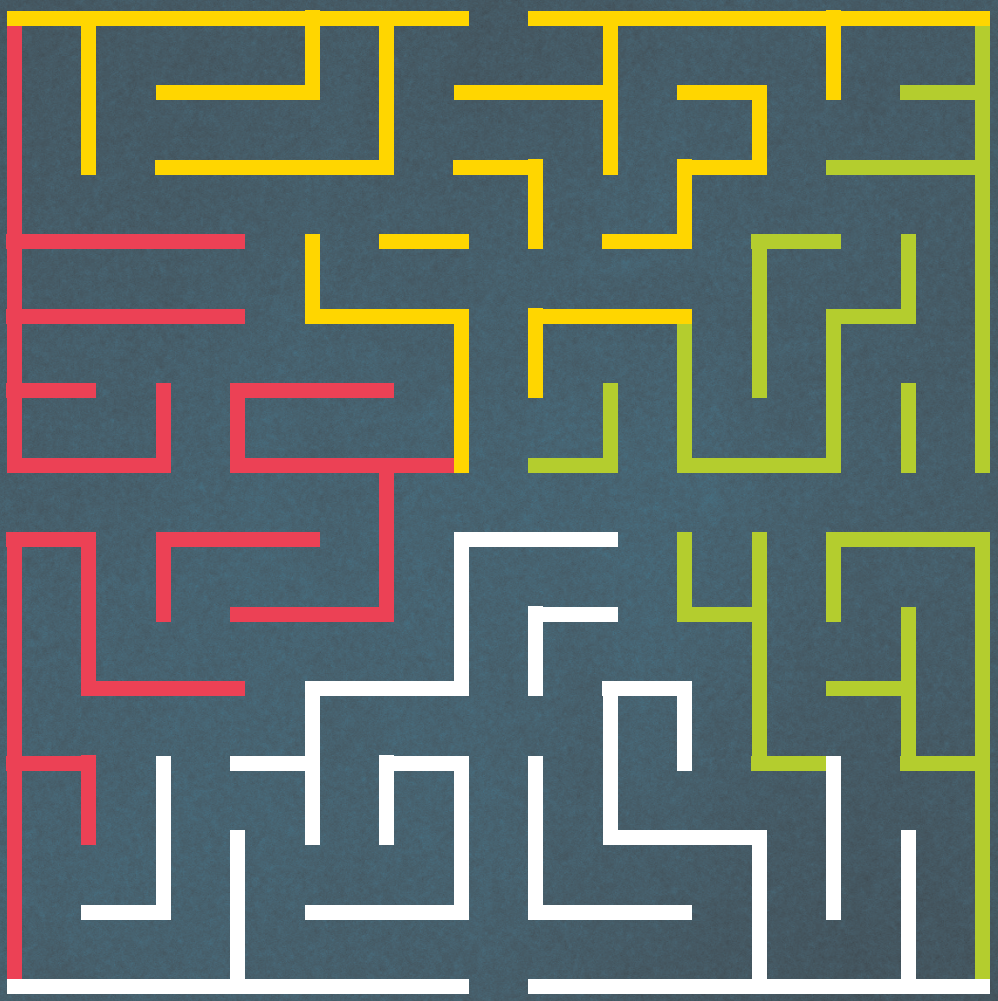
Adaptive testing for making unidimensional and multidimensional classification decisions



Maaike M. van Groen

ADAPTIVE TESTING FOR MAKING UNIDIMENSIONAL AND MULTIDIMENSIONAL CLASSIFICATION DECISIONS

Maaike M. van Groen

Graduation Committee

Chairman	prof. dr. ir. A.J. Mouthaan
Promotor	prof. dr. ir. T.J.H.M. Eggen
Copromotor	prof. dr. ir. B.P. Veldkamp
Members	prof. dr. C.A.W. Glas
	prof. dr. H.J.A. Hoijtink
	dr. A.W. Lazonder
	prof. dr. W. Van den Noortgate

ISBN 978-94-6259-416-6
Printed by Ipskamp Drukkers, Enschede
Cover designed by M. Brouwer (Cito)

This research was supported by Cito, Institute for Educational Measurement.

ADAPTIVE TESTING FOR MAKING UNIDIMENSIONAL AND MULTIDIMENSIONAL CLASSIFICATION DECISIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, November 21th, 2014 at 14:45

by

Maaike Margaretha van Groen
born on May 23th, 1984
in Woerden, the Netherlands

This dissertation has been approved by the promotors:
prof. dr. ir. T.J.H.M. Eggen
prof. dr. ir. B.P. Veldkamp

Contents

Chapter 1
Introduction

A large variety of test types exists. Some types of tests adapt their testing process to the characteristics of the individual student. A computerized adaptive test (CAT) tailors item selection and test length to the student's ability, but can also adapt the test content to the individual student.

Computerized adaptive testing can serve two different measurement goals: the tests can obtain efficient and precise ability estimates or can make efficient and accurate classification decisions. Both the testing goals are achieved while minimizing the test length. The majority of CAT research concerns the first goal, but a computerized classification test (CCT) serves the second goal. The focus of this thesis is on the second goal. These tests classify students into one of a limited number of mutually exclusive categories depending on the student's responses to the test items. The route toward the decision can be different for each student, but at the end of the test, an accurate and efficient decision is made for all students.

The testing procedure of a CCT requires two methods. One method selects the items based on some statistical criterion. The other method decides whether testing can be stopped and makes the classification decision. Both methods often use item response theory (IRT) to make the connection between the student's responses, the items, and the ability of the student, so that items can be selected and decisions can be made. The students, items, item selection method, classification method, and the item response theory model should be aligned with each other if a CCT has to result in efficient, but most importantly, accurate classification decisions. It is like walking in a maze in which at the end everyone meets the others in the center of the maze (see Figure 1.1). Although the way through the maze is different for everyone, at the end everyone reaches the same destination. The same applies to a CCT; the entities within the CCT all follow their own procedure, but at the end, everything is directed toward making efficient and accurate classification decisions.

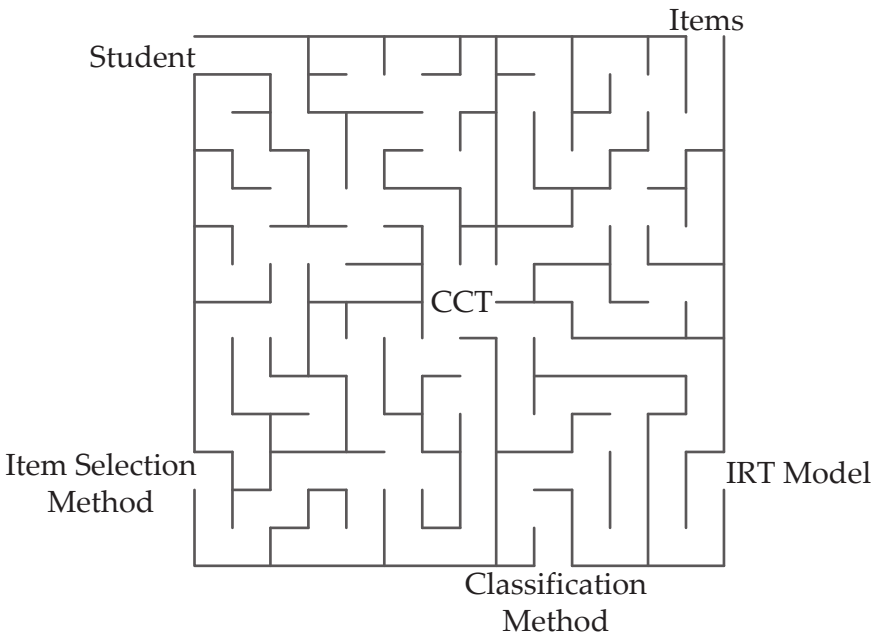


Figure 1.1. The components of a computerized classification test.

This introduction starts with a description of the design components of a CCT. The second section explores the contexts in which a CCT is used and the design of test environments for CCTs. Some characteristics of CCTs, and their current availability, are then discussed. The last part of the introduction introduces the research questions that are central to this thesis.

1.1 Components of CCTs

As described previously, a CCT consists of five separate design components that together form the basis for the design of the entire test development and test administration process. The five components of a CCT will be described next.

1.1.1 The Student

Computerized classification tests attempt to make an efficient and accurate classification decision for each student. The goal of most CCTs is to make a judgment about the student's ability. This implies that the student should be the focus in the test development process, during the testing, and after testing when the results of the test are reported to the student.

During the development process, test developers should have a clear mental picture of the intended testing population. Items should be written with this picture in mind because the test items should function the same for all groups

of students in the testing population; that is, there should be no differential item functioning or measurement invariance, items should have appropriate difficulty, and item content should be suitable for the intended students.

During the testing, the student should also be the central focus. The way items are presented should be appropriate for the students, and the navigation through the testing environment should be suitable for the students. Furthermore, also all procedures used should be evaluated for their functioning with regard to the students. One of these procedures selects the items. This procedure will be discussed in the fifth part of this section, but test developers should ensure that the procedure can not select items that are not suitable for the student.

After testing, the test results should be communicated to the student. A CCT can provide the classification decision, but it can also provide a knowledge profile for each student. The latter requires that multiple decisions are made for each student and that a classification is made into one of several levels per decision. Independent of the type of outcome, the classification decisions should be made with sufficient accuracy. One way to enhance learning as a result of testing is to provide feedback to the student (see, for example, Van der Kleij, 2013).

1.1.2 The Items

The items determine whether a CCT can make accurate and efficient decisions. In CAT, items are organized into an item bank. The item bank should be suitable for the specific testing situation (Van Groen, Eggen, & Veldkamp, 2014) and the intended testing population. In a calibrated item bank, model fit is established using item response theory, item parameter estimates are available, and items with inappropriate difficulty, fit, differential item functioning, a high lower asymptote, or low discrimination parameters are removed. Obviously, the items should be appropriate for the intended testing population. An important aspect when developing a CCT is that a sufficient number of items is available with optimal measurement properties at relevant positions on the ability scale.

To make valid inferences from the test, it is important that the construct validity of the test is established, preferably before the test is administered. Throughout the test development process, the test developers should monitor the validity of the test. The evidence-centered design framework (Mislevy, Steinberg, & Almond, 2003) can provide a guideline for test developers to ensure content validity during test development, but also to ensure validity afterwards. The argument-based approach can provide guidelines to make valid inferences based on test scores,

see Kane (2013). A procedure to evaluate validity based on the argument-based approach was developed by Wools, Eggen, and Sanders (2010).

1.1.3 The Item Response Theory Model

Item responses in CAT can be modeled using IRT. IRT specifies a relation between the score on an item, depending on the item parameters, and the student's ability (Van der Linden & Hambleton, 1997). The score on an item, $x_i = 1$ correct, $x_i = 0$ incorrect, and the ability, θ_j, of student j is modeled with a probability function. In unidimensional item response theory (UIRT), the probability of a correct response depends on just one ability parameter per student. For the two-parameter logistic model (Birnbaum, 1968/2008) the item probability is given by

$$\mathrm{P}_i\left(\theta, a_i, b_i\right) = \mathrm{P}\left(x_i = 1|\theta\right) = \frac{\exp\left(a_i\left[\theta - b_i\right]\right)}{1 + \exp\left(a_i\left[\theta - b_i\right]\right)}, \tag{1.1}$$

where a_i represents the discriminating power of item i, b_i difficulty, and θ ability. In CAT, the item parameters are considered to be estimated with precise enough to consider them known during testing (Veldkamp & Van der Linden, 2002).

In many tests, a vector of person abilities is required to describe the skills and knowledge necessary for answering the items (Reckase, 2009). The item responses can be modeled in these tests using multidimensional item response theory (MIRT). The multidimensional two-parameter logistic model is given by (Reckase, 1985)

$$\mathrm{P}_i(\boldsymbol{\theta}) = \mathrm{P}_i\left(x_i = 1|\mathbf{a}_i, d_i, \boldsymbol{\theta}\right) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}, \tag{1.2}$$

where $\mathbf{a}_i$ is the vector of the discrimination parameters, d_i denotes the easiness of the item, and $\boldsymbol{\theta}$ is the vector of the ability parameters. The number of elements in $\mathbf{a}_i$ is determined by the number of dimensions p, $l = 1, \cdots, p$.

Two types of multidimensionality can be distinguished. If more than one discrimination parameter is non-zero for each item, items are intended to measure multiple abilities (within-dimensionality; W.-C. Wang & Chen, 2004). If just one parameter is non-zero for each item in the test, between-dimensionality is present (W.-C. Wang & Chen, 2004). These tests consist of several related subtests.

When (M)IRT is used to model the student's responses, an IRT model should be selected that describes the data well and that is in coherence with the structure of the test.

Inferences about the student's ability can be drawn from the likelihood of the responses after k items with fixed item parameters are administered due to the local independence assumption:

$$\mathrm{L}\left(\boldsymbol{\theta};\mathbf{x}\right)=\prod_{i=1}^{k}\mathrm{P}_{i}\left(\boldsymbol{\theta}\right)^{x_{i}}\left[1-\mathrm{P}_{i}\left(\boldsymbol{\theta}\right)\right]^{1-x_{i}}, \tag{1.3}$$

where $\mathbf{x}=(x_1,\cdots,x_k)$ denotes the vector of responses to the administered items. If a unidimensional model is used, one element is imputed in Equation 1.3 for $\boldsymbol{\theta}$.

The vector of values $\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,\cdots,\hat{\theta}_p)$ that maximize the likelihood function in Equation 1.3 is taken as the ability estimate of $\boldsymbol{\theta}_j$. Unfortunately, the equations for finding maximum likelihood estimates have no closed-form solution (Segall, 1996). Several iterative procedures are available for finding the estimates, such as Newton-Raphson and the False Positioning Method. In addition to several estimation procedures, several types of estimates exist. In this thesis, weighted maximum likelihood estimates are used, instead of one of the Bayesian estimates or unweighted maximum likelihood estimates, because no prior is required and bias in the estimates is reduced compared to unweighted maximum likelihood.

1.1.4 The Classification Method

CAT requires a method that provides the outcome of the test and that determines whether testing can be stopped before the maximum test length is reached. CAT can provide two types of outcomes. An ability estimate is provided in CAT for ability estimation, and a classification decision in CAT for classification testing (CCT). Several stop criteria exist (Reckase, 2009; C. Wang, Chang, & Boughton, 2013; Yao, 2013), such as a specified number of items, when the ability estimate has reached a desired level of accuracy, a fixed testing time, or when a decision has been made with the desired level of confidence (Reckase, 2009). The focus in this thesis is on methods that provide one or more classification decisions with a fixed or flexible test length. The latter is possible if a classification method is used that stops testing when enough confidence is gained in the decision.

Two classification methods are often used in unidimensional classification testing, although other methods exist. The sequential probability ratio test (SPRT; Wald, 1947/1973) was first applied to CCT by Ferguson (1969) using classical test theory and by Reckase (1983) using IRT. The second method uses the confidence interval surrounding the ability estimates (Kingsbury & Weiss, 1979). These methods were applied to between-dimensionality MIRT for making a decision

for each dimension (Seitz & Frey, 2013a, 2013b). No classification methods are available for within-dimensionality. Two methods for making unidimensional decisions will be described in the next parts of this section for unidimensional CCT and a small comparison study is presented in the last part of this section.

Classification by the Sequential Probability Ratio Test

The SPRT has been applied to unidimensional classification testing by many scholars (Eggen & Straetmans, 2000; Finkelman, 2008; Spray, 1993; Thompson, 2009; Wouda & Eggen, 2009). A cutoff point is set on the ability scale with an indifference region surrounding it. This indifference region accounts for the uncertainty in the decisions, owing to measurement error, about students whose ability is close to the cutoff point (Eggen, 1999). Multiple cutoff points are set if a classification into one of several mutually exclusive categories is required with accompanying indifference regions (Eggen & Straetmans, 2000; Spray, 1993; Wouda & Eggen, 2009). Overlapping indifference regions should be avoided because that would imply that a classification into more than one category is acceptable.

Two hypotheses are formulated for each cutoff point θ_c based on the boundaries of the indifference region (Eggen, 2010):

$$\mathrm{H}_0 : \theta_j < \theta_c - \delta_c; \tag{1.4}$$

$$\mathrm{H}_\mathrm{a} : \theta_j > \theta_c + \delta_c, \tag{1.5}$$

in which δ_c denotes the width of the indifference region. In this thesis, δ is always set equal for all decisions. The likelihood ratio after k items are administered is used as the test statistic for the SPRT (Eggen, 2010):

$$\mathrm{LR}\left(\theta_c + \delta; \theta_c - \delta\right) = \frac{\mathrm{L}\left(\theta_c + \delta; \mathbf{x}_j\right)}{\mathrm{L}\left(\theta_c - \delta; \mathbf{x}_j\right)}. \tag{1.6}$$

Decision rules are then applied to decide whether to continue testing or to make a specific classification decision:

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) < (1-\beta)/\alpha; \\ &\text{ability below } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \leq \beta/(1-\alpha); \\ &\text{ability above } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \geq (1-\beta)/\alpha, \end{aligned} \tag{1.7}$$

where α and β define the acceptable decision errors (Eggen, 1999). They can be set to be symmetric or asymmetric, and can be set per cutoff point or decision. In this thesis, these are specified to be symmetric and equal for all cutoff points. If several cutoff points are specified, the decision rules are applied for each classification.

Van Groen, Eggen, and Verschoor (2010) and Van Groen and Verschoor (2010) conducted small studies on the characteristics of the SPRT in the case of unidimensional classification testing. They found that if δ was increased, test length decreased, and if α and β were increased, tests were slightly shorter, but test length was primarily influenced by the distance between the student's ability and the cutoff point. Varying α, β, and δ had a limited effect on accuracy, as determined by the proportion of correct decisions. A finding was that if student ability was close to the cutoff point accuracy decreased toward 50% correct decisions.

Classification by the Confidence Interval Method

Unidimensional classification tests can also use Kingsbury and Weiss's (1979) confidence interval method. This method stops testing as soon as the cutoff point is outside the confidence interval. The confidence interval is calculated using the t-distribution, specified by γ, and the standard error surrounding the estimates. As soon as the cutoff point falls outside the interval, $(\hat{\theta}_j - \gamma \cdot \text{se}(\hat{\theta}_j); \hat{\theta}_j + \gamma \cdot \text{se}(\hat{\theta}_j))$, testing is stopped using the following decision rules (Eggen & Straetmans, 2000)

$$\begin{aligned} &\text{administer another item if} && \hat{\theta}_j - \gamma \cdot \text{se}(\hat{\theta}_j) < \theta_c < \hat{\theta}_j + \gamma \cdot \text{se}(\hat{\theta}_j); \\ &\text{ability below } \theta_c \text{ if} && \hat{\theta}_j + \gamma \cdot \text{se}(\hat{\theta}_j) < \theta_c; \\ &\text{ability above } \theta_c \text{ if} && \hat{\theta}_j - \gamma \cdot \text{se}(\hat{\theta}_j) > \theta_c. \end{aligned} \tag{1.8}$$

The standard error of the ability estimate is (Hambleton, Swaminathan, & Rogers, 1991)

$$\text{se}(\hat{\theta}_j) = \frac{1}{\sqrt{\text{I}(\hat{\theta}_j)}}, \tag{1.9}$$

where $\text{I}(\hat{\theta}_j)$ denotes the Fisher information available in the observable variables for the estimation of θ_j (Mulder & Van der Linden, 2009). Fisher information is given by (Tam, 1992)

$$\text{I}(\hat{\theta}_j) = \sum_{i=1}^{k} a_i^2 \text{P}_i(\hat{\theta}_j)\text{Q}_i(\hat{\theta}_j). \tag{1.10}$$

Comparison of Classifications by the Sequential Probability Ratio Test and the Confidence Interval Method

Unfortunately, no method is available to make direct comparisons between classifications using different settings for the SPRT and the confidence interval method because no mathematical proof exists for linking the settings for the two methods. If one wants to apply one of the methods, simulation studies are required to investigate the effect of different settings of the classification methods on average test length and proportion of correct decisions. Van Groen et al. (2010) performed a small comparison study for the SPRT and the Kingsbury and Weiss (1979) method for different settings. They found that the influence of the settings on the average test length was larger for the SPRT than for the other method. In that study, the average test length was also higher for the SPRT, but this finding is probably caused by the specific settings in their study. The study is too limited to conclude that the confidence interval method always outperforms the SPRT. They also found that the SPRT resulted in more accurate decisions than the confidence interval method, but this can probably be explained by the longer tests. According to Eggen and Straetmans (2000), no general preference for one of the two approaches has been established yet.

1.1.5 The Item Selection Method

A large range of item selection methods is available and used for unidimensional CAT; see, for example, Eggen, 1999; Luecht, 1996; Stocking & Swanson, 1993; Thompson, 2009; Van der Linden, 2005, and Weissman, 2007. The majority of these methods focus on CAT for estimating ability. The focus of the item selection methods for CCTs is generally on tests with just one cutoff point. Eggen and Straetmans (2000), Spray (1993), and Wouda and Eggen (2009) investigated item selection for tests with multiple cutoff points. In this thesis, the discussion is limited to the two methods that form the basis of the methods that are used in Chapter 3.

The majority of the item selection methods for unidimensional CAT are based on Fisher information (Van der Linden, 2005). Fisher information is strongly related to the standard error of the ability estimate, which implies that if information is maximized, the standard error will be minimized. The objective function for item selection then becomes

$$\max \mathrm{I}_i(\hat{\theta}_j), \quad \text{for } i \in \mathrm{V}_a, \tag{1.11}$$

where V_a denotes the set of items still available for administration. The advantage of this method is that testing is tailored to the individual student's ability.

A second popular method maximizes information at the cutoff point, instead of at the current ability estimate. This implies that most information is available at the cutoff point to make the decision. The objective function is then

$$\max I_i(\theta_c), \quad \text{for } i \in V_a. \tag{1.12}$$

Other methods select items for tests with multiple cutoff points (Eggen, 1999; Spray, 1993). The problem with the second method is that, although test length in general will be shorter, item selection is not tailored to the individual student's ability and all students have identical, although of different length, tests. Unfortunately, no methods are available that combine the measurement efficiency and accuracy of maximization at the cutoff point with the tailoring to the student's ability of the maximization at the current ability estimate.

A large range of item selection methods exists for multidimensional CAT for estimating ability (Luecht, 1996; Mulder & Van der Linden, 2009; Reckase, 2009; Segall, 1996; Veldkamp & Van der Linden, 2002; C. Wang, Chang, & Boughton, 2011; Yao, 2012, 2013), but for multidimensional CCT (MCCT), no specialized methods are available. The discussion is limited here to Segall's (1996) method.

Segall (1996) developed an item selection method for multidimensional CAT analogous to the unidimensional method that maximizes information at the ability estimate. This method maximizes the determinant of the information matrix at the ability estimate. Fisher information for a p dimensional model is given by a $p \times p$ matrix, which is defined for dimensions l and m as (Tam, 1992)

$$I(\theta_l, \theta_m) = \frac{\frac{\partial}{\partial \theta_l} P_i(\boldsymbol{\theta}) \times \frac{\partial}{\partial \theta_m} P_i(\boldsymbol{\theta})}{P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})} = a_{il} a_{im} P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta}). \tag{1.13}$$

The item is then selected that has the largest determinant, which implies that the size of the confidence ellipsoid surrounding the ability estimate is minimized (Reckase, 2009). Again, the confidence region is approximated by the inverse of the information matrix; thus, the item is selected that has the largest determinant (Segall, 1996):

$$\max \det \left(\sum_{i=1}^{k} I(\hat{\boldsymbol{\theta}}_j, x_{ij}) + I(\hat{\boldsymbol{\theta}}_j, x_{k+1,j}) \right), \quad \text{for } k+1 \in V_{k+1}, \tag{1.14}$$

which is the determinant of the information matrix of the previous items and the potential item $k+1$. The left term denotes the information provided thus far. The right term is the information that potential item $k+1$ provides.

The described item selection methods result in tests that are expected to have optimal characteristics to obtain an efficient and accurate classification decision or have optimal items to tailor the test to the student's ability. These methods, however, ignore content validity and do not place restrictions on item usage.

Content validity can be reassured by using a content control method. A simple method was implemented in Eggen and Straetmans (2000). They used the Kingsbury and Zara (1989) approach, which selects the next item from the domain for which the difference between the desired and the achieved percentage of items selected thus far was the largest. This method is easily implemented, but can be used only if a limited number of content constraints is specified. If a large number of content restrictions is specified, more complicated methods have to be used (see Van der Linden, 2005).

Item usage can be controlled using an exposure control method. Especially, overexposure is a problem because the chance of the item content becoming known increases with each additional test administration. Item underexposure is mainly a problem for test developers due to the costs involved in item development. A simplified version (Eggen & Straetmans, 2000) of the Sympson and Hetter (1985) method was used for exposure control. For each selected item, a random number g is drawn from the interval (0,1). When g is larger than a specified reference value, the item is administered; if not, the item is not admissible for the remainder of the test. Although the method appears to be elegant, it does not place restrictions on administration of the item to large groups of students with similar ability.

Adaptivity as a Result of Item Selection

By tailoring the item selection, tests can be adapted for each student. Adaptive systems attempt to be different for different students by using available information about the student (Brusilovsky & Peylo, 2003). Wauters (2012) described four dimensions of adaptivity for adaptive learning environments of which three dimensions are considered applicable to CCTs. The medium of adaptivity is relevant only for adaptive learning environments since it relates to specific types of environments.

The first dimension of adaptivity concerns the form of adaptivity. Wauters (2012) distinguished three forms of adaptivity. Adaptive form representation

adapts the way items are presented to the student. This form of adaptivity is not frequently used in CCT because adapting the way an item is presented implies that multiple sets of item parameters are required for each item. Adaptive content representation provides intelligent help on each problem-solving step that is required for solving the item. This type of adaptivity is not often seen in CCTs, but perhaps it is possible to divide items into smaller items that each cover a part of the problem-solving steps. Depending on student ability, the entire item or parts of the item can be administered to the pupil. The third form of adaptivity concerns adaptive curriculum sequencing. This form selects the optimal question in order to learn certain knowledge efficiently and effectively (Wauters, 2012). Although CCTs are typically not designed to enhance learning during the test, this form of adaptivity can be used. If a logical flow of content can be established, content control can ensure that this flow is maintained.

The second dimension of adaptivity concerns the source of adaptivity. Again, Wauters (2012) distinguished three categories. The first category of features concerns item and course features. These include item difficulty and the topic of items. The former is commonly found in CCTs because item selection is often adapted using item difficulty parameter estimates. The latter can be included in content control. The second category of features concerns person features. In adaptive learning environments, these comprise the learner's knowledge level, motivation, cognitive load, interests, and preferences (Wauters, 2012). In CCTs, only a limited set of person features is used. The most important person feature concerns the ability estimate that is used to select items with appropriate difficulty. The third category of adaptivity concerns context features such as the time when, the place from which, and device on which the student works in the environment (Wauters, 2012). The possibility of adapting these features in CCTs depends on the stakes of testing and the capabilities of the assessment software.

The third dimension of adaptivity concerns the level of adaptivity. According to Wauters (2012), this concerns whether the source of adaptivity is considered static or dynamic. CCTs tend to have a static structure because adaptivity tends to be implemented consistently in one test administration using just one item selection method. Nevertheless, multi-segment CAT (Eggen, 2013) makes it possible to create different segments into one CAT with different selection and content control methods implemented in each segment. This enables creating dynamic CCTs.

1.2 The Context and Test Environment of CCTs

Thus far, attention was paid to the components of the CCTs. Obviously, a test is always administered within a certain context and within a specific testing environment. Two elements of that context are discussed. CCT is one of the many types of digital assessment. Some types are discussed in the next part of this section. In addition, test administration always takes place within a specific approach to assessment with specific consequences based on the test's results. Four approaches will be discussed in the second part of this section. The modules of a test environment for CCTs will be discussed in the third part of this section.

1.2.1 Digital Assessments

A large range of types of digital assessment exists besides CCTs. Six types of digital testing are discussed here including their ability to make classification decisions: linear tests (LTs), automatically generated tests (AGTs), computerized adaptive testing for estimating ability (CAT-E), adaptive learning environments (ALEs), educational simulations (ESs), and educational games (EGs).

The first type is linear testing. Test content, item order, and test length are the same for all students. Item selection is fixed before test administration, which implies that testing is not tailored to the individual student (Mellenberg, 2011). LTs can be used to estimate ability but also to make classification decisions.

The second type is automatically generated testing. Tests are assembled before administration using a set of test constraints and conditions (Parshall, Spray, Kalohn, & Davey, 2002). AGTs can also be used to make decisions.

The third type is computerized adaptive testing for estimating ability. This type of testing has already been discussed. The main difference between CAT-E and CCTs concerns the reported outcome: an ability estimate or a classification decision. CAT-E can be used to make classification decisions by setting a cutoff point at the ability scale and make a comparison between the estimated ability and the cutoff point.

The fourth type is adaptive learning environments. These systems optimize instruction to each student's individual needs, preferences, or context (Wauters, 2012). Typically, the focus is on providing instruction to the student as opposed to making a judgment based on the student's responses. ALEs can be used for making classification decisions in low-stakes test situations if IRT is used.

The fifth type is educational simulations. Educational simulations can be used for simulating real-world events. They typically report multiple aspects in proficiency for a wide range of abilities and skills.

The sixth type is educational gaming. EGs can facilitate learning but also keep the learner motivated and engaged (Novak, Johnson, Tenenbaum, & Shute, 2014). They can report real-time estimates of competencies across a range of skills and knowledge (Mislevy et al., 2014).

1.2.2 Test Approaches

Thus far, the components and modules of CCTs and several other types of digital assessments have been described. Administration of a CCT takes place with a certain goal in mind. Four types of assessment approaches will be discussed next, and the possibility of using CCTs in those contexts is explored.

The first approach is formative assessment. Formative assessment attempts to support and improve the learning process by making decisions at the level of the learner or the class (Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2013). CCTs can be used for formative assessment because their efficiency make it possible to assess a relatively large set of test domains while keeping the test length acceptable.

The second approach is formative evaluation. Formative evaluation focuses on making judgments about the school for developing educational policies within the school (Van der Kleij et al., 2013) and for improving education (Scheerens, Glas, & Thomas, 2003). CCTs can be used to make judgments about the school because the school can specify cutoff points that are related to the school goals. Using CCTs, the percentage of students who master the subject can be monitored over time, or before and after the innovation.

The third approach is summative assessment. This approach focuses on what has been learned by the end of the testing process (Stobart, 2008). A decision is then made about the student's mastery of a content domain (Van der Kleij et al., 2013). The focus of a CCT is on defining whether a student masters a topic; thus a CCT can make very efficient and accurate decisions for summative assessment.

The fourth approach is summative evaluation. In this approach judgments are made about the school (Van der Kleij et al., 2013) or about educational systems. CCTs can be used to make such judgments if individual test results are aggregated to the intended level. The percentage of students who master a topic or the decrease in the percentage of students who did not master the topic before and after intervention can be reported.

1.2.3 The Modules of a Test Environment for CCTs

A CCT is administered within a test environment. The test environment administers the test and keeps a record of all the information required for the digital administration of the test. Nwana (1990) and Wauters (2012) distinguished four modules of adaptive learning environments that are considered applicable to test environments for CCTs. These modules (student, tutor, knowledge, and user interface) will be described next.

In an adaptive learning environment, the student module contains all information concerning the student, such as the student's current knowledge level, student characteristics, and learning style (Wauters, 2012). According to Nwana (1990), the module forms a representation of the student's current knowledge with respect to the mastery of the knowledge in the domain module. This information can be used to select the items (Wauters, 2012).

The tutor module answers the learner's questions about goals and content, decides when a student needs help, and selects the items and tasks (Wauters, 2012). This module tries to enhance learning and tailors the test and instruction to the individual student's ability and characteristics.

In an adaptive learning environment, the knowledge module contains the knowledge the student is trying to acquire and the relationships between the knowledge elements (Paramythis & Loidl-Reisinger, 2004; Wauters, 2012). This implies that the module contains all information about the items, their item parameters, and content characteristics.

The user interface controls the interactions between the student and the testing system (Nwana, 1990), displays items, and retrieves the student's responses. A good user interface demonstrates consistency and clarity and reflects good interface design principles (Parshall et al., 2002). Although the user interface is the only part of the test that the student actually sees and interacts with (Parshall et al., 2002), it was not mentioned as a component of CCTs. The user interface was not included because it has limited influence on test content and test construction. The capabilities of the software that is used for developing the user interface can place restrictions on test content, but discussing those limitations falls outside the scope of this thesis.

1.3 Characteristics of CCTs and Their Availability

Thus far, a short overview was provided about computerized adaptive tests for making classification decisions. The introduction began with the components that should come together in a CCT to make efficient and accurate classification decisions: the student, the items, the item response theory model, classification methods, and item selection methods. The characteristics of CCTs that were investigated for this thesis will be discussed here.

When the research for this thesis started, computerized adaptive tests could be used only to obtain unidimensional ability estimates, to make unidimensional classification decisions, or to obtain multidimensional ability estimates, but not to make multidimensional classification decisions. While the research was being conducted, two methods became available for between-dimensional classification decisions (Seitz & Frey, 2013a, 2013b). Even with these two manuscripts, much more research can be conducted for multidimensional classification testing.

A CCT classifies students into one of two or more classification categories. For unidimensional CCTs, approaches for classifying into one of two and one of several categories are available for the SPRT (Eggen & Straetmans, 2000; Spray, 1993) and the confidence interval method (Eggen & Straetmans, 2000; Kingsbury & Weiss, 1979). The possibility of using a CCT for multiple multidimensional classifications with between- and within-dimensionality will be explored.

In the case of between-dimensionality, (Seitz & Frey, 2013a, 2013b) showed that the SPRT can be used to make a decision per dimension. If a decision is required on all dimensions simultaneously, an additional decision rule has to be used to combine the decisions per dimension in a decision on the entire test. It would be interesting to investigate whether a different solution is possible.

Spray, Abdel-Fattah, Huang, and Lau (1997) concluded that it was not possible to use the SPRT. Nevertheless, it could be interesting to explore the possibility to make multidimensional classification decisions with within-dimensionality.

Currently, two major types of item selection methods for CCTs select the item that has the most information at the current ability estimate or at the cutoff point. The former tailors item selection to the student's ability estimate; the latter results in the most accurate and efficient classifications. Ideally, a compromise between both would be used because that would result in accurate and efficient decisions while tailoring item selection to the student's ability. Such approaches to item selection could form an interesting topic for further study.

1.4 Research Questions and Thesis Outline

In the previous sections of this introduction, some of the topics that will be addressed have been mentioned. Four research questions were explored in this thesis. The answers to the research questions will be provided in the concluding epilogue of this thesis.

How can item selection by maximizing information at the cutoff point(s) and maximization at the ability estimate be combined to obtain accurate and efficient classification decisions, and tailoring item selection to the student's ability?

Currently existing item selection methods were previously discussed as a component of CCTs. The discussed methods for unidimensional classification testing select the item that provide the most information at the current ability estimate or at the cutoff point(s). Several methods will be described in Chapter 2 that take both the ability estimate and the cutoff points into account. One of the methods in Chapter 2 is extended for between-dimensionality in Chapter 3 and for within-dimensionality in Chapter 5.

How can multidimensional classification decisions be made on all dimensions simultaneously for tests with between- and within-dimensionality?

At the start of this research project, no classification methods were available for multidimensional IRT. During the project, two methods were developed for between-dimensionality (Seitz & Frey, 2013a, 2013b). Unfortunately, these methods cannot be used to make decisions on more than one dimension simultaneously. A method to make decisions on the entire test, but also on subtests, is discussed in Chapter 3. In Chapters 4 and 5, two methods to make multidimensional decisions are described for tests with within-dimensionality. The SPRT is applied in Chapter 4 and the confidence ellipsoid method in Chapter 5.

How can items be selected in tests with between- and within-dimensionality so that accurate and efficient decisions can be made?

To make accurate and efficient multidimensional classification decisions, items have to be selected with optimal characteristics. A method for selecting the items that have the largest determinant of the information matrix is available for MCAT for estimating ability (Segall, 1996). Since no classification method was available, no item selection methods were developed to make multidimensional classification

decisions. Such an item selection method is developed in Chapter 4 for within-dimensionality. In Chapters 3 and 5, an item selection method for between- and for within-dimensionality is described that takes both the cutoff point and the ability estimate into account (see research question 1).

In which contexts can computerized classification testing be used, and how should the test be designed in those contexts?

A small overview of assessment contexts was provided in earlier sections. The design, usability for different test approaches, and adaptivity is explored in Chapter 6 for different types of digital assessments. The types of digital assessments are linear testing, automatically generated testing, computerized adaptive testing for estimating ability, computerized classification testing, adaptive learning environments, educational simulations, and educational gaming.

The chapters in this thesis were written to be self-contained. Therefore, some overlap could not be avoided.

References

Birnbaum, A. (2008). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Charlotte, NC: Information Age. (Original work published 1968)

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, *13*(2-4), 156–169.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8

Eggen, T. J. H. M. (2013, October). *Computerized adaptive testing serving educational testing purposes*. Paper presented at the meeting of the IAEA, Tel Aviv, Israel.

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. doi: 10.1177/00131640021970862

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computerassisted branded test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh, PA.

Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442–463. doi: 10.3102/1076998607308573

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi: 10.1111/jedm.12000

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategie for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, *2*, 359–375. doi: 10.1207/s15324818ame0204_6

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a

certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404. doi: 10.1177/014662169602000406

Mellenberg, G. J. (2011). *A conceptual introduction to psychometrics*. Den Haag, the Netherlands: Eleven International.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., ... John, M. (2014). *Psychometric considerations in game-based assessment.* Redwood City, CA: GlassLab Research, Institute of Play.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62. doi: 10.1207/S15366359MEA0101_02

Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296. doi: 10.1007/S11336-008-9097-5

Novak, E., Johnson, T. E., Tenenbaum, G., & Shute, V. J. (2014). Effects of an instructional gaming characteristic on learning effectiveness, efficiency, and engagement: Using a storyline for teaching basic statistical skills. *Interactive Learning Environments*. doi: 10.1080/10494820.2014.881393

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, *4*(4), 251–277. doi: 10.1007/BF00168958

Paramythis, A., & Loidl-Reisinger, S. (2004). Adaptive learning environment and e-learning standards. *Electronic Journal of e-Learning*, *2*(1), 181–194.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. doi: 10.1177/014662168500900409

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

Scheerens, J., Glas, C. A. W., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring*. London, United Kingdom: Taylor & Francis.

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

Stobart, G. (2008). *Testing times: The uses and abuses of testing*. London, United Kingdom: Routledge.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277–292. doi: 10.1177/014662169301700308

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *In: Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

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*, *69*, 778–793. doi: 10.1177/0013164408324460

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Van der Kleij, F. M. (2013). *Computer-based feedback in formative assessment*. Unpublished doctoral dissertation, Twente University, Enschede, the Netherlands.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2013). Data-based decision making, assessment for learning, and diagnostic testing in formative assessment. In F. M. Van der Kleij (Ed.), *Computer-based feedback in formative assessment* (pp. 155–169). Unpublished doctoral dissertation, Twente University, Enschede, the Netherlands.

Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer. doi: 10.1007/0.387.29054.0

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*, *38*, 187–200. doi: 10.1177/0146621613509723

Van Groen, M. M., Eggen, T. J. H. M., & Verschoor, A. J. (2010, May). *Adaptive classification tests.* Paper presented at the Onderwijs Research Dagen [Educational Research Days], Enschede, the Netherlands.

Van Groen, M. M., & Verschoor, A. J. (2010, June). *Using the sequential probability ratio test when items and respondents are mismatched*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.

Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. doi: 10.1007/BF02295132

Wald, A. (1973). *Sequential analysis*. New York, NY: Dover. (Original work published 1947)

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

Wang, C., Chang, H.-H., & Boughton, K. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99–122. doi: 10.1007/S11336-011-9215-7

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295–316. doi: 10.1177/0146621604265938

Wauters, K. (2012). *Adaptive item sequencing in item-based learning environments*. Unpublished doctoral dissertation, KU Leuven, Belgium.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41–58. doi: 10.1177/0013164406288164

Wools, S., Eggen, T. J. H. M., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63–82. doi: 10.3280/CAD2010-001007

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*, *77*, 495–523. doi: 10.1007/S11336-012-9265-5

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3–23. doi: 10.1177/0146621612455687

Chapter 2
Item Selection Methods Based on Multiple Objective Approaches for Classifying Examinees into Multiple Levels

Abstract

Computerized classification tests classify examinees into two or more levels while maximizing accuracy and minimizing test length. The majority of currently available item selection methods maximize information at one point on the ability scale, but in a test with multiple cutting points, selection methods could take all these points simultaneously into account. If one objective is specified for each cutoff point, the objectives can be combined into one optimization function using multiple objective approaches. Simulation studies were used to compare the efficiency and accuracy of eight selection methods in a test based on the sequential probability ratio test. Small differences were found in accuracy and efficiency between different methods depending on the item pool and settings of the classification method. The size of the indifference region had little influence on accuracy, but considerable influence on efficiency. Content and exposure control had little influence on accuracy and efficiency.

This chapter was published as Van Groen, M.M., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement, 38(3)*, 187-200.

2.1 Introduction

Originally, computerized adaptive tests (CATs) were developed for obtaining an efficient estimate of an examinee's ability, but Weiss and Kingsbury (1984), Lewis and Sheehan (1990), and Spray and Reckase (1994) showed that CATs can also be used for classification problems (Eggen & Straetmans, 2000). In these computerized classification tests (CCTs), the main interest is not in obtaining an estimate, but in classifying the examinee into one of multiple categories (e.g., pass/fail or master/nonmaster). CCT can be used to find a balance between the number of items administered and the level of confidence in the correctness of the classification decision (Bartroff, Finkelman, & Lai, 2008). In CCT, the administration of additional items stops when enough evidence is available to make a decision. As in Eggen and Straetmans (2000), the focus in the current study is on classifying examinees into one of three (or even more) categories.

In adaptive classification testing, item selection is based on the examinee's previous responses, which tailors the item selection to the examinee's ability. Several item selection methods are described in the literature (see for example Eggen, 1999; Thompson, 2009). The design of the item selection method partly determines the efficiency and accuracy of the test (Thompson, 2009). Current methods select items based on one point on the scale and are often not adaptive in selecting items. However, if several cutoff points are specified, gathering as much information as possible at all cutoff points while considering the examinee's ability may be desirable. By doing so, information is gathered throughout a larger part of the ability scale. Especially at the beginning of the test, uncertainty exists about the ability of the examinee, which implies that gathering information at a range of points on the scale would be beneficial.

The study is organized as follows. First, details are given regarding computerized classification testing. Then some of the current and newly developed item selection methods are described. The performance of the methods was compared using simulation studies. The final section of this article gives concluding remarks.

2.2 Classification Testing

Computerized classification testing can be used if a classification decision has to be made about the level of an examinee in a certain domain. CCT was used to place students in one of three mathematics courses of varying difficulty in the Netherlands (Eggen & Straetmans, 2000), but can also be used if a decision

such as master/nonmaster is required. An advantage of classification testing is that shorter tests can be constructed, while maintaining the desired accuracy (Thompson, 2009). Reducing the number of items is important because the testing time is reduced, fewer items have to be developed, security problems are reduced, and item pools have to be replenished less often (Finkelman, 2008). Adaptive classification testing shares with CAT the advantage of adapting the test to the examinee's ability. This possibly reduces the examinee's frustration because fewer too easy or too hard items are administered and a larger set of items is selected from the item pool. However, examinees can experience that the items in a CAT are difficult (Eggen & Verschoor, 2006) when compared to a regular test in which an able student answers only relatively easy items. This drawback of CAT as well as CCT was overcome by Eggen and Verschoor (2006) by selecting relatively easy items. CCT also shares the drawback with CAT that examinees cannot change answers to previously administered items.

One part of the CCT procedure determines whether testing can be stopped and a decision can be made before the maximum test length is reached. Popular and well-tried methods are based on the sequential probability ratio test (SPRT). The SPRT (Wald, 1947/1973) was first applied to classification testing by Ferguson (1969) using classical test theory and by Reckase (1983) using item response theory (IRT). The SPRT has been applied to CAT and multistage testing (Luecht, 1996; Mead, 2006; Zenisky, Hambleton, & Luecht, 2010). Other available methods (Thompson, 2009) are not considered in this study.

In CCT, a cutoff point is specified between each pair of adjacent levels. The indifference regions are set around these points, which account for the uncertainty of the decisions due to measurement error, regarding examinees with ability close to the cutoff point (Eggen, 1999). If multiple cutoff points are specified, it would be strange if the indifference regions of different cutoff points overlapped. Overlapping indifference regions imply that classification into one of three levels is admissible for examinees with an ability within the overlapping regions and that uncertainty exists about decisions regarding the three levels. In this situation, test developers should reconsider the number of cutoff points and the size of the indifference regions. However, in practice, this is not always possible.

To apply the SPRT to a classification problem, two hypotheses are formulated for each cutoff point θ_c ($c = 1, \cdots, C$), based on the boundaries of the accompanying indifference region (Eggen, 2010):

$$\mathrm{H}_0 : \theta_j < \theta_c - \delta_{c1}; \tag{2.1}$$

$$\mathrm{H}_a : \theta_j > \theta_c + \delta_{c2}, \tag{2.2}$$

in which θ_j denotes ability for examinee j and $\delta_{c.}$ the widths of the indifference regions. These are set equal to δ. To avoid overlapping indifference regions, δ should be smaller than half the difference between adjacent cutoff points.

Item responses are modeled using IRT, in which a relation is specified for the score on an item depending on item parameters and the examinee's ability (Van der Linden & Hambleton, 1997). The relationship between a specific score on an item ($x_i = 1$ correct, $x_i = 0$ incorrect) and an examinee is modeled with a probability function. The model used here is the two-parameter logistic model (Birnbaum, 1968/2008), in which the probability of a correct response is given by

$$\mathrm{P}\left(x_i = 1|\theta\right) = \frac{\exp\left(a_i\left[\theta - b_i\right]\right)}{1 + \exp\left(a_i\left[\theta - b_i\right]\right)} = \mathrm{P}_i\left(\theta\right), \tag{2.3}$$

where a_i represents the discriminating power of item i, b_i difficulty, and θ ability.

A prerequisite for CCT is a calibrated item bank that is suitable for the specific testing situation. In a calibrated item bank, the fit of the model is established, and estimates of the item parameters are available, with items with inappropriate difficulty or low discrimination parameters removed.

In IRT, the probability of an examinee's responses to test items is conditionally independent given the latent ability parameter. Inference about the ability of an examinee can be drawn from the likelihood of the responses after k items are administered (Eggen, 1999) using

$$\mathrm{L}\left(\theta_j; \mathbf{x}_j\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\theta_j\right)^{x_{ij}} \left[1 - \mathrm{P}_i\left(\theta_j\right)\right]^{1-x_{ij}}, \tag{2.4}$$

in which $\mathbf{x}_j = (x_{1j}, \cdots, x_{kj})$ denotes the vector of responses to the administered items for examinee j.

When the SPRT is applied to classification testing, the likelihood ratio of both hypotheses after k items are administered (Eggen, 2010) is the test statistic:

$$\mathrm{LR}\left(\theta_c + \delta; \theta_c - \delta\right) = \frac{\mathrm{L}\left(\theta_c + \delta; \mathbf{x}_j\right)}{\mathrm{L}\left(\theta_c - \delta; \mathbf{x}_j\right)}. \tag{2.5}$$

Decision rules are applied to make the decision to continue testing or to make the decision that ability is at a level below or above the specific cutoff point:

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) < (1-\beta)/\alpha; \\ &\text{ability below } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \leq \beta/(1-\alpha); \\ &\text{ability above } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \geq (1-\beta)/\alpha, \end{aligned} \tag{2.6}$$

where α and β are small constants that specify acceptable decision errors (Eggen, 1999). In practice, a maximum test length is set to ensure that testing stops at some point. If the maximum test length is reached, the examinee is classified as performing above the cutoff point if the likelihood ratio is larger than the midpoint of the interval in Equation 2.6. If multiple cutoff points are specified and the decision is made that the ability is above cutoff point θ_c, the same procedure is applied for cutoff point θ_{c+1}.

If δ increases, the difference between the likelihoods is larger and thus, more uncertainty is allowed to make the decision, which implies less accurate decisions and shorter tests. Eggen (1999) found for the situation with one cutoff point that increasing α and β had little effect on the proportion of correct decisions, but increasing δ influenced classification accuracy.

2.3 Current Item Selection Methods

Several item selection methods can be used in CCT (see Eggen, 1999; Luecht, 1996; Stocking & Swanson, 1993; Thompson, 2009). Most methods were developed for tests with classification into one of two levels, but a few methods were proposed for tests with more levels. The majority of item selection methods are based on Fisher information (Van der Linden, 2005). Other types of information such as Kullback-Leibler information (Eggen, 1999) and mutual information (Weissman, 2007), can also be used but are not included here. If maximizing Fisher information is the objective, the optimization function becomes

$$\max \mathrm{I}_i(\theta), \quad \text{for } i \in \mathrm{V}_a, \tag{2.7}$$

where V_a denotes the set of items still available for administration. In Equation 2.7, the $k+1$th item is selected that has the most information I_i:

$$\mathrm{I}_i(\theta) = a_i^2 \mathrm{P}_i(\theta)\left[1 - \mathrm{P}_i(\theta)\right] \tag{2.8}$$

A method currently used maximizes information at the ability estimate $\hat{\theta}_j$. The accuracy of this estimate is related to the number of items available for estimation (Hambleton, Swaminathan, & Rogers, 1991), which causes the method to select items that are potentially not optimal at early stages of the test. The advantage of this method is that items are selected adaptively to the examinee.

A number of methods exists for classification tests with more than two levels (Eggen & Straetmans, 2000; Wouda & Eggen, 2009). Maximization of test information at the middle between the cutoff points and at the nearest cutoff point (Spray, 1993) are just two approaches (Eggen & Straetmans, 2000). The first method determines the middle of the cutoff points closest to the current estimate and maximizes information at that point. The second method optimizes at the cutoff point located nearest the ability estimate. Both methods base their item selection on the ability estimate, which is considered an advantage in educational settings.

As Weissman (2007) concluded, choosing an item selection method in conjunction with the SPRT is not straightforward. Spray and Reckase (1994) concluded that maximizing information at the cutoff score, for classifying into one of two levels, results, on average, in shorter tests than selecting items at the current ability estimate. Thompson (2009), however, concluded that this method is not always the most efficient option. Wouda and Eggen (2009) compared methods that maximize information at the middle of the cutoff points, at the nearest cutoff point, and at the ability estimate using simulations and found for the situation with two cutoff points that maximization at the middle of the cutoff points resulted in the most accurate, but also in the longest, tests.

The methods described thus far all select items based on some optimal statistical criterion. In practical testing situations, however, item exposure and content control also have to be considered. Item overexposure can be a security concern and several methods have been developed for dealing with it (for example, see Sympson & Hetter, 1985). Content control mechanisms can ensure that the assembled tests

meet test specifications; for example, 10 items should measure Domain A and at least 12 Domain B. For an extended overview of methods that deal with content restrictions, see Van der Linden (2005).

The described methods maximize information at one point on the latent scale using the ability estimate. One objective is formulated using this point and this objective is optimized. An alternative approach is to maximize information on all cutoff points simultaneously. If an objective is formulated for each cutoff point, they can be combined using multiple objective approaches. These approaches combine several objectives into one objective function. The developed methods all take the ability estimate into account. The advantage of these approaches is that measurement is more precise at all cutoff points. Multiple objective approaches were used for optimal test design in multidimensional testing (Veldkamp, 1999) and exposure control in CAT (Veldkamp, Verschoor, & Eggen, 2010).

2.4 Item Selection Based on Multiple Objective Approaches

Veldkamp (1999) described six approaches for combining multiple objectives: weighting methods (WM), ranking or prioritizing methods, goal programming (GP) methods, global-criterion (GC) methods, maximin (MA) methods, and constraint-based methods. These approaches are adapted for classification testing with multiple cutoff points. The methods are first described and then adapted.

2.4.1 Weighting Methods

A straightforward method for optimizing several objective functions involves combining them into one objective function to which weights can be added to give various objectives varying importance (Veldkamp, 1999). The weighted deviation model (Swanson & Stocking, 1993), in which the deviations from the goal values are combined using weights, is probably the most well-known application of WM to test construction problems. Instead of weighting the deviations, weighting is applied here to the objectives. The decision between Levels 1 and 2 could be considered to be more important than the decision between Levels 2 and 3. Specification of different weights for the two decisions ensures that more information is gathered at the first cutoff point. Varying weights while administering the test gives more weight to specific objectives at various testing stages. In this study, the weight for a specific cutoff point increases if the ability estimate is closer to

the cutoff point. This implies that item selection is adapted to the ability of the examinees. The resulting objective function is

$$\max \sum_{c=1}^{C} \frac{1}{|\hat{\theta}_j - \theta_c|} \mathrm{I}_i(\theta_c), \quad \text{for } i \in \mathrm{V}_a. \tag{2.9}$$

2.4.2 Ranking or Prioritizing Methods

If certain objectives are more important than others, ranking or prioritizing methods can be used. Ranking methods require all objectives to be ranked according to their perceived importance (Ignizio, 1982). In the first step, the most important objective is optimized. In the second step, a constraint is added that ensures that the value of the first objective remains close to the target value obtained and the second objective is optimized. This process continues until all objectives are optimized (Veldkamp, 1999). In most tests using the SPRT, ranking or prioritizing methods cannot be used because no differences in the importance of certain cutoff points are specified.

2.4.3 Goal Programming

In both methods discussed so far, the goal was to find the optimal solution. GP methods focus on achieving specific target values (Veldkamp, 1999). However, achieving all target values specified a priori is not always possible. In those situations, the preferred solution is calculated (Veldkamp, 1999). GP methods minimize the deviations between what was aspired to and what is actually accomplished (Ignizio, 1982). The combined objective function specifies the deviations from the targets and the priorities for achieving each objective (Mollaghasemi & Pet-Edwards, 1997). Several goal function approaches are described in the literature such as Van der Linden's (2005) framework for optimal test design and the normalized weighted absolute deviation heuristic (Luecht, 1996).

Veldkamp (1999) proposed that, in the absence of prespecified targets, the test assembler starts with an intuitive guess and uses the procedure iteratively. In CCT, no targets are available for information, but gathering as much information as possible is preferred. One possibility is to compute the sum of the information each available item can provide at each cutoff point and at the current ability estimate. The item with the largest sum is then selected. Weights can be added before calculating the sum. The resulting objective function then becomes what Luecht (1996) calls a composite objective function,

$$\max \sum_{s \in V_s} w_s I_i(\theta_s), \quad \text{for } i \in V_a, \tag{2.10}$$

where w_s denotes the weight for scale point s and $V_s = \{\theta_1, \cdots, \theta_C, \hat{\theta}_j\}$. In this study, all cutoff points were considered equally important so all weights were set equal. However, an utility function can be used to weight the objectives as well as weights can be based on policy decisions.

2.4.4 Global-Criterion Methods

GC methods optimize all objectives separately and combine the results into one global criterion. First, all objectives are optimized resulting in optimal values for every objective (Veldkamp, 1999). Second, the results are combined into a global criterion. The value of this global criterion is then optimized. The method for combining the results is specified a priori. When this method is applied to CCT, the first step is to optimize the objectives for each cutoff point separately. One possibility is to consider the items that provide the most information at the different cutoff points and the current ability estimate. Several combination methods are possible. The separate objectives were combined by calculating the sum of the information at the cutoff points for all items that provide the most information at one of the selected points. The combined objective then becomes

$$\max \sum_{c=1}^{C} I_i(\theta_c), \quad \text{for } i \in V_{max}, \tag{2.11}$$

where V_{max} denotes the set of available items, which provide the most information at one of the cutoff points or the current ability estimate. This differs from the previous method in that the optimal values for the objectives are combined and then the global optimum is used instead of using non-optimal values and then combining them into the goal function. Weights have not been included in this study, but can be added.

2.4.5 Maximin Methods

MA methods can be used if a maximum value has to be found on multiple points (Boekkooi-Timminga, 1989). A lower boundary is set on the target of the objectives. This boundary is then maximized (Van der Linden, 2005). If the objectives are on the same scale, the method ensures that unexpected extreme values for one or more objectives do not occur.

A starting value must be found for the boundary. This value should be low enough to ensure feasibility and high enough so that the calculations do not consume unreasonable amounts of time. In CCT, a lower boundary can be set on the information at the cutoff points and the ability estimate provided by the items that were administered thus far. The item was selected that maximized the boundary.

2.4.6 Constraint-Based Methods

Constraint-based methods require prioritizing the objectives (Veldkamp, 1999). One objective is optimized, and the other objectives are reformulated into constraints. Additional constraints can be added to specify other test characteristics such as the amount of testing time and content control using, for example Van der Linden's (2005) framework. In CCT, using a constraint-based method implies that one cutoff point is considered the most important. This cutoff point is formulated as an objective, and constraints are formulated for the remaining cutoff points. In the present study, no cutoff point was considered to be most important, so this method was not applied here.

2.5 Simulation Studies

Using simulation studies, the average test length (ATL) and classification accuracy were investigated for the various item selection methods. Classification accuracy was defined as the proportion of correct decisions (PCD). The influence of the size of the indifference region on the ATL and PCD is investigated in the second part. Simulations with different values for α and β are not reported here, because increasing the acceptable error rate had little effect on the PCD. This is in line with Eggen's (1999) findings; he reported the same finding for simulations with two cutoff points. The effects of content constraints and exposure control were investigated in the last section.

Methods based on WM, GP, GC, and MA were included in the simulation studies. Three existing item selection methods were also included: selecting the item that maximizes information at the current ability estimate (AE), selecting the item that maximizes information at the middle of the nearest set of cutoff points (MC; Eggen & Straetmans, 2000), and selecting the item that maximizes information at the nearest cutoff point (NC; Spray, 1993). Random item selection (RA) was included to serve as a baseline for the ATL and the PCD.

The characteristics of the item pool were expected to influence the ATL and the PCD of the eight item selection methods. Two item pools were investigated. The first pool was simulated with item parameters generated with $a \sim N(1.50, 0.50)$ with $a>0$ and $b \sim U(-3.00, 3.00)$. One thousand items were generated for the item pool, and maximum test length was set at 40 items. The specifications of this pool result into a rather "ideal" situation. One thousand examinees were randomly drawn from $\theta \sim N(0.00, 1.00)$. This was replicated 100 times for each item selection method. The ATL and PCD strongly depend on the number of defined cutoff points; thus, simulations using two, three, and four cutoff points are presented here. The cutoff points were set at the 33th, 66th percentiles of the population distribution for two cutoff points, 25th, 50th, 75th percentiles for three cutoff points and 20th, 40th, 60th, 80th percentiles for four cutoff points. In the study, $\delta = 0.10$, and $\alpha = \beta = 0.05$.

A second item pool consisted of 250 items from an operational test. The parameters of the items in the pool are from a mathematics test for adult education in the Netherlands (Eggen & Straetmans, 2000). The test was used to place students in one of three courses. Using a standard setting procedure, the cutoff points were set at -0.13 and 0.33. Testing was stopped after 40 items or less to ensure comparability with the simulations with the simulated item pool. The acceptable error rates (α, β) were set at 0.05 and δ at 0.10. The distribution of θ for generating examinee ability was set equal to the estimated population distribution. One thousand examinees were simulated with $N \sim (0.294, 0.522)$. The items had a mean item difficulty of 0.00, and the mean discrimination was 3.09. The simulations were executed for the eight item selection methods and were replicated 100 times.

2.5.1 Simulations with a Simulated Item Pool

The results for the simulated item pool simulations are summarized in Table 2.1. The ATL and the PCD are provided in the table. First, the table clearly indicates that on average at least 32 items were required before tests were terminated by the SPRT if two cutoff points were specified. Second, the PCD was just above the specified accuracy level. The differences in the PCD were rather small between the different item selection methods. However, the random method was 8% less accurate than the other methods.

Depending on the item selection method, almost 35 items were required with three cutoff points. AE was the most efficient method. The differences in the PCD were also small, but random item selection resulted, in an additional 8% more

Table 2.1
Results from Simulations with a Simulated Item Pool

	Two CP		Three CP		Four CP	
Item selection method	ATL	PCD	ATL	PCD	ATL	PCD
RA	39.533	0.820	39.776	0.745	39.868	0.676
AE	32.646	0.906	34.861	0.866	35.938	0.826
MC	32.694	0.902	34.989	0.862	36.038	0.827
NC	32.721	0.908	35.009	0.867	36.170	0.828
WM	34.153	0.907	37.201	0.867	38.359	0.830
GP	33.065	0.907	37.296	0.863	39.364	0.818
GC	35.602	0.902	39.275	0.855	39.961	0.809
MA	33.259	0.902	36.856	0.853	38.444	0.798

Note. CP = cutoff points; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutoff points; NC = nearest cutoff point; WM = method based on weighting; GP = goal programming; GC = global criterion; MA = maximin.

incorrect decisions. A minimum of 85% of the examinees were classified correct using one of the other item selection methods.

Even fewer tests were terminated by the SPRT with four cutoff points. AE, MC, and NC resulted in the lowest ATL. Depending on the method, 80-83% of the classifications were correct. RA classified accurately in 68% of the tests. A comparison of the simulations using two, three, and four cutoff points showed that more items were required before testing was terminated if more cutoff points were specified. In addition, the currently used methods tended to require fewer items if the number of cutoff points was higher. In addition, the classification accuracy decreased by 2% to 4% when an additional cutoff point was set. Specifying more cutoff points also implied that the multiple objective methods have to take more cutoff points into account, which could have resulted in longer tests if the distances between the current ability estimate and the cutoff points increased.

2.5.2 Simulations with the Mathematics Item Pool

The ATL was much smaller in the simulations with the mathematics pool (Table 2.2). RA was clearly outperformed by the other methods. Eleven additional items were required before a classification was made if RA was used instead of AE. The shortest tests were produced by AE, NC, GP, and WM.

Table 2.2
Results from Simulations with a Mathematics Item Pool

Item selection method	Average Test Length	Proportion of Correct Decisions
RA	31.599	0.875
AE	20.213	0.915
MC	20.666	0.912
NC	20.593	0.916
WM	20.923	0.917
GP	20.831	0.914
GC	22.571	0.909
MA	22.514	0.908

Note. ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutoff points; NC = nearest cutoff point; WM = method based on weighting; GP = goal programming; GC = global criterion; MA = maximin.

The proportion of correct decisions was the highest for WM. NC, AE, GP, and MC had a slightly lower PCD. RA resulted in the lowest PCD. Most methods classified more accurately than was specified by α and β.

2.5.3 Simulations with Various Delta Values

To investigate the effect of the size of the indifference region on the ATL and the PCD, the simulations were repeated with different values for δ. The effect of the size of the indifference region was investigated for the simulated and the mathematics item pool. The investigation of δ was limited to the range 0.050 to 0.400 for the simulations with the simulated pool and to the range 0.025 to 0.225 for the simulations with the mathematics pool. The simulations with the simulated item pool were performed with two cutoff points. As described previously, setting δ to larger values does not make any sense if that implies that the indifference regions of different cutoff points will overlap. The results are displayed for RA, AE, WM, and GC. The results of the methods that were not included here, were similar to the results of the presented methods, except for RA.

Simulations with a Simulated Item Pool

The PCD for the simulations with a simulated item pool is displayed in the left part of Figure 2.1. The difference in the PCD appeared to be related to δ. If δ was set at rather large values, the PCD decreased. The results also indicated that the

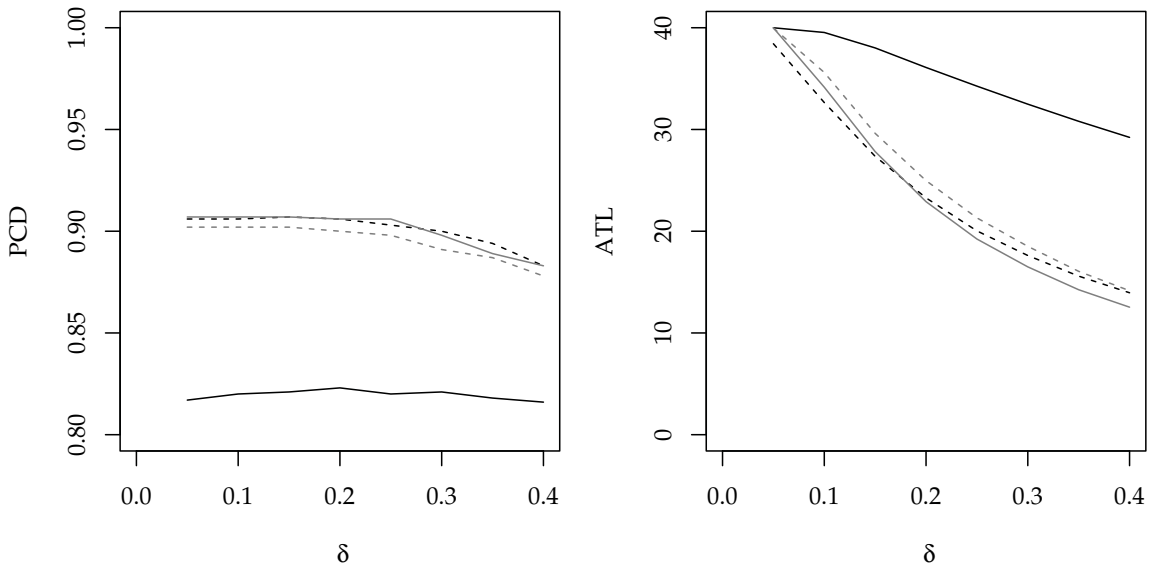


Figure 2.1. Results from simulations with a simulated item pool for different sizes of the indifference region.
Note. The solid black line denotes random item selection, the solid grey line method based on weighting, the dotted lack line ability estimate, and the dotted grey line global criterion. PCD = proportion of correct decisions; ATL = average test length; δ = width indifference region.

difference in accuracy between selecting items at random and other item selection methods was only slightly influenced by the value of δ.

The ATL of the simulations with a simulated item pool is plotted in the right part of the figure for different values of δ. The number of items administered before a classification decision was made, was clearly influenced by δ. ATL decreased if the δ was increased. The ATL decreased with 11 items if RA was used, but ATL decreased with up to 27 items if a different method was used.

Simulations with the Mathematics Item Pool

The PCD for the simulations with the mathematics item pool is displayed in the left part of Figure 2.2. In contrast to the simulations with a simulated item pool, the PCD decreased if the size of the indifference region was increased to 0.10. The difference in the PCD was rather small between different values of δ, but depending on the item selection method, if δ was set higher than 0.175, the PCD dropped below the desired accuracy level as specified by α and β. If RA was used, the PCD was for all investigated values of δ below the desired accuracy level.

As seen in Table 2.2, the choice for an item selection method influenced the PCD, but Figure 2.2 indicates that this also held for different values of δ. In the right part of Figure 2.2, the ATL is shown. The ATL decreased to 10 or 11

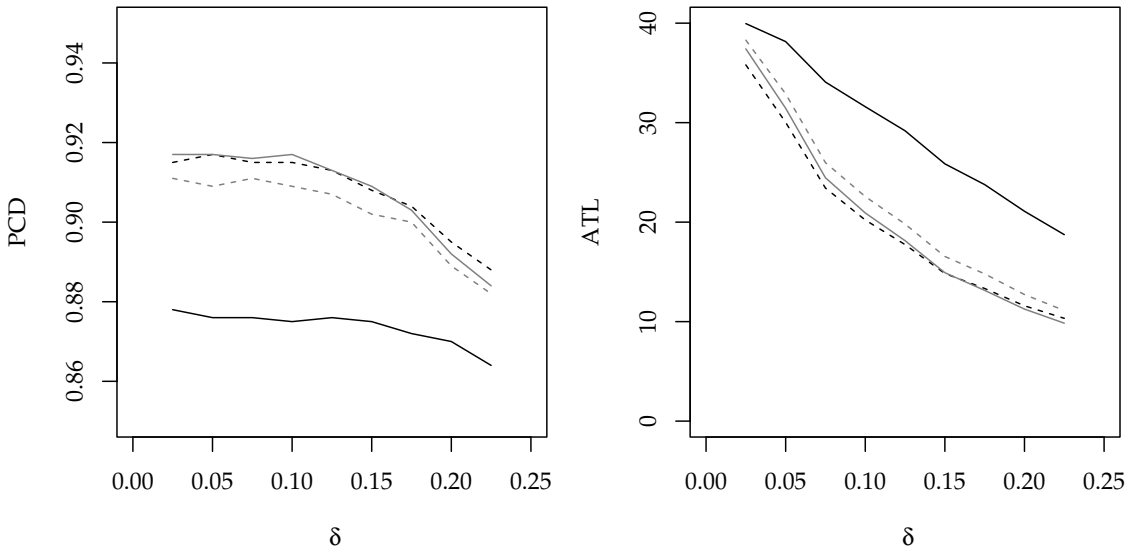


Figure 2.2. Results from simulations with the mathematics item pool for different sizes of the indifference region.
Note. The solid black line denotes random item selection, the solid grey line method based on weighting, the dotted lack line ability estimate, and the dotted grey line global criterion. PCD = proportion of correct decisions; ATL = average test length; δ = width indifference region.

items depending on the item selection method if δ was increased, except for RA. Although the test length decreased a lot, δ was only slightly increased, because the PCD had to remain above the specified accuracy level.

2.5.4 Simulations with Content and Exposure Control

Thus far, the simulations were limited to the situation in which no content specifications were specified and no action was taken to avoid overexposure. In actual testing programs, constraints have to be met for the content of the test, and attention is also paid to item exposure. In adaptive testing, implementing content or exposure control often results in longer tests. Eggen and Straetmans (2000) considered content and exposure control for the mathematics item pool. Their simulations were replicated in this study for the eight item selection methods.

The Kingsbury and Zara (1989) approach was used to select 16% of the items from the subdomain mental arithmetics/estimation, 20% from measuring/geometry, and the other items from the other domains in the curriculum. The item was selected from the domain for which the difference between the desired and achieved percentage of items selected thus far was the largest.

Exposure control was implemented using a simplified form (Eggen & Straetmans, 2000) of the Sympson and Hetter (1985) method. When an item was selected,

a random number g was drawn from the interval $(0, 1)$. When g>0.5, the item was administered; if not, another item was selected. The rejected item was not admissible for the respondent for the remainder of the test.

A different procedure was implemented to select the first three items. An examinee was presented a relatively easy item from the item pool. Fifty-four items were denoted as easy items. Depending on the implemented content control method, an easy item was selected for each domain, or three relatively easy items were selected at random.

Simulations were run implementing content (C) and exposure (E) control with the maximum test length set at 25, $\delta = 0.10$, and $\alpha = \beta = 0.05$. For random item selection, simulations were limited to the situation in which no content and exposure control was implemented, but items 1–3 were selected from the set of relatively easy items.

The results for these simulations are given in Table 2.3. Content control had limited influence on the ATL and the PCD. Exposure control had a slightly larger influence on the ATL, but also had a low impact on the PCD. Implementing content and exposure control resulted in the longest tests, the least accurate decisions, and had little influence on the ranking of the item selection methods.

Table 2.3
Results from Simulations with and without Content Constraints and Exposure Control

	No C, no E		C		E		C+E	
Selection	ATL	PCD	ATL	PCD	ATL	PCD	ATL	PCD
RA	23.103	0.838						
AE	17.122	0.896	17.343	0.895	18.218	0.885	18.439	0.886
MC	17.330	0.890	17.685	0.891	18.314	0.883	18.494	0.883
NC	17.667	0.896	17.701	0.897	18.561	0.887	18.648	0.888
WM	17.987	0.897	17.969	0.896	18.768	0.889	18.897	0.889
GP	17.451	0.893	17.675	0.895	18.392	0.884	18.580	0.884
GC	18.985	0.885	18.970	0.889	19.779	0.881	20.196	0.881
MA	18.529	0.885	18.845	0.885	19.256	0.878	19.609	0.877

Note. C = content constraints; E = exposure control; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutoff points; NC = nearest cutoff point; WM = method based on weighting; GP = goal programming; GC = global criterion; MA = maximin.

2.6 Discussion

Four item selection methods were developed and were compared with current methods for classifying examinees into multiple categories. The new methods consider the multiple cutoff points when selecting items. Simulations were used to investigate the effect of the item selection methods on the PCD and the ATL.

In the first series, a simulated item pool was used. Two, three, and four cutoff points were specified. Random item selection resulted in a lower PCD, but differences in ATL and PCD were small for the other methods. The differences in the PCDs for different item selection methods were also small. A second series of simulations was done using the mathematics item pool. The differences in the ATL and the PCD were also small except for random item selection.

It was expected that taking multiple cutoff points into account would decrease the ATL and increase the PCD. The results show that the available item selection methods classified as accurate and efficient as the multiple objective methods. This was probably caused by always taking all cutoff points into account. In the later stages we already know in which part of the scale a classification is likely to be made, but the methods also take the other parts of the scale into account. This can be solved by restricting the number of cutoff points that is considered by the item selection method after a number of items is administered. It would also be interesting to use a multiple objective approach in the starting phase of the test and then switch to one of the currently available methods. By using such an approach, the advantages of both types of methods are exploited; first a broad part of the scale is considered and then a more precise method can be used.

Comparing the simulations with a simulated pool and the mathematics item pool shows that characteristics of the item pool, distribution of ability, settings of the classification method, and the number of cutoff points all influenced the test length and the accuracy. This suggests that simulation studies should be executed during the development process for a classification test.

The test length and the accuracy were influenced by the size of the indifference regions. The simulations were repeated with different specifications for the indifference region. The ATL and the PCD decreased in the simulations with a simulated and a mathematics item pool. The test length and the accuracy were only slightly influenced by content and exposure control in the simulations with the mathematics item pool. This finding is in line with the findings of Eggen and Straetmans (2000).

In the present study, item selection methods were included that considered the examinee's ability. In addition to the obvious psychological and educational advantages of adaptive item selection, some initial simulations with item selection methods that were not adaptive showed that the number of items required before making a decision and the accuracy of the classifications was comparable or even better with adaptive methods than with methods that are not adaptive.

Investigating whether the current results can be replicated if other and larger item pools are used or if the examinees characteristics are changed would be interesting. Here, only one method was developed per multiple objectives approach, but other methods are possible. These should be investigated using simulations with different item pools, SPRT settings, and examinee characteristics.

References

Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*, *73*, 473–486. doi: 10.1007/s11336-007-9053-9

Birnbaum, A. (2008). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Charlotte, NC: Information Age. (Original work published 1968)

Boekkooi-Timminga, E. (1989). *Models for computerized test construction*. Unpublished doctoral dissertation, Twente University, Enschede, the Netherlands.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8

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. doi: 10.1177/00131640021970862

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*, 379–393. doi: 10.1177/0146621606288890

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computerassisted branded test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh, PA.

Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442–463. doi: 10.3102/1076998607308573

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Ignizio, J. P. (1982). *Linear programming in single and multiple objective systems*. Englewood Cliffs, NJ: Prentice-Hall.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, *2*, 359–375. doi: 10.1207/s15324818ame0204_6

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a

computerized mastery test. *Applied Psychological Measurement*, *14*, 367–386. doi: 10.1177/014662169001400404

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404. doi: 10.1177/014662169602000406

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*, 185–187. doi: 10.1207/s15324818ame1903_1

Mollaghasemi, M., & Pet-Edwards, J. (1997). *Making multiple objective decisions*. Los Alamitos, CA: IEEE Computer Society Press.

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.

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., & Reckase, M. D. (1994, April). *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.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277–292. doi: 10.1177/014662169301700308

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151–166. doi: 10.1177/014662169301700205

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *In: Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

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

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer. doi: 10.1007/0.387.29054.0

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*, 253–266. 10.1111/j.1745-3984.1999.tb00557.x.

Veldkamp, B. P., Verschoor, A. J., & Eggen, T. J. H. M. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicológica*, *31*(2), 335–355.

Wald, A. (1973). *Sequential analysis*. New York, NY: Dover. (Original work published 1947)

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41–58. doi: 10.1177/0013164406288164

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*.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8\18

Chapter 3
Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Between-Dimensionality

Abstract

A classification method is presented for multidimensional adaptive testing in which a decision is made on the entire test. The items, which measure one ability each, are modeled using multidimensional item response theory. One method exists for classification testing for between-dimensionality (Seitz & Frey, 2013). This method classifies based on the underlying dimensions, but not on the entire test. The method is extended to include decision making based on the entire test, on several dimensions, and on subsets of items. A measure is presented that provides information about the support for the decisions. Items were selected that maximize information at the current ability estimate or at a weighted combination of the cutoff points. An empirical example illustrates the methods discussed.

This chapter has been submitted as Van Groen, M.M. & Eggen, T.J.H.M. (2014). *Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Between-Dimensionality.* Manuscript submitted for publication.

3.1 Introduction

The majority of computerized adaptive tests (CATs) obtain a precise and efficient estimate of the examinee's ability. However, on some CATs, the goal is to make accurate and efficient classification decisions about the examinee's ability. In CATs for making classification decisions, the test length is minimized while the desired accuracy level is maintained. A large body of research is available for unidimensional CAT (UCAT), but knowledge about multidimensional CAT (MCAT) is still growing. Limited information is available about MCATs for making decisions.

Multidimensional CATs can be constructed using multidimensional item response theory (MIRT). In MIRT, a vector of person abilities describes the skills and knowledge required for answering the items (Reckase, 2009). The score on an item is modeled using the relationship between an examinee's ability and the item parameters (Van der Linden & Hambleton, 1997). With MIRT, two types of multidimensionality can be modeled. Between-dimensionality implies that items are intended to measure only one trait (Wang & Chen, 2004). Within-dimensionality implies that items measure several traits.

Several methods for making unidimensional classification decisions are available (Eggen, 1999; Spray, 1993; Weiss & Kingsbury, 1984), but few methods are available for multidimensional constructs. Van Groen, Eggen, and Veldkamp's (2014b) method can be used for within-dimensionality, and Seitz and Frey's (2013) method can be used for between-dimensionality. The latter makes decisions for each dimension but not a simultaneous decision on all dimensions.

In many testing situations, a global pass/fail or a classification into one of several mutually exclusive categories is required. Seitz and Frey's (2013) method cannot be used to make decisions on the entire test. An extension is proposed for making such decisions for between-dimensionality. The proposed method can also be applied to any subset of dimensions in the test, instead of all dimensions. This implies that if a test consists of several sections that are modeled with separate dimensions, a decision can be made for the different sections although some sections are modeled with multiple dimensions. The application of the method to subsets of items in the test is also explored. Seitz and Frey (2013) classified examinees into one of several levels. The proposed extensions can also classify examinees into multiple levels.

A measure is proposed that indicates the support for the decision that was made. After the settings are specified for the classification method, the measure indicates the relative support for the decisions that are made based on the test. This measure can be reported to the examinee in addition to the classification decision itself.

Multidimensional classification testing also requires a method for selecting the items. Several methods are available for selecting the items for MCATs that estimate ability (Reckase, 2009; Segall, 1996; Yao, 2012), but little knowledge is available for classification MCATs. In unidimensional classification testing, items are often selected based on the ability estimate or the cutoff point (Eggen, 1999). Similar approaches are explored here for MCATs with between-dimensionality.

The article is structured as follows. First, MIRT is described. Then, classification and item selection methods are presented. Fourth, a method is proposed for describing the support for the decisions. Fifth, the efficiency and effectiveness of the methods are investigated for an empirical example. Finally, remarks are made about multidimensional classification testing and directions for future research.

3.2 Multidimensional Item Response Theory

MCAT requires an item bank that is calibrated with a multidimensional item response theory model (Reckase, 2009). In a calibrated item bank, model fit is established, item parameter estimates are available, and items with undesired characteristics are removed (Van Groen, Eggen, & Veldkamp, 2014a). During the administration of the CAT, the item parameters are assumed to be estimated precisely enough to consider them known (Veldkamp & Van der Linden, 2002).

The two-parameter logistic model is used here (Reckase, 1985). In MIRT, a set of p abilities accounts for the examinee's responses to the items and the probability of a correct answer, $x_i = 1$, to item i is given by (Reckase, 2009)

$$\mathrm{P}_i(\boldsymbol{\theta}) = \mathrm{P}\left(x_i = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}\right) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}, \tag{3.1}$$

where $\mathbf{a}_i$ is the vector of the discrimination parameters, d_i is the scalar parameter representing the easiness of item, and $\boldsymbol{\theta}$ the vector of the ability parameters. If only one discrimination parameter is non-zero per item, for all items, the test is considered to have between-dimensionality (Wang & Chen, 2004), and items are intended to measure only one ability. If multiple discrimination parameters differ

from zero, within-dimensionality is present, and items are intended to measure multiple or all abilities.

MIRT can be used to model performance in complex domains, take into account several abilities simultaneously and represent different combinations of abilities for different items (Hartig & Höhler, 2008). If a MIRT model is fitted to the same data with the same number of dimensions, the within- and between-dimensional models are essentially equivalent and may be equally appropriate (Hartig & Höhler, 2008). Therefore, the choice for a within- or between-dimensional model depends on the intended interpretation of the model. In a within-dimensional model, the ability on one dimension can compensate the ability on another dimension. A between-dimensional test exists of several unidimensional subscales and does not specify whether abilities are completely different or share common elements. In a within-dimensional model, explicit assumptions are made about the relations between the abilities for different sets of items (Hartig & Höhler, 2008). If a test consists of several related unidimensional subscales, such as in the empirical example here, a between-dimensional model is a logical choice.

The likelihood of a vector of observed responses $\mathbf{x}_j$ to items $i = 1, \cdots, k$ for an examinee j with ability $\boldsymbol{\theta}_j$ equals the product of the probabilities of the responses to the test items because of the local independence assumption (Segall, 1996):

$$\mathrm{L}\left(\boldsymbol{\theta}_j|\mathbf{x}_j\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_j\right)^{x_{ij}} \mathrm{Q}_i\left(\boldsymbol{\theta}_j\right)^{1-x_{ij}}, \tag{3.2}$$

where $\mathrm{Q}_i(\boldsymbol{\theta}_j) = 1 - \mathrm{P}_i(\boldsymbol{\theta}_j)$ and item parameters are considered known. The vector of the values, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)$, that maximizes the likelihood function in Equation 3.2 is used as the ability estimate for $\boldsymbol{\theta}_j$ (Segall, 1996). The equations for finding the maximum likelihood estimates have no closed-form solution (Segall, 1996). Thus, an iterative search procedure, such as Newton-Raphson, is used for finding the maximum likelihood estimates. Weighted maximum likelihood (WML) estimation, which is used here, is an extension of maximum likelihood (ML) estimation and reduces the bias in the ML estimates. WML was developed by Tam (1992) as an extension of the unidimensional estimator by Warm (1989). The procedure used is described in Van Groen et al. (2014b).

3.3 Classification Methods

One part of the adaptive classification procedure determines whether testing can be stopped and which decision is made. Few methods are available for making multidimensional classification decisions (Seitz & Frey, 2013; Spray, Abdel-Fattah, Huang, & Lau, 1997; Van Groen et al., 2014b). Van Groen et al.'s (2014b) method uses a reference composite (M. Wang, 1985, 1986, as described by Reckase, 2009) for reducing the multidimensional space to a line. The eigenvectors of the largest eigenvalue of the $\mathbf{aa}'$ matrix for the entire construct determine the direction of the reference composite. Eigenvalues and eigenvectors make sense only if all elements of the matrix are non-zero. If items each measure only one, although different, dimensions, all non-diagonal elements are zero. The result is that the reference composite classifies on only one dimension. This implies that the method is not suitable to make classifications with between-dimensionality. Seitz and Frey (2013) developed a method to make classification decisions with between-dimensionality (see next section). The method is expanded in the second and third parts of this section to make decisions on the entire test and on parts of the test.

3.3.1 A Classification Method for Between-Dimensionality

Seitz and Frey (2013) developed a classification method for between-dimensional tests based on the fact that the multidimensional two-parameter logistic model is a combination of several unidimensional two-parameter logistic models (W.-C. Wang & Chen, 2004). This is a result of the fact that the likelihood for an item contains only one dimension, because the contributions of the other dimensions reduce to zero. Thus, the likelihood for an item is equal to the likelihood for the same item under the unidimensional two-parameter logistic model. Since items are combined that load on different dimensions, multiple unidimensional two-parameter logistic models are combined into one multidimensional two-parameter logistic model. Seitz and Frey (2013) implemented the unidimensional classification method using the sequential probability ratio test (SPRT) for each dimension separately.

The SPRT (Wald, 1947/1973) with unidimensional IRT was developed by Reckase (1983), following Ferguson (1969) for classical test theory, and used by, among others, Eggen (1999), Spray (1993), and Thompson (2009). A cutoff point, θ_{cl}, is set for each dimension, l, with an indifference region around cutoff point c. The indifference region accounts for the uncertainty of the decisions, owing to measurement error, about examinees with an ability close to the cutoff

point (Eggen, 1999). Hypotheses are formulated at both ends of the indifference regions:

$$\mathrm{H}_{0l} : \theta_{jl} < \theta_{cl} - \delta; \tag{3.3}$$

$$\mathrm{H}_{al} : \theta_{jl} > \theta_{cl} + \delta, \tag{3.4}$$

δ denotes half the size of the indifference region. The SPRT uses the likelihood ratio between the hypotheses for dimension l after k items as the test statistic:

$$\mathrm{LR}\left(\theta_{cl} + \delta; \theta_{cl} - \delta\right) = \frac{\mathrm{L}\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)}{\mathrm{L}\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)}, \quad l = 1, \cdots, p, \tag{3.5}$$

in which $L\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)$ and $L\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)$ for cutoff point c for dimension l are calculated using Equation 3.2 with only those items included that load on the dimension. This simple likelihood ratio is possible because the likelihood for items that measure the same ability reduces to a one-dimensional likelihood in the case of between-dimensionality (Seitz & Frey, 2013).

The SPRT applies decision rules per dimension to decide whether to continue testing or to make a decision:

$$\begin{aligned}
&\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) < (1-\beta)/\alpha; \\
&\text{ability below } \theta_c \text{ if} && \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \leq \beta/(1-\alpha); \\
&\text{ability above } \theta_c \text{ if} && \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \geq (1-\beta)/\alpha,
\end{aligned} \tag{3.6}$$

where α and β specify the acceptable classification error rates (Spray et al., 1997).

In practice, a maximum test length is set to ensure that testing stops at some point (Eggen, 1999). At this point, if the likelihood ratio for dimension l is larger than the midpoint of $\beta/(1-\alpha)$ and $(1-\beta)/\alpha$ for dimension l the examinee is classified as having an ability above the cutoff point.

Seitz and Frey (2013) also considered four cutoff points per dimension. They implemented the Armitage (1950) approach, in which the ratios are calculated for all combinations of levels (Spray, 1993). Since the likelihood has no additional local maxima besides the global maximum in the multidimensional two-parameter logistic model, the Sobel and Wald (1949) approach is used here. In this approach, the likelihood ratio is computed for adjacent levels (Eggen & Straetmans, 2000).

Seitz and Frey's (2013) method has three limitations. One, for tests with within-dimensionality the likelihood ratio does not reduce to Equation 3.5. Thus, this

method cannot be used then. Two, the method ignores information from items loading on other dimensions when classifying (Seitz & Frey, 2013). Three, the method requires an additional decision rule if one global decision has to be made. The method is adapted to make a decision on the entire test.

3.3.2 Extension for Making Decisions on the Entire Test

A fail/pass decision or a classification into one of multiple levels for the entire test is required in many testing situations. Seitz and Frey's (2013) method can be adapted to make a decision on the entire test. The likelihood in Equation 3.5 is extended to include all dimensions and all items:

$$\mathrm{LR}\left(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_c - \boldsymbol{\delta}; \mathbf{x}_j\right)}. \tag{3.7}$$

Here, $\boldsymbol{\theta}_c$ includes the cutoff points for all dimensions, and all elements of $\boldsymbol{\delta}$ are assumed to be equal to δ. If a test is modeled using three dimensions, Equation 3.7 becomes

$$\mathrm{LR}\left(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}\right) = \frac{\mathrm{L}\left(\theta_{c1} + \delta; \mathbf{x}_{j1}\right)}{\mathrm{L}\left(\theta_{c1} - \delta; \mathbf{x}_{j1}\right)} \cdot \frac{\mathrm{L}\left(\theta_{c2} + \delta; \mathbf{x}_{j2}\right)}{\mathrm{L}\left(\theta_{c2} - \delta; \mathbf{x}_{j2}\right)} \cdot \frac{\mathrm{L}\left(\theta_{c3} + \delta; \mathbf{x}_{j3}\right)}{\mathrm{L}\left(\theta_{c3} - \delta; \mathbf{x}_{j3}\right)}, \tag{3.8}$$

in which the cutoff point of interest is selected for each dimension. The items that load on each dimension are used to calculate the likelihood per dimension.

3.3.3 Extensions for Making Decisions on Parts of the Test

Two other extensions include making a decision using only part of the dimensions and making a decision on a subset of the items. A decision can be made on only part of the dimensions. The likelihood ratio of Equation 3.7 then consists of the dimensions that are relevant to the intended decision and becomes

$$\mathrm{LR}\left(\boldsymbol{\theta}_{cl} + \boldsymbol{\delta}; \boldsymbol{\theta}_{cl} - \boldsymbol{\delta}\right) = \prod \frac{\mathrm{L}\left(\boldsymbol{\theta}_{cl} + \boldsymbol{\delta}; \mathbf{x}_{jl}\right)}{\mathrm{L}\left(\boldsymbol{\theta}_{cl} - \boldsymbol{\delta}; \mathbf{x}_{jl}\right)}, \quad \text{for } l \in \mathrm{V}_c, \tag{3.9}$$

where V_c denotes the set of dimensions that are relevant for the current decision and $\mathbf{x}_{jl}$ the set of items for dimension l.

The second type of extension includes decisions that are based on only a part of the items that load on dimension l. The likelihood ratio is calculated for a subset of the items in the test and becomes for dimension l

$$\mathrm{LR}\left(\theta_{cl}+\delta;\theta_{cl}-\delta\right)=\frac{\mathrm{L}\left(\theta_{cl}+\delta;\mathbf{x}_{jls}\right)}{\mathrm{L}\left(\theta_{cl}-\delta;\mathbf{x}_{jls}\right)},\quad \text{for } x_{jls}\in \mathrm{V}_s, \tag{3.10}$$

where V_s denotes the set of items that form the basis of the decision.

Decisions can be made on several dimensions, several sets of items, and on multiple levels. This makes it possible to include an item in several decisions. An example of such an item is an item used in the decision about the entire test, but is also used to make a decision about a part of the test. Different dimensions, from which part of the items are selected, can be included in a decision. Seitz and Frey (2013) showed that it is possible to classify pupils into one of several mutually exclusive levels, which is also possible for the extensions that are proposed here.

3.4 Item Selection Methods

Another component of an MCAT selects the items (Reckase, 2009). The item selection method is important, because selecting items that are too hard or too easy or provide little information results in tests that do not function well (Reckase, 2009). Several methods are available for MCATs (Luecht, 1996; Mulder & Van der Linden, 2009; Reckase, 2009; Segall, 1996; Veldkamp & Van der Linden, 2002; Yao, 2012), but these methods all focus on obtaining an efficient ability estimate.

Two papers for multidimensional classification testing selected items using Segall's (1996) methods. Seitz and Frey (2013) applied Segall's (1996) Bayesian item selection method for between-dimensionality, and Van Groen et al. (2014b) applied Segall's (1996) ML method that maximizes the determinant of the information matrix for within-dimensionality. In the latter manuscript, Segall's (1996) method was adapted to select the items at the current projected ability estimate and the cutoff point based on the reference composite. Optimizing with respect to the ability estimate and the cutoff point are commonly used methods in unidimensional classification testing (Eggen, 1999; Spray & Reckase, 1994; Thompson, 2009). The same optimizations as in the unidimensional case are applied here for between-dimensionality.

3.4.1 Item Selection Based on the Ability Estimate

Segall (1996) developed a method analogous to the unidimensional method that maximizes Fisher information. Fisher information is a measure of the information in the observable variables on the ability parameters (Mulder & Van der Linden,

2009). The diagonal elements of this $p \times p$ matrix, $\mathbf{I}(\boldsymbol{\theta})$, for dimension l are (Tam, 1992)

$$\mathrm{I}(\theta_l) = \frac{\frac{\partial}{\partial\theta_l}\mathrm{P}_i(\boldsymbol{\theta}) \times \frac{\partial}{\partial\theta_l}\mathrm{P}_i(\boldsymbol{\theta})}{\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})} = a_{il}^2\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta}). \tag{3.11}$$

Information is calculated here for the current ability estimate, $\hat{\boldsymbol{\theta}}_j$. The off-diagonal elements of the information matrix are zero for items with between-dimensionality.

Segall's (1996) item selection method uses the relationship between the information matrix and the confidence ellipsoid around the estimates (Reckase, 2009). It obtains the largest decrement in the volume of the confidence ellipsoid (Segall, 1996). The inverse of the information matrix approximates the size of the confidence ellipsoid, thus the item is selected that maximizes (Segall, 1996)

$$\max \det \left(\sum_{i=1}^{k} \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij}) + \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j}) \right), \quad \text{for } k+1 \in \mathrm{V}_{k+1}, \tag{3.12}$$

in which V_{k+1} denotes the set of items that are available for administration. The maximization is over the determinant of the sum of the information matrix of the administered items and the information of potential item $k+1$. The selected item has the largest determinant. Thus, the volume of the confidence ellipsoid around the ability estimate is minimized (Reckase, 2009).

Segall's (1996) method is generally applicable to MCAT, but requires a non-singular matrix (Yao, 2012). A matrix is nonsingular if all diagonal elements are larger than zero. This implies that information is required for all dimensions in the test before Segall's (1996) method can be applied. Thus, Yao (2012) selected the first l items from each of the l dimensions. Another solution is to calculate the determinant for the part of the information matrix that contains non-zero elements.

If all available items measure the same dimension, Segall's (1996) method reduces to maximizing information at the current ability estimate. Thus, the unidimensional item selection method that maximizes information at the current ability estimate can be used instead. The quantity that is maximized becomes then

$$\max \mathrm{I}_i(\hat{\theta}_l), \quad \text{for } i \in \mathrm{V}_{al}, \tag{3.13}$$

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

3.4.2 Item Selection Based on the Cutoff Points

In unidimensional classification testing, information is often maximized at the cutoff point (Eggen, 1999; Spray, 1993). If all available items measure the same dimension, unidimensional item selection methods can be used. Van Groen et al. (2014a) described item selection methods based on the cutoff points and the ability estimate and found that the differences in accuracy and efficiency between the methods are very small. One of the most effective, but also very efficient, methods is the weighting method (Van Groen et al., 2014a). This method required a few more items on average than the method that selects at the nearest cutoff point, but it was the most accurate.

The weighting method combines several objective functions in one objective function with weights for the objectives (Van Groen et al., 2014a). The weight for a specific cutoff point increases if the ability estimate is closer to the cutoff point. Thus, the item selection is adapted to the examinee's ability. If all available items measure the same dimension l, the quantity that is maximized becomes

$$\max \sum_{c=1}^{C} \frac{1}{|\hat{\theta}_{jl} - \theta_{cl}|} \mathrm{I}_i(\theta_{cl}), \quad \text{for } i \in \mathrm{V}_{al}. \tag{3.14}$$

3.5 Measure for Reporting the Confidence in the Decision

The SPRT makes a decision but does not provide information about the support for the decision that was made. A measure is proposed that reflects the support for the decision. This measure can be reported as a test result, to indicate whether more or less support was available for the decision than the SPRT required.

The likelihood ratio in Equation 3.5 can provide the information, but reporting it has four issues. One, the value of the ratio depends on δ. Two, it is difficult for teachers and examinees to interpret. Three, with a different δ, the ratio has a different value. Four, the likelihood ratios for multiple decisions cannot be compared. The second is a problem because the comparison value depends on the decision rules. If the value is smaller than $\beta/(1-\alpha)$, this is the value to compare with. But what does a value smaller than the decision value imply? In addition, if the value is larger than $(1-\beta)/\alpha$, this value must be compared with.

A better alternative is to standardize the likelihood ratio in relation to the decision values. The likelihood ratio can be standardized by dividing the likelihood ratio by $(1-\beta)/\alpha$ for the support for the decision that the examinee's ability is

above the cutoff point. For the support for the decision that the ability is below the cutoff point, the ratio is inverted before the division and is divided by $\beta/(1-\alpha)$. Therefore, the support for the decision, DS, becomes

$$\begin{aligned} &\text{support for decision below } \theta_{cl}: && \text{DS}_- = \frac{\text{LR}(\theta_{cl}-\delta;\theta_{cl}+\delta)}{\beta/(1-\alpha)}; \\ &\text{support for decision above } \theta_{cl}: && \text{DS}_+ = \frac{\text{LR}(\theta_{cl}+\delta;\theta_{cl}-\delta)}{(1-\beta)/\alpha}. \end{aligned} \tag{3.15}$$

If DS_- or DS_+ equals one, enough support was gathered for the decision. If the value is smaller than one, the decision is made with less support than required. If DS_- or DS_+ is larger than one, then the decision is supported more than required. The measure can be reported as a test result. If an examinee is classified at a level with two adjacent levels, DS_+ and DS_- are reported for the adjacent levels.

The value of DS depends on δ, but since the combination of δ, α, and β is set beforehand so that decisions with the test are made with the required accuracy and efficiency, it is now possible to compare the gathered support with the required support for making the decision. The interpretation problem is fixed by using DS because a value smaller than one implies that the decision was made without enough support and a value larger than one implies that the support was above the required amount. It does not make sense to investigate the value of the DS before fixing δ, α, and β, because the value of the measure depends on the specified settings. At testing time, these values have been set by the test developer, thus interpreting DS now makes sense. A teacher can also compare the support for different decisions, because the interpretation of DS is the same for all decisions.

3.6 Empirical Example

The classification and item selection methods are illustrated with an empirical example. Seitz and Frey (2013) illustrated their method using a simulated item bank, but here response patterns and item parameters from an existing dataset were used. The End of Primary School Test (Cito, 2012) in the Netherlands provides an advise about the most suitable secondary education level for pupils. Dutch secondary education is divided into multiple levels: pre-vocational secondary education (VMBO), general secondary education (HAVO), and pre-university education (VWO). VMBO is divided even further into three levels: VMBO-BB, VMBO-KB, and VMBO-GT.

The test consists of three components: Language, Mathematics, and Study Skills. Language (LAN) is divided into Writing (WR), Text Comprehension (TC), Vocabulary (VO), and Spelling (SP). The latter consists of Spelling of Verbs (SPV) and Non-Verbs (SPN). Mathematics (MAT) consists of Measurement, Geometry, Time & Money (MGTM); Relations, Fractions & Percentages (RFP); and Numbers & Computations (NC). Study Skills (SS) consists of Study Texts (ST), Information Sources (IS), Geography (GE), and Tables & Graphs (TG).

The End of Primary School Test is calibrated using six unidimensional IRT scales. The fit of the scales for WR, TC, VO, SP, MAT, and SS is established annually. A scale score based on the pupil's proficiency is provided as the test result in addition to percentile scores for the three main components. Intervals for this scale score are related to the secondary education levels (Van Boxtel, Engelen, & De Wijs, 2011). This score and the primary school's recommendation are used by secondary education schools to classify pupils into different classes based on the education levels.

3.6.1 Study Design

In 2012, 147,099 pupils took the standard paper-and-pencil version of the End of Primary School Test. Other versions of the test, such as digital and easier versions, were not considered here. In the current study, classifications based on the SPRT were compared to classifications based on the scale score.

In addition to providing an overall classification decision, classification decisions were made on all components and subcomponents of the test. The lower-level classification decisions provide insight into the pupil's level on the underlying topics. A schematic overview of the decisions is shown in Figure 3.1. The decisions on the entire test and LAN were based on items measuring six and four dimensions. Classifications on the part of the items that measured one dimension were made for the components of SP, MAT and SS.

The dataset for the 2012 version of the Final Test was used here. The dataset was randomly divided into 10 datasets. The data of the first dataset were calibrated with a six between-dimensional two-parameter MIRT model using NOHARM 4 (Fraser & McDonald, 2012). The responses in the dataset and the resulting item parameters were used to make the classifications. The first dataset was also used to determine the cutoff points and settings for the SPRT. The intervals available for the five education levels (Van Boxtel et al., 2011) determined the cutoff points for the total test. The median of the ability estimates for each of the five intervals was

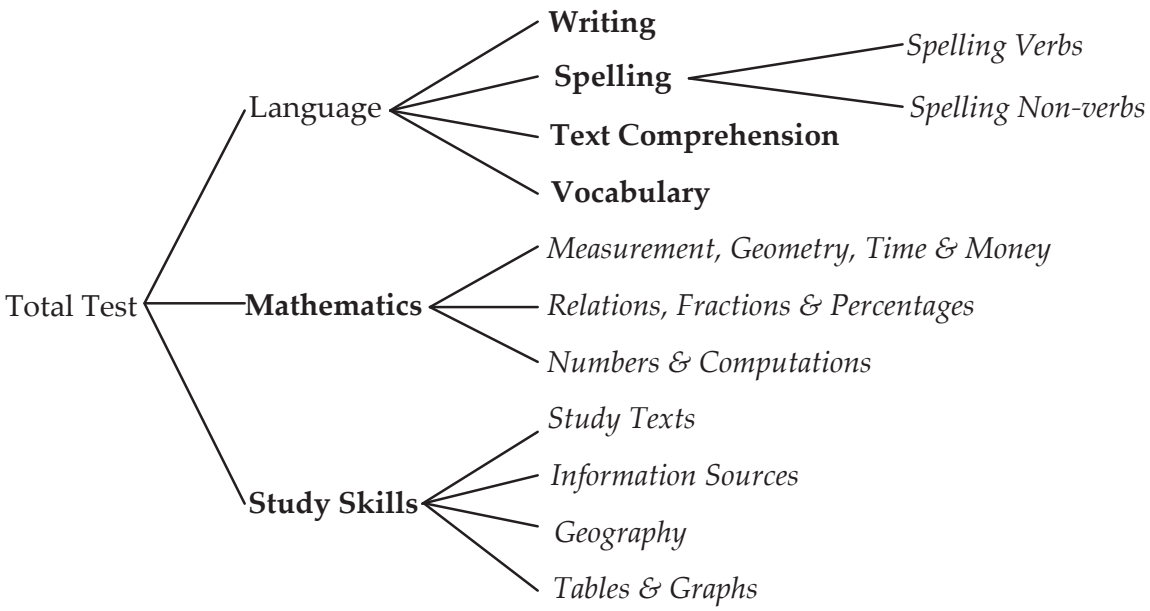


Figure 3.1. Decisions by the SPRT for the End of Primary School Test.
Note. One-dimensional decisions are printed in boldface. Decisions on subsets of a dimension are printed in italics.

calculated, and the midpoint between two adjacent medians was set as the cutoff point. The median is used instead of the mean due to the skewness of the ability distributions. The same procedure was applied for decisions based on parts of the scales. If a decision was based on multiple dimensions, the cutoff points on the separate dimensions were used for Equation 3.7.

The analyses consisted of two parts. In the first part, the consistency of the classification method was determined with all 200 items included but also with fewer items. Consistency was specified by the proportion of consistent decisions (PCD). The proportion of classifications by the SPRT and the classifications based on the intervals of the scale score that were equal denote the consistency (PCD1) for the total test. For the underlying decisions, no baseline classification decision was available. Since the pupils could also be classified based on their estimated ability, this classification was used for comparison for the lower-level decisions. This proportion of consistent decisions is denoted as PCD2. Unfortunately, this was possible only if just one dimension was involved in the decision, so this was not possible for the decision on LAN.

Test length was set at 150, 160, 170, 180, 190, and 200 items. Test length per component of the test was fixed in the analyses and shortened in proportion to the total test length. Components consisting of 10 items in the original test were not shortened. Two reasons existed for fixing the test length. One, a fixed test length is often considered more acceptable in high-stakes testing. Second, in initial analyses

with the SPRT the test length of the components of the End of Primary School Test could not be reduced much by the SPRT. The analyses with fewer than 200 items made it possible to investigate the influence of shortening the test on consistency.

By setting the test length at fewer than 200 items, it was also possible to investigate the influence of the item selection method on classification consistency. A third consistency measure, PCD3, compared the classification by the SPRT at the maximum and reduced test length. This measure could be used for investing the influence of reducing the test length and item selection on consistency.

The analyses were conducted using three item selection methods: maximization of information at the cutoff points using the weighting method, maximization of information at the current ability estimates, and random item selection. The components TC, WR, ST, and GE consisted of testlets. The item selection methods for testlet items had to administer groups of items around a common stimulus such as a text or map in the correct order. The information-based methods were adapted so that the testlet was selected that maximized the average testlet information at the specified point on the ability scale. Per testlet, the average amount of information per item was calculated for the items in the testlet and was used for selection. Random item selection selected one testlet at random.

The settings of the SPRT potentially influenced which decision was made. Analyses were conducted with $\alpha = \beta = 0.05$ or 0.10. The α and β values partly determined which decision was made, because they were used in the decision rules. A lower value could also have resulted in a shorter test, but the test length was specified beforehand. δ was set at 0.05. The size of δ also potentially influenced the decision that was made and the length of the test. In initial analyses, a higher value of δ resulted in less consistency between the classification by the SPRT and the classification based on the scale score, so δ was set at this small value.

Cross validity of the results was also investigated. The item bank was calibrated with 10% of the data. The remaining part of the data were divided into nine parts of equal size. The analyses were replicated for the remaining datasets with item parameters fixed at the values for the first dataset.

3.6.2 Results

The analyses consisted of two parts. In the first part, the consistency of the classifications with the maximum test length was investigated for the Dutch End of Primary School Test. In the second part, results of the first part were cross-validated with other datasets of the same test.

Classification Consistency at Maximum and Reduced Test Length

The goal of the first part of the analyses was to determine the consistency of the SPRT classifications and the decisions based on the scale score. The proportions of consistent decisions (PCDs) were used as a consistency measure. First, consistency was investigated at the maximum test length. Second, consistency was inspected at a reduced test length for the three item selection methods. Third, the differences in PCD between the selection methods are discussed. Fourth, the PCDs for the components of LAN, MAT, and SS were investigated.

Consistency at Maximum Test Length

For the total test with 200 items, the consistency (PCD1) between the decisions based on the scale score intervals and the SPRT decisions was 0.897, as shown in Table 3.1. The scale score classification and the SPRT decision contributed to the inconsistencies in four ways. One, the interval decisions contributed to the inconsistencies, because they were based on transformed raw scores. Two, the error in the item parameter estimates for the SPRT decisions contributed to the inconsistencies. Three, different dimensions compensated each other in the SPRT classifications but not in the classifications based on the intervals. Four, the ability estimates used for the cutoff points for the SPRT contained measurement error.

The classifications based on the scale score intervals were only available for the total test. For LAN, using the SPRT was the only possibility. This implied that no comparison decision was available at the maximum test length, which was required for calculating consistency. Classifying based on the ability estimates was not possible, because the component was modeled using multiple dimensions. The decisions for LAN could only be compared with decisions for shorter tests.

For all other decisions, the SPRT classifications and the classifications using the ability estimates were available. The resulting proportions of consistent decisions, PCD2, are presented in the table for MAT and SS. At 200 items, the consistency between the SPRT classifications and the classifications based on the ability estimates for MAT was 0.972. Classification consistency based on the PCD2 measure was lower for SS, 0.941, than for MAT. Inconsistencies in the decisions could be caused by estimation error in the item and ability parameters. Errors in the item parameters, however, influenced both classifications. Errors in the ability parameters affected only the classifications that were based on them.

Table 3.1
Proportion of Consistent Decisions for Different Test Lengths with Random Item Selection

Decision	α, β	Consistency	200	190	180	170	160	150
Total test	0.05	PCD1	0.897	0.890	0.882	0.871	0.860	0.847
Total test	0.10	PCD1	0.897	0.890	0.884	0.871	0.863	0.855
Mathematics	0.05	PCD2	0.972	0.970	0.966	0.962	0.962	0.955
Mathematics	0.10	PCD2	0.972	0.971	0.966	0.963	0.961	0.956
Study Skills	0.05	PCD2	0.941	0.941	0.941	0.941	0.941	0.941
Study Skills	0.10	PCD2	0.941	0.941	0.941	0.941	0.941	0.941
Total test	0.05	PCD3	1.000	0.952	0.935	0.914	0.895	0.880
Total test	0.10	PCD3	1.000	0.951	0.931	0.911	0.901	0.881
Language	0.05	PCD3	1.000	0.932	0.897	0.869	0.836	0.799
Language	0.10	PCD3	1.000	0.933	0.891	0.868	0.834	0.804
Mathematics	0.05	PCD3	1.000	0.912	0.886	0.847	0.826	0.805
Mathematics	0.10	PCD3	1.000	0.911	0.883	0.853	0.827	0.799

Note. Proportion of consistent decisions for classifications based on interval scores and for the SPRT decisions for the total test (PCD1), proportion of consistent decisions for classifications based on the ability estimates and for the SPRT decisions (PCD2) for decisions measured with one ability, and proportion of consistent decisions for classifications with 200 and shorter tests (PCD3). Study Skills was not shortened because its components were already short. $\alpha = \beta$ = acceptable error rates.

When the consistencies for the entire test, LAN, MAT, and SS for $\alpha = \beta = 0.05$ and 0.10 were compared, it seemed that the decisions for individual examinees were made independent of the size of α and β. This was not necessarily correct, because decisions at the maximum test length could be based on Equation 3.6, but they could also be calculated using the midpoint of $\beta/(1-\alpha)$ and $(1-\beta)/\alpha$. Which calculation was used depended on the possibility of stopping the test by the SPRT. Thus, on an individual basis, the calculation of the decision and the actual decision could have been different.

Consistency for Random Item Selection with Reduced Test Length

Table 3.1 also shows that if the test length was reduced, the classification consistency for the entire test decreased. The PCD with the SPRT and the classifications based on the intervals of the scale score decreased from 0.897 to 0.847. The consistency between the SPRT classifications at the maximum and reduced test length also decreased, from 0.952 to 0.880.

If a larger value was specified for the acceptable error rates, α and β, an unexpected result appeared. PCD1 increased in three out of five reduced test lengths. It was expected that the PCD1s would have remained equal or decreased. This increase was explained by the instability of the classifications of examinees whose ability was close to a cutoff point for the SPRT or close to a border value in the scale score classifications. The PCD3 measure decreased in three out of five test lengths. It was expected that the consistency would decrease because more decisions were calculated using the midpoint of $\beta/(1-\alpha)$ and $(1-\beta)/\alpha$ at a shorter test length. If a decision had to be forced, not enough evidence was available for using the decision rules in Equation 3.6. This suggested that the decisions would have differed more often. The increased PCDs were probably caused by the fact that if α and β were increased, the midpoint was used more often for calculating the decisions with 200 items.

The classification consistency between the SPRT at the maximum and reduced test length for LAN decreased from 0.933 to 0.799. This PCD also decreased a lot for MAT. The decrease in the PCD2 for MAT was much smaller than for the PCD3. This suggested that if test length was reduced much, classifications with the SPRT would become different. This could become a problem because the test stakes were high.

Test length was never reduced for SS because all of its components consisted of 10 items. A lower test length per component was considered undesirable in advance, so tests were never shortened. This implied that consistency could be determined only based on the classifications by the SPRT and the ability estimate. This also implied that simulations for SS were not repeated with the other item selection methods.

Consistency for Item Selection with the Ability Estimate with Reduced Test Length

If an information-based selection method was used instead of random selection, higher PCDs can be expected (Van Groen et al., 2014b, 2014a). As shown in Table 3.2, the PCD1 and the PCD3 on the entire test also decreased if the test length was decreased. If a higher value for α and β was specified, PCD1 decreased in two out of five test lengths. The classifications for the PCD3 measure were different only if the test length was reduced to 150 or 160. The PCDs of Language also dropped considerably. If the test length was reduced, the PCD3 measure started to increase if a higher value was set for α and β. This was not expected since a

Table 3.2
Proportion of Consistent Decisions for Different Test Lengths with Item Selection by Maximization at the Current Ability Estimate

Decision	α, β	Consistency	200	190	180	170	160	150
Total test	0.05	PCD1	0.897	0.890	0.879	0.872	0.859	0.848
Total test	0.10	PCD1	0.897	0.887	0.880	0.871	0.861	0.850
Mathematics	0.05	PCD2	0.972	0.970	0.965	0.965	0.966	0.959
Mathematics	0.10	PCD2	0.972	0.969	0.966	0.966	0.964	0.961
Total test	0.05	PCD3	1.000	0.969	0.951	0.935	0.921	0.902
Total test	0.10	PCD3	1.000	0.969	0.951	0.935	0.919	0.904
Language	0.05	PCD3	1.000	0.956	0.922	0.896	0.871	0.830
Language	0.10	PCD3	1.000	0.953	0.922	0.898	0.872	0.834
Mathematics	0.05	PCD3	1.000	0.947	0.918	0.889	0.867	0.849
Mathematics	0.10	PCD3	1.000	0.948	0.921	0.891	0.866	0.847

Note. Proportion of consistent decisions for classifications based on interval scores and for the SPRT decisions for the total test (PCD1), proportion of consistent decisions for classifications based on the ability estimates and for the SPRT decisions (PCD2) for decisions measured with one ability, and proportion of consistent decisions for classifications with 200 and shorter tests (PCD3). Consistency for Study Skills is reported in Table 3.1. $\alpha = \beta$ = acceptable error rates.

higher value should in theory result in less accurate decisions based on the SPRT. The SPRT decisions for MAT remained very consistent with the decisions based on the ability estimates if the test length was reduced. If the test length was reduced, the decisions by the SPRT tended to become less consistent (PCD3).

Consistency for Item Selection with the Weighting Method with Reduced Test Length

The decisions on the entire test became less consistent when the tests were shorter (see Table 3.3). The consistency between the SPRT decisions and the decisions based on the scale score became less consistent (0.897 to 0.848). The classifications also became less consistent when the test length was reduced (PCD3). PCD1 and PCD3 both decreased sometimes if α and β were increased but also increased sometimes. The consistency of the LAN classifications by the SPRT and based on the ability estimates decreased if fewer items were used. The differences in PCD3 between the analyses with $\alpha = 0.05$ and $\alpha = 0.10$ were small, but their direction varied. The consistency between the SPRT decisions and the classifications based on the ability estimates for MAT decreased slowly. The consistency of the decisions

Table 3.3
Proportion of Consistent Decisions for Different Test Lengths with Item Selection by Weighting

Decision	α, β	Consistency	200	190	180	170	160	150
Total test	0.05	PCD1	0.897	0.888	0.880	0.871	0.861	0.848
Total test	0.10	PCD1	0.897	0.888	0.877	0.871	0.862	0.850
Mathematics	0.05	PCD2	0.972	0.970	0.966	0.965	0.964	0.961
Mathematics	0.10	PCD2	0.972	0.969	0.967	0.966	0.964	0.960
Total test	0.05	PCD3	1.000	0.968	0.950	0.935	0.923	0.906
Total test	0.10	PCD3	1.000	0.969	0.951	0.935	0.920	0.908
Language	0.05	PCD3	1.000	0.955	0.924	0.899	0.874	0.834
Language	0.10	PCD3	1.000	0.953	0.924	0.899	0.871	0.836
Mathematics	0.05	PCD3	1.000	0.948	0.921	0.891	0.867	0.847
Mathematics	0.10	PCD3	1.000	0.946	0.922	0.890	0.868	0.849

Note. Proportion of consistent decisions for classifications based on interval scores and for the SPRT decisions for the total test (PCD1), proportion of consistent decisions for classifications based on the ability estimates and for the SPRT decisions (PCD2) for decisions measured with one ability, and proportion of consistent decisions for classifications with 200 and shorter tests (PCD3). Consistency for Study Skills is reported in Table 3.1. $\alpha = \beta$ = acceptable error rates

by the SPRT at the maximum test length and shorter test lengths declined much faster. If the acceptable error rates were increased, the differences in PCD2 and PCD3 were small, but the direction of the differences varied.

Consistency for Different Item Selection Methods with Reduced Test Length

The consistency of the decisions on the entire test, LAN, and MAT was also compared for different item selection methods. The differences in the consistency of the decisions on the entire test were very small for the PCD1 measure. At some test lengths, random item selection appeared to result in more consistent decisions than the other two methods. However, random selection could have resulted in the same incorrect classifications as the scale score classifications. In such situations, other item selection methods could have resulted in correct classifications, but they appeared to be inconsistent.

Random item selection was clearly outperformed by the other two methods on LAN. The consistency between the SPRT at the maximum test length and at the reduced test length was higher for the two information-based methods.

The differences in consistency on MAT were small for the PCD2. Depending on the test length, random or information-based methods resulted in more consistent

decisions. Random selection resulted in less consistent decisions, when the classifications at the maximum and reduced test lengths were compared (PCD3).

The differences between the weighting method and the item selection method based on the ability estimate were very small. The weighting method outperformed the other method in four out of five test lengths. It appeared that the weighting method resulted in the most consistent decisions, independent of test length.

Consistency on the Components with Reduced Test Length

Consistency was also investigated for decisions based on the LAN, MAT, and SS components (see Table 3.4). Per decision, the PCDs for the SPRT and the classifications based on the ability estimates (PCD2), and the consistency at 200 items and 180 items (PCD3), are reported. For each decision, the consistency is reported for $\alpha = 0.05$ and for three item selection methods.

The consistency, as measured with the PCD2 measure, of the LAN components with 200 items was always above 0.900. The consistency between the SPRT decisions and the classifications based on the ability estimates did not differ much among the item selection methods. The PCD3 measure for the LAN components clearly indicated that decisions were different for 180 and 200 items and for different item selection methods. The differences between the weighting and estimate methods were small, but when these methods were compared with random item selection the differences were larger.

The differences in the consistency of WR and TC for different item selection methods were smaller than for other components. This decrease was caused by the existence of testlets in these components. Testlet item selection resulted in the inclusion of items that would not have been selected if items had not been combined into testlets.

The classification consistency for MAT was around 0.900 with 200 items. Remarkably, the consistency of the SPRT and the ability estimates increased when the test length was reduced. The classifications by the SPRT at the maximum test length were more consistent with the classifications at reduced test length for the information-based item selection methods than for random item selection. The decisions for the SS components were relatively inconsistent (PCD2$<$0.900). This was probably caused by the low number of items per component.

Table 3.4
Proportion of Consistent Decisions for Decisions with a Reduced Test Length for Language, Mathematics, and Study Skills

	Test length					Random	Weighting	Estimate
Decision	200	180	Consistency	200	180	180	180
WR	30	26	PCD2	0.926	0.896	0.895	0.896
TC	30	26	PCD2	0.923	0.914	0.916	0.915
VO	20	17	PCD2	0.945	0.935	0.939	0.939
SP	20	20	PCD2	0.919	0.919	0.919	0.919
SPNV	10	10	PCD2	0.907	0.907	0.907	0.907
SPV	10	10	PCD2	0.903	0.903	0.903	0.903
MGTM	15	13	PCD2	0.889	0.924	0.926	0.942
RFP	20	17	PCD2	0.903	0.921	0.917	0.937
NC	25	21	PCD2	0.922	0.930	0.935	0.936
ST	10	10	PCD2	0.895	0.895	0.895	0.895
IS	10	10	PCD2	0.873	0.873	0.873	0.873
GE	10	10	PCD2	0.797	0.797	0.797	0.797
TG	10	10	PCD2	0.819	0.819	0.819	0.819
WR	30	26	PCD3	1.000	0.800	0.823	0.823
TC	30	26	PCD3	1.000	0.801	0.862	0.868
VO	20	17	PCD3	1.000	0.726	0.845	0.842
MGTM	15	13	PCD3	1.000	0.786	0.846	0.838
RFP	20	17	PCD3	1.000	0.807	0.867	0.866
NC	25	21	PCD3	1.000	0.811	0.868	0.866

Note. Proportion of consistent decisions for classifications based on the ability estimates and for the SPRT decisions (PCD2). WR: Writing; TC: Text Comprehension; VO: Vocabulary; MGTM: Measurement, Geometry, Time & Money; RFP: Relations, Fractions & Percentages; NC: Numbers & Computations; ST: Study Texts; IS: Information Sources; GE: Geography; TG: Tables & Graphs. Decisions for the components of Spelling and Study Skills were never shortened due to the already limited test length.

Cross-Validation of the Results

One important aspect of developing a test is to ensure that the results gathered during the analyses are valid for other sets of examinees as well. The End of Primary School Test was calibrated with 10% of the 2012 examinees. All analyses were repeated with nine other datasets consisting of 10% of the examinees each. The results for the entire and reduced test length (180) analyses with $\alpha = 0.05$ are presented in Table 3.5. The range in PCD was very small for the 10 datasets. These results implied that the item parameters from dataset 1 were representative

Table 3.5
Cross Validation of the Reported Proportions of Consistent Decisions

			Random	Estimate	Weighting
Decision	Consistency	200	180	180	180
Total test	PCD1	0.895-0.901	0.878-0.886	0.876-0.884	0.877-0.882
Mathematics	PCD2	0.969-0.973	0.965-0.968	0.965-0.971	0.966-0.970
Study skills	PCD2	0.934-0.941	0.934-0.941	0.934-0.941	0.934-0.941
Total test	PCD3	1.000-1.000	0.926-0.935	0.948-0.953	0.948-0.953
Language	PCD3	1.000-1.000	0.886-0.897	0.915-0.925	0.918-0.925
Mathematics	PCD3	1.000-1.000	0.880-0.890	0.917-0.924	0.917-0.926

Note. Proportion of consistent decisions for classifications based on interval scores and for the SPRT decisions on the entire test (PCD1), proportion of consistent decisions for classifications based on the ability estimates and for the SPRT decisions for decisions measured with one ability (PCD2), and proportion of consistent decisions for classifications with 200 and shorter tests (PCD3). Study Skills was never shortened due to the already limited test length of its components.

of the other datasets and that the consistency of the results was comparable over datasets. This last finding implies that the selected settings for the SPRT resulted in very consistent findings for different datasets.

3.7 Conclusions and Discussion

MCATs can be used to obtain an efficient ability estimate or a classification decision. In multidimensional classification testing, two methods are required. One method selects the items, and the second method determines whether testing can be stopped and which decision is made.

Several item selection methods are available for MCATs with ability estimation (Luecht, 1996; Mulder & Van der Linden, 2009; Reckase, 2009; Segall, 1996; Veldkamp & Van der Linden, 2002; Yao, 2012), but specialized methods for classification testing are scarce (Van Groen et al., 2014b). Seitz and Frey (2013) selected items using a method that was developed to obtain an efficient ability estimate (Segall, 1996). Here, items were also selected using a method that incorporated the ability estimate and the cutoff points into the selection procedure. This weighting method was developed for unidimensional classification testing (Van Groen et al., 2014a), but it could be used for MCAT with between-dimensionality. Using an empirical example, this method resulted in more consistent decisions than

selecting items using Segall's (1996) method, which was used by Seitz and Frey (2013). The differences in consistency were small except for random item selection.

A classification method to make multidimensional decisions was developed by Van Groen et al. (2014b) when each item measures multiple abilities. Seitz and Frey (2013) developed a method to make multidimensional decisions for tests in which each item measures one dimension. Their method makes a decision per dimension, but it cannot make a decision over all dimensions. Seitz and Frey's (2013) method was expanded here so that (a) a classification decision can be made on the entire test, (b) classification decisions can be made on several dimensions, and (c) classification decisions can be made on a part of the items for a dimension. The expansions make it possible to report an elaborate knowledge profile per examinee based on the examinee's answers. Per test, several decisions can be made, and per decision, several decision levels can be used. These knowledge profiles can be used by teachers to adapt their instruction to the examinee's ability.

A measure was developed for reporting the support for the decisions by the SPRT. The measure can be used for reporting, to examinees and their teachers, whether enough support was gained to make a decision or that the decision was forced by reaching the end of the test. The interpretation of the support measure is independent of the specifications that were used to make the decision, and the values of the measure can be compared between decisions. The support measure can be used for all SPRTs that are based on the decision rules, as specified here. This implies that the measure can also be used for unidimensional and within-dimensional classification decisions.

3.7.1 Future Directions and Further Remarks

The analyses were run with a fixed test length, although test length was often reduced to a fixed lower number. The reason for fixing the test length was that hardly any tests were shortened by the SPRT in initial analyses with a flexible test length. As Seitz and Frey (2013) reported, the average test length required to make decisions increases if the number of cutoff points increases. Four cutoff points were specified in the example, which is one explanation for the high number of tests that were stopped at maximum test length. A second explanation is that the test length of the components of the test had already been optimized by the test developers. In real tests, test length is a compromise between accurate scale scores and testing time. This makes it difficult to reduce test length even further.

Some directions for future research were specified. Other classification methods could be developed for MCAT because only methods based on the SPRT are available. Currently, only three item selection methods have been investigated for between-dimensionality classification testing. Other existing methods for UCAT and MCAT could also be investigated. Classification and item selection methods could also be developed for other multidimensional IRT models, such as models in which abilities cannot compensate each other or models with a guessing parameter.

Also more knowledge about the differences and similarities of within- and between-dimensionality is welcome. Comparisons could also be made between unidimensional and between-dimensional models. The accuracy of decisions with a unidimensional classification method for data simulated with a multidimensional model could also be investigated. Since real data were used here, no true classification was available. This implied that it was not possible to investigate it here. Current research on MIRT and MCAT uses Bayesian ability estimates. The prior required for Bayesian computations influences the ability estimates, especially at the start of the test. This implies that more research could be conducted on (weighted) maximum likelihood alternatives, in which no prior is required.

More studies are welcome on practical implementations of MCAT because few studies available concern applications, or use empirical data. Typically, simulations are run with an empirical or simulated item bank. Interesting complications arise if actual test data are used. Although no "true" classification is available, effectiveness of CAT methods can be studied only with real data. Furthermore, more knowledge is welcome about the way test results are reported to examinees and teachers. Test results can be used only for educational improvement if the methods of reporting them is closely aligned with the testing purpose.

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*, *B12*, 137–144.

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*, *23*, 249–261. doi: 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*, *60*, 713–734. doi: 10.1177/00131640021970862

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computerassisted branded test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh, PA.

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].

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*, *216*, 89–101. doi: 10.1027/0044-3409.216.2.89

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404. doi: 10.1177/014662169602000406

Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296. doi: 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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. doi: 10.1177/014662168500900409

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi: 10.1007/BF02294343

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*(1), 105–123.

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 Statistics*, *20*(4), 502–522.

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, April). *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*, *69*, 778–793. doi: 10.1177/0013164408324460

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

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 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*, *38*, 187–200. doi: 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*

within-dimensionality. Manuscript submitted for publication.

Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. doi: 10.1007/BF02295132

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: University of Iowa.

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*, *28*, 295–316. doi: 10.1177/0146621604265938

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi: 10.1007/BF02294627

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*, *77*, 495–523. doi: 10.1007/S11336-012-9265-5

Chapter 4
Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Within-Dimensionality

Abstract

A classification method is presented for adaptive classification testing with a multidimensional item response theory model in which items are intended to measure multiple traits, e.g., within-dimensionality. The reference composite is used in conjunction with the sequential probability ratio test (SPRT) to make decisions and decide whether testing can be stopped before the maximum test length is reached. Item selection methods are provided that maximize the determinant of the information matrix at the cutoff point or the projected ability estimate. A simulation study illustrates the efficiency and effectiveness of the classification method for different settings of the SPRT. Simulations were run with the two new item selection methods, random item selection, and maximization of the determinant of the information matrix at the ability estimate in the multidimensional space. The study also showed that the multidimensional SPRT has the same characteristics as the unidimensional SPRT and outperforms the unidimensional SPRT when applied to multidimensional data.

This chapter has been submitted as Van Groen, M.M., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). *Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Within-Dimensionality.* Manuscript submitted for publication.

4.1 Introduction

Computerized adaptive testing (CAT) estimates ability precisely or makes accurate decisions while minimizing test length. Much is known about unidimensional CAT (UCAT), but knowledge about multidimensional CAT (MCAT) has especially expanded over the last several years. Nevertheless, knowledge about MCAT to make classification decisions using item response theory (IRT) is scarce.

Several classification methods are available for constructs modeled with a unidimensional IRT model (Eggen, 1999; Spray, 1993; Weiss & Kingsbury, 1984). However, if the construct is modeled using a multidimensional IRT (MIRT) model, classification methods are available only for some situations. Seitz and Frey (2013) developed a multidimensional classification method that makes a decision per dimension using Wald's (1947/1973) sequential probability ratio test (SPRT). This method can be used if each item is assumed to measure only one trait. Spray, Abdel-Fattah, Huang, and Lau (1997) investigated the use of the SPRT with MIRT for items that are assumed to measure multiple traits. They concluded that using the SPRT is not feasible, because the likelihood ratio cannot be updated with unique probabilities after an additional item is administered. If a multidimensional decision is required for testing with items that measure multiple traits, no method is available. Therefore, a new method was developed to make such decisions.

The advantages of making multidimensional decisions are that the multidimensional structure of the data is respected, adaptive testing principles can be used, and test length is reduced even more than in MCAT for estimating ability.

IRT is often used for CAT and relates the score on an item based on the item parameters and the examinee's ability (Van der Linden & Hambleton, 1997). In MIRT, a vector of person abilities describes the skills and knowledge that the person brings to the test (Reckase, 2009). MIRT is discussed in the first part.

CAT for classification purposes requires two methods. One method decides whether testing can be finished and which decision can be made regarding the level of the examinee. The second method selects the items based on a statistical criterion or on the examinee's responses to previously administered items. A new method to make multidimensional classification decisions and some existing methods are discussed in the second part. Existing item selection methods for unidimensional classification testing, available methods for MCAT for estimating ability, and new methods for MCAT for classification testing are discussed in the third part.

The efficiency and effectiveness of the new classification and item selection methods are illustrated using simulation studies. In the last section, some remarks are made about multidimensional classification testing and directions for future research.

4.2 Multidimensional Item Response Theory

A prerequisite for CAT is a calibrated item pool suitable for the specific testing situation. In a calibrated item pool, the fit of the model is established, item parameter estimates are available, and items with undesired characteristics have been removed (Van Groen, Eggen, & Veldkamp, 2014). In MIRT, a set of p abilities is assumed to account for the examinee's responses to the items. The MIRT model used here is the dichotomous two-parameter logistic model (Reckase, 1985), in which the probability of a correct answer to item i is described by

$$\mathrm{P}_i(\boldsymbol{\theta}) = \mathrm{P}\left(x_i = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}\right) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}, \tag{4.1}$$

where $\mathrm{P}_i(\boldsymbol{\theta})$ is the probability of a correct answer $x_i = 1$, $\mathbf{a}_i$ is the vector of the discrimination parameters, d_i denotes the easiness of the item, and $\boldsymbol{\theta}$ is the vector of the ability parameters. Items that have multiple non-zero parameters $\mathbf{a}_i$ measure multiple abilities, a situation in which within-item dimensionality is present (W.-C. Wang & Chen, 2004). If just one discrimination parameter is non-zero for all test items, the test is considered to have between-item dimensionality. The item parameters are assumed to be estimated with enough precision to consider them known during test administration (Veldkamp & Van der Linden, 2002).

Owing to the assumption of local independence, the probability of a correct response to a set of items is a function of only the ability and item parameters (Segall, 1996). The likelihood of a vector of observed responses $\mathbf{x}_j$ to items $i = 1, \cdots, k$ for an examinee j with ability $\boldsymbol{\theta}_j$ equals the product of the probabilities associated with the responses to the administered items (Segall, 1996):

$$\mathrm{L}\left(\boldsymbol{\theta}_j | \mathbf{x}_j\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_j\right)^{x_{ij}} \mathrm{Q}_i\left(\boldsymbol{\theta}_j\right)^{1-x_{ij}}, \tag{4.2}$$

where $\mathrm{Q}_i(\boldsymbol{\theta}_j) = 1 - \mathrm{P}_i(\boldsymbol{\theta}_j)$. The vector of values $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)$ that maximize the likelihood function is taken as the estimate of ability $\boldsymbol{\theta}_j$ (Segall, 1996). The equations for finding maximum likelihood (ML) estimates have no closed-form solution

(Segall, 1996). An iterative search procedure, such as the Newton-Raphson method, must be used to find the estimates. This procedure is described for weighted maximum likelihood (WML) in the appendix. WML estimation, which was developed by Tam (1992) who extended Warm's (1989) unidimensional estimator, was applied because it reduces the bias in the ML estimates.

4.3 Classification Methods

Several termination methods exist for MCAT for estimating ability (C. Wang, Chang, & Boughton, 2013; Yao, 2013), but the focus here is on methods for MCAT for classification testing. These methods determine whether testing can be finished and which decision is made before the maximum test length is reached (Van Groen et al., 2014). Few methods are available to make classification decisions with MIRT. These approaches are described first. Due to the limitations of the available methods, in the final part of this section a new method is proposed to make decisions.

4.3.1 Existing Multidimensional Classification Methods

Two studies about making decisions using MIRT exist (Seitz & Frey, 2013; Spray et al., 1997). These studies concern MCAT with multiple unidimensional decisions for between-dimensionality (Seitz & Frey, 2013) and about the multidimensional SPRT for within-dimensionality (Spray et al., 1997).

MCAT for Between-Dimensionality

Seitz and Frey (2013) described an approach for MCAT in which multiple unidimensional decisions are made. They based their method on the fact that, in case of between-dimensionality, the multidimensional two-parameter logistic model is a combination of unidimensional two-parameter logistic models (W.-C. Wang & Chen, 2004). Seitz and Frey (2013) implemented the unidimensional classification method based on the sequential probability ratio test (SPRT) per dimension.

The SPRT (Wald, 1947/1973) was applied to classification testing by Ferguson (1969) using classical test theory and Reckase (1983) using IRT and used by Eggen (1999), Spray (1993), and Thompson (2009), among others. If the SPRT is applied, a cutoff point is set between adjacent levels with a surrounding indifference region. This region accounts for the uncertainty of the decisions, owing to measurement error, regarding examinees with ability values close to the cutoff point (Eggen,

1999). Two hypotheses are formulated for the cutoff point, θ_c, using the boundaries of the indifference region (Eggen, 2010):

$$\mathrm{H}_0 : \theta_j < \theta_c - \delta; \tag{4.3}$$

$$\mathrm{H}_a : \theta_j > \theta_c + \delta, \tag{4.4}$$

in which δ denotes the distance between the cutoff point and the boundary of the indifference region. The likelihood ratio between the hypotheses after k items are administered is calculated for the unidimensional SPRT (Eggen, 2010):

$$\mathrm{LR}\left(\theta_c + \delta; \theta_c - \delta\right) = \frac{\mathrm{L}\left(\theta_c + \delta; \mathbf{x}_j\right)}{\mathrm{L}\left(\theta_c - \delta; \mathbf{x}_j\right)}, \tag{4.5}$$

in which $\mathrm{L}\left(\theta_c + \delta; \mathbf{x}_j\right)$ and $\mathrm{L}\left(\theta_c - \delta; \mathbf{x}_j\right)$ are calculated using the unidimensional version of Equation 4.2. Decision rules are used to decide to continue testing or to decide that the student's ability is below or above the cutoff point (Eggen, 1999):

$$\begin{aligned}
&\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) < (1-\beta)/\alpha; \\
&\text{ability below } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \leq \beta/(1-\alpha); \\
&\text{ability above } \theta_c \text{ if} && \mathrm{LR}(\theta_c + \delta; \theta_c - \delta) \geq (1-\beta)/\alpha,
\end{aligned} \tag{4.6}$$

where α and β specify the acceptable classification error rates (Spray et al., 1997). A maximum test length is set in practical testing situations to ensure that testing stops at some point (Eggen, 1999). At the maximum test length, the examinee is classified as having ability above the cutoff point if the likelihood ratio is larger than the midpoint of the interval of Equation 4.6.

Seitz and Frey (2013) implemented the SPRT by setting cut scores, θ_{cl}, for all dimensions $l = 1, \cdots, p$, with surrounding indifference regions. The SPRT is calculated for each dimension p using

$$\mathrm{LR}\left(\theta_{cl} + \delta; \theta_{cl} - \delta\right) = \frac{\mathrm{L}\left(\theta_{cl} + \delta, \hat{\boldsymbol{\theta}}_{jc-l}, \mathbf{x}\right)}{\mathrm{L}\left(\theta_{cl} - \delta, \hat{\boldsymbol{\theta}}_{jc-l}, \mathbf{x}\right)}, \quad l = 1, \cdots, p, \tag{4.7}$$

in which $\theta_{cl} - \delta$ and $\theta_{cl} + \delta$ are imputed for dimension l and $\hat{\boldsymbol{\theta}}_{jc-l}$ denote the provisional estimates for all dimensions except dimension l. Since no decision is required on the other dimensions when making the decision for dimension l, ability estimates are imputed for the other dimensions (Seitz & Frey, 2013). In the case of between-dimensionality, Equation 4.7 reduces to

$$\text{LR}\left(\theta_{cl}+\delta;\theta_{cl}-\delta\right)=\frac{\text{L}\left(\theta_{cl}+\delta;\mathbf{x}_j\right)}{\text{L}\left(\theta_{cl}-\delta;\mathbf{x}_j\right)},\quad l=1,\cdots,p. \tag{4.8}$$

If items load on multiple dimensions, Seitz and Frey's (2013) method cannot be used because the ratio does not reduce to Equation 4.8. Furthermore, the method requires the specification of an additional decision rule if one global decision is to be obtained for the test. This implies that Seitz and Frey's (2013) method can be used only for between-dimensional tests in which no overall decision is required.

MCAT for Within-Dimensionality

Spray et al. (1997) investigated the possibility of applying the SPRT to MCAT. They specified a passing rate on a reference test with a standard setting method and obtained an equivalent latent passing score by solving for $\boldsymbol{\theta}$. The ability values that result in the passing rate define the curve in the multidimensional space that divides the space into two mutually exclusive regions (Spray et al., 1997). Surrounding this curve, the curves denoting the indifference region are formed. According to Spray et al. (1997), the ability values that satisfy these curves do not necessarily result in constant probability values for each item. This implies that the likelihood ratio cannot be updated with a unique value for each item; thus, the SPRT cannot be extended to the multidimensional setting.

4.3.2 A Classification Method for Within-Dimensionality

Since the likelihood ratio requires unique values for updating the ratio, a method should be developed that results in unique values if the SPRT is to be applied. The reference composite, which was developed by M. Wang (1985, 1986) and described by Reckase (2009), reduces the multidimensional space to an unidimensional line. By using the reference composite, the likelihood ratio can be updated with unique values after an additional item is administered.

Reference Composite

The reference composite can be used to relate the multidimensional abilities to a unidimensional line in the multidimensional space (Reckase, 2009). This line describes the characteristics of the matrix of the discrimination parameters for the set of items. All $\boldsymbol{\theta}$ points can be projected on the reference composite. Using projection, examinees are ranked on the reference composite. A higher value on the reference composite denotes a students who is more able than a student with

a lower value on the reference composite. To distinguish between ability in the $\boldsymbol{\theta}$ space and ability as projected on the reference composite, the latter will be called proficiency ξ.

The direction of the reference composite is given by the eigenvector of the $\mathbf{aa}'$ matrix that corresponds to the largest eigenvalue of this matrix (Reckase, 2009). The p elements of the eigenvector are the direction cosines $\alpha_{\xi l}$ for the angle between the reference composite and the p dimension axes. The line is drawn in the multidimensional space through the origin with the direction cosines specifying the precise position of the line.

To calculate the proficiency, an additional line is drawn through the $\boldsymbol{\theta}$-point and the origin (Reckase, 2009). The length of this line L_j for an examinee j from the origin to $\hat{\boldsymbol{\theta}}_j$ point is given by (Reckase, 2009)

$$\mathrm{L}_j = \sqrt{\sum_{l=1}^{p} \hat{\theta}_{jl}^2} \tag{4.9}$$

and the direction cosines for the line are calculated using (Reckase, 2009)

$$\cos \alpha_{jl} = \frac{\hat{\theta}_{jl}}{\mathrm{L}_j}, \quad l = 1, \cdots, p, \tag{4.10}$$

in which α_{jl} is the angle between axis l and line L_j. The angle, $\alpha_{j\xi} = \alpha_{jl} - \alpha_{\xi l}$, between L_j and the reference composite is used to calculate the estimated proficiency $\hat{\xi}_j$ on the reference composite (Reckase, 2009):

$$\hat{\xi}_j = \mathrm{L}_j \cos \alpha_{j\xi}; \tag{4.11}$$

The calculation of $\hat{\xi}_j$ is illustrated in Figure 4.1 for two dimensions.

Multidimensional Decision Making Using the Reference Composite

Using the reference composite, the examinees' ability can be ranked on a unidimensional line. The position of the reference composite is fixed before test administration based on all items in the item pool. By fixing the reference composite, ability is measured on the same scale for all examinees and that it is possible to set cutoff points.

The SPRT requires specifying a cutoff point, ξ_c, and the surrounding indifference region. The cutoff point and δ^{ξ} are set on the reference composite. The

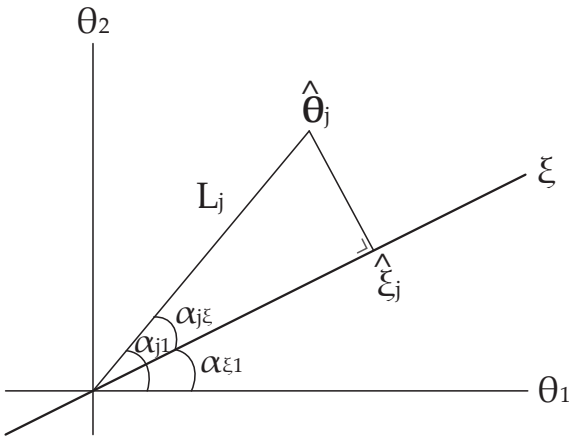


Figure 4.1. Projection of ability on the reference composite.

boundaries of the indifference region for the SPRT can be transformed to their respective $\boldsymbol{\theta}$ points using

$$\boldsymbol{\theta}_{\xi_{c+\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c + \delta^{\xi}); \tag{4.12}$$

$$\boldsymbol{\theta}_{\xi_{c-\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c - \delta^{\xi}), \tag{4.13}$$

where $\boldsymbol{\alpha}_{\xi}$ includes all angles between the reference composite and the dimension axis. The likelihood ratio in Equation 4.5 now becomes

$$\mathrm{LR}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c-\delta}}; \mathbf{x}_j\right)}, \tag{4.14}$$

which can be used to make multidimensional classification decisions with the following decision rules

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) < (1-\beta)/\alpha; \\ &\text{ability below } \xi_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \leq \beta/(1-\alpha); \\ &\text{ability above } \xi_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \geq (1-\beta)/\alpha. \end{aligned} \tag{4.15}$$

4.4 Item Selection Methods

The item selection method is a critical component of MCAT, because selecting items that are too hard or too easy or provide little information results in tests that do not function well (Reckase, 2009). Several methods are available for MCAT for obtaining an efficient ability estimate (e.g., Luecht, 1996; Mulder & Van der Linden, 2009; Reckase, 2009; Segall, 1996; Veldkamp & Van der Linden, 2002; C. Wang, Chang, & Boughton, 2011; Yao, 2012, 2013). The literature also provides

several methods for UCAT for classification testing (e.g., Eggen, 1999; Spray & Reckase, 1994; Thompson, 2009) such as maximizing information at the cutoff point or at the ability estimate. Nevertheless, methods to select items for MCAT to make classification decisions are scarce. Seitz and Frey (2013) selected items using Segall's (1996) method for MCAT for estimating ability. This method is discussed in the next section. Unidimensional item selection methods for classification testing are described in the second section. In the third section, these methods are adapted for application to MCAT using Segall's (1996) item selection method.

4.4.1 An Item Selection Method for MCAT for Ability Estimation

The method that maximizes the determinant of the Fisher information matrix was developed for MCAT for estimating ability (Segall, 1996). Fisher information is a measure of the information in the observable variables on the ability parameters (Mulder & Van der Linden, 2009). The elements of $p \times p$ matrix $\mathbf{I}(\boldsymbol{\theta})$ for dimensions l and m are defined as (Tam, 1992)

$$\mathrm{I}(\theta_l, \theta_m) = \sum_{i=1}^{k} \frac{\frac{\partial}{\partial \theta_l}\mathrm{P}_i(\boldsymbol{\theta}) \times \frac{\partial}{\partial \theta_m}\mathrm{P}_i(\boldsymbol{\theta})}{\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})} = \sum_{i=1}^{k} a_{il}a_{im}\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta}). \tag{4.16}$$

Segall's (1996) item selection method is based on the relationship between the information matrix and the confidence region around the estimates (Reckase, 2009). In the multidimensional case, the confidence interval forms an ellipsoid surrounding the estimate (Reckase, 2009). The method selects the item that results in the largest decrement in the volume of the confidence ellipsoid (Segall, 1996). Since the size of the confidence ellipsoid can be approximated by the inverse of the information matrix, the item can be selected that maximizes (Segall, 1996)

$$\max \det \left(\sum_{i=1}^{k} \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij}) + \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j}) \right), \quad \text{for } k+1 \in \mathrm{V}_{k+1}, \tag{4.17}$$

which is the determinant of the information matrix of the administered items and the potential item $k+1$. The left term denotes the information provided by the items that were previously administered. The right term is the information a potential item $k+1$ can provide. The next item is administered that, when added to the information matrix, will result in the largest determinant of the matrix. This implies that the volume of the confidence ellipsoid around the ability estimate is minimized (Reckase, 2009).

4.4.2 Item Selection Methods for UCAT for Classification Testing

In unidimensional classification testing, two methods are commonly used in addition to random item selection. The first method maximizes Fisher information at the ability estimate. Segall (1996) adapted this method for MCAT. The unidimensional version minimizes the confidence interval surrounding the ability estimate using

$$\max \mathrm{I}_i(\hat{\theta}_j), \quad \text{for } i \in \mathrm{V}_a, \tag{4.18}$$

where V_a denotes the set of items in the item bank still available for administration.

The second method maximizes Fisher information at a different point on the ability scale. The method maximizes information at the cutoff point, which results in the following objective function:

$$\max \mathrm{I}_i(\theta_c), \quad \text{for } i \in \mathrm{V}_a. \tag{4.19}$$

In unidimensional IRT with the SPRT, maximizing information at the cutoff point is considered to be the most efficient (Eggen, 1999; Spray & Reckase, 1994).

4.4.3 Item Selection Methods for MCAT for Classification Testing

Segall's (1996) item selection method selects the item with the largest determinant of the information matrix at the current ability estimate. When making multidimensional classification decisions with the SPRT, this method can be adapted to select items that maximize on some fixed point on the reference composite, analogous to the item selection methods for unidimensional classification testing.

The first item selection method for multidimensional classification testing maximizes the determinant of the information matrix at the projected ability estimate. The rationale behind this method is that interest is limited here to the points that fall on the reference composite, but not in all other points in the multidimensional space. Thus maximizing at the reference composite seems to make sense. The current ability estimate is estimated using weighted maximum likelihood estimation (see Appendix). The estimate can be projected onto the reference composite using Equation 4.11. To calculate the information matrix, $\hat{\xi}_j$ is transformed to its corresponding point in the multidimensional space using $\boldsymbol{\theta}_{\hat{\xi}_j} = \cos \boldsymbol{\alpha}_{\xi} \times \hat{\xi}_j$. The objective function for the item selection method then becomes

$$\max \det \left(\sum_{i=1}^{k} \mathrm{I}(\boldsymbol{\theta}_{\hat{\xi}_j}, x_{ij}) + \mathrm{I}(\boldsymbol{\theta}_{\hat{\xi}_j}, x_{k+1,j}) \right), \quad \text{for } k+1 \in \mathrm{V}_{k+1}. \tag{4.20}$$

The second item selection method for multidimensional classification testing maximizes the determinant of the information matrix at the cutoff point on the reference composite. This value is already on the reference composite, but has to be transformed to the multidimensional $\boldsymbol{\theta}$ space using

$$\boldsymbol{\theta}_c = \cos \boldsymbol{\alpha}_{\tilde{\xi}} \times \tilde{\xi}_c. \tag{4.21}$$

The resulting objective function is

$$\max \det \left(\sum_{i=1}^{k} \mathrm{I}(\boldsymbol{\theta}_c, x_{ij}) + \mathrm{I}(\boldsymbol{\theta}_c, x_{k+1,j}) \right), \quad \text{for } k+1 \in \mathrm{V}_{k+1}. \tag{4.22}$$

4.5 Simulation Study

The effectiveness and the efficiency of the developed classification and item selection methods were investigated using a simulation study. The results with the multidimensional SPRT were evaluated using well-known characteristics of the unidimensional SPRT. One well-known characteristic of the unidimensional SPRT is that an increase in α and β often does not influence the accuracy of the classifications, but shortens the test considerably (Eggen & Straetmans, 2000). An increase in the size of δ also results in shorter tests, but does not influence accuracy (Eggen & Straetmans, 2000). The item selection methods were compared with random item selection. It was expected that maximizing the determinant of the information matrix at the cutoff point, the projected ability estimate, or the ability estimate would result in shorter and more accurate tests than selecting items at random.

4.5.1 Simulation Design

An item pool from the ACT Assessment Program, which was used by Ackerman (1994) and Veldkamp and Van der Linden (2002), was used to evaluate MCAT for classification testing. The item pool consisted of 180 items, previously calibrated with a two-dimensional compensatory IRT model with within-dimensionality using NOHARM 4 (Fraser & McDonald, 1988). The fit of the MIRT model with two dimensions was established (Veldkamp & Van der Linden, 2002). The items

in the pool were previously classified according to the six content and three skill categories used in the ACT Assessment Program (Veldkamp & Van der Linden, 2002), but a two-dimensional model fitted the data. The means of the discrimination parameters were 0.422 and 0.454 with standard deviation 0.268 and 0.198. The observed correlation between the parameters was 0.093, which can be explained by the orthogonal constraint in the calibration. The mean of the easiness parameter was -0.118 with standard deviation 0.568. The matrix of the discrimination parameters resulted in angles between the dimension axis 1 and 2 with the reference composite of 44.621 and 45.379 degrees.

Simulations were run for four item selection methods: random selection and maximization of information at the cutoff point, the projected ability estimate, and the ability estimate. The maximum test length was set at 50 items, following Veldkamp and Van der Linden (2002). The acceptable decision error rates α and β were set equal in each simulation. α and β were set at 0.05 and 0.10 with $\delta = 0.1$, 0.2, and 0.3. The chosen values for α and β are commonly found in unidimensional CCT. In each simulation, 1,000 simulees were generated from a multivariate standard normal distribution. The correlation between the dimensions was varied, $\rho = 0.0, 0.3$, and 0.6. The cutoff point was set at 0.0 at the reference composite, which implied $\boldsymbol{\theta}_{\xi_c} = \{0.0, 0.0\}$ and was the midpoint of the ability distribution. Each simulation condition was replicated 100 times.

A well-known characteristic of the unidimensional SPRT is that as ability becomes closer to the cutoff point, the test length increases (Eggen & Straetmans, 2000), and the proportion of correct decisions nears 0.5 (Van Groen & Verschoor, 2010). Additional simulations were run to investigate the effect of the distance between ability and the cutoff point. This study used 372,100 simulees: 100 at each of 61 evenly spaced points on θ_1 from -3 to 3 combined with the same number of points on θ_2. The maximum test length was again set at 50 items. $\alpha = \beta = 0.10$, $\delta = 0.20$, the cutoff point was set at 0.0, and the items were selected by the method that had the largest determinant of the information matrix at the cutoff point.

The multidimensional classifications were compared with unidimensional classifications in a third series of simulations. Although a two-dimensional model was required for model fit, which implied the use of a multidimensional classification method, a comparison was made with unidimensional classification testing. One hundred datasets were generated consisting of 180 items and 10,000 simulees each. The datasets were calibrated using NOHARM (multidimensional) and BILOG (unidimensional). The cutoff points were set for each dataset and calibration at the

median of the observed ability distribution. The indifference regions were set 0.1, 0.2, and 0.3 times the standard deviation of the observed ability distribution. This ensured that the unidimensional and multidimensional classification decisions were made using the same criteria, $\alpha = \beta = 0.05$ and 0.10. Simulations were run with a minimum test length of 3 and a maximum test length of 50 in addition to a fixed test length of 50 items. To exclude the influence of item selection methods, items were selected at random.

4.5.2 Dependent Variables

Efficiency of the multidimensional SPRT was evaluated with the average test length (ATL), which was calculated per condition as the mean test length over 1,000 simulees of 100 replications. Although reducing the test length reduces respondent burden, test development costs, and test administration costs, effectiveness was considered to be more important. Effectiveness was investigated using the proportion of correct decision (PCD). This was calculated per condition as the mean of the PCD for each simulation over 100 replications. The PCD compared the classification based on the reference composite used to generate the data with the decisions by the SPRT. The decision based on the SPRT was compared to the classification based on the $\boldsymbol{\theta}$ used to generate data.

4.5.3 Simulation Results

Table 4.1 presents the ATL for different settings for the SPRT and with four item selection methods. The performance of random item selection (RA), maximization of the determinant of the information matrix at the projected ability estimate (PA), ability estimate (AE), or the cutoff point (CP) was evaluated. RA resulted in the longest tests. CP resulted in the shortest tests. An increase in α and β decreased ATL with several items. An increase in δ also resulted in shorter tests. If the correlation between the abilities was increased, the test length also decreased.

The effectiveness of the classification method is shown in Table 4.2. PCD is given for simulations with different settings for the SPRT and with four item selection methods. RA was the least accurate method. The PCD was lower for the simulations with $\alpha = \beta = 0.05$ than was specified beforehand. The simulations with the other three item selection methods were more accurate, and the differences between these were negligible. α, β, and δ appeared to have no influence on the PCD. If the correlation between the abilities was higher, a more accurate classification decision was made.

Table 4.1
Average Test Length for Different SPRT Settings and Item Selection Methods

	$\alpha = 0.05$			$\alpha = 0.10$		
Condition	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$
Random item selection						
0.0	50.000	48.404	43.598	49.916	45.303	38.431
0.3	49.999	47.778	42.383	49.850	44.217	36.949
0.6	49.998	47.186	41.324	49.752	43.218	35.725
Item selection by maximization at the ability estimate						
0.0	49.996	43.165	35.065	48.792	37.411	28.624
0.3	49.993	41.819	33.541	48.341	35.918	27.318
0.6	49.989	40.785	32.310	47.887	34.674	26.188
Item selection by maximization at the projected estimate						
0.0	49.990	43.362	35.359	48.898	37.673	28.893
0.3	49.992	43.911	35.965	49.165	38.261	29.824
0.6	49.981	43.045	34.890	48.952	37.160	28.739
Item selection by maximization at the cutoff point						
0.0	49.531	40.851	32.337	47.269	34.733	25.840
0.3	49.234	39.264	30.660	46.268	33.022	24.230
0.6	48.901	37.815	29.120	45.399	31.541	22.956

Note. The correlation between the abilities is provided in the column labeled Condition. $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

Simulations were run to investigate whether the ATL and the PCD depended in the same way on the distance between ability and the cutoff point as in the UCCT. In Figure 4.2, the ATL and the PCD are shown for different combinations of ability. The ATL increased considerably if the combination of both ability values was close to the cutoff point. The PCDs for simulations with a combination of ability values that was close to the cutoff point decreased considerably. If ability was close to zero, the PCDs became close to 0.50. The PCDs were even lower than 0.50 for other combinations of $\boldsymbol{\theta}$ values with a mean value near the cutoff point.

The ATL is shown in Table 4.3 for the UCCT and MCCT simulations with a flexible test length, but not for the fixed 50-item tests. The ATL was almost equal for the simulations with no correlation between abilities. If a correlation was specified between the abilities, the ATL was much shorter for the MCCTs than for the UCCTs. The differences in the ATL increased if a higher value was specified for the correlations, α, β, or δ.

Table 4.2
Proportion of Correct Decisions for Different SPRT Settings and Item Selection Methods

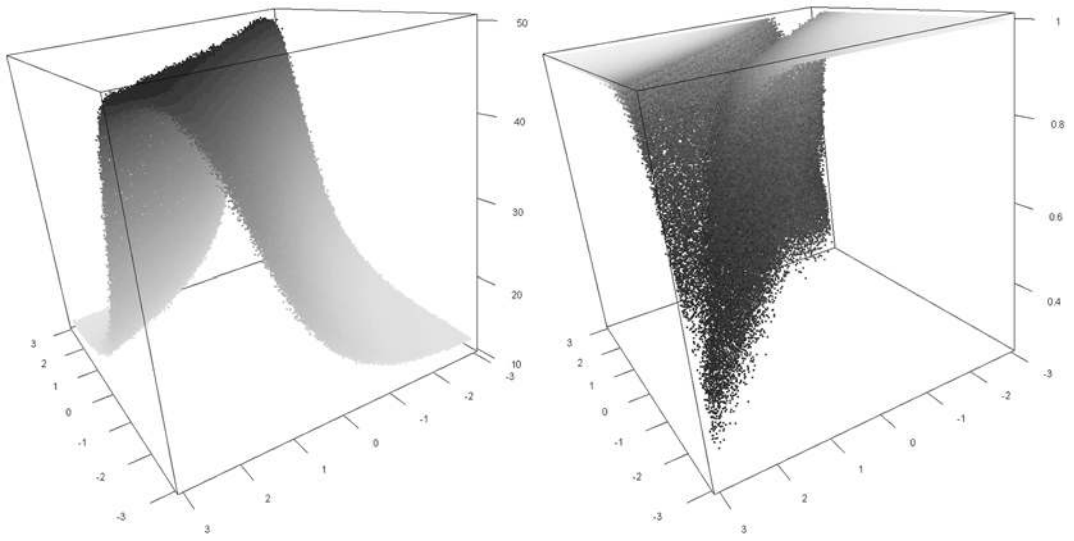
	$\alpha = 0.05$			$\alpha = 0.10$		
Condition	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$
Random item selection						
0.0	0.865	0.866	0.868	0.866	0.867	0.867
0.3	0.880	0.882	0.882	0.883	0.882	0.882
0.6	0.892	0.894	0.893	0.893	0.893	0.890
Item selection by maximization at the ability estimate						
0.0	0.895	0.896	0.897	0.895	0.896	0.895
0.3	0.908	0.907	0.906	0.909	0.908	0.908
0.6	0.918	0.917	0.918	0.916	0.915	0.915
Item selection by maximization at the projected estimate						
0.0	0.896	0.895	0.895	0.897	0.895	0.894
0.3	0.903	0.901	0.904	0.904	0.904	0.901
0.6	0.910	0.912	0.912	0.912	0.912	0.911
Item selection by maximization at the cutoff point						
0.0	0.898	0.896	0.898	0.897	0.898	0.897
0.3	0.909	0.910	0.909	0.908	0.910	0.909
0.6	0.919	0.920	0.918	0.917	0.919	0.919

Note. The correlation between the abilities is provided in the column labeled Condition. $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

The PCDs for the unidimensional versus multidimensional comparison study are shown in Table 4.4. The differences between the PCDs for unidimensional and classification testing were negligible. The only notable difference was if a large δ was specified in conjunction with a high correlation and a high α.

4.5.4 Discussion of the Results

The main aim of the simulation study was to investigate whether typical characteristics of the unidimensional SPRT also applied to the multidimensional SPRT using a reference composite. The effect of the settings for α and β on the unidimensional SPRT is that an increased value resulted in shorter tests, but accuracy was not influenced (Eggen & Straetmans, 2000). The simulations with the multidimensional SPRT demonstrated similar effects on the PCD and the ATL. Another characteristic of the unidimensional SPRT is that if the indifference region is increased, the ATL



(a) Average test length (b) Proportion correct decisions

Figure 4.2. Average test length and proportion of correct decisions with maximization on the cutoff point.

Table 4.3
Average Test Length for Different SPRT Settings for Unidimensional and Multidimensional Classifications

	$\alpha = 0.05$			$\alpha = 0.10$		
Condition	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$
UCCT flexible test length						
0.0	49.999	48.242	43.312	49.892	45.010	37.964
0.3	49.975	46.165	39.744	49.426	41.791	33.816
0.6	49.834	43.858	36.523	48.539	38.731	30.415
MCCT flexible test length						
0.0	49.999	48.251	43.325	49.896	45.012	37.938
0.3	49.534	42.937	35.381	47.895	37.677	29.084
0.6	44.986	32.209	23.441	40.191	25.877	17.321

Note. Simulations for unidimensional (UCCT) and multidimensional classification testing (MCCT) with a flexible test length and random item selection. The correlation between the abilities is provided in the first column. $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region

decreases, and the PCD is not influenced (Eggen & Straetmans, 2000). The same was found for the multidimensional SPRT. A third characteristic typical of the SPRT is inaccuracy if ability approaches the cutoff point (Van Groen & Verschoor, 2010). The simulations showed that the tests were considerably longer if the distance between the cutoff point and the combination of the ability values became

Table 4.4
Proportion of Correct Decisions for Different SPRT Settings for UCCT and MCCT

	$\alpha = 0.05$			$\alpha = 0.10$		
Condition	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$
UCCT flexible test length						
0.0	0.867	0.867	0.886	0.867	0.866	0.866
0.3	0.882	0.881	0.881	0.882	0.881	0.881
0.6	0.889	0.889	0.889	0.889	0.889	0.888
UCCT fixed test length						
0.0	0.867	0.867	0.867	0.866	0.866	0.867
0.3	0.881	0.881	0.881	0.881	0.881	0.881
0.6	0.889	0.889	0.889	0.889	0.889	0.889
MCCT flexible test length						
0.0	0.867	0.865	0.866	0.866	0.866	0.866
0.3	0.881	0.881	0.881	0.881	0.880	0.879
0.6	0.888	0.888	0.881	0.888	0.884	0.867
MCCT fixed test length						
0.0	0.866	0.866	0.866	0.866	0.866	0.866
0.3	0.881	0.881	0.881	0.881	0.881	0.881
0.6	0.888	0.888	0.888	0.888	0.888	0.888

Note. Simulations for unidimensional (UCCT) and multidimensional classification testing (MCCT) with a fixed (50) and flexible test length and random item selection. The correlation between the abilities is provided in the first column.

very small. In the multidimensional case, this finding applied to all possible combinations of ability whose mean was close to the cutoff point.

The results of the simulation study were also in line with previous unidimensional findings by Spray and Reckase (1994), Eggen (1999), and Thompson (2009), in which selecting the items by maximizing the information at the cutoff point was the most efficient. Selecting the items that had the largest determinant of the information matrix at the cutoff point on the reference composite also resulted in the multidimensional case in the shortest tests. As expected, the other methods outperformed random item selection.

In the third part of the simulation study, the multidimensional SPRT resulted in much shorter tests than the unidimensional SPRT without influencing classification accuracy. If a multidimensional IRT model improves model fit for a specific dataset, the use of a multidimensional classification procedure not only makes sense from a philosophical point of view but also reduces test length considerably.

4.6 Conclusions and Discussions

A classification method was developed to make classification decisions in tests with items that are intended to measure multiple traits. The method can be used in testing situations in which the construct of interest is modeled using a multidimensional item response theory model. A reference composite is constructed in the multidimensional space. This reference composite is used to make multidimensional classification decisions with the sequential probability ratio test.

Segall's (1996) item selection method was adapted to select items that had the largest determinant of the information matrix at the cutoff point or the current projected ability estimate. The methods use the $\boldsymbol{\theta}$ point that corresponds to the intended point on the reference composite. The method based on the (projected) ability estimates uses the weighted maximum likelihood estimator developed by Tam (1992). The Newton-Raphson method was used for finding the values of the weighted maximum likelihood estimates (see Appendix).

Simulations were used to investigate the average test length, the proportion of correct decisions, and the characteristics of the classification method. The efficiency and the accuracy were compared for different item selection methods and different settings for the classification method.

The differences in efficiency and effectiveness between the item selection methods appeared to be small. The settings of the classification method had more influence on the average test length than on the proportion of correct decisions. Tests could be shortened considerably without much effect on the accuracy of the decisions. It was established that the multidimensional classification method had the same characteristics as the unidimensional version of the same method. The settings of the SPRT had the same influence as on the unidimensional SPRT. It was also shown that if the unidimensional SPRT was used for multidimensional data, the test length increased considerably.

4.6.1 Future Directions and Further Remarks

The current simulations were run with an item pool in which each item measured multiple dimensions. If the items load on one dimension, the new classification method cannot be used. If each item measures just one dimension, the non-diagonal elements of the $\mathbf{a}\mathbf{a}'$ matrix are zero. The eigenvalues and the eigenvectors of such a matrix do not make sense, and the resulting classification decisions are solely based on the dimension that discriminates the most.

As an example, simulations were run with an item pool that had been calibrated with a two-dimensional model. The classification method can be applied to models with more dimensions. A fixed test length can also be specified when using the SPRT. The method makes a classification decision after testing has ended.

In the current study, decisions were made based on the total set of items administered. Reckase (2009) showed that reference composites can be constructed for underlying domains as well. Investigating whether it is possible to classify at these domains as well would be interesting. Such classifications can provide information regarding the level of the examinees for the underlying domains.

The current version of the SPRT is used to classify into one of two levels. It is expected that an adaptation to the multidimensional classification method similar to that of Spray (1993) or Eggen and Straetmans (2000) for unidimensional classification testing can be made. This would enable test developers to classify examinees into one of multiple levels, such as basic, proficient, and advanced.

The simulations in the current study used an item bank in which dimensions were restricted to be orthogonal at each other. Although it has been shown in the past (Ackerman, 1994; Veldkamp & Van der Linden, 2002) that the model fits the data, fit of a model with orthogonality constraints cannot always be established for other datasets. The SPRT as described here can also be used if orthogonality is not assumed. The effects of fitting an orthogonal model and a not-orthogonal model to the same dataset should be investigated, and their resulting classification decisions should be compared.

A weighted maximum likelihood estimator was used in the current study. The effectiveness and efficiency of the estimator has not been intensively studied and should be compared with other estimators. If the estimator is used for other studies, the researcher should investigate the appropriateness of using the estimator for the intended study.

In actual testing programs, constraints have to be met for the content of the test, and attention has to be paid to item exposure. In adaptive testing, implementing content or exposure control often results in longer tests. The effects of content and exposure control should be investigated before the multidimensional classification method is applied in actual testing programs.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255–278. doi: 10.1207/s15324818ame0704_1

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8

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. doi: 10.1177/00131640021970862

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computerassisted branded test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh, PA.

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]. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Hambleton, R. K., & Swaminatan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Kale, B. K. (1962). On the solution of likelihood equations by iteration processes. The multiparametric case. *Biometrika*, *49*, 479–486. doi: 10.1093/biomet/49.3-4.479

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404. doi: 10.1177/014662169602000406

Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296. doi: 10.1007/S11336-008-9097-5

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes: The art of scientific computing (Fortran version)*. Cambridge, United Kingdom: University Press.

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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. doi: 10.1177/014662168500900409

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi: 10.1007/BF02294343

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*(1), 105–123.

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, April). *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*, *69*, 778–793. doi: 10.1177/0013164408324460

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*, *38*, 187–200. doi: 10.1177/0146621613509723

Van Groen, M. M., & Verschoor, A. J. (2010, June). *Using the sequential probability ratio test when items and respondents are mismatched*. Paper presented at the conference of the International Association for Computerized Adaptive

Testing, Arnhem, the Netherlands.

Van Ruitenburg, J. (2006). *Algorithms for parameter estimation in the Rasch model* (Report No. 2005-4). Arnhem, the Netherlands: Cito.

Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. doi: 10.1007/BF02295132

Wald, A. (1973). *Sequential analysis*. New York, NY: Dover. (Original work published 1947)

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

Wang, C., Chang, H.-H., & Boughton, K. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99–122. doi: 10.1007/S11336-011-9215-7

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: University of Iowa.

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*, *28*, 295–316. doi: 10.1177/0146621604265938

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi: 10.1007/BF02294627

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*, *77*, 495–523. doi: 10.1007/S11336-012-9265-5

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3–23. doi: 10.1177/0146621612455687

Appendix: Weighted Maximum Likelihood Estimation

Ability can be estimated using the weighted maximum likelihood estimator. Tam (1992) developed a weighted maximum likelihood estimator for multidimensional IRT similar to Warm's (1989) weighted maximum likelihood estimator for unidimensional IRT. This estimator reduces the bias in the estimate (Tam, 1992). In weighted maximum likelihood estimation, the following set of equations has to be solved (Tam, 1992)

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right)\right] + \frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln w(\boldsymbol{\theta})\right] = \mathbf{0}, \tag{4.23}$$

in which the first part denotes the derivatives of the natural logarithm of Equation 4.2 and the second part the weights that reduce the bias in the estimates. The natural logarithm of both parts is used because it simplifies the calculations. The set of likelihood equations is (Segall, 1996)

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right)\right] = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right) \\ \frac{\partial}{\partial \theta_2} \ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right) \end{bmatrix}. \tag{4.24}$$

In the two-parameter MIRT model, the partial derivatives for θ_l reduce to

$$\frac{\partial}{\partial \theta_l} \left[\ln \mathrm{L}(\boldsymbol{\theta}|\mathbf{x})\right] = \sum_{i=1}^{k} a_{il}[x_i - \mathrm{P}_i(\boldsymbol{\theta})] \quad l = 1, \ldots, p, \tag{4.25}$$

in which a_{il} denotes the discrimination parameter for dimension l for item i.

The weighting function that Tam (1992) developed is given by

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln w(\boldsymbol{\theta})\right] = -\mathbf{I}(\boldsymbol{\theta}) \times \mathbf{B}(\boldsymbol{\theta}), \tag{4.26}$$

in which $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information that the administered items provide for estimating the ability parameters (see Equation 4.16), and $\mathbf{B}(\boldsymbol{\theta})$ denotes the factor that reduces the bias in the estimates. The factor that reduces the bias in the WML estimate for dimension l is given by (Tam, 1992)

$$\mathrm{B}(\theta_l) = \frac{-\mathrm{J}(\theta_l)}{2\mathrm{I}(\theta_l, \theta_l)^2} \quad l = 1, \ldots, p, \tag{4.27}$$

where $\mathrm{J}(\theta_l)$ is an element of a $p \times 1$ matrix $\mathbf{J}$ (Tam, 1992):

$$\mathrm{J}(\theta_l) = \sum_{i=1}^{k} \frac{\left(\frac{\partial}{\partial \theta_l} \mathrm{P}_i(\boldsymbol{\theta}) \times \frac{\partial^2}{\partial \theta_l^2} \mathrm{P}_i(\boldsymbol{\theta}) \right)}{\mathrm{P}_i(\boldsymbol{\theta}) \mathrm{Q}_i(\boldsymbol{\theta})} = \sum_{i=1}^{k} a_{il}^3 \mathrm{P}_i(\boldsymbol{\theta}) \mathrm{Q}_i(\boldsymbol{\theta})^2 - a_{il}^3 \mathrm{P}_i(\boldsymbol{\theta})^2 \mathrm{Q}_i(\boldsymbol{\theta}). \quad (4.28)$$

Using Equations 4.16, 4.25, 4.26, and 4.27, the sets of equations 4.23 that have to be solved become

$$\sum_{i=1}^{k} a_{il}[x_{ij} - \mathrm{P}_i(\boldsymbol{\theta})] - \sum_{m=1}^{p} \left[\mathrm{I}(\theta_l, \theta_m) \times \frac{-\mathrm{J}(\theta_m)}{2\mathrm{I}(\theta_m, \theta_m)^2} \right] = 0 \quad l = 1, \ldots, p. \quad (4.29)$$

The equations to find the (weighted) maximum likelihood estimates have no closed-form solution, so an iterative numerical procedure has to be used (Segall, 1996). Several procedures can be used to find the estimates, such as Fisher's method of scoring (Kale, 1962), the Newton-Raphson method, and the false positioning method. The Newton-Raphson method was used in this study to find the estimates. Segall (1996) used this procedure to find maximum likelihood estimates. To find the weighted maximum likelihood estimates, the procedure was adapted to include the weighting part of Equation 4.29. The Newton-Raphson method does not converge when the second derivatives of the functions are infinite (Hambleton & Swaminatan, 1985). As an indication of a possible lack of convergence, the difference between iterations can be used. A small comparison study showed that the Newton-Raphson method resulted in more accurate estimates than the false positioning method. However, if the Newton-Raphson method did not appear to converge, the estimation algorithm switched toward the false positioning method. The estimation method also changed if the difference between iterations is very large. Both methods are described in the following sections.

WML Estimation Using the Newton-Raphson Method

The update function for the Newton-Raphson method for iteration $j + 1$ has the general form (Segall, 1996)

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{\Delta}^{(j)}, \quad (4.30)$$

in which $\boldsymbol{\Delta}^{(j)}$ is described by Segall (1996) as

$$\boldsymbol{\Delta}^{(j)} = \frac{f(\boldsymbol{\theta})}{\frac{\partial}{\partial \boldsymbol{\theta}} [f(\boldsymbol{\theta})]}, \quad (4.31)$$

in which

$$f(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln \mathrm{L}\left(\mathbf{x}|\boldsymbol{\theta}\right)\right] + \frac{\partial}{\partial \boldsymbol{\theta}} \left[\ln w(\boldsymbol{\theta})\right], \tag{4.32}$$

and

$$\frac{\partial}{\partial \boldsymbol{\theta}}[f(\boldsymbol{\theta})] = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \left[\ln \mathrm{L}\left(\mathbf{x}|\boldsymbol{\theta}\right)\right] + \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \left[\ln w(\boldsymbol{\theta})\right]. \tag{4.33}$$

All elements of Equation 4.32 were provided in the previous section. The elements of the second partial derivative for dimension l of the likelihood part in Equation 4.33 are given by

$$\frac{\partial^2}{\partial \theta_l^2}[\ln \mathrm{L}(\boldsymbol{\theta}|\mathbf{x})] = \frac{\partial}{\partial \theta_l}\left[\sum_{i=1}^{k} a_{il}[x_{ij} - \mathrm{P}_i(\boldsymbol{\theta})]\right] = \sum_{i=1}^{k} -a_{il}^2 \mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta}) \quad l = 1, \ldots, p. \tag{4.34}$$

The elements of the second partial derivative of the weighting part of Equation 4.33 become

$$\begin{aligned}
\frac{\partial^2}{\partial \theta_l^2}\left[w(\boldsymbol{\theta})\right] &= \frac{\partial}{\partial \theta_l}\left[\sum_{m=1}^{p}\left(\frac{\mathbf{I}(\theta_l, \theta_m) \times -\mathbf{J}(\theta_m)}{2\mathbf{I}(\theta_m, \theta_m)^2}\right)\right] \\
&= \sum_{m=1}^{p} \frac{2\mathbf{I}(\theta_l, \theta_m)\mathbf{J}(\theta_m)\frac{\partial}{\partial \theta_l}\left[\mathbf{I}(\theta_m, \theta_m)\right]}{2\mathbf{I}(\theta_m, \theta_m)^3} \\
&- \sum_{m=1}^{p} \frac{\frac{\partial}{\partial \theta_l}[\mathbf{I}(\theta_l, \theta_m)]\mathbf{J}(\theta_m)\mathbf{I}(\theta_m, \theta_m)}{2\mathbf{I}(\theta_m, \theta_m)^3} \\
&- \sum_{m=1}^{p} \frac{\mathbf{I}(\theta_l, \theta_m)\frac{\partial}{\partial \theta_l}\left[\mathbf{J}(\theta_m)\right]\mathbf{I}(\theta_m, \theta_m)}{2\mathbf{I}(\theta_m, \theta_m)^3}.
\end{aligned} \tag{4.35}$$

The remaining elements of Equation 4.35 are specified by

$$\begin{aligned}
\frac{\partial}{\partial \theta_l}\left[\mathbf{I}(\theta_l, \theta_m)\right] &= \sum_{i=1}^{k} \frac{\partial}{\partial \theta_l}[a_{il}a_{im}\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})] \\
&= \sum_{i=1}^{k} a_{il}^2 a_{im}\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})^2 - a_{il}^2 a_{im}\mathrm{P}_i(\boldsymbol{\theta})^2\mathrm{Q}_i(\boldsymbol{\theta}),
\end{aligned} \tag{4.36}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \theta_l}\left[\mathbf{I}(\theta_m, \theta_m)\right] &= \sum_{i=1}^{k} \frac{\partial}{\partial \theta_l}[a_{im}^2\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})] \\
&= \sum_{i=1}^{k} a_{il}a_{im}^2\mathrm{P}_i(\boldsymbol{\theta})\mathrm{Q}_i(\boldsymbol{\theta})^2 - a_{il}a_{im}^2\mathrm{P}_i(\boldsymbol{\theta})^2\mathrm{Q}_i(\boldsymbol{\theta}),
\end{aligned} \tag{4.37}$$

and

$$\frac{\partial}{\partial\theta_l}\left[\mathbf{J}(\theta_m)\right] = \sum_{i=1}^{k} a_{il}a_{im}^3\mathrm{P}_i(\theta)\mathrm{Q}_i(\theta)^3 - 4a_{il}a_{im}^3\mathrm{P}_i(\theta)^2\mathrm{Q}_i(\theta)^2 + a_{il}a_{im}^3\mathrm{P}_i(\theta)^3\mathrm{Q}_i(\theta). \tag{4.38}$$

The iterations of the Newton-Raphson procedure continue until the differences between the estimates for different iterations become very small (for example, 0.0001). If the Newton-Raphson method does not converge, the false positioning method can be used instead.

WML Estimation Using False Positioning Method

Another numerical iterative procedure to find the weighted maximum likelihood estimates is the false positioning method, also known as regula falsi. This method searches iteratively on an interval consisting of a set of two reasonable values for $\boldsymbol{\theta}$ (Van Ruitenburg, 2006) for each dimension; for example, the vector $\boldsymbol{\theta_l} = -\mathbf{5}$ contains reasonable values for the left boundary of the interval and $\boldsymbol{\theta_r} = \mathbf{5}$ for the right boundary. The derivative of the weighted maximum likelihood equation Equation 4.32 is calculated for each dimension m using

$$\frac{\partial}{\partial\theta_{ml}}f(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_{ml}}\left[\ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right)\right] + \frac{\partial}{\partial\theta_{ml}}\left[\ln w(\boldsymbol{\theta})\right] \quad m = 1, \cdots, p, \tag{4.39}$$

and

$$\frac{\partial}{\partial\theta_{mr}}f(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_{mr}}\left[\ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right)\right] + \frac{\partial}{\partial\theta_{mr}}\left[\ln w(\boldsymbol{\theta})\right] \quad m = 1, \cdots, p. \tag{4.40}$$

In each iteration, a straight line is drawn through the points $(\theta_{ml}; \frac{\partial}{\partial\theta_{ml}}f(\boldsymbol{\theta}))$ and $(\theta_{mr}; \frac{\partial}{\partial\theta_{mr}}f(\boldsymbol{\theta}))$ for each dimension m (Van Ruitenburg, 2006). A new replacement point $\boldsymbol{\theta}_s$ is determined per dimension based on the point where the line meets the dimension axis using (Press, Flannery, Teukolsky, & Vetterling, 1989)

$$\theta_{ms} = \theta_{ml} - \frac{\frac{\partial}{\partial\theta_{ml}}\left[f(\theta_{ml})\right](\theta_{mr} - \theta_{ml})}{\frac{\partial}{\partial\theta_{mr}}\left[f(\theta_{mr})\right] - \frac{\partial}{\partial\theta_{ml}}\left[f(\theta_{ml})\right]} \quad m = 1, \cdots, p. \tag{4.41}$$

The slopes are calculated at point $\boldsymbol{\theta}_s$ for all dimensions:

$$\frac{\partial}{\partial\theta_{ms}}f(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_{ms}}\left[\ln \mathrm{L}\left(\boldsymbol{\theta}|\mathbf{x}\right)\right] + \frac{\partial}{\partial\theta_{ms}}\left[\ln w(\boldsymbol{\theta})\right] \quad m = 1, \cdots, p. \tag{4.42}$$

If the slope for dimension m is positive, point θ_{ms} replaces the left boundary on the interval for dimension m. If the slope is negative, the right boundary is replaced. After replacement, a new point θ_{ms} is calculated. Iteratively, the procedure is repeated until the size of the interval becomes very small (for example, < 0.0001) for each dimension. The point θ_{ms} is then used as the ability estimate.

Chapter 5
Multidimensional Computerized Adaptive Testing for Classifying Examinees with the SPRT and the Confidence Interval Method

Abstract

Computerized adaptive tests (CATs) were developed to obtain efficient estimates of the examinee's abilities, but they can also classify examinees into one of two or more levels. Several methods are available to make the classification decisions for constructs modeled with a unidimensional item response theory model. These methods stop testing when enough confidence has been reached to make a decision. However, if the construct is multidimensional, few classification methods are available. Classification methods based on Wald's Sequential Probability Ratio Test are available for application to CAT with a multidimensional item response theory model in which items measure one or multiple abilities. It was investigated in the current study whether it was possible to adapt the popular unidimensional confidence interval method to make decisions on the entire test if each item measures only one dimension and to make decisions if items measure multiple dimensions. Simulation studies were used to investigate and compare the efficiency and effectiveness of the classification methods. Comparisons were made between different item selection methods, between different classification methods, and between different settings for the classification methods.

This chapter has been submitted as Van Groen, M.M., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). *Multidimensional Computerized Adaptive Testing for Classifying Examinees with the SPRT and the Confidence Interval Method.* Manuscript submitted for publication.

5.1 Introduction

Traditionally, the goal of a computerized adaptive test (CAT) was to obtain a precise and efficient ability estimate for the examinee. In one type of CAT, computerized classification testing (CCT), the goal is to obtain an accurate and efficient classification decision. Test length is minimized in CCT, while the desired accuracy of the decisions is maintained. A large body of literature is available on CATs, but knowledge about multidimensional CATs (MCATs) to make classification decisions is still scarce (Seitz & Frey, 2013a, 2013b; Spray, Abdel-Fattah, Huang, & Lau, 1997; Van Groen & Eggen, 2014; Van Groen, Eggen, & Veldkamp, 2014b).

MCATs can be constructed using multidimensional item response theory (MIRT). A vector of person abilities is used in MIRT to describe the skills and knowledge required for answering an item (Reckase, 2009). Two types of multidimensionality can be modeled using MIRT: between- and within-dimensionality (W.-C. Wang & Chen, 2004). Between-dimensionality implies that each item measures only one ability, and a between-dimensional test contains several unidimensional subscales (Hartig & Höhler, 2008). Items are intended to measure multiple abilities on tests with within-dimensionality (W.-C. Wang & Chen, 2004).

An MCAT for making classification decisions requires a method that decides whether enough evidence is available to make the decision and which decision is made. Two well-known unidimensional methods, the sequential probability ratio test (SPRT) (Eggen, 1999; Reckase, 1983; Spray, 1993) and the confidence interval method (Kingsbury & Weiss, 1979), were adapted for MCAT.

For between-dimensionality, Seitz and Frey (2013a) developed a method to make decisions per dimension using the SPRT. This method was extended to make classification decisions on all dimensions simultaneously, on several dimensions, and on subsets of items by Van Groen and Eggen (2014). Seitz and Frey (2013b) also adapted Kingsbury and Weiss's (1979) confidence interval method. This method is extended in the current study to make decisions on the complete multidimensional test and on subtests.

Only one classification method exists for within-dimensionality. Van Groen et al. (2014b) developed a classification method using the SPRT. The multidimensional space has to be reduced to a unidimensional line, the reference composite (M. Wang, 1985, 1986 as described in Reckase (2009)), before a classification decision can be made. A second method is developed here to make classification decisions using the confidence interval that surrounds the ability estimate.

In this study, an overview of the various multidimensional classification methods is given and their efficiency and accuracy are compared for between- and within-dimensionality.

An MCAT also requires a method that selects the items. Several item selection methods exist for MCAT for estimating ability (Reckase, 2009; Segall, 1996; Yao, 2012). Unidimensional item selection methods can often be used for between-dimensionality. For MCAT with within-dimensionality, only a few methods are available. These methods are discussed in the third section.

Multidimensional item response theory is discussed next. The newly developed and the existing classification methods are discussed in the third section. After these methods are discussed, item selection methods are described. The simulation study is included in the fourth section. In the final section, remarks are made about multidimensional classification testing and directions for future research.

5.2 Multidimensional Item Response Theory

An important part of a multidimensional computerized adaptive test is the item bank that has been calibrated with a MIRT model (Reckase, 2009). In a calibrated item bank, model fit is established, item parameter estimates are available, and items with undesired characteristics are removed (Van Groen, Eggen, & Veldkamp, 2014a). During testing, the item parameters are assumed to be estimated with enough precision to consider them known (Veldkamp & Van der Linden, 2002).

The MIRT model used here is the dichotomous two-parameter logistic model (Reckase, 1985). A set of p abilities accounts for the examinee's responses to the items, and the probability of a correct answer, $x_i = 1$, to item i is given by (Reckase, 2009)

$$\mathrm{P}_i(\boldsymbol{\theta}) = \mathrm{P}\left(x_i = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}\right) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}, \tag{5.1}$$

where $\mathbf{a}_i$ is the vector of discrimination parameters, d_i is the easiness of the item, and $\boldsymbol{\theta}$ is the vector of ability parameters.

In a computerized classification test, ability estimates are sometimes used by the item selection or classification method. Ability is estimated using the likelihood function. The likelihood of a vector of observed responses $\mathbf{x}_j = (x_{1j}, ..., x_{kj})$ to items $i = 1, \cdots, k$ for an examinee j with ability $\boldsymbol{\theta}_j$ equals the product of the probabilities of the responses to the administered items (Segall, 1996):

$$\mathrm{L}\left(\boldsymbol{\theta}_j|\mathbf{x}_j\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_j\right)^{x_{ij}} \mathrm{Q}_i\left(\boldsymbol{\theta}_j\right)^{1-x_{ij}}, \tag{5.2}$$

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

The vector of values, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)$, that maximizes Equation 5.2 is used as the ability estimate $\boldsymbol{\theta}_j$ (Segall, 1996). Since the equations 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 (WML) estimation is an extension of maximum likelihood estimation that reduces the bias in the ML estimates. WML for MIRT was developed by Tam (1992) based on Warm's (1989) unidimensional estimator. The WML procedure used here was described in Van Groen et al. (2014b).

Two types of multidimensionality exist: within- and between-dimensionality. If more than one discrimination parameter is non-zero for one or more items, within-dimensionality is present (W.-C. Wang & Chen, 2004), and items are intended to measure multiple or all abilities. Using a within-dimensional model, complex domains can be modeled while taking into account several abilities simultaneously and representing different combinations of abilities for different items (Hartig & Höhler, 2008). If just one discrimination parameter is non-zero per item in the test, the test is considered to have between-dimensionality (W.-C. Wang & Chen, 2004), and items are intended to measure just one ability. Between-dimensional models are typically used if a test consists of several unidimensional subscales.

5.3 Classification Methods

Four methods for making multidimensional classification decisions are described here. The SPRT can be used for tests with between-dimensionality to make decisions for each dimension (Seitz & Frey, 2013a). This method was expanded by Van Groen and Eggen (2014) to make decisions on the entire test, on several dimensions simultaneously, and on subsets of items. Seitz and Frey (2013b) adapted Kingsbury and Weiss's (1979) classification method for between-dimensionality. This method is extended here to make classification decisions on the entire test, on several dimensions, and on subsets of items.

The SPRT can also be used for within-dimensionality (Van Groen et al., 2014b). The method will be described for making multiple decisions and for making decisions on subsets of items. In the last part of this section, it will be shown

that it is also possible to adapt Kingsbury and Weiss's (1979) confidence interval method to make multidimensional decisions with within-dimensionality.

5.3.1 The SPRT for Between-Dimensionality

Seitz and Frey (2013a) developed a classification method for between-dimensionality based on the unidimensional classification method (Eggen, 1999; Reckase, 1983; Spray, 1993; Thompson, 2009) using the sequential probability ratio test (Wald, 1947/1973). The SPRT is used to make a classification decision per dimension.

The multidimensional SPRT requires that a cutoff point, θ_{cl}, is set for each dimension l, $l = 1, ..., p$, with an indifference region around the cutoff point c. The indifference region accounts for the measurement error in the decisions for examinees with an ability close to the cutoff point (Eggen, 1999). The SPRT compares two hypotheses (Van Groen et al., 2014b):

$$\mathrm{H}_{0l} : \theta_{jl} < \theta_{cl} - \delta; \tag{5.3}$$

$$\mathrm{H}_{al} : \theta_{jl} > \theta_{cl} + \delta, \tag{5.4}$$

where δ specifies the distance between the cutoff point and the boundary of the indifference region.

The likelihood ratio between the hypotheses for dimension l after k items are administered is calculated for the SPRT (Van Groen et al., 2014b):

$$\mathrm{LR}\left(\theta_{cl} + \delta; \theta_{cl} - \delta\right) = \frac{\mathrm{L}\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)}{\mathrm{L}\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)}, \quad l = 1, \cdots, p, \tag{5.5}$$

in which $\mathrm{L}\left(\theta_{cl} + \delta; \mathbf{x}_{jl}\right)$ and $\mathrm{L}\left(\theta_{cl} - \delta; \mathbf{x}_{jl}\right)$ are calculated using Equation 5.2 with only the items included that load on the dimension.

Decision rules are applied to the likelihood ratio to decide whether to continue testing or to make a classification decision:

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) < (1-\beta)/\alpha; \\ &\text{ability below } \theta_{cl} \text{ if} && \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \leq \beta/(1-\alpha); \\ &\text{ability above } \theta_{cl} \text{ if} && \mathrm{LR}(\theta_{cl} + \delta; \theta_{cl} - \delta) \geq (1-\beta)/\alpha, \end{aligned} \tag{5.6}$$

where α and β specify the acceptable classification error rates (Spray et al., 1997).

In practical settings, a maximum test length is specified at which the examinee is classified as having an ability above the cutoff point if the likelihood ratio is larger than the midpoint of the interval in Equation 5.6.

An examinee can be classified into one of multiple levels using the SPRT. Following Van Groen and Eggen (2014), the Sobel and Wald (1949) approach was used in which only adjacent levels are compared (Eggen & Straetmans, 2000).

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 & Eggen, 2014):

$$\mathrm{LR}\left(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_c - \boldsymbol{\delta}; \mathbf{x}_j\right)}, \tag{5.7}$$

where $\boldsymbol{\theta}_c$ includes the cutoff points for all dimensions and decisions and $\boldsymbol{\delta}$ includes the δ for all decisions. The decision rules for the entire tests then become

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}) < (1-\beta)/\alpha; \\ &\text{ability below } \boldsymbol{\theta}_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}) \leq \beta/(1-\alpha); \\ &\text{ability above } \boldsymbol{\theta}_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_c + \boldsymbol{\delta}; \boldsymbol{\theta}_c - \boldsymbol{\delta}) \geq (1-\beta)/\alpha. \end{aligned} \tag{5.8}$$

Van Groen and Eggen (2014) also investigated the use of the SPRT to make decisions using only a subset of the dimensions or a subset of the items. The first extension implies that Equation 5.7 includes only the selected dimensions and items that load on those dimensions. The second extension is relevant if only some of the items are selected that load on dimension l. The likelihood ratio of Equation 5.7 is then calculated for a subset of the items. Van Groen and Eggen's (2014) extensions make it possible to include an item in several decisions. This implies that a knowledge profile can be created for each student, which can be used to adapt the instruction to the student based on the student's current knowledge.

5.3.2 The Confidence Interval Method for Between-Dimensionality

Seitz and Frey (2013b) also adapted Kingsbury and Weiss's (1979) unidimensional classification method to between-dimensionality tests. The method stops testing as soon as the cutoff point is outside the confidence interval.

The likelihood function of a between-dimensional test reduces to one unidimensional likelihood function per dimension. This result can be used to adapt

the confidence interval method to between-dimensionality. The method uses the following decision rules:

$$\begin{aligned} &\text{administer another item if} && \hat{\theta}_{jl} - \gamma \cdot \text{se}(\hat{\theta}_{jl}) < \theta_{cl} < \hat{\theta}_{jl} + \gamma \cdot \text{se}(\hat{\theta}_{jl}); \\ &\text{ability below } \theta_{cl} \text{ if} && \hat{\theta}_{jl} + \gamma \cdot \text{se}(\hat{\theta}_{jl}) < \theta_{cl}; \\ &\text{ability above } \theta_{cl} \text{ if} && \hat{\theta}_{jl} - \gamma \cdot \text{se}(\hat{\theta}_{jl}) > \theta_{cl}, \end{aligned} \tag{5.9}$$

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

$$\text{se}(\hat{\theta}_{jl}) = \frac{1}{\sqrt{\text{I}(\hat{\theta}_{jl})}}, \tag{5.10}$$

where $\text{I}(\hat{\theta}_{jl})$ denotes the Fisher information available in the observable variables for estimating θ_{jl} (Mulder & Van der Linden, 2009). Fisher information is given by (Tam, 1992)

$$\text{I}(\hat{\theta}_{jl}) = \sum_{i=1}^{k} a_{il}^2 \text{P}_i(\hat{\theta}_{jl}) \text{Q}_i(\hat{\theta}_{jl}). \tag{5.11}$$

This method can be used only to make decisions per dimension, but not to make decisions on all dimensions simultaneously. Three possible solutions for this problem were found for reducing multiple abilities into one global ability (Van der Linden, 1999; Veldkamp & Van der Linden, 2002; Yao, 2012). Veldkamp and Van der Linden (2002) defined a composite ability as a linear combination of $\boldsymbol{\theta}$ so that all remaining ability parameters become nuisance parameters. Those parameters remain a part of the likelihood function, which implies that application of the Kingsbury and Weiss (1979) method is still not possible. Yao (2012) combined the abilities using optimal weights. This approach appears to be attractive, but cannot be used because the weights would then vary during testing. This implies that it is not possible to specify a fixed cutoff point for the SPRT. Van der Linden's (1999) approach can be used as a starting point for application of the confidence interval method to between-dimensionality classification testing.

Van der Linden (1999) assumed that the parameter of interest is a linear combination of the abilities, $\boldsymbol{\lambda}'\boldsymbol{\theta}$, where $\boldsymbol{\lambda}$ is a combination of non-negative weights. The weights are chosen to reflect the relative importance of the dimensions but these are considered to be of equal importance. However, the dimensions vary in

number of items; thus, the weights are given by 1 divided by the number of items for each dimension. The composite ability is given by (Van der Linden, 1999)

$$\zeta = \sum_{l=1}^{p} \theta_l \lambda_l. \tag{5.12}$$

The confidence interval method requires that the standard error of measurement is known for the ability estimate. The standard error of measurement is, following Yao (2012),

$$\text{se}(\zeta) = \sqrt{\text{V}(\zeta)}, \tag{5.13}$$

and

$$\text{V}(\zeta) = \boldsymbol{\lambda}' \text{V}(\boldsymbol{\theta}) \boldsymbol{\lambda}, \tag{5.14}$$

and $\text{V}(\boldsymbol{\theta}) = \text{I}(\boldsymbol{\theta})^{-1}$. The non-diagonal elements of matrix I are zero, and the diagonal elements can be calculated using Equation 5.11.

Following Eggen and Straetmans (2000), the decision rules then become

$$\begin{aligned} &\text{administer another item if} && \hat{\zeta}_j - \gamma \cdot \text{se}(\hat{\zeta}_j) < \zeta_c < \hat{\zeta}_j + \gamma \cdot \text{se}(\hat{\zeta}_j); \\ &\text{ability below } \zeta_c \text{ if} && \hat{\zeta}_j + \gamma \cdot \text{se}(\hat{\zeta}_j) < \zeta_c; \\ &\text{ability above } \zeta_c \text{ if} && \hat{\zeta}_j - \gamma \cdot \text{se}(\hat{\zeta}_j) > \zeta_c, \end{aligned} \tag{5.15}$$

where ζ_c denotes the cutoff point for the composite ability. The cutoff point can be determined with Equation 5.12 using the cutoff points for the dimensions.

If a decision on several dimensions simultaneously is required, but not on all dimensions, only those dimensions should be included in Equation 5.12. The information matrix as used for Equation 5.14 should include only the items that are relevant for those dimensions. If only some of the items that measure a specific dimension is used to make a decision, only those items should be included in the information matrix.

5.3.3 The SPRT for Within-Dimensionality

The SPRT can be used for tests with within-dimensionality, but requires an additional step (Van Groen et al., 2014b). Reckase (2009) described a method for reducing the multidimensional space to a unidimensional line (M. Wang, 1985, 1986). This line, the reference composite, is then used to make classification decisions with the SPRT.

The reference composite through the multidimensional space describes the characteristics of the matrix of discrimination parameters for the items in the item bank (Reckase, 2009). The direction of the line is described by the eigenvector of the $\mathbf{a}\mathbf{a}'$ 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.

$\boldsymbol{\theta}$-points can be projected on the reference composite (Reckase, 2009). A higher value on the reference composite, denoted as ξ_j, denotes a more proficient student than a lower value (Van Groen et al., 2014b). Proficiency can be calculated using an additional line through the $\boldsymbol{\theta}_j$-point and the origin. The length of this line is given by (Reckase, 2009)

$$\mathrm{L}_j = \sqrt{\sum_{l=1}^{p} \hat{\theta}_{jl}^2}, \tag{5.16}$$

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

$$\cos \alpha_{jl} = \frac{\hat{\theta}_{jl}}{\mathrm{L}_j}, \quad l = 1, \cdots, p. \tag{5.17}$$

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:

$$\hat{\xi}_j = L_j \cos \alpha_{j\xi}. \tag{5.18}$$

The reference composite can now be used to make classification decisions with the SPRT (Van Groen et al., 2014b). The cutoff point for the SPRT, ξ_c, and δ^{ξ} are specified on the reference composite. The boundaries of the indifference region are then transformed to their $\boldsymbol{\theta}$-points using

$$\boldsymbol{\theta}_{\xi_{c+\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c + \delta^{\xi}); \tag{5.19}$$

$$\boldsymbol{\theta}_{\xi_{c-\delta}} = \cos \boldsymbol{\alpha}_{\xi} \times (\xi_c - \delta^{\xi}), \tag{5.20}$$

where $\boldsymbol{\alpha}_{\xi}$ includes the angles between the reference composite and all dimension axes. The likelihood ratio for the SPRT becomes (Van Groen et al., 2014b)

$$\mathrm{LR}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}\right) = \frac{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c+\delta}}; \mathbf{x}_j\right)}{\mathrm{L}\left(\boldsymbol{\theta}_{\xi_{c-\delta}}; \mathbf{x}_j\right)}, \tag{5.21}$$

which can be used to make multidimensional classification decisions with the following decision rules

$$\begin{aligned} &\text{administer another item if} && \beta/(1-\alpha) < \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) < (1-\beta)/\alpha; \\ &\text{ability below } \xi_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \leq \beta/(1-\alpha); \\ &\text{ability above } \xi_c \text{ if} && \mathrm{LR}(\boldsymbol{\theta}_{\xi_{c+\delta}}; \boldsymbol{\theta}_{\xi_{c-\delta}}) \geq (1-\beta)/\alpha. \end{aligned} \tag{5.22}$$

Decisions can be made using the reference composite for different subsets of items, subsets of dimensions, and more than two categories.

5.3.4 The Confidence Interval Method for Within-Dimensionality

Until now, Kingsbury and Weiss's (1979) confidence interval method has never been applied to within-dimensional classification testing. However, the method can be applied to within-dimensionality tests using the reference composite.

After an item is administered, the examinee's ability is estimated, the estimate is projected onto the reference composite, and the corresponding point on the reference composite in the multidimensional space is calculated using

$$\boldsymbol{\theta}_{\hat{\xi}_j} = \cos \boldsymbol{\alpha}_\xi \times \hat{\xi}_j. \tag{5.23}$$

The reference composite can be considered to be a combination of abilities that are combined using weights, $\boldsymbol{\lambda}_\xi$. The weights are based on the angles between the reference composite and the dimension axes, $\lambda_{\xi l} = 1/\alpha_{\xi l}$.

The standard error of measurement is required for the confidence interval method and follows from (Yao, 2012):

$$\mathrm{se}(\xi) = \sqrt{\mathrm{V}(\xi)}, \tag{5.24}$$

with

$$\mathrm{V}(\xi) = \boldsymbol{\lambda}_\xi' \mathrm{V}(\boldsymbol{\theta}_\xi) \boldsymbol{\lambda}_\xi. \tag{5.25}$$

The variance in the point in the multidimensional space that corresponds to the ability estimate is approximated by the inverse of the information matrix of the same point. The decision rules then become

$$\begin{aligned} &\text{administer another item if} && \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c < \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j); \\ &\text{ability below } \xi_c \text{ if} && \hat{\xi}_j + \gamma \cdot \text{se}(\hat{\xi}_j) < \xi_c; \\ &\text{ability above } \xi_c \text{ if} && \hat{\xi}_j - \gamma \cdot \text{se}(\hat{\xi}_j) > \xi_c, \end{aligned} \tag{5.26}$$

where ξ_c denotes the cutoff point for the reference composite. The cutoff point can be determined using the cutoff points specified for each dimension.

5.4 Item Selection Methods

Computerized classification testing requires a method to select the items during test administration. Multiple methods are available for unidimensional classification testing (Eggen, 1999; Spray & Reckase, 1994; Van Groen et al., 2014a). Several methods are available for multidimensional computerized adaptive testing for ability estimation (Reckase, 2009; Segall, 1996; Yao, 2012), but limited knowledge is available for selecting items in multidimensional classification testing (Seitz & Frey, 2013a; Van Groen et al., 2014b; Van Groen & Eggen, 2014).

In unidimensional classification testing, items are often selected based on maximizing information at the ability estimate or the cutoff point (Eggen, 1999; Spray, 1993; Thompson, 2009). The same approach is followed here for between- and within-dimensionality. One specific method, the weighting method, will be used to select based on the cutoff points. Selection methods for multidimensional computerized adaptive testing often use Bayesian ability estimates, but weighted maximum likelihood estimates were used here (Van Groen et al., 2014b).

5.4.1 Item Selection Methods for Between-Dimensionality

If all available items measure the same ability, unidimensional item selection methods can be used (Van Groen & Eggen, 2014), but items must be selected per dimension. In the first part of this section, the unidimensional item selection method that maximizes information at the ability estimate is described. Then the unidimensional weighting method is described.

Item Selection Based on the Ability Estimate

In unidimensional classification testing, Fisher information is often maximized at the current ability estimate (Eggen, 1999; Spray, 1993). Since information is

related to the standard error of measurement, maximizing information at the current ability estimate tries to reduce the standard error of the estimates. In between-dimensional testing, this boils down to (Van Groen & Eggen, 2014)

$$\max \mathrm{I}_i(\hat{\theta}_l), \quad \text{for } i \in \mathrm{V}_{al}, \tag{5.27}$$

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

Item Selection Using the Weighting Method

In unidimensional classification testing, information is often maximized at the cutoff point (Eggen, 1999; Spray, 1993). Since item selection for between-dimensionality can be done using unidimensional methods, items can be selected that provide the most information at the cutoff point using

$$\max \mathrm{I}_i(\theta_{cl}), \quad \text{for } i \in \mathrm{V}_{al}. \tag{5.28}$$

This method can be used only if just one cutoff point is specified per dimension. Although several methods have been developed for item selection with multiple cutoff points (Eggen & Straetmans, 2000; Wouda & Eggen, 2009), Van Groen et al. (2014a) developed the weighting method. This method combines the objective functions per cutoff point into one weighted objective function. Item selection is adapted to the examinee's ability by adapting the weight for the cutoff points depending on the distance of the cutoff point to the current ability estimate (Van Groen et al., 2014a). The quantity maximized for item selection for dimension l becomes

$$\max \sum_{c=1}^{C} \frac{1}{|\hat{\theta}_{jl} - \theta_{cl}|} \mathrm{I}_i(\theta_{cl}), \quad \text{for } i \in \mathrm{V}_{al}. \tag{5.29}$$

5.4.2 Item Selection Methods for Within-Dimensionality

Item selection for within-dimensionality has to take the multidimensional structure of the MIRT model into account. This implies that multidimensional item selection methods have to be used, but these methods have been developed for multidimensional computerized adaptive testing to obtain an efficient and precise ability estimate. One of Segall's (1996) methods selects the items at the current ability estimate and was adapted to select at a weighted combination of the cutoff points (Van Groen et al., 2014b).

Item Selection Based on the Ability Estimate

Segall's (1996) method selects the item that results in the largest decrement of the volume of the confidence ellipsoid surrounding the ability estimate. The size of the ellipsoid can be approximated by the inverse of the information matrix, so the item is selected that has the highest (Segall, 1996)

$$\max \det \left(\sum_{i=1}^{k} \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij}) + \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j}) \right), \quad \text{for } k+1 \in V_{k+1}, \tag{5.30}$$

where $\sum_{i=1}^{k} \mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{ij})$ denotes the information that has been gained thus far and $\mathrm{I}(\hat{\boldsymbol{\theta}}_j, x_{k+1,j})$ the information that a potential item provides. The item is administered that will result in the largest determinant of the information matrix, which implies that the volume of the confidence ellipsoid is minimized (Reckase, 2009).

Item Selection Using the Weighting Method

The second item selection method for within-dimensional classification testing is based on maximization at a weighted combination of the cutoff points (Van Groen et al., 2014b). For adaptation of Segall's (1996) method, the multidimensional points corresponding to the cutoff points must be calculated. The weighting method (Van Groen et al., 2014a) then selects the item that corresponds to

$$\max \sum_{c=1}^{C} \frac{1}{|\xi_{\hat{\theta}_j} - \xi_c|} \mathrm{I}_i(\theta_{\xi_c}), \quad \text{for } i \in \mathrm{V}_{al}. \tag{5.31}$$

5.5 Simulation Studies

The efficiency and the accuracy of the classification and item selection methods were investigated using simulation studies. Two series of simulations were conducted, one series with between-dimensionality and one with within-dimensionality. The goal of these simulations was to compare the accuracy and efficiency of the SPRT and the confidence region methods for each type of multidimensionality. The simulations were conducted with various settings for the classification methods and with several item selection methods.

5.5.1 Design of the Simulation Study with Between-Dimensionality

The End of Primary School Test (Cito, 2012) in the Netherlands provides a recommendation about the most suitable level of secondary education for pupils. As a test result, a scale score based on the pupil's ability is given. This scale score can be related to a advise for the level of secondary education (Van Boxtel, Engelen, & De Wijs, 2011). The scale score and the primary school's recommendation are used by secondary schools to classify pupils into different classes based on the educational levels.

The test consists of four components: Language, Mathematics, Study Skills, and Environmental Studies. However, only the Language component was included in the simulations. Language (100 items) consists of four components: Writing (30 items), Text Comprehension (30 items), Vocabulary (20 items), and Spelling (20 items). Spelling is divided even further into Spelling of Verbs (10 items) and Spelling of Non-Verbs (10 items). Decisions were made in the current study about the entire Language component, the four Language components, and both Spelling components. The decisions for Language and Spelling were made without consideration of the decisions on underlying components and whether decisions had already been made for those components.

The test is always calibrated using four independent unidimensional scales for Language. A between-dimensional model with the same number of dimensions was used here. The responses of the 147,099 pupils in the dataset for the 2012 End of Primary School Test were used to obtain item parameters for the between-dimensional model with NOHARM 4 (Fraser & McDonald, 2012).

The 2012 data were used to estimate the pupils' ability distribution. The correlations between the dimensions for the between-model were determined using the ability estimates of the observed data. The observed correlations were then used to generate respondents. Ten datasets were simulated, each consisting of records for 1,000 simulees and all 100 Language items. The response patterns of these datasets were used for all simulations. By replicating the simulations, random fluctuations were mediated. The ability parameters used to generate the data can be used to determine the "true classifications" for each student.

Settings for the Classification Methods

Cutoff points are available for the scale scores that are reported for the test (Van Boxtel et al., 2011). Three levels were used in the simulations, which were

related to the levels within the Dutch educational system. As in Van Groen et al. (2014b), the median of the ability estimates for each of the three intervals was calculated using the observed data. The midpoint between two adjacent medians was set as the cutoff point. The same procedure was applied to calculate the cutoff points for decisions based on parts of the scales. If a decision was based on multiple dimensions, the cutoff points on the separate dimensions were used.

The settings for the SPRT potentially influence which decision is made. Analyses were conducted with $\alpha = \beta = 0.05$ and 0.10. The size of δ for the SPRT also influences test length and accuracy. The constant for δ varied between 0.10, 0.20, and 0.30 multiplied with the standard deviation of the observed abilities. The reason for the multiplication with the standard deviation is that this relates the size of delta to the observed ability distribution. For the decision on Language, the average standard deviation over the four dimensions was used. If δ became larger than half the distance between the cutoff points, δ was set at half the distance to prevent overlap in indifference regions for adjacent cutoff points.

The settings of the confidence interval method also influence the number of items that are required before a classification can be made and which classification is made. The confidence intervals were set at 80%, 85%, 90%, and 95%. The corresponding values of the t-distribution for the intervals were used for γ.

Item Selection in the Simulations

Three item selection methods were used in the simulations: random item selection, maximization of information at the current ability estimate, and maximization of information at a weighted combination of the cutoff points. It was expected that the last two methods would be more efficient than random item selection.

The Text Comprehension and Writing components consist of testlets. A testlet is a set of items that share a common stimulus, such as a text, which must be administered in the correct order. As in Van Groen and Eggen (2014), the information-based methods selected the testlet that maximized the average information. The random item selection method selected a testlet at random.

Simulation Design

Two series of simulations were conducted with the between-dimensional datasets. The first series of simulations included all items for Language (100 items). The idea behind the simulations was that it would then be possible to compare the accuracy of the SPRT and the confidence interval method. Accuracy is reported here as

the proportion of consistent decisions (PCD) between the classifications by the classification method and the classifications based on the abilities used to generate the data. Unfortunately, no baseline classification is available if several dimensions are involved in the decision. This implies that the decisions for Language must be compared, if possible, with the decisions made by the classification methods with 100 items. Those decisions were generated with $\alpha = 0.05$ and $\delta = 0.05$.

The second series of between-dimensional simulations were conducted without a fixed test length. Testing was stopped as soon as this was possible based on the specifications of the SPRT or the confidence interval method. Two outcome measures were reported per simulation: the average test length (ATL) and the proportion of consistent decisions. The ATL and the PCD were compared between the classification methods.

A remark must be made about the between-dimensional simulations with a flexible test length. Content distribution over the Language and Spelling components had to be controlled, because content validity was considered important. The content distribution was controlled using the Kingsbury and Zara (1989) approach. This approach selected the items from the component for which the difference between the desired and achieved percentage was the largest.

5.5.2 Design of the Simulation Study with Within-Dimensionality

The within-dimensional simulations were conducted with the ACT item pool, which was also used in Ackerman (1994), Van Groen et al. (2014b), and Veldkamp and Van der Linden (2002). The item pool consists of 180 items that were previously calibrated with a two-dimensional compensatory IRT model using NOHARM II (Fraser & McDonald, 1988). An orthogonal solution was used for the calibration. Items are intended to measure both dimensions, which implies that the test has within-dimensionality. The items were previously classified into six content and three skill categories (Veldkamp & Van der Linden, 2002). The content categories include Coordinate Geometry (CG; 27 items), Elementary Algebra (EA; 30 items), Intermediate Algebra (IA; 27 items), Pre-Algebra (PA; 42 items), Plane Geometry (PG; 42 items), and Trigonometry (TG; 12 items), and the skill categories are Analysis (AN; 17 items), Application (AP; 89 items), and Basic Skills (BS; 74 items). More information about the item pool can be found in Van Groen et al. (2014b) and Veldkamp and Van der Linden (2002).

Decisions were made in the current study on the three skill categories. Decisions on the entire Mathematics scale were made independent of the decisions for the categories and whether a decision had already been made for the categories.

Ten datasets were generated, each consisting of 1,000 simulees from a multivariate standard normal distribution. By replicating the simulations with 10 datasets, random fluctuations were mediated. The resulting 10 datasets were used for all simulations with the SPRT and the confidence interval method.

Settings for the Classification Methods

The cutoff points for the classification methods were specified on the reference composite. They were positioned at -0.609 and 0.609. These points are based on the 33th and 66th percentiles of the distributions for the underlying dimensions. The α and β for the SPRT were set to 0.05 and 0.10. δ was set at the reference composite and varied: 0.10, 0.20, and 0.30. Since no actual ability distribution was known, δ was not tailored to the ability distribution. The same values as before were used for γ for the confidence interval method.

Item Selection in the Simulations

Three item selection methods were used in the simulations: random item selection, the within-dimensional version of the weighting method, and the method that maximizes the determinant of the information matrix at the current ability estimate. The last method will often be referred to as the method based on the ability estimate in the remainder of the simulation studies.

Simulation Design

Two series of simulations were run for within-dimensionality. The first series consisted of runs in which test length was fixed at the maximum number of items that was available for the decision in the item pool. The simulation studies are provided here to illustrate the discussed methods. This also implies that no practical restrictions were specified for the test length. With a fixed test length the accuracy of the classifications with the SPRT and the confidence interval method can be compared. The PCD is defined here as the consistency of the classifications based on the (composite of the) ability estimates and the decisions by the classification methods. The second series of simulations did not place

Table 5.1
Proportion of Consistent Decisions with Maximum and Flexible Test Length for the SPRT

$\alpha = \beta$	δ	LAN	WR	SP	SPNV	SPV	TC	VO
Maximum Test Length								
-	0.1 sd	0.995	0.947	0.940	0.987	0.909	0.938	0.956
-	0.2 sd	0.993	0.947	0.938	0.987	0.904	0.936	0.956
-	0.3 sd	0.989	0.942	0.936	0.987	0.895	0.935	0.954
Flexible Test Length								
0.05	0.1 sd	0.995	0.947	0.940	0.987	0.909	0.938	0.956
0.05	0.2 sd	0.993	0.947	0.938	0.987	0.904	0.936	0.956
0.05	0.3 sd	0.987	0.942	0.936	0.987	0.895	0.935	0.954
0.10	0.1 sd	0.995	0.947	0.940	0.987	0.909	0.938	0.956
0.10	0.2 sd	0.992	0.947	0.938	0.987	0.904	0.936	0.956
0.10	0.3 sd	0.978	0.942	0.936	0.987	0.895	0.935	0.954

Note. The proportion of consistent decisions (PCD) for Language (LAN) was based on the SPRT classification with 100 items and $\alpha = \beta = \delta = 0.05$. The PCDs using the classification based on the true ability for Writing (WR), Spelling (SP), spelling Non-Verbs (SPNV), Spelling Verbs (SPV), Text Comprehension (TC), and Vocabulary (VO). $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

restrictions on the test length. The efficiency and accuracy of the simulations with the SPRT and the confidence interval method can now be investigated.

5.5.3 Results for the Example with Between-Dimensionality

The simulations for between-dimensionality consisted of two parts. In the first part, the test length was fixed at 100 items. No constraints were set on test length in the second part. Testing was stopped at the moment that enough confidence was gathered to make a decision. The PCDs for the SPRT are given in Table 5.1.

For the fixed length simulations, the item selection methods had no effect on accuracy because all items were administered. The size of α and β was omitted from the table for the simulations with a fixed test length because the PCDs were equal. The small inconsistencies in the decisions for Language with the SPRT were caused by the different specifications for δ that were used for the original classification decisions and the SPRT decisions. The PCDs for all other decisions were based on the comparison between the SPRT decisions and the classifications based on the abilities that were used to generate the data. These decisions were less consistent than the decisions on Language, but were based on fewer items and on a different classification criterion.

Table 5.2
Proportion of Consistent Decisions with Maximum and Flexible Test Length for the Confidence Interval Method

Method	γ	LAN	WR	SP	SPNV	SPV	TC	VO
Maximum Test Length								
-	-	1.000	0.996	1.000	1.000	1.000	0.996	0.995
Flexible Test Length								
RA	1.9600	0.952	0.995	1.000	1.000	1.000	0.995	0.995
RA	1.6449	0.927	0.988	0.998	1.000	1.000	0.990	0.994
RA	1.4395	0.905	0.978	0.993	1.000	1.000	0.985	0.993
RA	1.2816	0.880	0.965	0.987	1.000	1.000	0.973	0.990
AE	1.9600	0.969	0.991	1.000	1.000	1.000	0.995	0.995
AE	1.6449	0.944	0.982	0.998	1.000	1.000	0.992	0.994
AE	1.4395	0.919	0.974	0.994	1.000	1.000	0.986	0.992
AE	1.2826	0.901	0.967	0.988	1.000	1.000	0.976	0.988
WM	1.9600	0.965	0.995	1.000	1.000	1.000	0.996	0.995
WM	1.6449	0.943	0.988	0.999	1.000	1.000	0.993	0.994
WM	1.4395	0.921	0.977	0.996	1.000	1.000	0.987	0.994
WM	1.2816	0.901	0.968	0.992	1.000	1.000	0.978	0.991

Note. The proportion of consistent decisions (PCD) for Language (LAN) was based on the combined ability classification. The PCDs for Writing (WR), Spelling (SP), spelling Non-Verbs (SPNV), Spelling Verbs (SPV), Text Comprehension (TC), and Vocabulary (VO) were based on the true ability classification for random item selection (RA), maximization at the ability estimate (AE), and the weighting method (WM). γ denotes the constant for the confidence interval.

The proportions of consistent decisions for the simulations with a flexible test length are also presented in the table. The results for the SPRT are based on the simulations with all three item selection methods. The differences in PCD between the item selection methods were negligible. Therefore, the results have not been specified per method. If α or δ increased, the PCD decreased. All decisions appeared to be consistent.

The PCDs for the simulations with the confidence interval method are given in Table 5.2 for the simulations with a fixed and flexible test length.

For the fixed length simulations, the decisions used for the Language PCD were calculated using the combined ability that is also used for the classification method. All other PCDs were based on the classifications using the abilities that were used to generate data. The close relationship between the abilities and the confidence interval method became apparent in the high values for the PCD.

The table also contains the PCDs for the simulations with a flexible test length. The value of γ clearly influenced the consistency of the decisions. The PCDs were very high, except for the Language decisions. The decisions apparently changed when the test length was decreased. The PCDs were influenced by the item selection methods. Random item selection resulted in less consistent decisions for Language than the other two methods.

When comparing the decisions with a fixed and a flexible test length, reducing the test length with the SPRT had a small influence on consistency. The influence of the decreases by the confidence interval method on the PCDs was larger.

Testing was stopped in the second part of the simulations when enough confidence was available in the decisions. The average test length of the SPRT simulations is presented in Table 5.3. The test length of the SPRT simulations for Language could be reduced with almost 36 items. The test length for the other decisions could be reduced much less, but this could probably be explained by the already optimized test length of the components of the test. If α and β were increased, the test length was reduced. If the value of δ was increased, the tests became much shorter. The simulations with the weighting method and the method that maximized the determinant of the information matrix at the ability estimate were more efficient than the simulations with random item selection.

Table 5.4 contains the average test length of the simulations with the confidence interval method. The Language test length could almost be reduced to half the original length when the confidence interval method was used. An increase in the size of γ resulted in shorter tests. Tests with random item selection were often a bit longer than tests with one of the other item selection methods.

When the simulations with the SPRT and the confidence interval method were compared, it became apparent that the confidence interval method resulted on average in shorter Language tests than the SPRT. The majority of the tests for the components were also shorter for the confidence interval method.

The test length of the simulations with the SPRT could be reduced without much influence on the consistency of the decisions. With the confidence interval, the PCDs decreased more than with the SPRT if the test length was decreased.

5.5.4 Results for the Example with Within-Dimensionality

Simulations were also run with several item selection and classification methods for the item pool with within-dimensionality. Test length was not decreased in the

Table 5.3 *Average Test Length for Simulations with Three Item Selection Methods for the SPRT*

$\alpha = \beta$	δ	LAN	WR	SP	SPNV	SPV	TC	VO
Random Item Selection								
0.05	0.1 sd	95.968	29.988	20.000	10.000	10.000	29.996	20.000
0.05	0.2 sd	86.577	29.335	19.654	10.000	10.000	29.439	19.974
0.05	0.3 sd	77.922	28.260	18.928	9.965	9.943	28.309	19.662
0.10	0.1 sd	92.666	29.881	19.960	10.000	10.000	29.905	20.000
0.10	0.2 sd	80.879	28.603	19.160	9.979	9.973	28.687	19.792
0.10	0.3 sd	67.522	27.150	18.124	9.823	9.774	27.161	19.057
Maximization at the Ability Estimate								
0.05	0.1 sd	95.069	29.986	20.000	10.000	10.000	29.995	20.000
0.05	0.2 sd	84.922	29.259	19.688	9.998	10.000	29.270	19.946
0.05	0.3 sd	75.422	28.091	18.751	9.940	9.892	28.033	19.422
0.10	0.1 sd	91.588	29.865	19.976	10.000	10.000	29.865	20.000
0.10	0.2 sd	78.926	28.482	19.068	9.961	9.950	28.428	19.619
0.10	0.3 sd	64.599	26.984	17.856	9.757	9.646	26.813	18.637
Weighting Method								
0.05	0.1 sd	95.014	29.985	20.000	10.000	10.000	29.993	20.000
0.05	0.2 sd	85.048	29.258	19.502	9.998	10.000	29.265	19.929
0.05	0.3 sd	75.314	28.109	18.593	9.935	9.885	28.019	19.408
0.10	0.1 sd	91.527	29.860	19.938	10.000	10.000	29.874	20.000
0.10	0.2 sd	78.922	28.480	18.886	9.959	9.936	28.429	19.618
0.10	0.3 sd	64.035	26.978	17.770	9.758	9.660	26.775	18.626

Note. Average test length for Language (LAN), Writing (WR), Spelling (SP), spelling Non-Verbs (SPNV), Spelling Verbs (SPV), Text Comprehension (TC), and Vocabulary (VO) with random item selection (RA), maximization at the ability estimate (AE), and the weighting method (WM). $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

first part of the simulation study, but testing was stopped in the second part of the simulations if enough confidence was gathered for the decisions.

The proportions of consistent decisions for the simulations with the SPRT with a fixed test length are presented in Table 5.5. Since all items were administered, no effects of the use of a different item selection method were possible. The size of α and β did not influence the PCDs for the SPRT. The size of δ had a negligible influence on the PCDs.

The second part of the simulations with the SPRT had a flexible test length. The PCDs for those simulations are also given in the table. The item selection

Table 5.4 *Average Test Length for Simulations with Three Item Selection Methods for the Confidence Interval Method*

γ	LAN	WR	SP	SPNV	SPV	TC	VO
Random Item Selection							
1.9600	64.598	27.534	19.428	10.000	10.000	28.461	19.590
1.6449	57.350	25.404	18.349	10.000	10.000	26.572	18.689
1.4395	53.772	23.730	17.381	9.903	9.970	24.925	17.756
1.2816	49.814	22.080	16.384	9.683	9.765	23.262	16.770
Maximization at the Ability Estimate							
1.9600	64.088	26.850	19.196	10.000	10.000	27.638	19.386
1.6449	57.405	24.772	17.999	9.998	10.000	25.476	18.266
1.4395	52.506	23.024	17.012	9.871	9.924	23.740	17.269
1.2816	48.922	21.597	16.125	9.677	9.619	22.306	16.213
Weighting Method							
1.9600	63.962	27.271	19.399	10.000	10.000	27.856	19.489
1.6449	56.608	25.156	18.272	10.000	10.000	25.783	18.483
1.4395	52.381	23.401	17.316	9.917	9.979	24.149	17.488
1.2816	48.584	21.892	16.440	9.717	9.728	22.626	16.530

Note. Average test length with for Language (LAN), Writing (WR), Spelling (SP), spelling Non-Verbs (SPNV), Spelling Verbs (SPV), Text Comprehension (TC), and Vocabulary (VO) with random item selection (RA), maximization at the ability estimate (AE), and the weighting method (WM). γ denotes the constant for the confidence interval.

methods had a negligible influence on the consistency for the SPRT. The sizes of α, β, and δ had little influence on the consistency for the simulations with the SPRT.

The settings for γ did not influence the PCDs for the second classification method with a fixed test length (see Table 5.6). The Mathematics decisions were the most consistent decisions and the PCDs for Analysis were the lowest. The latter finding can probably be explained by the low number of available items (17).

The size of γ for the confidence interval method with a flexible test length did influence the PCDs, except for Analysis. Depending on the decision, the weighting method or the method based on the ability estimate resulted in the most consistent decisions with the confidence interval method.

The decisions at maximum test length for Analysis were less consistent with the confidence interval method than those with the SPRT. The differences between the classification methods were smaller for the other decisions. The decisions with the SPRT were more consistent than the decisions with the confidence interval method if test length was not fixed.

Table 5.5
Proportion of Consistent Decisions with Maximum and Flexible Test Length for the SPRT

$\alpha = \beta$	δ	MATH	AN	AP	BS
Maximum Test Length					
-	0.1 sd	0.872	0.592	0.817	0.816
-	0.2 sd	0.871	0.592	0.817	0.815
-	0.3 sd	0.871	0.592	0.817	0.815
Flexible Test Length					
0.05	0.1 sd	0.872	0.592	0.817	0.816
0.05	0.2 sd	0.871	0.592	0.817	0.815
0.05	0.3 sd	0.869	0.592	0.817	0.815
0.10	0.1 sd	0.872	0.592	0.817	0.816
0.10	0.2 sd	0.871	0.592	0.817	0.815
0.10	0.3 sd	0.862	0.592	0.816	0.814

Note. Proportion of consistent decisions (PCD) for the decisions with maximum and reduced test length for Mathematics (MATH), Analysis (AN), Application (AP), and Basic Skills (BS). γ denotes the constant for the confidence interval. $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

The ATL of the SPRT simulations is presented in Table 5.7. Except for Analysis, on average, the tests were shorter for the SPRT than the specified maximum. If α, β, or δ was increased, the ATL decreased for the SPRT. Tests with random item selection for the SPRT were on average longer than for the other methods. On average, the weighting method resulted in the shortest tests for the SPRT.

If γ was decreased in the simulations with the confidence interval method (see Table 5.8), the ATL decreased. In contrast to the SPRT simulations, the Analysis tests are sometimes shortened by the confidence interval method. Except for Analysis, random item selection was always outperformed by the other selection methods for the confidence interval method simulations.

The ATLs of the confidence interval method were often lower than those for the SPRT, but their consistency was also lower.

5.6 Conclusions and Discussion

Two classification methods for two types of multidimensionality were investigated in this study. Wald's sequential probability ratio test (SPRT) was applied by Seitz and Frey (2013a) to between-dimensional tests. The method was adapted to make classifications on the entire test and on parts of the test by Van Groen and Eggen

Table 5.6
Proportion of Consistent Decisions with Maximum and Flexible Test Length for the Confidence Interval Method

Method	γ	MATH	AN	AP	BS
Maximum Test Length					
-	-	0.873	0.566	0.816	0.816
Flexible Test Length					
RA	1.9600	0.749	0.563	0.711	0.716
RA	1.6449	0.725	0.561	0.683	0.690
RA	1.4395	0.703	0.557	0.663	0.668
RA	1.2816	0.677	0.560	0.634	0.653
AE	1.9600	0.776	0.565	0.726	0.733
AE	1.6449	0.740	0.563	0.699	0.713
AE	1.4395	0.708	0.564	0.677	0.690
AE	1.2826	0.688	0.565	0.644	0.671
WM	1.9600	0.762	0.566	0.725	0.738
WM	1.6449	0.729	0.566	0.693	0.717
WM	1.4395	0.705	0.565	0.676	0.701
WM	1.2816	0.687	0.566	0.656	0.691

Note. Proportion of consistent decisions (PCD) for the decisions with with maximum and reduced test length for Mathematics (MATH), Analysis (AN), Application (AP), and Basic Skills (BS). γ denotes the constant for the confidence interval.

(2014). The SPRT has been applied to within-dimensional tests by Van Groen et al. (2014b). The confidence interval method was originally developed by Kingsbury and Weiss (1979) for unidimensional item response theory, but has been applied by Seitz and Frey (2013b) to between-dimensional testing. However, their method could only make decisions for each dimension independently. Thus, the method was adapted to make decisions on all items in within-dimensional tests.

For unidimensional classification testing, it is well-known that both methods and the settings that were used for them influence test length and accuracy. It has also been established that these methods have a similar influence in the case of multidimensional computerized classification testing (Van Groen et al., 2014b; Van Groen & Eggen, 2014), but those studies did not include all four classification methods and all the discussed item selection methods.

Simulation studies were used to investigate the efficiency and effectiveness of the classification methods. Comparisons were made between different item selection methods, between different classification methods, and between different

Table 5.7 *Average Test Length for Simulations with Three Item Selection Methods for the SPRT*

$\alpha = \beta$	δ	MATH	AN	AP	BS
Random Item Selection					
0.05	0.1 sd	173.413	17.000	88.983	73.973
0.05	0.2 sd	145.088	17.000	86.106	71.130
0.05	0.3 sd	114.373	17.000	80.292	65.812
0.10	0.1 sd	165.995	17.000	88.641	73.543
0.10	0.2 sd	123.560	17.000	82.185	67.802
0.10	0.3 sd	89.777	16.997	71.954	57.957
Maximization at the Ability Estimate					
0.05	0.1 sd	170.393	17.000	88.980	73.967
0.05	0.2 sd	130.136	17.000	84.936	70.095
0.05	0.3 sd	93.670	17.000	77.598	63.055
0.10	0.1 sd	160.515	17.000	88.475	73.367
0.10	0.2 sd	104.208	17.000	80.040	65.825
0.10	0.3 sd	68.043	16.996	65.983	52.745
Weighting Method					
0.05	0.1 sd	169.733	17.000	88.966	73.953
0.05	0.2 sd	128.886	17.000	84.622	69.732
0.05	0.3 sd	92.221	17.000	76.973	62.618
0.10	0.1 sd	159.506	17.000	88.361	73.241
0.10	0.2 sd	102.716	17.000	79.520	65.275
0.10	0.3 sd	66.335	16.997	65.061	52.313

Note. Average Test Length for Mathematics (MATH), Analysis (AN), Application (AP), and Basic Skills (BS). $\alpha = \beta$ = acceptable error rates; δ = distance between cutoff point and boundary of indifference region.

settings for the classification methods. Based on the simulations, one could conclude that the SPRT is a more accurate method than the confidence interval method to make a decision on Language; however, the latter is more efficient for between- and within-dimensionality. The current findings are based on just one dataset per type of dimensionality and on specific settings for the classification methods. If a different dataset had been used, the results could have been different. In addition, a clear interaction exists between the settings for the classification methods, and the consistency and efficiency of the decisions. If different settings had been specified, the outcomes could have been different.

In the current study, only maximum test length and flexible test length were included in the simulations. In order to compare the decisions made by the SPRT

Table 5.8 *Average Test Length for Simulations with Three Item Selection Methods for the Confidence Interval Method*

γ	MATH	AN	AP	BS
Random Item Selection				
1.9600	115.261	16.870	56.734	56.444
1.6449	96.659	16.794	50.464	52.014
1.4395	83.179	16.672	44.590	47.968
1.2816	70.185	16.465	39.110	44.155
Maximization at the Ability Estimate				
1.9600	108.244	16.948	51.715	53.321
1.6449	80.874	16.903	42.003	49.512
1.4395	62.853	16.864	33.995	45.561
1.2816	50.057	16.666	27.086	39.606
Weighting Method				
1.9600	106.564	16.899	50.941	53.640
1.6449	81.560	16.875	40.724	49.908
1.4395	63.011	16.828	32.000	45.821
1.2816	49.392	16.750	25.834	41.309

Note. Average Test Length for Mathematics (MATH), Analysis (AN), Application (AP), and Basic Skills (BS).

and confidence interval method, one should fix the test length at a value lower than the number of items available for the decisions in the item pool. This way, one could compare the classification consistency between both methods for different item selection methods and different settings for the classification method.

The efficiency of the SPRT and the confidence interval method can be investigated by matching the PCDs of simulations with the methods. This implies that a large range of settings for the classification methods must be used in order to find which settings result in precisely the same consistency. Then the average test length of the simulations can be compared. This requires that the classifications used for determining the PCDs are identical. If a decision is based on more than one dimension for between-dimensionality, this implies that a classification based on an external criterion has to be used. If no matching values for the PCDs can be established, this implies that the comparison cannot be made.

Another possibility for comparing the SPRT and the confidence interval method would be to investigate whether it is possible to make a comparison based on a mathematical proof. However, as far as the authors are aware, no proof has yet been presented that matches the SPRT to the confidence interval method.

In the current study, a dataset was used for the between-dimensionality simulations and another dataset for the within-dimensionality simulations. This implies that it is not possible to compare the decisions between the models. It would be interesting to investigate whether different decisions are made if a different dimensionality type has been specified for a dataset. Such a comparison was made by Hartig and Höhler (2008), who compared ability estimates based on between- and within-dimensional models. Such an analyses would imply, however, that decisions are compared that were based on different ideas about the structure of the data. Typically, the structure of the test determines which type of dimensionality is used for modeling. However, for research purposes it could be interesting to compare the classification decisions that resulted from the two models.

When the underlying results of the simulations were examined closely, it became apparent that (weighted) maximum likelihood estimation for estimating ability for tests with within-dimensionality can be troublesome. The estimated abilities sometimes diverged from their true values, but compensation between the abilities also occurred. These preliminary findings suggest that more knowledge about ability estimation in the within-dimensional case would be welcome.

The item selection method for CATs with within-dimensionality also needs attention. The method that maximizes the determinant of the information matrix at some point in the ability scale was originally developed for MCAT for estimating ability. This implies that the method is not developed for CCTs using a reference composite. To give an example, the method that maximizes the determinant at the cutoff point does not necessarily select an item whose difficulty is close to the point in the multidimensional space that corresponds to the cutoff point. This item has a value on the reference composite that is close to the cutoff point on the reference composite, but the actual point of most information for the item can be far away from the cutoff point in the multidimensional space.

It would also be interesting to investigate other item selection methods from unidimensional classification testing or multidimensional computerized adaptive testing for estimating ability for application to multidimensional classification testing. Other methods could result in slightly more consistent or efficient decisions. Nevertheless, it is expected that the differences between the methods will be small.

Finally, it would be interesting to investigate whether it is possible to use other classification methods for multidimensional classification testing. Currently, only the SPRT and the confidence interval method have been applied. The effectiveness and efficiency of other methods could be an interesting topic for further study.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255–278. doi: 10.1207/s15324818ame0704_1

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*, *23*, 249–261. doi: 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*, *60*, 713–734. doi: 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]. Armidale, Australia: University of New England, Centre for Behavioral Studies.

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*, *216*, 89–101. doi: 10.1027/0044-3409.216.2.89

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategie for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, *2*, 359–375. doi: 10.1207/s15324818ame0204_6

Mulder, J., & Van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296. doi: 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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. doi: 10.1177/014662168500900409

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

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

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 Statistics*, *20*(4), 502–522.

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, April). *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*, *69*, 778–793. doi: 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*, *24*,

398–412. doi: 10.3102/10769986024004398

Van Groen, M. M., & Eggen, T. J. H. M. (2014). *Multidimensional computerized adaptive testing for classifying examinees on tests with between-dimensionality*. Manuscript submitted for publication.

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*, *38*, 187–200. doi: 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 within-dimensionality*. Manuscript submitted for publication.

Veldkamp, B. P., & Van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. doi: 10.1007/BF02295132

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: University of Iowa.

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*, *28*, 295–316. doi: 10.1177/0146621604265938

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi: 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*, *77*, 495–523. doi: 10.1007/S11336-012-9265-5

Chapter 6
Assessment Approaches and Types of Digital Assessments

Abstract

It is only through assessment that one can determine whether instruction has resulted in the intended learning outcomes (Wiliam, 2011). Assessment serves different goals, and a single test can have multiple goals. Tests have to be designed so that outcomes are sufficiently accurate and so that they are suitable for the testing goals, have the correct grain size, and have a test length suitable for the testing purpose and testing population.

Four types of assessment approaches are explored: formative assessment, formative evaluation, summative assessment, and summative evaluation. Depending on the type of assessment, different types of tests can be used. Seven types of digital assessments are explored: linear testing, automatically generated tests, computerized adaptive testing, computerized classification testing, adaptive learning environments, educational simulations, and educational games. The types of tests vary in the design of their modules (student, tutor, knowledge, and user interface) and in the way testing is adapted.

This chapter has been submitted as Van Groen, M.M., & Eggen, T.J.H.M. (2014). *Assessment Approaches and Computer-Based Testing.* Manuscript submitted for publication.

6.1 Introduction

It is only through assessment that one can determine whether instruction has resulted in the intended learning outcomes (Wiliam, 2011). The relations between learning, teaching, and assessment are increasingly being recognized (Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2013). Interdependencies also exist between different approaches, types, and assessment design. The current article explores the interdependencies between approaches, types of digital assessments, and design, with some attention paid to the relations between assessment, teaching, and learning.

First, four approaches are explored: formative assessment, formative evaluation, summative assessment, and summative evaluation. Second, seven types of digital assessment are investigated: linear tests, automatically generated tests, computerized adaptive tests, computerized classification tests, adaptive learning environments, educational simulations, and educational games. Third, four design modules are described for different test types: student, tutor, knowledge, and user interface. While discussing design, attention is paid to how testing is adapted to the individual examinee. After describing the elements of the interdependencies, the interdependencies between test approaches and test types are explored.

6.2 Test Approaches

A wide range of assessment approaches and definitions of these approaches exist. Four approaches are described: formative assessment, formative evaluation, summative assessment, and summative evaluation. A complication when approaches are distinguished is that a specific test can serve several purposes (Stobart, 2008). The primary goal before and during test development determines which approach is the most eminent for an assessment.

The stakes of testing differ in the approaches. A high stakes assessment has substantial consequences for some or all parties involved (Stobart, 2008). A low stakes test has limited consequences for some or all parties involved. An assessment can be high stakes for one party, but low stakes for another party. The stakes of the test have a major influence on test design.

6.2.1 Formative Assessment

The first approach is formative assessment. The focus is here on supporting and improving the learning process to facilitate learning, by making decisions at the levels of the learner and the class (Van der Kleij et al., 2013). Well-designed and implemented formative assessment suggests how instruction should be modified and tells the teacher what pupils know and can do (Bennett, 2011). The use of assessment for instructional guidance offers a powerful way for improving schooling (Wiliam, 2013). Assessment outcomes have to be precise enough at the individual level, so that instruction can be adapted. Feedback is considered important for enhancing learning from formative assessments (Van der Kleij et al., 2013). Van der Kleij et al. (2013) described three types of formative assessment: data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT).

Data-based Decision Making

The first type is DBDM. Schildkamp and Kuiper (2010) define it as "systematically analyzing existing data sources within the school; applying outcomes of analyses to innovate teaching, curricula, and school performance; and implementing (e.g., genuine improvement actions) and evaluating these innovations" (p. 482). Using the data, teachers can then set learning goals given the current student knowledge level (Van der Kleij et al., 2013).

Assessment for Learning

The second type is AfL, which focuses on the quality of the learning process (Van der Kleij et al., 2013). AfL attempts to make testing a part of the learning process (Stobart, 2008) and focuses "on what is being learned and on the quality of classroom interactions and relationships" (p. 145).

Diagnostic Testing

The third type is DT, which assumes that how a task is solved indicates the learner's developmental stage (Van der Kleij et al., 2013). Testing concerns individual educational needs; DT is not meant for comparing learners, but for promoting learning and development (Van der Kleij et al., 2013).

6.2.2 Formative Evaluation

The second approach is formative evaluation. Harlen (2007) and Shepard (2005) distinguish formative assessment and formative evaluation. Assessment focuses on making judgments about pupils and their class, while evaluation judges programs or schools (Harlen, 2007; Stobart, 2008).

The use of assessment data makes formative evaluation a form of testing (Haertel, 2013). Individual results are aggregated to get group or school information (Sanders, 2013). Test outcomes have to be sufficiently precise at this level. The focus is on improvement and the day-to-day running of educational systems and organizations (Scheerens, Glas, & Thomas, 2003).

6.2.3 Summative Assessment

In summative assessment, tests make inferences about individual students (Haertel, 2013) and focus is on what has been learned by the end of the process. The test results play a role in making a decision about the mastery of a content domain by the pupil or class (Van der Kleij et al., 2013), guiding decisions about student ability grouping, determining entry into and exit from education, making college admission decisions (Haertel, 2013), and deciding whether a student graduates. Focus is on differences among students at a given time point (Haertel, 2013). Sanders (2013) defined summative assessment as making judgments about students and specified several goals: selection, classification, certification, or placement.

6.2.4 Summative Evaluation

The fourth approach, summative evaluation, focuses on the use of assessment data for making judgments about schools (Van der Kleij et al., 2013) or educational systems. As a result, educational systems can focus their energy and effort towards improvement in measured schooling outcomes (Haertel, 2013). Measurement has to be precise at the group level and the test design can be adapted accordingly.

Accountability of schools is one form of summative evaluation in which schools provide information on their performance and functioning to outside parties so that schools are open to public review (Scheerens et al., 2003).

The other form of summative evaluation supersedes the school level. Judgments can be made at the regional or (inter)national level and are typically based on large scale assessments. The goal is to inform relevant parties about the state

of the educational system (Sanders, 2013). Tests are used that are low-stakes at the individual level for monitoring national standards by testing a representative sample of students (Stobart, 2008). The assessments intend to show what students can do at one time, but changes across years are also monitored (Harlen, 2007). Adjustments in policy and standards are made based on the findings.

6.3 Types of Tests

Besides different test approaches, several types of digital assessments exist. The possibilities while testing and with the assessment outcome depend on the type of test. Seven types are discussed here: linear tests, automatically generated tests, computerized adaptive tests, computerized classification tests, adaptive learning environments, educational simulations, and educational games. The discussion is limited to the types of tests that can be scored automatically by a computer.

6.3.1 Linear Tests

The first type is linear testing (LT) in which the content, item order, and test length are the same for everyone. Items are selected before test administration, with or without pretesting, although without is generally not advisable. Testing is inefficient because testing is not tailored to the individual student (Mellenberg, 2011). As a test result, a number correct score, percentile score in a distribution, classification, or ability estimate is reported.

6.3.2 Automatically Generated Tests

The second type is automatically generated testing (AGT) in which fixed length tests are produced that satisfy a number of constraints or conditions (Parshall, Spray, Kalohn, & Davey, 2002). Constraints include content restrictions and psychometric properties (Parshall et al., 2002). Test forms are assembled before administration using heuristics or linear programming (Parshall et al., 2002; Van der Linden, 2005). The same results can be reported as for linear testing.

6.3.3 Computerized Adaptive Tests

The third type is computerized adaptive testing (CAT) in which items are selected that best fit the examinee (Mellenberg, 2011). This results in a more precise ability estimate. After each response, the examinee's ability is estimated and the next item is selected that has optimal properties at the new estimate (Van der Linden

& Glas, 2010). Testing can be stopped when the desired accuracy level has been obtained or after a fixed number of items. CAT requires an item bank that is calibrated with item response theory (IRT). The item bank is assembled after a pretest and contains items that have desired characteristics, such as appropriate difficulty. After testing, an ability estimate, percentile, or classification based on the estimate can be reported.

Mixture forms of CAT and LT are multistage testing (MST) and multi segment computerized adaptive testing (MSCAT; Eggen, 2013). In MST, item sets are used as building blocks for the test (Zenisky, Hambleton, & Luecht, 2010). These blocks are assembled beforehand but are selected during testing based on the examinee's ability. In MSCAT, the CAT consists of several segments and branching rules between the segments (Eggen, 2013). The segments form their own CAT. Since MST and MSCAT are a form of CAT, they are not discussed further.

6.3.4 Computerized Classification Tests

The fourth type is computerized classification testing (CCT). CCT is a form of CAT, but the purpose and intended use of the tests differs. CCT classifies examinees in one of multiple levels instead of estimating ability (Eggen & Straetmans, 2000). The difference in testing goal implies that some parts of the assessment procedure differ. Testing is stopped when sufficient confidence is available to make a decision. As a test result, the examinee's classification decision is reported.

6.3.5 Adaptive Learning Environments

The fifth type is item-based adaptive learning environments (ALE). These systems provide instruction that is optimized to each learner's individual needs, preferences, and/or the context (Wauters, 2012). The interaction between the learner and the learning tasks is optimized by adapting the sequencing of learning content and feedback, which leads to more efficient and effective learning (Wauters, 2012).

Brusilovsky (1999) distinguished two types of environments: adaptive hypermedia systems and intelligent tutoring systems. The first type contains knowledge in the form of hypertexts (Wauters, 2012). The presentation of the information and the overall link structure is adapted based on registered user actions. The second type of ALE supports the learner in the problem solving process and provides help on every part of problem solving based on discovered knowledge gaps. The discussion is limited here to intelligent tutoring systems in which the learning environment selects the items for the learner.

6.3.6 Educational Simulations

The sixth type is educational simulations (ES). In ES, dynamic or interactive features are present, such as viewing an animation, that act in ways that prompt a change or response from a system (Levy, 2013). They often contain complex tasks that require multiple steps, capture multiple features of task performance, produce products in a relatively unconstrained manner, and relate task features with aspects of performance (Levy, 2013). Educational simulations allow for observation of student behavior in environments that approximate relevant real-world situations where it would be impractical, cost prohibitive, or unethical to place students in the actual situation for assessment purposes (Levy, 2013). Education simulations can report multiple aspects of proficiency in which proficiency can include a wide range of abilities and skills.

6.3.7 Educational Games

The last type is educational games (EG). Although EG are used often for instructional purposes, the focus is here on assessment environments. Novak, Johnson, Tenenbaum, and Shute (2014) summarized some potential benefits of game-based learning: it facilitates hands-on student-centered learning and it encourages integration of knowledge from different areas to make decisions and to examine outcomes. Games facilitate learning because playing not only affects the learning outcomes, but they keep the learner engaged and motivated (Novak et al., 2014). Evidence-centered design can be used to make valid inferences from games (Mislevy, Steinberg, & Almond, 2003; Shute & Ke, 2012). The systems can report real-time estimates of competencies across a range of knowledge and skills (Mislevy et al., 2003; Shute & Ventura, 2014). Educational games can provide useful information to students, teachers and the system itself (Mislevy et al., 2014). This requires "reasoning from the specific things that students do, to what they know and do more broadly, and what the system, the teacher, the students themselves might do next" (p. 9) to develop skills and knowledge.

6.4 Test Design and Adaptivity

Depending on the test type, test design will be different. The architecture of an ALE consists of four modules: student, tutor, knowledge, and user interface (Nwana, 1990; Wauters, 2012). Other modules of other scholars, for example group and adaptation modules (Paramythis & Loidl-Reisinger, 2004), can be redefined

into the four modules. These modules for ALE are considered applicable to other types of assessments, but the design varies between the test types. Test design has major consequences on the precision of the test outcomes and possible test use. This also implies that the stakes of testing have an influence on test design. As Gee (2008) states it, a game's design is inherently connected to designing good learning for students. According to Mislevy et al. (2014) "a better design process jointly addresses the concerns of game design, instructional design, and assessment as required, so that key considerations of each perspective are taken into account from the beginning" (p.13).

One way in which tests differ is in how they are adapted for individual students. Adaptive systems attempt to be tailored for different students by taking into account information accumulated in the student model (Brusilovsky & Peylo, 2003). The adaptation model defines what can be adapted and when and how it is to be adapted (Paramythis & Loidl-Reisinger, 2004). Items are selected so that testing is tailored to the individual students or so that learning is optimized.

Adaptivity was defined by Wauters, Desmet, and Van den Noortgate (2010) as "the adjustment of one or more characteristics of the learning environment in function of the learner's needs and preferences and/or the context". Wauters (2012) describes four dimensions of adaptivity in ALE: medium, form, source, and level of adaptivity. The majority of these dimensions are considered applicable to other types of testing as well. The features of the environment involved in the adaptivity, the form dimension, are classified into three categories by Wauters (2012): course/item features, person features, and context features. These features are described in the next sections on the design modules. The level and context of adaptivity cannot be attributed to specific modules so they will be discussed in the final parts of this section.

LT and AGT are by definition not adaptive, but some degree of adaptation is possible if different examinees receive different (assembled) test forms based on previous performance. The possibilities of adaptation for CAT, CCT, ALE, ES, and EG are discussed for each module. In some types of tests, the learner selects his own learning path or items, but discussion here is limited to adaptation by the computer except for EG.

6.4.1 Student Module

The student module contains all information about the student, such as the student's current knowledge level, student characteristics, and learning style

(Wauters, 2012). The module often forms a representation of the student's current knowledge with respect to the mastery of the knowledge in the domain module (Nwana, 1990) and maintains a live account of the user's actions within the system (Paramythis & Loidl-Reisinger, 2004). This information can be used by teachers to monitor the student's knowledge and progress or for item selection (Wauters, 2012). The item selection method requires the ability estimate, which items were already administered, the content of the administered items, and so on. The student's knowledge level can be estimated using IRT. Other characteristics can be estimated based on the way the student navigates through the system or can be specified.

The student modules of LT and AGT can be very simple. The module contains information about the administered tests and their results. This information makes it possible to select a new test and to monitor student performance over time.

The student module in CAT contains at a minimum, the ability estimate along with the information that is required for LT. The module uses this estimate for item selection and monitors which items have been answered. Student background information can also be included and used for adapting the item selection.

The student module of CCT resembles the module for CAT. If items are selected that measure optimal at the cutoff point, no ability estimate needs to be included.

The student module is a central component in ALE and contains the learner's knowledge level, characteristics, learning style, and so on (Wauters, 2012). The tutor module uses this information to tailor the selection of learning material to the student.

The student module in ES articulates in the student model what is being measured about the student (Levy, 2013). The student model tends to be multidimensional because inferences have to be made about multiple, distinct aspects of proficiency (Levy, 2013).

The student module in EG contains information about the student's achievements during the game. Such achievements include the completion of levels and the number of points that have been scored. Information required for the assessment is also stored here.

6.4.2 Tutor Module

The tutor regulates the instructional interactions with the student (Nwana, 1990), selects the learning material, answers the learner's questions about goals and content, and decides when a learner needs help (Wauters, 2012). The module's

importance strongly depends on the type of test. In some assessments, no tutor module is required because staff fulfills the role of the module and items were selected beforehand. In other testing systems, the module selects the items, which ensures that measurement is efficient. The necessity of providing information about the learning or testing goals depends on the test because this information often has to be available before testing begins.

In LT and AGT, the tutor module can be very simple or even absent. The tutor module can provide the learning goals and information about the test.

The tutor module is the central module of CAT. The module tailors the test to the examinee by selecting those items that are optimal for obtaining the ability estimate. In one item selection method, the optimal item maximizes the information in the test for ability estimation by maximising information at the ability estimate (Chang & Ying, 1996). Selecting the item that matches the ability estimate results in tests that are not too easy or too difficult (Reckase, 2009). If necessary, easier items can be selected (Eggen & Verschoor, 2006). Many item selection methods are available, see, for example, Van der Linden and Glas (2010). At the beginning of the test, a starting procedure is required because no ability estimate is available for selecting the items (Eggen & Straetmans, 2000). This procedure can be based on student characteristics or previous ability estimates.

The tutor module in CCT can be identical to the module in CAT, but tends to be less efficient. Popular selection method in CCT selects the item that maximizes information at the cutoff point(s) (Eggen, 1999; Eggen & Straetmans, 2000; Spray & Reckase, 1994). The information about the examinee from the student module is not used by these methods. These methods result in tests that are not tailored to the examinee's ability, so alternative methods were proposed that use both the ability estimate and the cutoff points (Van Groen, Eggen, & Veldkamp, 2014).

The tutor module is an important part of ALE. The module tailors instruction based on the learner's educational needs and motivation (Wauters, 2012). The item selection method can use a mixture of the ability estimate and student background characteristics to select the items (Wauters, 2012).

In ES, the tutor module selects the next task that enhances learning the most. The module uses the information in the student module for selection. The module can use, for example, multidimensional ability estimates to select the next task.

In EG, the games are designed to set up certain goals, but they often leave students free to achieve these goals in their own ways (Gee, 2008). Players can design their own goals, but are limited by the rule space designed into the game

(Gee, 2008) and the level structure in the game. This implies that students can select their own learning and assessment path. The module provides the next situation for the student based on the learning goals. Feedback can be given at each moment and is often summarized at the end of the level (Gee, 2008).

6.4.3 Knowledge Module

The knowledge module represents the knowledge that the learner is trying to acquire and the relationships between the knowledge elements (Paramythis & Loidl-Reisinger, 2004; Wauters, 2012). The steps for solving the problem are also stored here (Nwana, 1990). For each item, all elements, such as pictures, question text, and so on, are stored. Items and course features, such as difficulty and topic (Holt, Dubs, Jones, & Greer, 1994; Wauters, 2012), are also stored. Item difficulty can be estimated using IRT or items can be classified into difficulty levels by experts. Features related to the topic are based on content classifications by experts. Items and learning tasks are selected by the tutor module from the knowledge module. In practice, tests tend to have an elaborate set of content specifications. A domain specification is available for each item, which can be used by the item selection method to control test content. If certain item combinations are required or not allowed, these relations are also specified here.

Wauters (2012) defined three forms of adaptivity: adaptive form representation, adaptive content representation, and adaptive curriculum sequencing. The information for these forms of adaptivity is stored here. Adaptive form representation determines whether pictures and video are visible and whether and how links are presented (Wauters, 2012). Adaptive content representation adapts the help a learner receives in the problem solving steps based on the learner's knowledge (Brusilovsky, 1999; Wauters, 2012). The feedback a learner receives when or after answering is a second form of adaptive content representation. Adaptive curriculum sequencing selects the optimal question to learn certain knowledge in an efficient and effective way (Wauters, 2012). The system helps the student to find an optimal path through the learning material (Brusilovsky, 1999).

In LT, the knowledge module contains the items. The test developers specify item order beforehand so that content distribution within the assessment is guaranteed.

The knowledge module for AGT contains the item pool and item characteristics. Tests are assembled by the tutor module that meet the specified content constraints. If psychometric constraints were set, this information is stored here.

In CAT, the knowledge module contains the calibrated item pool. The items and item parameters are stored in the item pool. Some items share a stimulus, such as a common text, and have to be administered in the corresponding order. This information is also stored here. In some tests, elaborate content specifications are required, with items having additional attributes that are related to the specifications. These attributes are used for content control in item selection. In CAT, item content can be adapted during testing. Adaptive form representation is only allowed in CAT if the representation differences are included in the calibration model because adding a figure could make an item easier. In CAT, it is assumed that the examinee's true ability remains constant during testing. If help or feedback is offered during testing, the examinee's ability is influenced, which is a violation of IRT assumptions. Feedback can always be given after testing. Caution should be taken while modeling items that were administered in a test with adaptive curriculum sequencing because IRT assumptions are possibly violated.

The knowledge module of CCT is very similar to the module of CAT. In CCT, it is also possible to adapt the curriculum sequencing after making a preliminary decision. A final decision is then made after enough confidence has been gained for the second decision.

The knowledge module of ALE contains the knowledge a learner is trying to acquire (Wauters, 2012). The content organization and the way content is offered is based on knowledge about the topic (Nwana, 1990). Information about the items has to be stored, such as difficulty and topic related features. ALE try to enhance learning during testing. Adaptive form representation ensures that each learner learns in a way matching to his learning preferences. By adapting the curriculum sequence in ALE, learning can be enhanced. An item can be selected that follows the previous items content-wise. Content adaptation is possible in ALE. Items can be selected so that a logical content flow is realized based on the learner's knowledge. By providing help and feedback, learning is also enhanced (Van der Kleij et al., 2013) but these should be modeled in the statistical model.

The knowledge module contains the task model for ES if evidence centered design (Mislevy et al., 2003) was used for test development. The task model specifies the situations in which evidence can be collected for estimating proficiency (Levy, 2013). The model specifies the features of the task, activities, or situations and which information is collected using the task (Levy, 2013). Adaptive form representation, adaptive curriculum sequencing, and adaptation in terms of content can all be applied in ES.

Wainess, Koenig, and Kerr (2011) described the CRESST methodology for creating EG in which instruction and assessment are embedded within the player interaction framework. Instruction, domain content, and assessment are presented to the student using presentation objects that are directly presented to the student, background objects that the student should actively pursue, person-to-object or person-to-character/player interaction in which something is visually emphasized, and storage or workshop objects that summarize instruction. This information is included in the module.

6.4.4 User Interface Module

The user interface module is the only part of the test that the examinee actually sees and interacts with (Parshall et al., 2002). It is the communication component that controls the interactions between the student and the system (Nwana, 1990). The module displays items and retrieves the responses from the student. A good user interface demonstrates consistency and clarity and reflects good interface design principles, see Parshall et al. (2002) and Bartram and Hambleton (2006).

The design of the user interface partly determines the student's reactions to taking the test, and their performance may be affected (Parshall et al., 2002). The user interface of LT, AGT, CAT, and CCT should be easy to use and understand.

The design principles for the user interface are even more important in ALE, ES, and EG than in the other test types. The user interface should be developed so that the learner is intrinsically motivated to continue learning. In other test types, motivation is often external. In ALE, ES, and EG, the design should also enhance motivation. In EG, it is important that the learning follows from the game play, so instruction and game mechanics should be integrated (Wainess et al., 2011).

6.4.5 Level of Adaptivity

In the previous parts of this section, attention was paid to the test modules and the way they can be adapted for individual students. Wauters (2012) discussed two levels of adaptivity: static and dynamic. In static environments, the source of adaptivity remains constant while working in the environment (Wauters, 2012). In dynamic environments, the source of adaptivity changes while testing.

CAT is often a static environment because adapting the difficulty of the items that are administered is almost always the source of adaptivity. The set of items from which the most suitable item is selected can be varied during testing. If this occurs, CAT can be a dynamic environment.

CCTs are almost always static environments because the source of adaptivity remains the same during the testing. The item content can be varied throughout the test, which could make the test dynamic.

In ALE and ES, the level of adaptation can be dynamic or static depending on the way adaptation is specified (Wauters, 2012). Nowadays, ALE tend to be dynamic and adaptive to make learning more efficient (Wauters, 2012).

EG are always dynamic because the students choose their own learning path through the game. Everything is adapted based on the students play.

6.4.6 Context of Adaptation

Besides adapting the test itself, the context of the test administration can also be adapted based on context features, such as the time when, place from, and device on which the learner interacts with the system (Wauters, 2012). Assessments are nowadays made on a variety of devices, at several locations, and at several points in time. Adaptation of the context can be complicated in CAT and CCT. If the stakes are high, it is questionable whether the time and place of assessment should be varied between examinees. If a test is administered on multiple devices, it should be investigated whether this results in the same item parameters. Since the stakes are much lower in ALE, ES, and EG, the location and time of the learning can be easily adapted to suit the learner's preferences. The device on which the environment is used is only limited by whether the device can display the environment.

6.5 Assessment Approaches and Types of Tests

The type of test and its design imply that certain test types are more or less suitable for use with a specific assessment approach. For each assessment approach it is explored whether it is possible to use LT, AGT, CAT, CCT, ALE, ES, or EG.

6.5.1 Formative Assessment for Different Types of Tests

Van der Kleij et al. (2013) described three types of formative assessment: DBDM, AfL, and DT. The focus in these types is on improving the learning process to facilitate learning of the pupil and class.

Data-based Decision Making

In DBDM, existing data sources are systematically analyzed for innovation of teaching, curricula, and school performance (Schildkamp & Kuiper, 2010). The implemented innovations are then evaluated. The results of LT and AGT can be used for DBDM. The pupil's ability can be estimated, used as input for innovation, and used as a baseline before implementing the innovations. After implementation, LT and AGT can be used for evaluation. The precision of the outcomes differs between students.

CAT can provide the same information for DBDM as LT but it provides more precision. Testing is tailored to the examinee's ability, which results in a more precise estimate or results in shorter tests (Parshall et al., 2002). One potential practical problem of CAT for DBDM and other assessment approaches is that item parameters have to be known before inclusion in the item pool. This implies that the construction of such tests is challenging for schools. If a standardized CAT is available, this type of assessment results in efficient and precise estimates that can be used for innovating education.

CCT can have the same limitations as CAT, but if items are selected at the cutoff point that was set by the school, they can be very useful. In DBDM, the school can set learning goals (Van der Kleij et al., 2013) and, based on those goals, cutoff points. Before innovation, which and how many students master the learning goal can be determined. After, innovation, which and how many students now master the learning goal can be determined.

ALE and ES are less suitable for DBDM. ALE and ES tailor the instruction and assessment to the pupil's knowledge level, so each pupil is offered an individual learning path (Wauters, 2012). At the level of the pupil, it is possible to provide additional instruction based on the test outcomes. However, it is difficult to use the reported learning outcomes for innovating the instruction at the class or school level because some students have learned a specific topic but others haven't.

EG are not suitable for DBDM because the student chooses his own path through the game. This makes it very difficult to aggregate results.

Assessment for Learning

In AfL, the focus is on the quality of the learning process (Van der Kleij et al., 2013). The quality of the assessment process depends largely on the possibility to use the test results to make inferences about student learning and accompanying

adaptations to the instruction (Bennett, 2011). LT and AGT can be used to check whether the student has understood a certain topic, but tests can become very long if information is required about a lot of components of the topic. A second, related problem is that a large number of tests is required if the student's knowledge level is measured frequently.

A more efficient solution is CAT, which requires on average fewer items to obtain the same precision. It is also possible to use the system more frequently by excluding previously administered items. The estimates in CAT can be very precise. The teacher then knows which students need additional instruction and on which topics. If estimates are based on subsets of items, it is possible to retrieve a knowledge profile of the student. The information about the knowledge gaps becomes even more precise when multidimensional IRT (MIRT) is used.

Traditionally, CCT could provide limited guidance for use in the day-to-day process of learning. Since the development of CCT with multiple levels per decision (Eggen & Straetmans, 2000; Spray, 1993) more guidance can be provided by CCT. The use of MIRT can provide even more information about the student's knowledge level by making decisions on multiple dimensions, parts of tests, or on multiple levels (Seitz & Frey, 2013; Van Groen & Eggen, 2014). Furthermore, the same possibilities are available with CCT as with CAT.

The joint focus on adapting the instruction to the learner's knowledge in AfL, ALE, and ES means that these environments can be used for AfL. The instruction can be adapted to the test outcomes. The system also detects changes in the learner's knowledge level that results from the teacher's instruction and adapts its instruction accordingly. Feedback can also be used to enhance learning.

EG are especially interesting for AfL. The student can choose his own path, which makes it suitable to assess the learning process. The test developer should make sure that it is beneficial for the student to choose the path that enhances learning the most, for example, by giving incentives for choosing difficult levels.

Diagnostic Testing

DT is based on the assumption that how a task is solved is indicative of the learner's developmental stage (Van der Kleij et al., 2013). To obtain the precise level of information that is required, fine grained assessment data about the process have to be obtained (Rupp, Gushta, Mislevy, & Shaffer, 2010). In LT and AGT, a large number of items is required to get detailed information at different

abilities, which implies that only a small part of a content domain can be measured or tests have to consist of a large number of items.

CAT has the advantage of tailoring the test to the student's ability. This implies that if fine grained information is required, items can be administered at the ability estimate instead of testing at a wide ability range as in LT. The content of the next item can be determined based on the specific misconceptions in the process of the student solving the item (Van der Kleij et al., 2013).

CCT can reduce the test length even further because precision only has to be obtained at the cutoff points. If several cutoff points are used per misconception, a detailed report can be provided to the teacher. DT often measures multiple abilities at the same time, which implies that a MIRT model is required. The same item can be used for testing multiple misconceptions, so testing can be done efficiently using a multidimensional classification test (Van Groen & Eggen, 2014).

ALE and ES can be used for DT. DT focuses on obtaining information about the student's misconceptions in solving tasks. ALE attempt to find these weaker points in the student's knowledge and offer additional learning material. ES can also offer more tasks and help for improving weaker points (Wauters, 2012).

EG can provide the required information. The test developer can provide information about the problem solving steps that were taken, or not, by the student based on the student's actions in the game. While playing the game, students produce action sequences which draw on the skills or competencies that are assessed (Shute & Ventura, 2014). These action sequences can provide insight into the problem solving steps that the student used. The system has to be developed in such a way that accurate information can be obtained at the correct grain size.

6.5.2 Formative Evaluation for Different Types of Tests

Formative evaluation focuses on making judgments about programs or schools, rather than individuals (Harlen, 2007; Stobart, 2008), to develop educational policies in schools (Van der Kleij et al., 2013), and improving education (Scheerens et al., 2003). Precision of the outcomes is required at the school level, but not on the individual level. LT can be used for formative evaluation. LT and AGT are relatively easy to develop and can be developed by the school. The test results can show areas for possible improvement. Individual assessment results are aggregated for formative evaluation (Sanders, 2013). If different test forms are administered to groups of students, a wider range of topics can be covered.

CAT is a more efficient solution for retrieving the same or more precise results. If CAT is available for the intended domain, the results can be used for formative evaluation. If such a test is not available, it will be difficult to develop new CAT or CCT for the intended purposes.

ALE and ES can be used for formative evaluation, but the results have to be manually aggregated to see which areas of the educational program require improvement. In ALE and ES, students make their own test, which requires more effort to draw conclusions at the school level.

The test results of EG are even more difficult to aggregate. Students choose their own path through the environment, which means that no structured data are available for aggregation.

6.5.3 Summative Assessment for Different Types of Tests

Summative assessment focuses on what has been learned by the end of the process (Stobart, 2008). A decision is made about the mastery of a content domain by the pupil or class (Van der Kleij et al., 2013). LT and AGT can be used for summative assessment, because students' estimated ability can be ranked, and based on the ranking, the best performing students can be selected. Another possibility is to determine a cutoff point on the ability scale that is used to decide whether an examinee passes the test.

CAT can be used to make judgments about examinees. CAT provides ability estimates just as LT, but requires fewer items on average (Mellenberg, 2011). This means that more topics can be measured, topics can be measured in more detail, or tests can become shorter.

CCT can be used to make judgments about examinees because they classify examinees into different levels. CCT can be more efficient than CAT, LT, and AGT when a cutoff point is specified before test administration.

ALE, ES, and EG cannot be used for summative assessment. Since all students follow their own learning path, comparing individual students and deciding which students have enough knowledge becomes troublesome. An additional problem is that the judgments about students are often made about a large range of topics. These environments have a small grain size, which is in contrast to the requirements of the judgments for summative assessment.

6.5.4 Summative Evaluation for Different Types of Tests

In summative evaluation, judgments are made about schools (Van der Kleij et al., 2013) or educational systems. LT and AGT are often used for summative evaluation because precision is required at the group level, but not on the individual level. Thus, if a test form with limited accuracy is administered, the overall outcome can still be reliable at the group level. Different test forms can be used to cover a wide range of topics.

CAT tailors the test to the individual examinee's ability. Since measurement is intended to be at the group level in summative evaluation, the choice for tailoring tests to the individual level seems strange.

In summative evaluation, information is gathered about how much groups of students know about a topic. With this information, judgments are made about the school or about the educational system. CCT determines whether students pass or fail the test, but based on the test, it is not known how much the students know. This implies that CCT is not a logical choice for summative evaluation.

ALE, ES, and EG appear to be unsuitable for summative evaluation because of their focus on the individual learning path. It is difficult to make a judgment about the school if just individual results are available.

6.6 Discussion

Testing can serve different goals, and one test can have several goals. The assessment should be suitable for all those goals. Thus, tests have to be designed so that their outcomes are sufficiently accurate, have the correct grain size, are matched to the stakes of testing, and have a test length suitable for the testing purpose and population.

Not all assessment types are suitable for all assessment approaches. Before development of an assessment can start, the precise test goals and the intended use of the test results should be determined. Based on these goals, which type of test is the most suitable and feasible for the assessment can be determined. After specification of the type of test, it can be determined how the different modules of the test should look like and how testing can be adapted. Test design and adaptivity of the assessment should be closely aligned to gain the intended test outcomes and implications of those outcomes.

The current article was limited to types of digital assessments that can be scored by a computer. This means that tests that are paper-based or that need to be judged by evaluators were not included in the study. Item selection was limited to selection by the computer, except for educational gaming, but tests also exist in which the student selects the items.

References

Bartram, D., & Hambleton, R. K. (Eds.). (2006). *Computer-based testing and the internet: Issues and advances*. Chichester, United Kingdom: John Wiley & Sons.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*, 5–25. doi: 10.1080/0969594X.2010.513678

Brusilovsky, P. (1999). Adaptive and intelligent technologies for web-based education. *Künstliche Intelligenz* [Artificial Intelligence], *13*(4), 19–25.

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, *13*(2-4), 156–169.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229. doi: 10.1177/014662169602000303

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M. (2013, October). *Computerized adaptive testing serving educational testing purposes*. Paper presented at the meeting of the IAEA, Tel Aviv, Israel.

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. doi: 10.1177/00131640021970862

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*, 379–393. doi: 10.1177/0146621606288890

Gee, J. P. (2008). Learning and games. In K. Salen (Ed.), *The ecology of games: Connecting youth, games, and learning* (pp. 21–40). Cambridge, MA: The MIT Press. doi: 10.1162/dmal.9780262693646.021

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*, 1–18. doi: 10.1080/15366367.2013.783752

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education.* (Primary Review Research Survey 3/4). United Kingdom: University of Cambridge.

Holt, P., Dubs, S., Jones, M., & Greer, J. E. (1994). The state of student modeling. In J. E. Greer & G. I. McCalla (Eds.), *Student modeling: The key to individual knowledge-based instruction* (pp. 3–35). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-662-03037-0_1

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment*, *18*, 182–207. doi: 10.1080/10627197.2013.814517

Mellenberg, G. J. (2011). *A conceptual introduction to psychometrics*. Den Haag, the Netherlands: Eleven International.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment.* Redwood City, CA: GlassLab Research, Institute of Play.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62. doi: 10.1207/S15366359MEA0101_02

Novak, E., Johnson, T. E., Tenenbaum, G., & Shute, V. J. (2014). Effects of an instructional gaming characteristic on learning effectiveness, efficiency, and engagement: Using a storyline for teaching basic statistical skills. *Interactive Learning Environments*. doi: 10.1080/10494820.2014.881393

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, *4*(4), 251–277. doi: 10.1007/BF00168958

Paramythis, A., & Loidl-Reisinger, S. (2004). Adaptive learning environment and e-learning standards. *Electronic Journal of e-Learning*, *2*(1), 181–194.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

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

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, *8*(4), 1–48.

Sanders, P. (2013). Het doel van toetsen [The purpose of testing]. In P. Sanders (Ed.), *Toetsen op school* [Testing at schools] (pp. 9–20). Arnhem, the Netherlands: Stichting Cito Instituut voor Toetsontwikkeling.

Scheerens, J., Glas, C. A. W., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring*. London, United Kingdom: Taylor & Francis.

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which

data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*, 482–496. doi: 10.1016/j.tate.2009.06.007

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*(1), 105–123.

Shepard, L. A. (2005, October). *Formative assessment: Caveat emptor*. Paper presented at the ETS Invitational Conference, The Future of Assessment: Shaping Teaching and Learning, New York, NY.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 43–58). New York, NY: Springer Science+Business.

Shute, V. J., & Ventura, M. (2014). *Stealth assessment: Measuring and supporting learning in video games.* Cambridge, MA: Massachusetts Institute of Technology.

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., & Reckase, M. D. (1994, April). *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.

Stobart, G. (2008). *Testing times: The uses and abuses of testing*. London, United Kingdom: Routledge.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2013). Data-based decision making, assessment for learning, and diagnostic testing in formative assessment. In F. M. Van der Kleij (Ed.), *Computer-based feedback in formative assessment* (pp. 155–169). Unpublished doctoral dissertation, Twente University, Enschede, the Netherlands.

Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer. doi: 10.1007/0.387.29054.0

Van der Linden, W. J., & Glas, C. A. W. (2010). Preface. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. v–vii). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8

Van Groen, M. M., & Eggen, T. J. H. M. (2014). *Multidimensional computerized adaptive testing for classifying examinees on tests with between-dimensionality*. Manuscript submitted for publication.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item selection

methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*, *38*, 187–200. doi: 10.1177/0146621613509723

Wainess, R., Koenig, A., & Kerr, D. (2011). *Aligning instruction and assessment with game and simulation design* (CRESST report 780). Los Angeles, CA: University of California.

Wauters, K. (2012). *Adaptive item sequencing in item-based learning environments*. Unpublished doctoral dissertation, KU Leuven, Belgium.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*, 549–562. doi: 10.1111/j.1365-2729.2010.00368.x

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14. doi: 10.1016/j.stueduc.2011.03.001

Wiliam, D. (2013). How is testing supposed to improve schooling? Some reflections. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 55–59. doi: 10 .1080/15366367.2013.784165

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer. doi: 10.1007/978-0-387-85461-8\18

Chapter 7
Epilogue

Computerized adaptive tests (CATs) were developed to obtain efficient estimates of examinees' abilities, but Lewis and Sheehan (1990), Reckase (1983), Spray and Reckase (1994), and Weiss and Kingsbury (1984) showed that CATs can be used for classification testing (Eggen & Straetmans, 2000). In computerized classification tests (CCTs), the goal is not to obtain an estimate, but to classify the examinee into one of multiple categories. A CCT attempts to find a balance between the test length and the level of confidence in the accuracy of the decision (Bartroff, Finkelman, & Lai, 2008). Testing stops when enough evidence is available to make a decision.

Computerized classification testing requires a method that selects the items and a method that decides whether testing can be stopped and which decision is made. Both methods influence the accuracy and efficiency of a CCT (Thompson, 2009). Several classification methods are available for unidimensional classification testing, including the method based on Wald's (1947/1973) sequential probability ratio test (SPRT; Reckase, 1983) and the method that stops testing if the confidence interval surrounding the ability estimate no longer includes the cutoff points (Kingsbury & Weiss, 1979).

A large variety of item selection methods is available for unidimensional classification testing (Eggen, 1999; Thompson, 2009). These include maximization of information at the ability estimate or at the cutoff point. If multiple cutoff points are specified, alternative methods are available (Eggen & Straetmans, 2000; Spray, 1993; Wouda & Eggen, 2009).

A wide range of classification and item selection methods are available for CAT in which constructs are modeled with unidimensional item response theory (IRT). For multidimensional IRT, however, knowledge about CAT for estimating ability is still building. Literature about CAT to make classification decisions was almost entirely absent when the research for this thesis started.

Four research questions were formulated in the introduction of this thesis. Some answers to these questions will be provided next. After discussing the research questions, some remarks will be made about the research in this thesis. The last section of the epilogue discusses a few directions for further research.

7.1 Discussion of the Research Questions

The first research question concerned adaptive item selection methods for tests with multiple cutoff points:

How can item selection by maximizing information at the cutoff point(s) and maximization at the ability estimate be combined to obtain accurate and efficient classification decisions, and tailoring item selection to the student's ability?

Current item selection methods can be divided into two types: adaptive and non-adaptive methods. The first group of methods adapts the selection of the items to the student's ability. An example is the method that maximizes the information in the test at the student's current ability estimate. The second group of methods does not tailor the test to the student's ability, but uses a different criterion for item selection. The methods that maximize information at the cutoff point or some combination of the cutoff points are examples of such methods. These methods tend to be more efficient when making classification decisions.

Four item selection methods were developed in Chapter 2 that form a hybrid between the types of item selection methods. These methods can be used when a classification decision is made into one of more than two categories. The new item selection methods use multiple objective approaches to combine the efficiency of the methods that maximize information at the cutoff points with the tailoring of the test to the examinee's ability by the method that maximizes information at the current ability estimate.

The efficiency and accuracy of the four new and four existing methods were investigated using simulations with a simulated and an existing item pool. Random item selection was found to be the least efficient and accurate method. The differences in the average test length and the proportion of correct decisions were small for the other methods. The weighting method was found to be the most accurate method in the majority of the simulations and it also required fewer items than the other new methods. When compared to existing methods, this method was as effective and efficient as the majority of the other methods. The weighting

method was also applied to multidimensional item response theory (MIRT) in Chapters 3 and 5. The findings of Chapter 2 imply that item selection can be tailored to the student's ability without administering many additional items.

One conclusion results from the simulations in Chapter 2, but similar findings were reported in Chapters 3, 4, and 5: differences in the efficiency and the accuracy between item selection methods are very small at the population level, except for random item selection. Nevertheless, in a practical setting, item selection should be investigated at the level of individual students. Differences in the test length and the classifications at different points on the ability scale were not addressed.

The second research question concerns making classification decisions in the context of multidimensional item response theory:

How can multidimensional classification decisions be made on all dimensions simultaneously for tests with between- and within-dimensionality?

Spray, Abdel-Fattah, Huang, and Lau (1997) concluded that the SPRT procedure could not be easily adapted to make multidimensional classification decisions. They investigated the possibility of using the SPRT in the context of what is nowadays called within-dimensionality.

However, in the context of between-dimensionality, both the SPRT and the confidence interval method can be used to make classification decisions. Seitz and Frey (2013a, 2013b) showed that it was possible to apply these methods to tests with between-dimensionality. In their methods, decisions can be made for each dimension using unidimensional classification methods.

When educational assessments are used to determine whether a student masters a topic or not, a decision is required on the entire test, instead of on each of the underlying subscales. Making such decisions was addressed in Chapters 3 and 5.

In Chapter 3, it was shown that decisions can be made on all dimensions simultaneously. This implies that it is possible to determine whether a student masters a topic or not. The same method can also be used to decide at which of the multiple levels a student performs.

A second classification method, Kingsbury and Weiss's (1979) confidence interval method was extended in Chapter 5 to make decisions on tests with between-dimensionality. The ability estimates based on the subscales have to be combined first before a classification decision can be made. After combining the estimates, the confidence interval around the combined estimate can be used with simple decision rules to classify students.

Although Spray et al. (1997) concluded that the SPRT could not be extended straightforwardly to multidimensional CCT, the SPRT was applied to classification testing in case of within-dimensionality in Chapter 4. The multidimensional space has to be reduced to a unidimensional line, the reference composite (Wang, 1985, 1986, as described in Reckase, 2009), before the SPRT can be used. The SPRT then uses the reference composite to make the decisions.

In Chapter 5, the reference composite was used to make classification decisions with Kingsbury and Weiss's (1979) confidence interval method on tests with within-dimensionality. The confidence interval surrounding the proficiency, as estimated on the reference composite, is used to make the classification decision.

The third research question dealt with selecting the items in multidimensional classification testing:

How can items be selected in tests with between- and within-dimensionality so that accurate and efficient decisions can be made?

A large variety of item selection methods exists for MCAT. Unfortunately, these methods were developed to obtain an efficient and precise ability estimate. Obviously, this is caused by the fact that until recently, no classification methods were available in the context of MIRT. Some of the item selection methods that were developed to obtain an ability estimate can be used when the goal of the multidimensional CAT (MCAT) is to make a classification decision.

It was concluded in Chapter 3 that unidimensional item selection methods can be used to select the items in a test with between-dimensionality. It is only possible to use these methods if the non-diagonal elements of the information matrix are zero. If (weighted) maximum likelihood estimates are used, these elements are zero and unidimensional methods can be used. However, if an estimate is used that results in non-zero non-diagonal elements, multidimensional methods have to be used to select the items.

Items were selected in Chapters 3 and 5 for tests with between-dimensionality using unidimensional methods that maximize information at the current ability estimate or at a weighted combination of the cutoff points (the weighting method).

One important remark has to be made regarding item selection for between-dimensional tests: the test developer should ensure that items are selected for all subscales, because if not, classification decisions will be contrived. A simple content control method, such as Kingsbury and Zara's (1989)-approach, can be used to ensure this.

In Chapter 4, items were selected for tests with within-dimensionality. Segall's (1996) item selection method that maximizes the determinant of the information matrix at the current ability estimate was used. Since simulees were classified into one of two categories, Segall's (1996) method could be adapted to maximize the determinant of the information matrix at the cutoff point or at the projected ability estimate.

In Chapter 5, multiple cutoff points were set for the simulations with within-dimensionality. This implied that the method that maximized the determinant of the information matrix at the cutoff point could not be used. Thus a method that takes multiple cutoff points into account had to be used. The weighting method of Chapter 2 was adapted to select items that had the largest determinant of the information matrix at a weighted combination of the cutoff points.

The fourth research question concerned the context of CCT and the design of systems for administering CCTS:

In which contexts can computerized classification testing be used, and how should the test be designed in those contexts?

Four assessment approaches were explored in Chapter 6: formative assessment, formative evaluation, summative assessment, and summative evaluation. Assessment focuses on the individual student or class, whereas evaluation focuses on the school or program (Harlen, 2007). If the focus is on enhancing learning, it is formative assessment. If the focus is on making a decision after the learning process is finished, it is summative assessment.

Computerized classification tests can be used for formative assessment if the level of decision making is fine-grained enough for the testing goal. CCTs can be used to create a knowledge profile of the student. Teachers can use this profile to detect weak areas in the student's knowledge and adapt instruction accordingly. CCTs can also provide information for developing school policy and can be used for innovation at the program or school level. Individual test results have to be aggregated to the correct level in order to detect content areas for improvement.

Summative assessments are used to determine whether a student masters a content domain (Van der Kleij, 2013). A CCT can provide this information because making judgments about mastery is the main goal of a computerized classification test. If judgments have to be made about schools or educational programs, as in summative evaluation, the results have to be aggregated to the required level.

CCTs were originally developed to make judgments which implies that the summative functions of assessments were supported. Nowadays, formative functions are also supported but the additional requirements for formative functions have to be dealt with. One such solution was the development of knowledge profiles that can be used to distinguish the student's weaker knowledge areas.

In the second part of the question, the focus is on the design of systems for administering CCTs. Wauters (2012) and Nwana (1990) distinguished four different modules for assessment systems: student, tutor, knowledge, and user interface module. A short overview is given here for CCT and in Chapter 6 for other types of digital tests.

The student module contains all information concerning the student (Wauters, 2012). For a CCT, the module monitors which items were administered to the student, and if necessary, also the current ability estimate. The latter can be used to select the items or to make the classification decision.

The tutor module regulates the instructional interactions with the student (Nwana, 1990). In a CCT, the module selects the items. This implies that the module interacts with the student and knowledge modules to select the items.

The items are stored within the knowledge module. This includes item information like item parameters, content specifications, but also the item content. The knowledge module contains the calibrated item bank including the item text, answer alternatives, and scoring keys.

The user interface module is the part of the test that the student sees and interacts with (Parshall, Spray, Kalohn, & Davey, 2002). The user interface should be easy to use and easy to understand, because otherwise, the student's reactions to taking the test and his or her performance may be affected (Parshall et al., 2002).

7.2 Further Remarks

As a result of the research for this thesis, some remarks can be made. First, some general remarks will be made about the studies and computerized classification testing. Then, remarks will be made about the research for Chapters 2 to 6.

7.2.1 General Remarks About the Research in this Thesis

One of the things that was noted about the research into computerized adaptive (classification) testing and psychometrics is that the majority of research starts with a psychometric model for which data are then simulated or gathered. The

other way around, selecting a measurement problem and then developing a psychometric model for solving the measurement problem, is less common. By using real item pools and data, psychometric researchers can be challenged to develop models that can be used in actual testing.

The majority of the research into computerized classification testing, and computerized adaptive testing in general, is based on simulation studies. Typically, one or more item pools are simulated that are then used for drawing conclusions about the characteristics, accuracy, and efficiency of classification and item selection methods. As an alternative, operational item pools or datasets can be used. Although the use of operational item pools or datasets sometimes places limits on the research that is possible, it enables the investigation of characteristics, accuracy, and efficiency in a more realistic and therefore often more valuable setting.

One of the topics that must be addressed when developing an educational CCT is content validity. The administered items must be representative of the intended construct. Content control methods can be used to ensure that the intended construct is measured. Nevertheless, the impact of content control is often not included in studies about CAT and CCT.

Another topic that should be addressed more often when CCTs and tests in general are developed is the selection of an item response theory model. Depending on the model, different possibilities with the model and the test results exist. Nevertheless, the majority of educational tests are still based on unidimensional item response theory models. In the past, this made a great deal of sense due to the lack of software for multidimensional item response theory. However, over the last couple of years, following the publication of Reckase (2009), a lot of research has been done and knowledge will be able to grow even more if test developers start applying MIRT more often to actual test development problems.

If a decision is made with the SPRT, three values are set: α, β, and δ. In the simulations in Chapters 2 to 5, decisions were made with equal values for α and β, but also with identical values for all cutoff points. α and β can be set to values that represent the loss caused by making a type I or a type II error. Different values can also be set for cutoff points for the same decision when a classification decision is made into one of three or more levels. In the multidimensional case, the values can also be varied over dimensions. In the simulation studies, the indifference regions were set to be symmetric. It is possible to specify asymmetric indifference regions or to specify different values for δ for different classification decisions or dimensions. In the simulations with between-dimensionality, δ was varied over

decisions. In the simulations with unidimensionality or within-dimensionality, δ was equal for all SPRTs. Unfortunately, research about decision criteria for δ is not available yet.

In this thesis, the γ for the confidence interval method was also set to be equal over dimensions and decisions. If needed γ can be varied over decisions and dimensions.

A well-known characteristic of the SPRT is that accuracy and efficiency depend on the position of the student on the ability scale relative to the position of the cutoff point. The studies in Chapters 2 to 5 addressed primarily the average test length and the proportion of consistent decisions for the entire set of simulees. As Van Groen and Verschoor (2010) showed for the unidimensional SPRT, the average test length of simulees whose ability is close to the cutoff point is much higher than for those whose ability is further away from the cutoff point. The same result was presented in Chapter 4 for within-dimensionality. The proportion of consistent decisions is also influenced by the distance between the ability and the cutoff point. If the difference becomes very small, the proportion goes to 0.5 (Eggen & Straetmans, 2000; Van Groen & Verschoor, 2010). Chapter 4 showed that in the within-dimensional case, this applies to all ability combinations close to the cutoff point on the reference composite. Test developers should consider these problems with the SPRT when applying the method to high-stakes testing.

Differences in the average test length and the proportion of consistent decisions were inspected in Chapters 2 to 5 based on ideas about the relevance of the size of the difference. An alternative approach would be to use formal statistical tests; it would then have been clear whether differences are significant or not. Effect size could also have been used to determine whether a difference in test length or consistency was relevant.

The simulation studies in this thesis showed that the differences in the average test length and the proportion of consistent decisions between adaptive and non-adaptive item selection methods tended to be negligible. This implies that, in contrast to the popular idea that maximization at the cutoff point results in the shortest and most accurate tests, adaptive item selection methods can be used to tailor item selection to the student's ability. The advantages of tailoring item selection are, just as in CATs for estimating ability, that all students are challenged at their own level, a larger set of the items in the item pool are used, and a much larger variety of test forms within a group of students.

One remark can be made about the performance of random item selection in the simulation studies for Chapters 2 to 5. The method required more items in the majority of the simulations and also resulted in the least accurate decisions. The choice for an information-based method results on average in shorter tests with more accurate decisions.

7.2.2 Remarks About Chapter 2: Item Selection Methods for Multiple Level Classifications

Two remarks are made here about Chapter 2. The objective functions that were developed for Chapter 2 include all cutoff points that are set for the decision in addition to the current ability estimate. One could wonder whether it is necessary to include all cutoff points for all students for all test lengths. If at some point during the testing process it is determined that a student's ability is higher than the second cutoff point, is it then necessary to include the first cutoff point in the objective function? Also, if it is save to assume that ability is probably not higher than the second cutoff point, is it then necessary to include the third cutoff point in the objective function? An alternative approach could be to use adaptive selection to select which cutoff points are included in the objective function.

The second remark that could be made about the simulations of Chapter 2 is that the results of the simulations with the exposure method are probably less influenced by the exposure control than is typically found in actual testing situations. In such situations, test developers also want to ensure that groups of students with a relatively similar ability get different test forms. The modified Sympson and Hetter (1985) method, which was used by Eggen and Straetmans (2000), does not prevent that such groups of students receive very similar tests. An alternative approach would be to implement an exposure approach that takes the exposure into account per range of the ability scale.

7.2.3 Remarks About Chapter 3: Between-Dimensional Classification Testing

The first remark regarding Chapter 3 is that it would have been interesting if a flexible test length had been specified in the simulations. By specifying a flexible test length, it can be investigated which test length is required for maintaining the consistency of the decisions. Specifying a fixed test length that is lower than the number of relevant items in the item pool enables the comparison of the proportion of consistent decisions for different item selection methods; however, investigation of the required number of items can be interesting for practitioners.

The choice to use post-hoc simulations in which existing data are used to compare item selection methods and different settings for the SPRT has one major problem attached to it: no true classification is available. This implies that no baseline classification can be determined. The End of Primary School Test (Cito, 2012) gives a recommendation about the level of secondary education. As a solution to this problem, the decisions on the entire test were compared with this recommendation. Obviously, both methods contribute to classification inconsistencies. Nevertheless, it would be interesting to repeat the simulations with the item pool with generated simulees because that would allow an investigation into classification accuracy.

The third remark about Chapter 3 concerns the limited knowledge about multidimensional item response theory in practical applications. In the preparation phase of the post-hoc simulations, it became apparent that a lot of practical points concerning the application of MIRT in CAT are not described. To give an example for tests with between-dimensionality, the fact that information is only available if items regarding that domain have been administered is not described in the literature. Seitz and Frey (2013a) selected items using maximization of the determinant of the information matrix. This method suggests that information has been gathered at all dimensions, but in fact, information has only been gathered at dimensions for which items have been selected. Such practical points are well described for unidimensional IRT, but not for multidimensional IRT.

7.2.4 Remarks About Chapter 4: Within-Dimensional Classification Testing

Multiple remarks can be made regarding Chapter 4. In the simulation study, only one dataset was used to investigate the efficiency and the consistency of the classification and item selection methods. If another dataset had been used, different results would have been possible. The study also contained only one cutoff point. No conclusions can be drawn for tests with within-dimensionality in which a student has to be classified into one of more than two levels. The cutoff point that was used was set in the center of the ability distribution. In actual testing programs, the cutoff can be at a different point in the ability distribution; this implies that the reported results cannot be generalized to other situations.

The ACT item pool that was used was previously calibrated with a two-dimensional model. No simulation results are available for tests modeled with more dimensions. This implies that no knowledge is available about the performance of the methods on tests with more than two dimensions.

One problem became apparent when simulations were conducted with multidimensional data and the classifications with the unidimensional and multidimensional SPRT were compared. It was expected that the multidimensional SPRT would provide more consistent and more efficient classifications. If items were selected at random, the expected results were found. However, if a method was used that incorporates the current ability estimate, unexpected results were found. The proportion of consistent decisions of the unidimensional SPRT was sometimes slightly higher than the consistency of the multidimensional decisions. At the time, it was thought that the results were caused by differences in the specifications of the SPRTs because δ and the cutoff points had to be matched between the unidimensional and multidimensional simulations. However, when the simulation study for Chapter 5 was conducted, it became apparent that the results could also have been caused by inaccuracies or instabilities in the ability estimates.

In the last part of the simulation study of Chapter 4, classifications by the unidimensional and the multidimensional SPRT were compared. As a research topic, this comparison makes sense. When developing educational tests, the test developers should first determine which model; a unidimensional or a multidimensional model, fits the data. Based on differences in model fit, a model can be selected for the test application. After the model has been selected, the performance of the item selection and classification methods for the selected model can be investigated. Thus, investigating the performance of unidimensional and multidimensional item selection and classification methods for the same datasets does not make sense from a model selection perspective.

A remark was already made in the conclusion of Chapter 4 about the use of the classification method with a non-orthogonal calibration. The SPRT and item selection methods can be used if a non-orthogonal model has been used in the calibration phase. Nevertheless, test developers should select an orthogonal or non-orthogonal model based on model fit criteria and then investigate the performance of the classification and item selection methods; however, they should not select the model based on differences in the performance of the classification methods.

7.2.5 Remarks About Chapter 5: Multidimensional Classification Testing

In Chapter 5, weighted maximum likelihood estimates were used in the simulations for between- and within-dimensionality. The estimates based on the data that were generated with a between-dimensional model appeared to be precise, but the estimates for the data with a within-dimensional model appeared to be less precise.

Estimates sometimes diverged from their true values; and for other examinees, compensation between the estimates was found. It was determined that the deviations increased if more dimensions had to be estimated or the number of items decreased. Almost all studies about multidimensional computerized adaptive testing use a Bayesian estimator. Here, a weighted maximum likelihood estimator was used. Although the latter has the advantage that the estimates are not influenced by the choice of a prior distribution, Bayesian estimators have the advantage that more knowledge is available about their characteristics. Unfortunately, only a few very small comparison studies are available. This implies that more research is needed into the characteristics of (weighted) maximum likelihood estimation; in addition, a substantial comparison study would be welcome.

The consistency and efficiency of the SPRT and the confidence interval method could be compared based on Chapter 5 for three item selection methods for two types of multidimensionality. One item pool was investigated in Chapter 5 with a limited set of settings for the classification methods. The simulations were too limited to enable conclusions about which method outperforms the other method. When comparing the results, researchers should ensure that either the test length or the proportion of consistent decisions is equal for both methods. By fixing test length at a lower length than the number of relevant items in the item pool, one can compare the proportion of consistent decisions. It is important that the classifications with which the decisions are compared for the proportion of consistent decisions are the same. Although that seems logical, this was found to be challenging for decisions that were based on multiple dimensions. Fixing the proportion of consistent decisions can be achieved by matching, but this requires simulations with a large range of settings for the classification methods. The average test length can then be compared, in addition to investigation of differences in test length and local classification consistency at different points on the ability scale. The most interesting solution to make a comparison between the methods would be based on a mathematical relationship between the two methods; however, this mathematical relationship has not been determined yet.

In Chapter 5, the classifications based on a between- and a within-dimensional model could not be compared because different models had been used in the simulation studies. It could be interesting to compare those classifications from a research point of view. Hartig and Höhler (2008) investigated the differences in ability estimates between the two models, but a similar study that is focused

on classification testing would be interesting. In test development, however, test developers should determine which model fits the data before the performance of the classification methods can be investigated.

In the simulations for between-dimensionality in Chapter 5, one sequence of dimensions was used. However, another sequence would have been possible. The first item was selected from the first dimension, the second item from the second dimension, and so on. In theory, this could have influenced the number of items that were required to make a decision. The probability that the decision itself was influenced is very small, but small effects on the test length are not ruled out. The inclusion of testlet items can potentially increase the sequencing effect. In another simulation study, the effects of dimension sequencing could be investigated.

7.2.6 Remarks About Chapter 6: Assessment Approaches and Types of Digital Assessments

Although many remarks are possible about Chapter 6, the discussion here is limited to two remarks about the implications for computerized classification testing. The first remark concerns the implementation of the dimensions of adaptivity, as defined by Wauters (2012), for CCTs. Some dimensions of adaptivity have been a research topic for many years, but other dimensions and aspects of dimensions are still open for investigation. For example, curriculum sequencing is an aspect that has received little attention. Curriculum sequencing can be useful if a CCT is to be used for formative assessment. The combination of knowledge profiles with selecting the most relevant topics from the item bank could be a contributing factor in enhancing the learning from the assessment.

The second remark concerns the depth of the discussion of assessment approaches and test design for computerized classification testing. It would be interesting to explore these topics further and it would be useful for the practice of developing educational tests if more practical guidelines would be developed for the use of CCT for different assessment approaches and the design of the test modules for those approaches. Currently, limited practical guidelines are available for developing a CCT (Bartram & Hambleton, 2006; Parshall et al., 2002; Thompson, 2007), but these concern design or psychometric guidelines instead of focusing on the educational implications of the choices that are being made.

7.3 Future Directions

Several possible lines of further research have already been described in this epilogue or in the discussion sections of earlier chapters, but some additional directions are discussed next.

In the current thesis, two methods were included to make the decisions: the SPRT and the confidence interval method. Other methods, such as the stochastically curtailed SPRT (Finkelman, 2008) and the generalized likelihood ratio (Thompson, 2011), were not included. This implies that these methods currently are not available to make multidimensional classification decisions. It would be interesting to investigate whether these methods can also be applied to MCCT.

For some tests, the application of IRT models is not possible or is considered to be too complicated. As an alternative, curtailed and stochastic curtailed methods were developed that are based on sumscores (Finkelman, Smits, Kim, & Riley, 2012). It would be interesting to investigate whether these methods can be applied to educational assessments and how they perform in comparison to CCT. Finkelman et al. (2012) compared the methods with polytomous IRT models, but it is not known yet whether the same results hold for dichotomous IRT models.

Another interesting possibility would be to apply tree-based item response methods (De Boeck & Partchev, 2012) to make classification decisions in the context of educational assessment. A possibility to explore would be the development of adaptive tests, but also to compare tree-based item response methods with item response theory-based classification testing.

One major concern during test development is that items function the same for students with similar ability (Holland & Wainer, 1994); thus, test developers investigate differential item functioning (DIF). If it is determined that DIF is not present, this supports the construct validity of the test (Van Groen, Ten Klooster, Taal, Van de Laar, & Glas, 2010). It would be interesting to investigate whether decisions that were made by a CCT are sensitive to DIF.

Computerized classification testing requires the specification of cutoff points, but how these cutoff points were specified received little attention in this thesis. The cutoff points in Chapters 3 and 5 were determined using existing cutoff points on the scale score that was already available. Standard setting methods (Cizek & Bunch, 2007) are a well developed area of research, but their application to CAT to obtain ability estimates or to make classification decisions has received little attention. It would be an interesting topic for further study.

Typically, the differences in classification consistency and efficiency between different item selection methods, excluding selecting the items at random, tend to be small. The newly developed expected log-likelihood ratio (ELR) method by Nydick (2014) resulted in a lower average test length than maximization of information at the cutoff point or Kullback-Leibler divergence (Chang & Ying, 1996) in simulations with a three-parameter logistic IRT model. The method is based on optimizing the expected SPRT given the current ability estimate. It would be interesting to explore whether these findings can be replicated with a two-parameter IRT model and also to investigate whether it is possible to extend the method to the generalized likelihood ratio method (Thompson, 2011). The method can, in its current form, only be used when students are classified into one of two categories for unidimensional IRT. It is an open question whether the method can be expanded to make classification decisions into one of more than two categories and whether the model can be applied to multidimensional IRT.

The last discussed direction for future research that is discussed here concerns the use of other multidimensional item response theory models than the multidimensional two-parameter logistic model. Other models, such as non-compensatory or three-parameter models, could also be used to make classification decisions. It would be interesting to see whether the developed classification and item selection methods have to be adapted and how they perform with these models.

References

Bartram, D., & Hambleton, R. K. (Eds.). (2006). *Computer-based testing and the internet: Issues and advances*. Chichester, United Kingdom: John Wiley & Sons.

Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*, *73*, 473–486. doi: 10.1007/s11336-007-9053-9

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229. doi: 10.1177/014662169602000303

Cito. (2012). *Eindtoets Basisonderwijs 2012* [End of Primary School Test 2012]. Arnhem, the Netherlands: Cito.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. doi: 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*, *60*, 713–734. doi: 10.1177/00131640021970862

Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442–463. doi: 10.3102/1076998607308573

Finkelman, M. D., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement*, *36*, 632–658. doi: 10.1177/0146621612451647

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education.* (Primary Review Research Survey 3/4). United Kingdom: University of Cambridge.

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*, *216*, 89–101. doi: 10.1027/0044-3409.216.2.89

Holland, P. W., & Wainer, H. (1994). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategie for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, *2*, 359–375. doi: 10.1207/s15324818ame0204_6

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 367–386. doi: 10.1177/014662169001400404

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, *4*(4), 251–277. doi: 10.1007/BF00168958

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics*, *39*, 203–230. doi: 10.3102/1076998614524824

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

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. doi: 10.1007/978-0-387-89976-3

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi: 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, April). *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.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *In: Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, *12*(1), 1–13.

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

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation*, *16*(4), 1–7.

Van der Kleij, F. M. (2013). *Computer-based feedback in formative assessment*. Unpublished doctoral dissertation, Twente University, Enschede, the Netherlands.

Van Groen, M. M., Ten Klooster, P. M., Taal, E., Van de Laar, M. A. F. J., & Glas, C. A. W. (2010). Application of the health assessment questionnaire disability index to various rheumatic diseases. *Quality of Life Research*, *19*, 1255-1263. doi: 10.1007/s11136-010-9690-9

Van Groen, M. M., & Verschoor, A. J. (2010, June). *Using the sequential probability ratio test when items and respondents are mismatched*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.

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: University of Iowa.

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.

Wauters, K. (2012). *Adaptive item sequencing in item-based learning environments*. Unpublished doctoral dissertation, KU Leuven, Belgium.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

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*.

Summary

Computerized adaptive tests (CATs) were developed to obtain an efficient estimate of a student's ability, but they can also be used to classify an examinee into one of multiple levels. The advantages of using computerized classification tests (CCTs) are that item selection can also be tailored to the student's ability, but tests are often shorter than other CATs, and accuracy can be maintained.

Unidimensional CAT with multiple cutoff points was investigated in **Chapter 2**. A CCT requires a method to select the items. Besides random selection, methods use some statistical criterion with or without the current ability estimate for selecting the items. A second method decides whether testing can be stopped and classifies the student into one of two or more mutually exclusive categories. Item selection for unidimensional classification testing with multiple cutoff points was investigated in Chapter 2. Important advantages of dynamically assembling tests during test administration are that item selection can be tailored to the individual student's ability, a larger part of the item pool is used, and a much larger spread in test forms within groups of students is realized. The majority of currently available item selection methods in CCT maximize information at one point on the ability scale, but in a test with multiple cutoff points, the selection methods can take all these points simultaneously into account. If one objective is specified for each cutoff point, the objectives can be combined into one optimization function using multiple objective approaches. Several methods were developed in Chapter 2 that incorporate the measurement efficiency of maximizing information at the cutoff points with the tailoring of the method that maximizes information at the current ability estimate. Simulation studies were used to compare the efficiency and accuracy of eight selection methods using a classification method based on the sequential probability ratio test (SPRT). Small differences were found in accuracy and efficiency between different methods depending on the item pool and the settings of the classification method. The size of the indifference region had little influence on accuracy but considerable influence on efficiency. Content and exposure control had little influence on accuracy and efficiency.

Unidimensional classification decisions were made in Chapter 2, but multidimensional decisions were addressed in Chapters 3 to 5. Several methods are available to make classification decisions for constructs modeled using a unidimensional item response theory model. These methods stop testing when enough confidence has been reached to make the decision. However, few classification methods are available for multidimensional constructs. A classification method was presented in **Chapter 3** for adaptive classification testing with a multidimensional item response theory model in which items are intended to measure one trait each, e.g., between-dimensionality. In the case of between-dimensionality, an existing method can be applied to classification testing. This method makes classification decisions per dimension using the SPRT. The method was extended to include decision making based on all the dimensions in the test simultaneously, on several dimensions, and on subsets of items. A measure is presented that provides information about the support for the decisions. Items were selected that provided the most information at the current ability estimate or at a weighted combination of the cutoff points. The latter method was developed in Chapter 2 for unidimensional CCT. An empirical example illustrates the discussed methods.

Chapter 3 focused on making classification decisions on tests in which each item is intended to measure one ability. In contrast, in **Chapter 4** items were intended to measure multiple abilities. A popular unidimensional classification method, which was also used in Chapter 2, was applied to tests with within-dimensionality. The reference composite was used in conjunction with the sequential probability ratio test to make decisions and decide whether testing could be stopped before the maximum test length was reached. Item selection methods were provided that maximize the determinant of the information matrix at the cutoff point or the projected ability estimate. A simulation study illustrated the efficiency and effectiveness of the classification method for different settings of the SPRT. Simulations were run with the two item selection methods, random item selection, and maximization of the determinant of the information matrix at the ability estimate in the multidimensional space. The study also showed that the multidimensional SPRT had the same characteristics as the unidimensional SPRT and outperformed the unidimensional SPRT when applied to multidimensional data with random item selection.

Two classification methods were applied in **Chapter 5** to tests with between- and within-dimensionality. The SPRT was applied in Chapters 3 and 4 to tests with between- and within-dimensionality. A second popular unidimensional classification method uses the confidence interval surrounding the ability estimates. This method also exists for between-dimensionality, but can make only one decision per dimension. This method was extended in Chapter 5 to make decisions on the entire multidimensional test, on combinations of the test dimensions, and on subsets of the items that load on a specific dimension. This method cannot be used for tests with within-dimensionality, but as shown in Chapter 4, decisions can also be made with the reference composite. The confidence region method was adapted in Chapter 5 for tests with within-dimensionality. The current ability estimate was projected at the reference composite, and testing could be stopped if the confidence regions surrounding the projected ability estimates no longer included a cutoff point. Items were selected in Chapter 5 based on maximization of the (determinant of the) information matrix at the current ability estimate or based on the weighting method of Chapter 2. This method was adapted in Chapter 5 for item selection in the case of within-dimensionality. Existing item banks were used to investigate the performance of the extended classification methods for between-dimensionality and the classification methods for within-dimensionality that were developed in Chapters 4 and 5. The required test length and the consistency of the classifications were investigated for multiple item selection methods and for several settings of the classification methods.

The context of computerized classification testing was explored in **Chapter 6**. The usability of computerized classification tests for formative assessment was investigated, formative evaluation, summative assessment, and summative evaluation. In addition to CCT, six other types of digital assessments were discussed: linear testing, automatically generated tests, computerized adaptive testing for estimating ability, adaptive learning environments, educational simulations, and educational games. The types of tests vary in the design of their modules (student, tutor, knowledge, and user interface), in the way testing is adapted to the individual student's current knowledge level, and in the possibility of using the test for different test approaches. It was concluded that tests should be designed in a way that outcomes are sufficiently accurate and that the tests are suitable for the testing goals, have the correct grain size, and have a test length suitable for the testing purpose and testing population.

Samenvatting Adaptieve toetsen voor het nemen van unidimensionele en multidimensionele classificatie beslissingen

Computergestuurde adaptieve toetsen (CAT) werden ontwikkeld om een efficiënte schatting van de vaardigheid van een student te verkrijgen. CAT kan echter ook worden gebruikt om studenten te classificeren in één uit twee of meer niveaus. Voordelen van computergestuurde classificatietoetsen (CCT) zijn dat de item selectie afgestemd kan worden op de vaardigheid van de student, de toetsen veelal korter zijn dan bij andere adaptieve toetsen het geval is en dat de accuraatheid van de classificatiebeslissingen bewaakt kan worden.

Computergestuurde classificatietoetsen vereisen een methode die de items voor de toets selecteert. Naast aselecte trekking zijn er methoden die gebaseerd zijn op een statistisch criterium met of zonder gebruik van de lopende vaardigheidsschatting. Itemselectie voor unidimensionele classificatietoetsen is onderzocht in **hoofdstuk 2**. Een groot voordeel van het dynamisch samenstellen van toetsen gedurende de toetsafname is dat de itemselectie daarbij afgestemd kan worden op de vaardigheid van de individuele leerling, er effectiever gebruik gemaakt wordt van de itembank en er sprake is van een grote spreiding in toetsinhoud binnen groepen leerlingen. Het merendeel van de beschikbare itemselectiemethoden maximaliseert informatie op een specifiek punt op de vaardigheidsschaal. Indien een toets meerdere grenspunten heeft dan zou de itemselectiemethode al deze punten gelijktijdig mee moeten nemen bij het bepalen van het volgende item. Als een doelfunctie is gespecificeerd voor ieder grenspunt, dan kunnen de doelfuncties worden gecombineerd in een optimalisatiefunctie met behulp van bestaande methodieken voor het combineren van meerdere doelfuncties. De itemselectiemethoden die ontwikkeld zijn voor hoofdstuk 2 combineren de meetefficiëntie van de methoden die informatie maximaliseren op de grenspunten met de afstemming op het individu door de methode die informatie maximaliseert op de lopende vaardigheidsschatting. Een andere methode bepaalt wanneer de toets

gestopt kan worden en welke beslissing er genomen mag worden over het niveau van de student. Met behulp van simulaties zijn de efficiëntie en accuraatheid van beslissingen die zijn genomen met behulp van Wald's sequential probability ratio test (SPRT) vergeleken voor acht verschillende itemselectiemethoden. Afhankelijk van de gebruikte itembank en de specificaties voor de classificatiemethoden zijn er kleine verschillen gevonden in accuraatheid en efficiëntie. De omvang van het indifferentiegebied bleek weinig invloed te hebben op de accuraatheid maar had aanzienlijke invloed op de efficiëntie. Controlemechanismen voor het bewaken van de inhoudelijke samenstelling van de toets en het reguleren van het itemgebruik bleken weinig invloed te hebben op de accuraatheid en efficiëntie.

De beslissingen die genomen werden in hoofdstuk 2, hadden betrekking op unidimensionele toetsen. In de hoofdstukken 3 tot en met 5 worden beslissingen genomen voor multidimensionele constructen. In het verleden zijn er diverse methoden ontwikkeld om classificatiebeslissingen te nemen voor constructen die gemodelleerd zijn met unidimensionele itemresponstheorie. Deze methoden stoppen de afname van de toets indien er voldoende vertrouwen bestaat in de te nemen beslissing. Er zijn echter weinig classificatiemethoden beschikbaar voor multidimensionele constructen. In **hoofdstuk 3** is een classificatiemethode ontwikkeld die gebruikt kan worden voor constructen die gemodelleerd zijn met behulp van multidimensionele itemresponstheorie voor items die ieder laden op slechts één dimensie. Er is dan sprake van een simpele dimensionaliteitsstructuur. In dat geval kunnen bestaande methoden worden gebruikt voor classificatietoetsing. De SPRT is één van de bestaande methoden voor unidimensionele toetsen die gebruikt kan worden voor toetsen met een simpele dimensionaliteitsstructuur. Deze methode neemt dan een beslissing voor iedere afzonderlijke dimensie. De methode is in hoofdstuk 3 uitgebreid, zodat er ook beslissingen genomen konden worden op alle dimensies tezamen, op meerdere dimensies, en op subsets van items. Daarnaast is in dit hoofdstuk een maat gepresenteerd die informatie geeft over het vertrouwen in de beslissingen die genomen zijn met behulp van de SPRT. Items werden geselecteerd op basis van maximalisatie van informatie op de lopende vaardigheidsschatting of op een gewogen combinatie van de grenspunten. De laatste methode was al ontwikkeld in hoofdstuk 2 voor unidimensionele CCT. Een voorbeeld met behulp van empirische data illustreerde de besproken methoden.

In hoofdstuk 3 lag de focus op het nemen van classificatiebeslissingen bij toetsen waarin ieder item één vaardigheid meet. In tegenstelling tot hoofdstuk 3, meten items in **hoofdstuk 4** meerdere of alle dimensies per item. Als items meerdere afzonderlijke vaardigheden meten, is er sprake van een complexe dimensionaliteitsstructuur. De SPRT is in hoofdstuk 4 toegepast op toetsen met een complexe dimensionaliteitsstructuur. Er wordt een referentielijn gebruikt om beslissingen te nemen met de SPRT, waarbij de SPRT bepaalt of de toetsing gestopt kan worden. Items werden in de simulaties voor het hoofdstuk geselecteerd middels maximalisatie van de determinant van de informatiematrix op het grenspunt, op de lopende vaardigheidsschatting, of op de vaardigheidsschatting die geprojecteerd was op de referentielijn. De bevindingen zijn vervolgens vergeleken met die voor aselecte itemselectie. De accuraatheid en de efficiëntie van de classificatiemethoden werd onderzocht met behulp van simulaties voor verschillende specificaties voor de SPRT. De simulaties lieten zien dat de multidimensionele SPRT dezelfde kenmerken heeft als de unidimensionele SPRT en dat deze beter functioneerde op multidimensionele data dan de unidimensionele SPRT als items aselect werden gekozen.

In **hoofdstuk 5** werden classificatiebeslissingen genomen voor toetsen met een simpele en een complexe dimensionaliteitsstructuur met behulp van een tweetal classificatiemethoden. De eerste is de SPRT, die in hoofdstukken 3 en 4 ook werd gebruikt. Een andere unidimensionele classificatiemethode gebruikt het betrouwbaarheidsinterval rond de vaardigheidsschatting voor het nemen van beslissingen. Deze methode kan ook gebruikt worden voor toetsen met een simpele dimensionaliteitsstructuur en neemt dan een beslissing per dimensie. Deze methode is in hoofdstuk 5 uitgebreid om beslissingen te nemen op alle dimensies gelijktijdig, op meerdere dimensies en op een deel van de items die laden op een dimensie. De methode kan niet rechtstreeks worden toegepast als er sprake is van een complexe dimensionaliteitsstructuur. Als een referentielijn wordt gebruikt, net als in hoofdstuk 4, is het wel mogelijk om een aangepaste versie van de methode te gebruiken voor toetsen met een complexe dimensionaliteitsstructuur. De geschatte vaardigheid moet dan worden geprojecteerd op de referentielijn. Toetsing wordt dan beëindigd als de betrouwbaarheidsregio rondom de geprojecteerde vaardigheidsschatting niet langer het grenspunt bevat. De items werden in hoofdstuk 5 geselecteerd op basis van maximalisatie van de (determinant van de) informatiematrix op de lopende vaardigheidsschatting of op basis van de wegingsmethode uit hoofdstuk 2. Deze methode is in hoofdstuk 5

aangepast voor toetsen met een complexe dimensionaliteitsstructuur. De prestaties van de bestaande, maar uitgebreide, classificatiemethoden voor toetsen met een simpele dimensionaliteitsstructuur en de nieuwe methoden voor toetsen met een complexe dimensionaliteitsstructuur zijn onderzocht met behulp van bestaande itembanken. De vereiste toetslengte en de consistentie van de beslissingen zijn bekeken voor een drietal itemselectiemethoden en voor diverse specificaties voor de classificatiemethoden.

In **hoofdstuk 6** is de context van CCT bekeken. Er is verkend wat de mogelijkheden zijn om CCT te gebruiken voor formatief assessment, formatieve evaluatie, summatief assessment en summatieve evaluatie. Dit is ook bekeken voor een aantal andere typen digitale toetsen: lineaire toetsen, automatisch samengestelde toetsen, computergestuurde adaptieve toetsen voor het schatten van de vaardigheid, adaptieve leeromgevingen, educatieve simulaties en educatieve spellen. De toetstypen vereisen ieder een eigen ontwerp voor de modulen (student, kennis, begeleiding en gebruikersinterface) waaruit zij zijn opgebouwd. Daarnaast verschillen zij in de wijze waarop toetsing wordt afgestemd op het vaardigheidsniveau van de individuele student en in de mogelijkheden om de toets te gebruiken voor verschillende toetsdoelen. In het hoofdstuk werd geconcludeerd dat toetsen zodanig ontworpen dienen te worden dat de uitkomsten voldoende nauwkeurig zijn, dat ze geschikt zijn om te worden gebruikt voor het gespecificeerde toetsdoel, het juiste detailniveau omvatten en dat ze een lengte hebben die geschikt is voor het toetsdoel en de studentpopulatie.

Dankwoord

Een proefschrift over toetsing is enerzijds een vorm van toetsing, maar dient anderzijds ook bij te dragen aan de kennis over toetsing. Een assessment kan dus meerdere doelen hebben zonder dat deze in alle gevallen tegenstrijdig hoeven te zijn. Naar aanleiding van dit proefschrift zal een classificatiebeslissing worden genomen maar hopelijk is dat de eerste beslissing op basis van de in dit proefschrift gepresenteerde methoden voor het nemen van classificatiebeslissingen.

Net als tijdens de afname van een adaptieve classificatietoets staat in een promotietraject het pad dat gevolgd wordt vooraf niet vast. Het gevolgde pad, de uitkomst en de opbrengsten van een promotie zijn alle afhankelijk van gemaakte keuzes maar ook van de feedback die gegeven wordt bij het maken van die keuzes. Theo, bedankt voor je vele kritische kanttekeningen bij mijn werk, de inhoudelijke discussies en al je scherpe inzichten als ik weer eens iets bedacht had. Bernard, bedankt voor het meedenken over wat uiteindelijk het tweede hoofdstuk werd, en al je feedback en het meedenken in de afrondende fase van het proefschrift.

Mijn werkgever, Cito, wil ik bedanken voor het bieden van de mogelijkheid om onderzoek te doen naar adaptief toetsen en daarbij de vrijheid te geven om zelf dat onderzoek in te richten. Mijn collega's van de afdeling Psychometrisch Onderzoeks- en Kenniscentrum en de promovendi van het RCEC wil ik bedanken voor hun continue bereidheid tot meedenken en het beantwoorden van vragen. Daarnaast wil ik Angela bedanken voor alle discussies over simulaties en het overdragen van kennis over tal van onderwerpen. Marie-Anne, bedankt dat je mijn paranimf wil zijn maar vooral voor de gezelligheid en (inhoudelijke) steun in de afgelopen tijd. Ron en Anke, bedankt voor de gezelligheid en het delen van kennis over methodologie. Matthieu, bedankt voor je tips over LaTeX en R. Floor, Patricia en Rianne, bedankt voor jullie organisatorische hulp waar dat nodig was.

Mijn ouders wil ik bedanken voor hun geduld en interesse de afgelopen jaren, en hun stimulans om me verder te ontwikkelen. Elvira, bedankt dat je mijn paranimf wilt zijn. Tenslotte wil ik Edward bedanken voor al zijn geduld, steun, hulp met het uitvoeren van de simulaties en alle mooie momenten samen.

Curriculum Vitae

Maaike van Groen was born on May 23, 1984 in Woerden, the Netherlands. After completing the Atheneum, she studied Educational Design, Management & Media at Twente University. For her bachelor thesis assignment, she did a quantitative study about staff, organization, and management in the Dutch Secondary Education for the Education Inspectorate. She completed the Twente University's Research Master Social Systems Evaluation & Survey Research in 2009. Her internship concerned the analysis of the Health Assessment Questionnaire using item response theory for patients with rheumatic diseases. Her master thesis assignment was at Cito in Arnhem, the Netherlands, and consisted of a study that was focused on the calibration, item classification, and respondent classification of a computerized adaptive test for geography.

In 2009 she started with a PhD project at Cito in Arnhem, the Netherlands on computerized adaptive testing under the supervision of prof. dr. ir. Theo Eggen and prof. dr. ir. Bernard Veldkamp. The studies presented in this thesis are the result of this PhD project. In addition to doing research, she works as a research scientist at Cito's Psychometric Research Center since 2009. She advises content experts about methodological issues in all phases of the test development process. Her work focuses on item response theory, sampling, and developing research designs for primarily the Dutch primary education.

Research Valorisation

Manuscripts

Van Groen, M. M. (2012). Computerized Classification Testing and Its Relationship to the Testing Goal. In *Psychometrics in Practice at RCEC*. Eggen, T. J. H. M. & Veldkamp, B. P. (Eds.). doi: 10.3990/3.9789036533744.ch11

Van Groen, M. M. & Eggen, T. J. H. M. (2014). *Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Between-Dimensionality.* Manuscript submitted for publication.

Van Groen, M. M., Eggen & T. J. H. M. (2014). *Assessment Approaches and Computer-Based Testing.* Manuscript submitted for publication.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2013). *Multidimensional Computerized Adaptive Testing for Classifying Examinees on Tests with Within-Dimensionality.* Manuscript submitted for publication.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classification of respondents into multiple levels. *Applied Psychological Measurement*, 38, 187-200. doi: 10.1177/0146621613509723

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). *Multidimensional Computerized Adaptive Testing for Classifying Examinees with the SPRT and Confidence Regions.* Manuscript submitted for publication.

Presentations

Van Groen, M. M. & Eggen, T. J. H. M. (2013, July). *Multidimensional Computerized Adaptive Testing for Classifying Examinees.* Paper presented at the International Meeting of the Psychometric Society, Arnhem, The Netherlands.

Van Groen, M. M. & Eggen, T. J. H. M. (2013, October). *Multidimensional Computerized Adaptive Testing for Classifying Examinees on Several Constructs.* Paper presented at the RCEC Workshop on IRT and Educational Measurement, Enschede, The Netherlands.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2011, December). *Item Selection Methods based on Multiple Objective Approaches for Classification of Respondents into Multiple Levels.* Paper presented at the 21st IOPS Winter Conference, Leiden, the Netherlands.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2011, October). *Item Selection Methods Based on Multiple Objective Approaches for Classification of Respondents into Multiple Levels.* Paper presented at the IACAT Conference, Monterey, California, USA.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2011, October). *Item Selection Methods based on Multiple Objective Approaches for Classification of Respondents into Multiple Levels.* Paper presented at the RCEC workshop on IRT and Educational Measurement, Enschede, the Netherlands.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014, June). *Computerized Adaptive Testing for Classifying Examinees Using MIRT for Items that Measure One or Multiple Abilities.* Paper presented at the 29th IOPS Summer Conference, Tilburg, the Netherlands.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014, July). *Computerized Adaptive Testing for Classifying Examinees using MIRT for Items that Measure One or Multiple Abilities*. Paper presented at the International Meeting of the Psychometric Society, Madison, Wisconsin, USA.

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014, October). *Computerized Adaptive Testing for Classifying Examinees for Multidimensional Constructs.* Paper presented at the 2014 Computerized Adaptive Testing Summit, Princeton, New Jersey, USA.

Van Groen, M. M., Eggen, T. J. H. M., & Verschoor, A. J. (2010, May). *Adaptive Classification Tests.* Paper presented at the Onderwijs Research Dagen [Educational Research Days], Enschede, The Netherlands.

Van Groen, M. M., & Verschoor, A. J. (2010, June). *Using the Sequential Probability Ratio Test When Items and Respondents are Mismatched.* Paper presented at the IACAT conference, Arnhem, The Netherlands.