than the confidence interval method for these settings of the SPRT and confidence interval method. Random item selection results in the longest tests. Maximization of information at the ability estimate results in the shortest tests.

The plot for the proportion of correct decisions for the SPRT indicates that 80–87% of the classifications are accurate. The plot for the confidence interval method indicates that the settings for γ have a limited influence on classification accuracy. The plot also shows that random selection results in slightly less accurate decisions. The SPRT results in more accurate decisions than the confidence interval method.

Based on these simulations, one might conclude that the SPRT results in longer tests with more accurate decisions. However, this might be caused by the settings of the SPRT and the confidence interval method in the example.

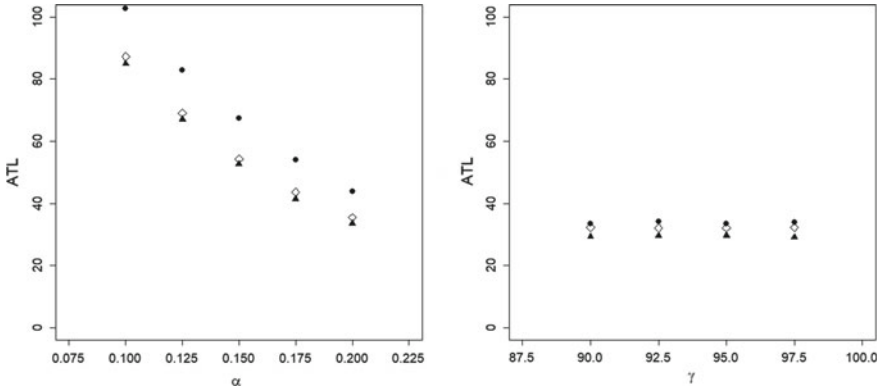


Fig. 14.1 Average test length, example 1. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

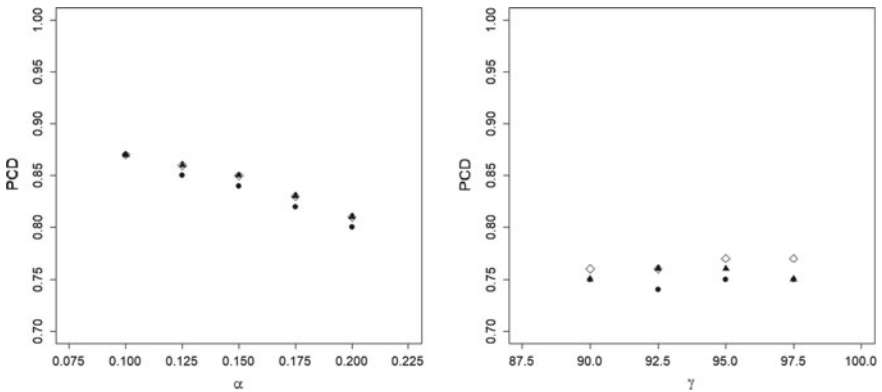


Fig. 14.2 Proportion of correct decisions, example 1. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

14.5.2 Example 2: Within-Item Multidimensionality

An within-item multidimensional ACT mathematics item pool, as in Ackerman (1994) and Veldkamp and Van der Linden (2002), was used in the second example. The item parameters had been obtained previously using an orthogonal two-dimensional NOHARM II (Fraser and McDonald 1988) calibration. Items are intended to measure coordinate geometry, elementary algebra, intermediate algebra, pre-algebra, plane geometry, and trigonometry, and they require three skill categories; analysis, application, and basic skills (Veldkamp and Van der Linden 2002).

14.5.2.1 Simulation Design, Example 2

The standard normal multivariate ability distribution was used to generate 10,000 students. The cutoff points were specified on the reference composite based on the 33th and 66th percentiles of the underlying dimensions. Simulations were generated with $\alpha = \beta$ between 0.025 and 0.10, and δ was set at 0.40. γ varied between values corresponding to 90–97.5% confidence intervals. A maximum of 50 items was selected using random selection, maximization at the ability estimate, and the weighting method. No content restrictions were made in the simulations.

14.5.2.2 Simulation Results, Example 2

The results of the simulations for example 2 can be seen in Figs. 14.3 and 14.4. The plot for average test length for the SPRT shows that higher values for α and β result in a shorter test. As expected, the average test length increased when a larger confidence interval was required. The plots also suggest that test length is often shorter for the confidence interval method. However, this depends on the settings for the SPRT and the confidence interval method. Random selection results in the longest tests, and the weighting method results in the shortest test. The effect of the item selection method appears to be consistent regardless of the classification method.

The plot for the proportion of correct decisions for the SPRT indicates that the values for α and β did not influence the accuracy of the decisions. In contrast, the size of the confidence interval did influence the accuracy. As expected, random selection resulted in the least accurate decisions. The confidence interval method resulted in less accurate decisions than did the SPRT. This might be caused by the shorter tests for the confidence interval method.

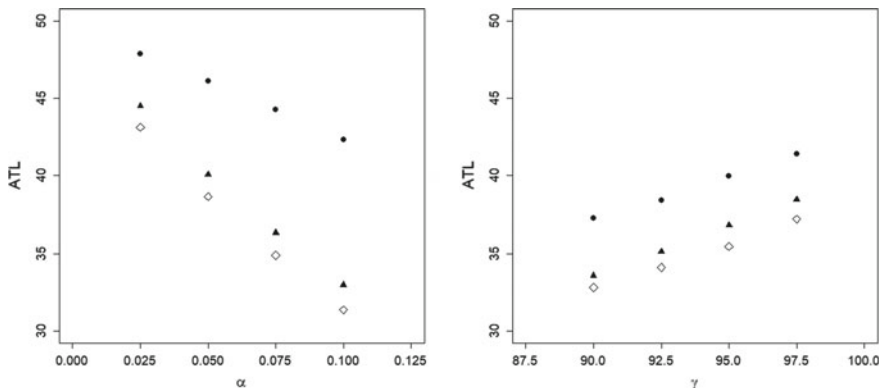


Fig. 14.3 Average test length, example 2. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

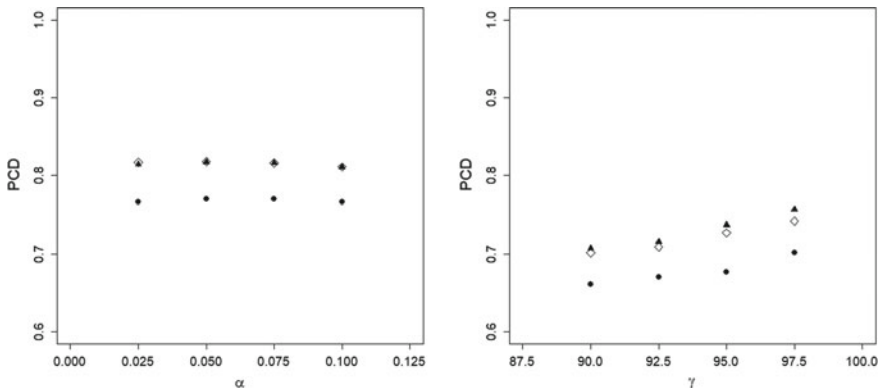


Fig. 14.4 Proportion of correct decisions, example 2. Left plot: SPRT; right plot: confidence interval method. $\delta = 0.4$. Dots: random selection; triangles: maximization at the ability estimate; diamonds: weighting method

Based on the current settings for the SPRT and confidence interval method, one might conclude that the SPRT resulted in longer tests in this example dataset. However, those longer tests result in more accurate classification decisions. This finding clearly indicates the well-known trade-off between accuracy and test length.

14.6 Conclusions and Discussion

Multidimensional computerized classification testing can be used when classification decisions are required for constructs that have a multidimensional structure. Here, two methods for making those decisions were included here for two types of multidimensionality. In the case of between-item multidimensionality, each item is intended to measure just one dimension. By contrast, in case of within-item multidimensionality, items are intended to measure multiple or all dimensions.

Wald's (1947) sequential probability ratio test has previously been applied to multidimensional classification testing by Seitz and Frey (2013a) and Van Groen et al. (2014b) for between-item multidimensionality and by Van Groen et al. (2016) for within-item multidimensionality. Kingsbury and Weiss's (1979) confidence interval method was applied to between-item multidimensionality by Seitz and Frey (2013b) and Van Groen et al. (2014c), and by Van Groen et al. (2014c) to within-item multidimensionality. The present study's implementation of the confidence interval method for between-item dimensionality differs from the implementation in Van Groen et al. (2014c). Here, different weights were used that reflect the number of items per dimension relative to the test length.

Three methods were included for selecting the items: random item selection, maximization at the current ability estimate, and the weighting method. The last

method maximizes information based on a combination of the cutoff points weighted by their distance to the ability estimate. Two examples illustrated the use of the classification and item selection methods.

Classifications were made into one of three classification levels on the entire test. Van Groen et al. (2014c) showed classifications on both the entire test as well as on some of the items or dimensions. They also required that decisions had to be made on all partial classification decisions before a classification decision could be made on the entire test. Here, only one decision was made per test.

Examples were included to demonstrate the use of the classification and item selection methods. Simulations included only two different item banks, a limited number of settings for the classification methods, and a limited number of students. More thorough studies are needed in order to draw conclusions regarding the most effective settings and classification methods. Van Groen et al. (2014c) provided some further suggestions for comparison of the classification methods. Moreover, only three item selection methods were included here. Many more item selection methods (Reckase 2009; Segall 1996; Yao 2012) exist for multidimensional adaptive testing. These should be investigated as well. Finally, α , β , δ , and γ were set equal for all cutoff points. The use of different values for different cutoff points should be investigated as well.

Two classification methods were included here. A third unidimensional classification method, the generalized likelihood ratio test developed by Bartroff, Finkelman, and Bartroff et al. (2008), has never been applied to multidimensional classification testing. This method is based on the sequential probability ratio test. This suggests that it should be feasible to expand the method to multidimensional classification testing.

Research on multidimensional classification testing is still limited to the work of (Seitz and Frey 2013a, b), Spray et al. (1997), and the current authors (Van Groen et al. 2014b, c, 2016). Further studies will be required before multidimensional classification testing can be applied in practice. To date, too little is known about the classification methods, the item selection methods in this context, and the settings for the classification methods to administer multidimensional classification tests with confidence.

References

- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame0704_1.
- Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*. <https://doi.org/10.1007/s11336-007-9053-9>.
- Cito. (2012). *Eindtoets Basisonderwijs 2012 (End of primary school test 2012)*. Arnhem, The Netherlands: Cito.
- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*. <https://doi.org/10.1177/01466219922031365>.

- Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*. <https://doi.org/10.1177/00131640021970862>.
- Fraser, C., & McDonald, R. P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory* (Computer Software).
- Fraser, C., & McDonald, R. P. (2012). *NOHARM 4: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* (Computer Software).
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*. <https://doi.org/10.1027/0044-3409.216.2.89>.
- Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategy for mastery decisions (Research Report 79-5)*. Minneapolis, M.N.: University of Minnesota Press.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame0204_6.
- Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*. <https://doi.org/10.1007/S11336-008-9097-5>.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York, NY: Academic Press.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*. <https://doi.org/10.1007/BF02294343>.
- Seitz, N.-N., & Frey, A. (2013a). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, 55(1), 105–123.
- Seitz, N.-N., & Frey, A. (2013b). *Confidence interval-based classification for multidimensional adaptive testing* (Manuscript submitted for publication).
- Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Report No. ACT-RR-93-7). Iowa City, IA: American College Testing.
- Spray, J. A., Abdel-Fattah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized adaptive test when the item pool and latent space are multidimensional* (Report No. 97-5). Iowa City, IA: American College Testing.
- Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans, LA.
- Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation, Columbia University, New York, NY).
- Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*. <https://doi.org/10.1177/0013164408324460>.
- Van Boxtel, H., Engelen, R., & De Wijs, A. (2011). *Wetenschappelijke verantwoording van de Eindtoets 2010 (Scientific report for the end of primary school test 2010)*. Arnhem, The Netherlands: Cito.
- Van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*. <https://doi.org/10.3102/10769986024004398>.
- Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014a). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*. <https://doi.org/10.1177/0146621613509723>.

- Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014b). Multidimensional computerized adaptive testing for classifying examinees on tests with between-dimensionality. In M. M. van Groen (Ed.), *Adaptive testing for making unidimensional and multidimensional classification decisions* (pp. 45–71) (Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands).
- Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014c). Multidimensional computerized adaptive testing for classifying examinees with the SPRT and the confidence interval method. In M. M. van Groen (Ed.), *Adaptive testing for making unidimensional and multidimensional classification decisions* (pp. 101–130) (Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands).
- Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement*. <https://doi.org/10.1177/0146621616648931>.
- Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*. <https://doi.org/10.1007/BF02295132>.
- Wald, A. (1973). *Sequential analysis*. New York, NY: Dover (Original work published 1947).
- Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effect of latent space misspecification on the application of IRT (Research Report MW: 6–24–85)*. Iowa City, IA: University of Iowa Press.
- Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.
- Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*. <https://doi.org/10.1177/0146621604265938>.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*. <https://doi.org/10.1007/BF02294627>.
- 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*.
- Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*. <https://doi.org/10.1007/S11336-012-9265-5>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

