Comparing a Multistage and a Linear Summative Test on Ability Estimate Precision and Classification Accuracy

Researcher:
Michel Lamoré (s1006193), m.lamore@student.utwente.nl

Date: 19-05-2017

Internal supervisor (University of Twente):
Prof. Dr. Bernard Veldkamp, b.p.veldkamp@utwente.nl
Prof. Dr. Theo Eggen, t.j.h.m.eggen@utwente.nl

External supervisors (Cito):
Dr. Maaike van Groen, Maaike.vanGroen@cito.nl
Dr. Hendrik Straat, Hendrik.Straat@cito.nl

Keywords:
summative test, multistage testing, linear testing, item response theory

Abstract

At the end of primary education in the Netherlands, it has to be decided what level of secondary school the primary school students will attend. The initial advice for the level of secondary education that is most suitable for a pupil is given by the school. Next to that, all pupils take a test, which offers an independent advice on the most suitable level of secondary education. One of those tests is the Centrale Eindtoets, which is developed by Stichting Cito under the direction of the College voor Toetsen en Examens. This test provides two measures. Firstly, it classifies pupils into categories, which are linked to the levels of secondary education in the Netherlands, based on their performance on a test. Secondly, it offers an estimate of a pupil's ability in the form of a standardized score. Accurate classification in this test is important, because misclassifications can lead to pupils attending a level of secondary education that is too high or too low compared to their ability.

The test is currently administered in a linear format, which implies that all pupils respond to the same items, regardless of their ability. Therefore, it is likely that pupils have to respond to items that are too easy or too hard relative to their ability. Responding to items that are too easy results in a lack of challenge, while responding to items that are too hard results in frustration. Both emotions can negatively impact a pupil's performance on the test. Secondly, items that are too easy or too hard relative to a pupil's ability provide less than optimal information about the ability of the pupil. This is because specific test items provide optimal information about a pupil's ability on a small range of the ability scale, which implies that an item selection with a mismatch in item difficulty for a particular pupil results in suboptimal information about that pupil's ability.

For low measurement precision, it is more likely that two (or more) adjacent school advices are within the pupil's ability confidence interval and thus the probability for misclassification is higher.

To increase classification accuracy on a test, there are two common approaches: increasing the amount of items that measure optimally around the cut-off point between two classification categories, or increasing the amount of items on a test. Both options are impractical in this case: test items have already been carefully chosen as to optimize the amount of test information available around the cut-off points between the classification categories, and the test already takes three mornings. Another option is to make use of adaptive testing, in which the pupils receive test items based on their performance on the test. Currently, an adaptive, multistage, version of the Centrale Eindtoets is under development. This version of the Centrale Eindtoets consists of three stages. In the first stage, it presents all pupils an initial block of items, or module, to gather an initial set of responses. Based on the responses on the first stage, pupils are routed to one of three modules with different difficulty levels based on their ability. After the second stage, the pupil is again routed to one of three modules, based on the performance on the first and second stage. As the items are adapted to the pupil's estimated ability, it becomes possible to administer items that provide more information in the range of classification categories to which a pupil will likely belong. Therefore, measurement precision can be increased by opting for adaptive testing instead of linear testing.

Although the advantages of the multistage the Centrale Eindtoets over a linear variant are evident from the literature, it is unknown to what extent the choice of the test design influences the measurement precision and the classification accuracy of the test.

In that light, a simulation study was performed with two configurations of the multistage version of the Centrale Eindtoets, and one configuration of the linear version of the Centrale Eindtoets. The two variants of the multistage version of the Centrale Eindtoets differ with respect to the placement of the test items across the three different stages of the test. With the results of this simulation study, the linear and multistage version of the Centrale Eindtoets are compared with respect to the precision of the ability estimates and classification accuracy. Furthermore, the influence of different classification methods on classification accuracy is investigated. Lastly, the influence of different module designs on the precision of the ability estimates and classification accuracy is examined.

The results show that a multistage version of the Centrale Eindtoets outperforms the linear version of the Centrale Eindtoets on both measurement precision and classification accuracy. Furthermore, "the sum of the estimated probability on all items" classification method consistently provides the highest classification accuracy, regardless of the test variant. Finally, the second variant of the multistage the Centrale Eindtoets outperforms the first variant of the multistage the Centrale Eindtoets, both in terms of measurement precision and classification accuracy.

Based on the results from this study, one can conclude that the multistage the Centrale Eindtoets will indeed be an improvement compared with a linear the Centrale Eindtoets. Keeping in mind the limitations of the study, and the fact that the test design in the present study does not conform to all requirements of the 2018 version of the multistage the Centrale Eindtoets, it can be stated that adaptive testing will indeed be an improvement over the current linear way of testing.

Contents

Like in many countries, the Dutch educational system consists of three levels: primary, secondary and tertiary education. Primary education is intended for all pupils between four and twelve years old, and is compulsory from the age of five (EP-Nuffic, 2015). After primary education, pupils can choose from three main levels of secondary education: pre-vocational secondary education (vmbo), general secondary education (havo), and pre-university education (vwo). Vmbo is further divided into vmbo-bb, vmbo-kb, and vmbo-gt. Secondary education lasts between four and six years, depending on the selected level. The initial advice for the level of secondary education that is most suitable for a pupil's ability is given by the school. Next to that, schools are obliged to let their pupils take a test at the end of primary education. This test is known as the final test for primary education. Based on a pupil's performance on this test, the pupil receives an independent advice on the most suitable level of secondary education. Three different tests are available to the schools (Rijksoverheid, 2016). The first test is offered by the College voor Toetsen en Examens (CvTE), who offers the test on behalf of the Dutch government. The other two tests are offered by private organizations, who offer the test with approval from the Dutch government.

The final test for Dutch primary education that is offered on behalf of the Dutch government is known as the Centrale Eindtoets, which is developed by Stichting Cito under the direction of CvTE. The test is administered in three mornings. In the test, the pupils are assessed on their knowledge of mathematics, reading, language skills, and optionally environmental studies. At the end of the Centrale Eindtoets, pupils are classified into one of eight overlapping levels, based on their performance on this test: (1) vmbo-bb, (2) vmbo-bb/kb, (3) vmbo-kb, (4) vmbo-gt, (5) vmbo-gt/havo, (6) havo, (7) havo/vwo, or (8) vwo. The classification decision forms the basis for the independent advice on the most suitable level of secondary education. Next to that, the Centrale Eindtoets provides pupils with an estimate of their ability, in the form of a scale score, ranging from 501 to 550. Therefore, in the Centrale Eindtoets, classification accuracy and a precise measurement of the pupils' abilities, or measurement precision, are very important.

As in all tests, both classification accuracy and measurement precision in the Centrale Eindtoets can never be perfect. Measurement precision and classification accuracy are both linked to the concept of test information. Test information is defined as the amount of information that the items on a test provide for the estimation of a pupil's ability. As the Centrale Eindtoets is currently administered in a linear format, not all items provide much information for the estimation of a pupil's ability. This is due to the fact that, in a linear test, every pupil responds to the same items, regardless of the ability of the pupil. This implies that it is likely that a pupil has to respond to items that are too easy or too hard. There are at least two negative consequences.

Firstly, having to respond to items that are too easy results in a lack of challenge, while having to respond to items that are too hard results in frustration. Both emotions can have a negative effect on a pupil's performance on a test (Linacre, 2000). Secondly, specific test items provide optimal information on a small range of the ability scale and thus also provide optimal information for a small proportion of all pupils. This implies that items that are too easy or too hard relative to a pupil's ability provide less than optimal information about the ability of the pupil. Optimal information about a pupil's ability estimate is obtained when the item difficulty matches the pupil's ability. This leads to a smaller measurement error, as well as higher measurement precision. From a measurement perspective it is desirable to let pupil's respond to many items with a lot of information and few items with less information, to obtain a precise estimate of the pupil's ability. When high measurement precision is not obtained, it is more likely that two (or more) adjacent school advices are within the pupil's confidence interval and thus the probability for misclassification is higher. In other words: a reduced measurement precision results in less accurate classification decisions.

Classification accuracy is important in this test because a misclassification may lead to an advice for a lower or higher level of education than most appropriate for the pupil's ability. When the advice is incorrect, a pupil might be advised a level of secondary education that this suboptimal for his or her ability.

Test information, which is a sum of item information, is positively related to measurement precision: the more test information is available from the items a pupil has responded to, the more precise a measurement is, and the more precise the ability estimate will be. Test information is also positively related to classification accuracy. When classifying a pupil at the end of the test, an incorrect classification decision leads to one of two possible outcomes. The first possible outcome is a false

positive. This error occurs when the pupil's ability estimate lies above a cut-off point for a certain classification level, while the pupil's true ability is below this cut-off point. This pupil will be erroneously classified as having an ability above the cut-off point. The second possible outcome, a false negative, occurs when the pupil's ability estimate falls below a cut-off point for a certain classification level, while the pupil's true ability is above this cut-off point. This pupil will be erroneously classified as having an ability below the cut-off point.

To increase the amount of test information around a cut-off point, there are two common options. Firstly, increasing the amount of items that measure around this cut-off point (Hambleton, Swaminathan, & Rogers, 1991). Secondly, selecting items that are better at distinguishing between pupils that have an ability close to this cut-off point (Hambleton et al., 1991). However, in the case of the Centrale Eindtoets, both ways to increase test information are not realistic. Firstly, as the Centrale Eindtoets is administered over the course of three mornings, it is impractical to prolong the test. Although increasing the amount of test items generally increases test information, having too many items on a test will lead to pupil fatigue. When pupil fatigue is a factor, one does not just measure the test construct, but also how well pupils deal with fatigue. This is called construct-irrelevant variance (Huff & Sireci, 2005), and it can decrease the measurement precision of a test. Secondly, the goal of the Centrale Eindtoets is both to classify pupils into the classification levels that correspond with the most appropriate level of secondary education for these pupils, and to provide pupils with an estimation of their ability. Because of this two-fold goal of the Centrale Eindtoets, the items on the Centrale Eindtoets need to be (a) good at distinguishing between pupils that have an ability close to the cut-off points in the test, and (b) good at distinguishing between pupils with different abilities in general.

Currently, Stichting Cito is developing an adaptive version of the Centrale Eindtoets under the direction of CvTE. This version presents all pupils an initial block of items, or module, to gather an initial set of responses. Based on the previous responses, pupils are routed to one of three modules, which differ in difficulty, that best suits their ability. After this second module, the pupils' ability estimates are updated, and another module most suitable for their ability is presented. This form of testing is known as multistage testing. As the items to which a pupil responds are tailored to the pupil's ability, the measurement precision is increased. Moreover, pupils are less like to receive items that are too easy or too hard for their ability. Consequently, the test will challenge the pupils, while reducing frustration.

Measurement precision is important in the Centrale Eindtoets, because the cut-off points for the classification levels are close together. In a test with a limited number of classification levels that lie far apart, measurement precision is less important for accurate classification decisions. Even when the ability estimate of a pupil is somewhat higher or lower than it should be due to measurement error, this is unlikely to influence the final classification decision for that pupil. However, in a test like the Centrale Eindtoets, in which the cut-off points for the classification levels are close together, measurement error is of greater influence. In these kinds of tests, measurement error is more likely to result in a misclassification. This concept is illustrated in Figure 1.

In this figure, it is illustrated what the effect is of moving from a test with two classification levels to test with four classification levels. As can be seen in the figure, the higher the amount of classification levels, the higher the measurement precision should be to avoid misclassifications.

In the current linear the Centrale Eindtoets, measurement precision is held back by the fact that the classification cut-off points are distributed along a broad range of abilities. Imagine one wants to increase the measurement precision for one classification cut-off point. This can be achieved by replacing some existing items in the test with new items that are more discriminatory around this classification cut-off point. However, doing so would decrease the measurement precision for the other classification points. Therefore, this approach is not viable. The same procedure would be possible with adaptive testing. Given an estimation of the ability level of a pupil, it can be determined in which range of classification levels the pupil will likely belong. This pupil can then receive items that are more discriminatory around these classification points. Thus it can be said that measurement precision can be increased by opting for adaptive testing instead of linear testing.

Although the advantages of a multistage the Centrale Eindtoets over a linear variant are evident from the literature, it is unknown to what extent the choice of the test design influences the measurement precision and the classification accuracy of the test. As Verschoor and Eggen (2014) state, the decisions made when developing a multistage test are interdependent. Moreover, an optimal

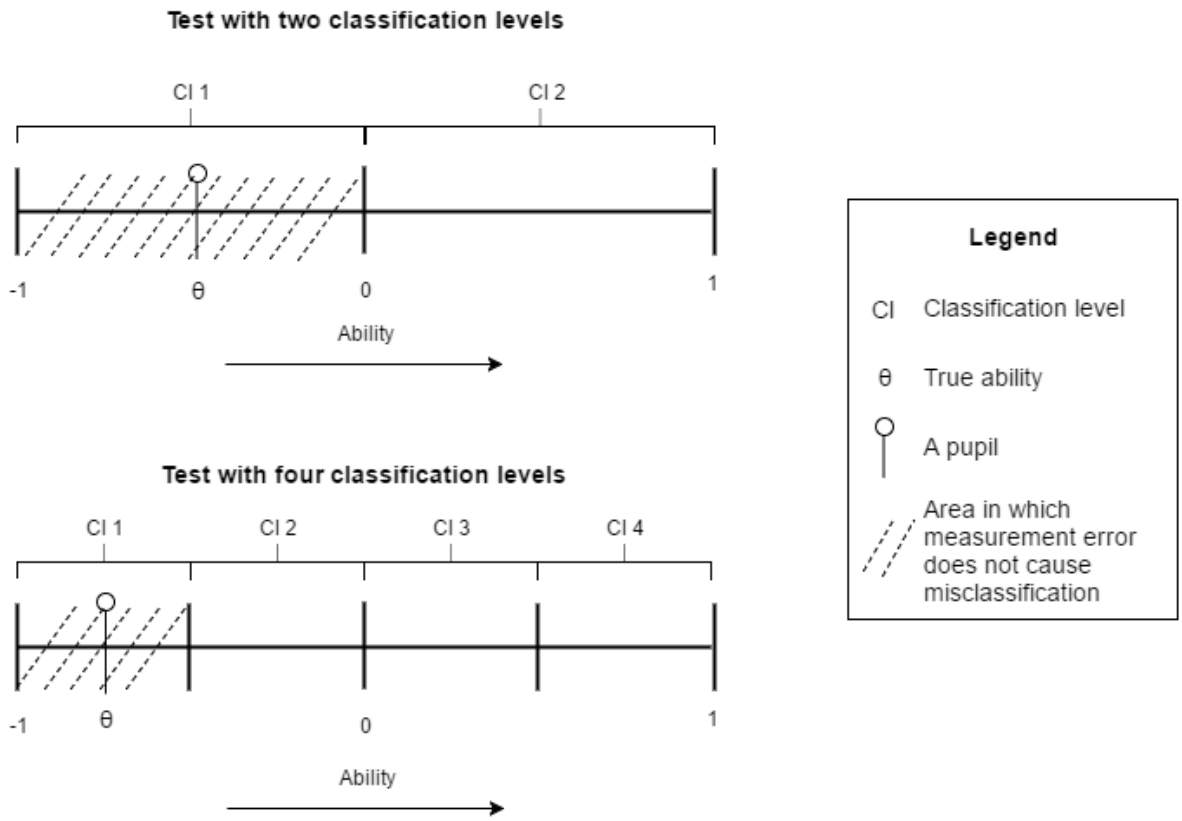

Figure 1. Comparison of a test with two classification levels and a test with four classification levels on the size of the area in which measurement error does not cause misclassification

way to construct multistage tests has not yet been established (Verschoor & Eggen, 2014). As creating a multistage test design from scratch is a complex task, this will not be done in the present study. Instead, the design of this simulation study will be based on the design for adaptive version of the Centrale Eindtoets in 2018. The test contains the domains reading, mathematics, language skills, vocabulary and writing. The first three domains are tested in an adaptive way. The last two domains are tested in a linear way.

This means that all pupils have to respond to the same items, regardless of their ability. It must be noted that the test domain vocabulary will not be present in the Centrale Eindtoets 2018, but is included in the present study to facilitate a comparison with the current version of the Centrale Eindtoets, which contains this test domain. As will be discussed in theoretical framework, there are several options for the composition of the modules. In this study, two possible compositions will be used.

The present study can contribute to the knowledge base on multistage testing (MST) by investigating what the optimal test design is for this specific adaptive multi-category classification test. Therefore, the goal of this research is to investigate the effect of several configurations of the multistage and linear versions of the Centrale Eindtoets on measurement precision and classification accuracy. This comparative research will be performed by means of a simulation study.

The present study will be discussed in several chapters. Firstly, in chapter one, a theoretical framework describes the knowledge base for the remainder of thesis. The second chapter details the methodology employed in the present study. The third chapter presents the results of the simulation study. Finally, the last chapter provides a conclusion and discussion.

1 Theoretical Framework

This chapter lays theoretical foundation for this study. In the present study, a comparison was made between several configurations of the multistage and linear versions of the Centrale Eindtoets. In order to make this comparison, a psychometrical theory is needed. Specifically, a theory is needed that

makes it possible to make comparisons at item level, without limiting your conclusion to one specific population. This is because the different versions of the Centrale Eindtoets do not share the same items. Furthermore, the measures linked to this theory should be independent of a specific test. This makes it possible to the generalise conclusions of this study to similar tests. Therefore, this chapter introduces item response theory (IRT). In IRT, the item properties are specified independently of the specific test in which they are contained. Furthermore, item properties are specified independently of the population that has taken the test. As such, IRT makes it possible to compare different test designs independently of the population and the specific test items.

IRT consists of a large collection of models. Four of those models will be presented in the first part of this chapter. Secondly, the assumptions underlying those models are discussed. Thirdly, the item information function, which aids in determining how suitable an item is for the intended population of the test, is introduced. Fourthly, ability estimation makes it possible to estimate a pupil's ability from his or her responses to a test's items. Ability estimation is discussed in section 1.4. Fifthly, for this ability estimation process, the item parameters must be known. When this is not the case, the ability and item parameters both have to be estimated, as discussed in section 1.5. Sixthly, two types of adaptive testing – computerized adaptive testing (CAT) and MST – will be discussed. Seventhly, MSTs can be designed with different specifications depending on the purpose of test they facilitate. Some considerations that are made when designing a MST are discussed in section 1.7. The chapter ends by discussing classification methods, which are used to classify pupils after the test.

1.1 IRT Models

As stated in the introduction to this chapter, IRT consists of a large collection of models, which specify the relation between the probability of correctly answering an item, the ability of the pupil, and the item's properties (Hambleton et al., 1991). This relationship is captured in a formula known as the item characteristic curve (ICC). The models typically differ with respect to assumptions about item parameters. Four of these models will be discussed: the Rasch model, the two-parameter logistic model, the three-parameter logistic model, and the one-parameter logistic model.

1.1.1 The Rasch Model

In the Rasch model, the probability that a pupil with ability θ responds to an item i correctly is defined as (Hambleton et al., 1991):

$$\mathrm{P_i}(U_i = 1|\theta) = P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} i = 1, 2, \dots, n, \tag{1.1}$$

where n is the number of test items, U_i is the response of a pupil to item i, and b_i is the difficulty parameter. For more than one item, the ICCs for this model manifest themselves as parallel S-shaped curves with values between 0 and 1. The probability to answer an item correctly increases with an increase in the ability of the pupil, and decreases with an increase in the difficulty parameter. An example of four ICCs for items of varying difficulty can be seen in Figure 1.1. In this figure, item 3 has the lowest value for b_i (i.e. is the easiest to answer correctly), while item 2 has the highest value for b_i (i.e. is the hardest to answer correctly).

An attractive property of this model is the fact that the sum score is a sufficient statistic for a pupil's ability (H. G. Fischer, 1995). In other words, the sum score provides all information that is required to estimate a pupil's ability. This is in contrast to more complex IRT models, for which the sum score does not correspond to distinct abilities.

1.1.2 The two-parameter logistic model

A limitation of the Rasch model is the assumption that items are equally discriminating: each item is equally effective in distinguishing among different abilities. However, this assumption does not always hold. To model items that are not equally discriminating, the two-parameter logistic model is used. This model is similar to the Rasch model, with the addition of the item discrimination parameter a:

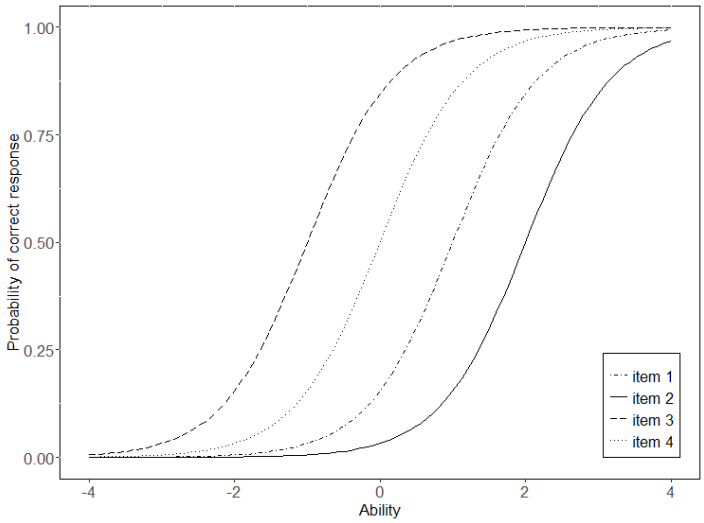

Figure 1.1. ICCs for four items with different difficulty parameters. Adapted from Fundamentals of Item Response Theory (p. 14), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc.

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} i = 1, 2, \ldots, n. \tag{1.2}$$

In this model, the parameter a_i specifies the steepness of the slope of the ICC at the point where $P(\theta) = 0.5$ for item i. Items with a steeper slope have a higher discriminatory power than items with a less steep slope (Hambleton et al., 1991). Figure 1.2 shows four ICCs for items with different difficulty and discrimination parameters. In this figure item 2 has the least discriminatory power, while item 3 has the most discriminatory power.

1.1.3 The three-parameter logistic model

The three-parameter logistic model extends the two-parameter logistic model with the pseudo-chance-level parameter *c*. This extension facilitates a nonzero lower asymptote in the ICC. This asymptote is used to represent less able pupils, who answer selected-response items, such as multiple-choice items, correctly through guessing. The model is defined as:

$$P_i(\theta) = \mathrm{c_i} + (\mathrm{i} - \mathrm{c_i}) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} i = 1, 2, \ldots, n, \tag{1.3}$$

(Hambleton et al., 1991). Figure 1.3 shows the ICCs for six typical items modelled under the three-parameter logistic model. It can be observed that item 3 has a lower asymptote of 0.25, which indicates that pupils with an ability below -0.5 have a 25% chance of answering the item correctly. In contrast, less able pupils have no chance of answering items 1, 2, and 4 correctly.

1.1.4 The one-parameter logistic model

When the Rasch model is used to model the items in a test, but goodness-of-fit statistics show that this model does not fit, there are three choices: either the pupils who cause the poor goodness-of-fit are removed from the sample, the items that show a poor goodness-of-fit are removed from the test, or a different model is chosen (Verhelst & Glas, 1995). All options must be considered carefully. Firstly, when removing pupils from the sample, the generalizability of the results might be comprised. Secondly, when removing the items that are poorly modelled under Rasch, content validity might be compromised. Thirdly, when choosing a different model, some of the attractive properties of the Rasch model, like the sum score as sufficient statistic, might be lost. In order to retain the sum score as sufficient statistic like in the Rasch model, while gaining the flexibility of the two-parameter logistic model, the one-parameter logistic model (OPLM) can be used (Verhelst, Glas, & Verstralen, 1995).

As the sum score is a sufficient statistic for a pupil's ability in this model, it can be used as a representation of the ability of a pupil. With a process known as imputing, the difference in discriminatory power between items can be taken into account (Verhelst & Glas, 1995). In the case of

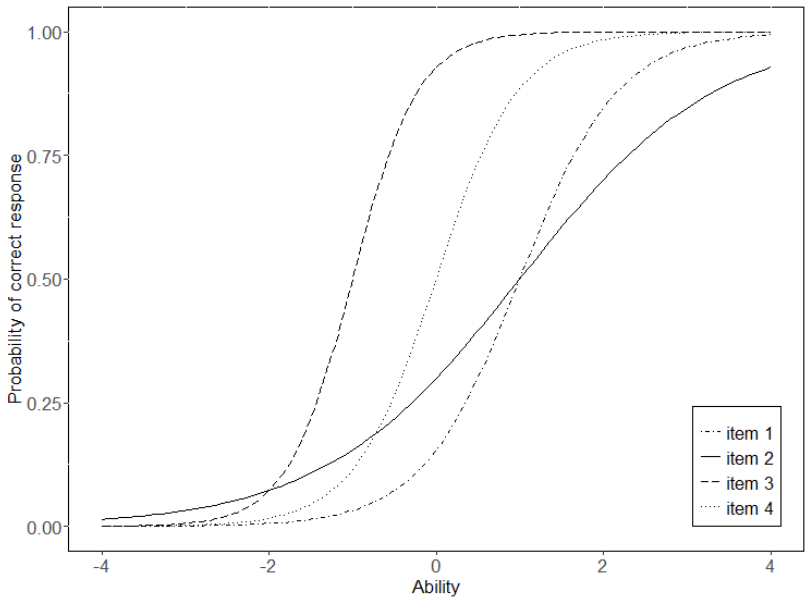

Figure 1.2. Two-parameter ICCs for four typical items. Adapted from Fundamentals of Item Response Theory (p. 16), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc.

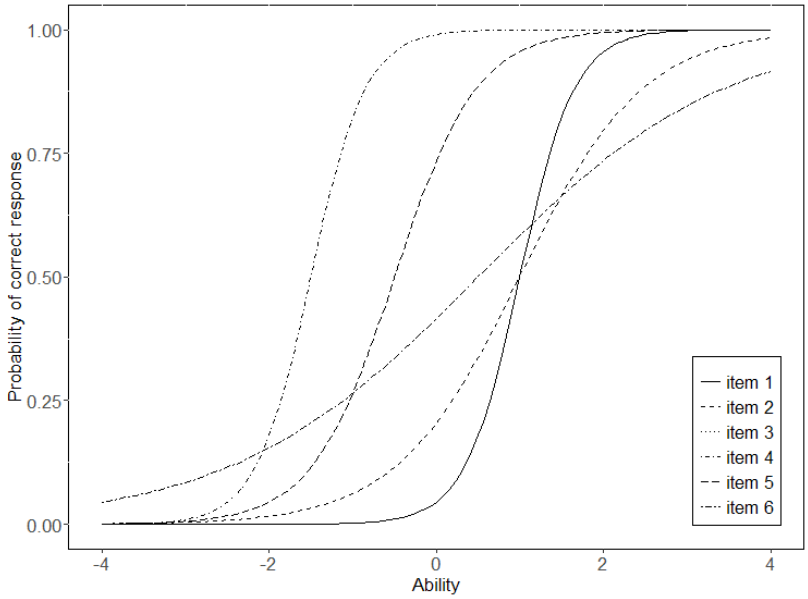

Figure 1.3. Three-parameter ICCs for six typical items. Adapted from Fundamentals of Item Response Theory (p. 18), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc.

the OPLM this implies that the discrimination parameter is imputed. Using either regression analysis or a two parameter logistic model, the most suitable value for the discrimination parameter can be found. If the latter is used, the discrimination parameter is rounded towards its nearest integer value. The OPLM is then estimated with the discrimination parameters fixed to the rounded values. This property makes it possible to model items that are not equally effective in distinguishing among different abilities, unlike the Rasch model.

1.2 Model Assumptions

Recall that IRT models describe the relation between item properties and pupil's performance using an ICC. In order to model this relation assumptions are made. It must be noted that when these assumptions are not met, the validity of the IRT estimates might be compromised. Four assumptions underlying the models described above – unidimensionality, local independence, monotonicity, and parameter invariance – will be explained in the next sections.

1.2.1 Unidimensionality

In IRT, it is assumed that the test's items measure a predefined set of underlying abilities (Hambleton & Swaminathan, 1985). The most widely used models assume that only one ability is measured by the test items, which is referred to as *unidimensionality* (Hambleton et al., 1991). As Hambleton et al. (1991) state, a model can never be unidimensional in a strict sense, because factors like test anxiety and motivation influence performance on a test. However, the assumption holds when there is a clear dominant factor that explains test performance (Hambleton et al., 1991). Other models assume multiple factors are measured by the test's items, but those fall outside of the scope of the present study.

1.2.2 Local independence

According to Hambleton et al. (1991), local independence means that, given the same ability, the pupils' responses to any set of items are statistically independent. In other words, the pupils' responses are only dependent on their ability. Local independence is defined as:

$$P(U_1, U_2, \dots, U_n|\theta) = \prod_{i=1}^{n} P(U_i|\theta). \tag{1.4}$$

However, local dependence only holds when the abilities that the test items measure have been correctly defined. For example, if an item on an English test contains a clue to the answer, the ability to detect the clue is being tested next to the pupil's proficiency in English (Hambleton et al., 1991).

1.2.3 Monotonicity

The ICC, which describes the probability of responding correctly to an item, is a monotone increasing function of the ability. In other words, the higher the ability of a pupil, the higher the probability that this pupil answers the item correctly. This property does not hold when there is a negative relation between ability and the probability of answering an item correctly, or the ICC for an item is not continuous.

1.2.4 Parameter invariance

Item and ability parameters are invariant: parameters that characterize an item are not dependent on the ability distribution of the pupils. In other words, item parameters do not change from one group of pupils to another (Hambleton & Jones, 1993). This assumption makes it possible to estimate item parameters that will hold regardless of the group of pupils taking the test. When items in a test do not adhere to this assumption one speaks of differential item functioning (DIF). DIF occurs when pupils with the same ability, but from different populations, have a different probability of correctly responding to an item (Hambleton et al., 1991). One method to detect DIF is hypothesis testing, in which the null hypothesis states that the item parameters for one item are equal in two different groups (Hambleton et al., 1991).

1.3 Information

In test construction, the item information function is used to determine how informative an item is, with regards to ability estimation, for the intended population. Specifically, this function describes how much information is provided by the item at a given ability (Hambleton et al., 1991). In this case, information is defined as "the contribution items make to ability estimation at points along the ability continuum" (Hambleton et al., 1991, pp. 91-92). Item information for item *i* is defined as:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \tag{1.5}$$

where $Q_i(\theta)$ = 1 - $P_i(\theta)$.

In Table 1.1, the item parameters of five test items are given. Figure 1.4 shows the item information functions of these five test items. In this figure, items 1 and 3 have the highest discrimination parameters, and as such, they provide the most information and have the steepest slopes of all item information functions in the graph. However, whether these items should be selected for a test depends on the expected ability of the intended population. For example, if one expects the intended population

Table 1.1

Item Parameters for Five Typical Test Items

Test Item	Item Parameter		
	a_i	b_i	c_i
1	1.80	1.00	0.00
2	1.80	1.00	0.25
3	1.80	-1.50	0.00
4	1.20	-0.50	0.10
5	0.40	0.50	0.15

Note. Adapted from *Fundamentals of Item Response Theory* (p. 18), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc. Copyright 1991 by SAGE Publications, Inc.

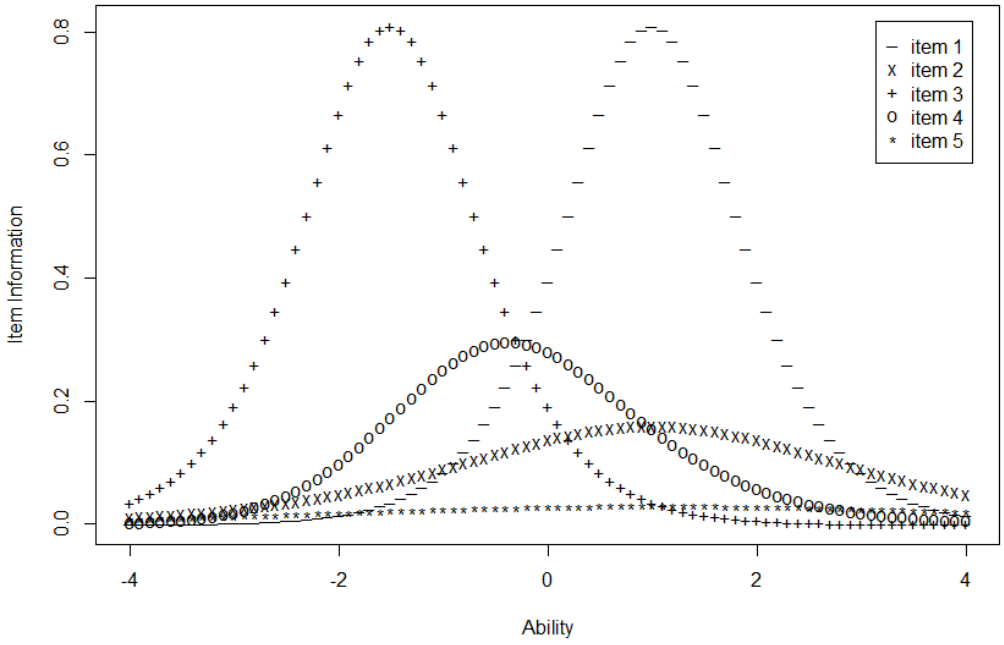

Figure 1.4. Item information functions for five typical test items. Adapted from Fundamentals of Item Response Theory (p. 93), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc.

to have an ability around 0, items 1 and 3 would not be the first choice despite their high discrimination parameters, because item 4 provides more information around this point. Items 2 and 5 are unlikely to be selected for a test at all, because they do not provide a lot of information at any point on the ability scale. To find out how informative a test is at a certain ability, one sums the item information functions, at that ability, for all items in the test:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{1.6}$$

(Hambleton et al., 1991). Equation 1.6 shows that the items independently contribute to the information a test provides. This makes it possible to construct a test from individual items, with a target for the test information in mind. For example, in a test for classification purposes, one usually aims to provide the most test information around the classification cut-off points. The amount of information that a test provides at a certain ability is inversely related to the precision of the ability estimate at that point:

$$SE(\hat{\theta}) \approx \frac{1}{\sqrt{I(\hat{\theta})}}, \tag{1.7}$$

where SE is the standard error of estimation (Hambleton et al., 1991). In other words, the more test information a test provides, the higher the measurement precision.

1.4 Ability Estimation

To estimate a pupil's ability based on his or her responses on the test items, ability estimation methods are used. One of the methods to estimate the pupil's ability is maximum likelihood estimation. This procedure for the estimation of the pupil's ability assumes that the item parameters for the test are known. Maximum likelihood estimation is in a sense analogous to how a medical doctor uses the clinical inferences process to diagnose the disease of a patient (Jehangir, 2005). In the clinical inference process, the symptoms of the patient are used to find a diagnosis that has the highest likelihood of being true, given those symptoms. Much in the same way, maximum likelihood estimation uses the responses of a pupil to find the ability that has the highest likelihood of being true, given those responses.

Given the item parameters, and the pupil's responses, the likelihood of observing those responses under the IRT model for each ability is determined. The ability that has the highest likelihood, given the responses of the pupil, is defined as the maximum likelihood estimate of the ability for that pupil (Hambleton et al., 1991). The relation between the item parameters, the responses of a pupil on the items, and the underlying ability is formalized in the likelihood function:

$$L(\boldsymbol{u}|\theta) = \prod_{i=1}^{n} P_j^{u_i} Q_j^{1-u_i}, \tag{1.8}$$

where u_i is the observed response to item i and $\boldsymbol{u}$ the vector of responses.

The likelihood function (see equation 1.8) is usually converted to a logarithmic function. To indicate the fact that this conversion has been performed, the name log-likelihood function is used. The conversion changes the likelihood function from a product function to a sum function, which is less computationally expensive, and improves the scaling (Hambleton et al., 1991).

An example of the log-likelihood function for pupils of various abilities can be seen in Figure 1.5. As this figure shows, each pupil has only one maximum for the likelihood function, at which point the slope of the line becomes zero. Unfortunately, it is not possible to set the derivative of the likelihood function to zero, and solve this equation. This is because there is no solution for this equation. Therefore, an iterative mathematical search procedure must be used to find a pupil's ability estimate. One example of such an iterative search method is the Newton-Raphson procedure (Segall, 1996). Using this approach the corresponding ability can be estimated for almost all response patterns.

When performing ability estimation, it is of interest how precise these ability estimates are. One measure of ability estimate precision is the root mean square error (RMSE; Willmott & Matsuura, 2005). In contrast with the SE, which specifies the error of estimation for one particular ability, the RMSE is a measure of ability estimate precision over the whole ability range. The RMSE is defined as:

$$RMSE = \sqrt{\sum_{j=1}^{m} \frac{\left(\hat{\theta}_j - \theta_j\right)^2}{m}}, \tag{1.9}$$

where $\hat{\theta}$ is the estimated ability for pupil j, and m is the number of pupils who take the test. The RMSE can only be calculated when the true ability of the pupils is known. Therefore, this measure is only relevant for simulation studies, and cannot be used in real life tests.

1.5 Item Parameter Estimation

In the previous section, the procedure for the estimation of the ability of a pupil was described. In this procedure, it is assumed that the parameters of the items the pupil has responded to are already available. However, this will not always be the case. When the item parameters are not yet available, both the ability of the pupil and the parameters of the items have to be estimated.

During estimation there is no unique solution. In order to eliminate this problem, an arbitrary scale for the ability values and item difficulty values must be chosen. A common choice is to assume a standard normal distribution for (a) the ability of all pupils, or (b) the item difficulty

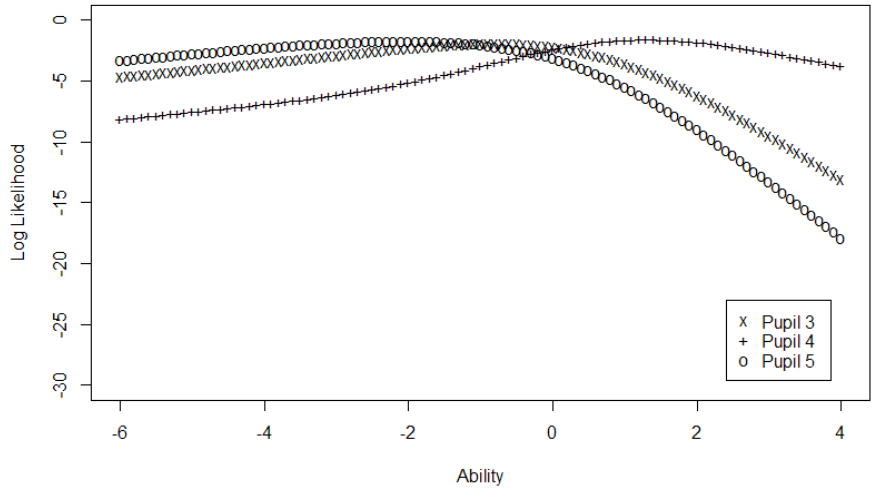

Figure 1.5. Log-likelihood functions for three pupils. Adapted from Fundamentals of Item Response Theory (p. 36), by R.K. Hambleton et. al., 1991, California: SAGE Publications, Inc.

(Hambleton et al., 1991). Once a scale has been fixed for the ability values or the item difficulty values, it is no longer possible to obtain multiple solutions, and the estimation procedure can begin. In order to perform item parameter estimation, there are at least three common ML methods: joint maximum likelihood, marginal maximum likelihood, and conditional maximum likelihood. In the first method, the ability and item parameters are estimated simultaneously (Hambleton et al., 1991). However, the results obtained using this procedure are inconsistent, as this procedure might converge towards incorrect values (Ghosh, 1995). In the second method, the ability parameters are first integrated out, after which the item parameters are estimated (Hambleton et al., 1991). To accomplish this, the assumption is made that the ability parameters are sampled from a larger distribution (Johnson, 2007). With the item parameters known, the ability parameters are estimated (Hambleton et al., 1991). The last method treats the ability parameters of the pupils as given. To accomplish this, the IRT model is separated into a conditional part, which only depends on the item parameters, and a part for the sufficient statistics (Eggen, 2000). As such, this method can only be used for the Rasch or OPLM model. Of the three models, only marginal maximum likelihood can currently be used for multistage testing. For more information on those three models, the reader is referred to literature (e.g. Baker & Kim, 2004).

1.6 Adaptive Testing

The precision of the pupil's estimated ability can be increased compared to linear testing by opting for adaptive testing. In adaptive testing, the item selection is adapted based on the responses of the pupil. Two types of adaptive testing – CAT and MST – are discussed.

1.6.1 Computerized adaptive testing

The earliest application of adaptive testing can be traced back to the work of Binet on intelligence testing in early 1900s (Weiss, 1985). However, using adaptive testing was hardly feasible until the advent of more powerful computers (Hambleton et al., 1991). Since the 1990s CAT has been a popular test administration model (Becker & Bergstrom, 2013). With a tailored selection of items the pupil's ability can be estimated more precisely than with a linear test, given the same test length, without the need to increase the amount of items on the test (Wainer, Kaplan, & Lewis, 1992, in Yan, von Davier, & Lewis, 2014).

In CAT the pupil's previous responses determine the selection of each following item. The pupil's ability is estimated after every response using all previous responses. Using this estimate, the item contributing most to a more precise measurement, given the test constraints, is selected. A common test constraint is the fact that enough items to adequately assess each tested construct are presented. This item selection procedure ensures that the test is tailored to the pupil, and items that are too easy or too hard for the pupil according to the estimated ability are not presented (Becker & Bergstrom, 2013).

CAT will not be investigated in this comparative simulation study. Three important reasons for this decision are the need for a large pre-test sample, content inspectability, and cost of item construction. Firstly, the need for a large pre-test sample. Before a test can be used, it has to be pretested. In the case of the CAT, it can only be tested on Dutch primary school pupils, as this is the target group of the test. Furthermore, as a CAT contains a large number of items due to its adaptive nature, a large sample of pupils is required for this pre-test. However, it is hard to recruit large amounts of pupils for this pretest in the Netherlands. Secondly, CATs are constructed on-the-fly, with a specific combination of test items for each individual pupil. As this process leads to many possible test variants, it is virtually impossible to inspect the content specifications of every test variant (Kim, Chung, Park, & Dodd, 2013). However, one of the requirements of the Centrale Eindtoets is the fact that its content must be inspectable by CvTE. Furthermore, the Centrale Eindtoets has a complex set of content specifications. To formulate all these content specifications as constraints in the item selection procedure of a CAT would be meticulous task. Lastly, CAT requires a larger item bank than MST. Increasing the size of the item bank leads to an increase in item development efforts, and therefore increases the required amount of pupils in the pre-test.

1.6.2 Multistage testing

An alternative to CAT is MST. At the start of an MST, a pupil is administered an initial module, known as the *routing test*, which is used to estimate the pupil's proficiency (Yan et al., 2014). After this routing test, the pupil is presented a module that contributes most to a precise measurement, given the pupil's performance so far. For tests consisting of multiple consecutive modules, information on the pupil's proficiency is updated after each module, using all previous responses, after which the most appropriate next module is selected.

An example of a MST consisting of multiple consecutive modules can be seen in Figure 1.6. In this example, the pupil is first presented with a routing test. After responding to the items in the routing test, the pupil's performance so far is determined, and the module in stage two that is most suitable, based some criterion, is selected. This criterion depends on the purpose of the test. In the selected stage three module, the process is repeated: the pupil responds to the module's items, after which a proficiency estimate is obtained, and the appropriate stage 3 module is selected.

As each module has a fixed item set, each module can be constructed to cover all predetermined content specifications, while retaining the adaptive property. However, given the same item pool, pupils' ability estimates from a MST are slightly less precise than from a CAT (Hambleton & Xing, 2006, in Yan et al., 2014).

1.7 MST Design

When designing a MST, the number of stages, modules, and items have to be considered (Veldkamp, 2014). The choices will influence the characteristics of the test. The exact settings depend on the purpose of the test and the desired measurement precision (Veldkamp, 2014). For example, if the purpose of the test is ability testing, the MST design should facilitate a high estimation accuracy for a range of abilities (Yan et al., 2014). By contrast, tests for classification purposes should focus more on measurement accuracy near the cut-off points for the classification levels (Yan et al., 2014). The design decisions for the number of stages, module design, and routing are discussed.

1.7.1 Number of stages

In deciding on the number of stages in an MST, a trade-off is made between simplicity and flexibility. In an MST with only two stages, the complexity of test assembly is the lower than in MSTs with more stages. However, there is a higher likelihood of routing error, as there is only routing point (Yan et al., 2014). Especially for pupils with abilities near the cut-off points, routing errors are likely. One way to guard against routing error is to create an overlap between modules, or to increase the amount of stages (Weiss & Betz, 1974). Opting for more stages gives more flexibility in tailoring the test to pupils' abilities. However, it also increases the complexity of test assembly and test analyses, while not necessarily increasing measurement precision of the test (Luecht & Nungester, 1998).

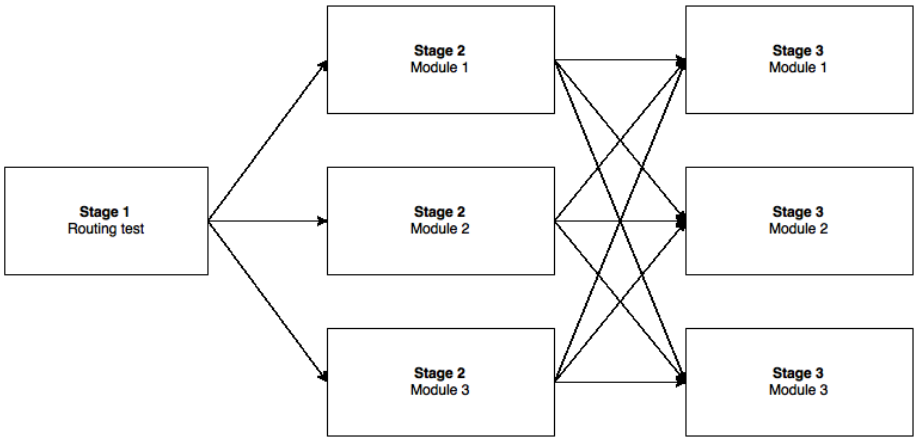

Figure 1.6. Example of a three stage MST with three modules in the second and third stage.

1.7.2 Module design

Similar to deciding on the number of stages, the decision of the number of modules in an MST is a trade-off between simplicity and flexibility. When dealing with fixed-length modules, a maximum of four modules per stage is desirable (Armstrong, Jones, Koppel, & Pashley, 2004). Several factors need to be taken into account when designing a module, such as the range of difficulty parameters and range of discrimination parameters for the items in the module (Yan et al., 2014). Especially the characteristics of the routing test have a major influence on the measurement precision of the whole test. According to Kim and Plake (1993), the characteristics of the first stage module highly influence the measurement precision of the whole test. Furthermore, Kim and Plake (1993) found that increasing the length of the first stage module has the largest effect in reducing ability estimation errors.

1.7.3 Routing

There are many methods that can be used to decide how to route a pupil to the next module in a MST. Two possible ways to make routing decisions are (a) selecting the path that provides the most information for the pupil, given the pupil's current ability estimate, and (b) using the sum score up until that point to select the best path. When the IRT model under which the test items are modeled does not imply the unweighted sum score as a sufficient statistic, some information is lost when opting for (weighted) sum score routing. However, Luecht and Nungester (1998) demonstrated that sum score routing is probably sufficiently accurate for path selection purposes. Furthermore, by opting for sum score routing, the module selection routines that the test delivery software has to support are simplified (Luecht, Brumfield, & Breithaupt, 2006).

To perform routing in an MST, routing cut-off points have to be determined. Two ways to determine those routing cut-off points are (a) the maximum information method, and (b) the defined population intervals method (Luecht et al., 2006). In the first method, the test information function is used, after each stage, to determine the next module. In this process, for each possible module the amount of provided information, given a pupil's current ability estimate, is calculated. The module that contains the items that provide the most information for the pupil is selected. In the second method, it is predetermined which proportions of pupils in the population are required to take each route. For example, take a MST design with one module in the first stage, and two modules in second stage. In this design, the pupils are split into two groups by performance. The lowest performing group will be routed to the first module in the second stage, while the highest performing group will be routed to the second module in the second stage.

1.8 Classification methods

Tests can be broadly categorized as tests for ability estimation purposes, and tests for classification purposes. In tests for ability estimation purposes, the resulting measure of the test is the estimated ability of the pupils. However, in tests for classification purposes, the result of the test is the appropriate category for each pupil. To make this classification decision, classification methods are

used. The accuracy of the resulting classification decisions are measured with the proportion of correct decisions (PCD). The PCD is calculated by dividing the amount of correct classification decisions by the total amount of pupils participating in the test. Three classification methods relevant to this study – the sequential probability ratio test (SPRT), the sum of the probability of correct responses on all items, and the estimated ability classification method using the Rasch model – are discussed.

1.8.1 The sequential probability ratio test

The SPRT was originally developed by Wald (1973). In tests with only two classification levels, such as pass or fail tests, an ability cutoff point θ_c is placed between the two levels, with an indifference region around this point. The indifference region controls for uncertainty in the classification decision, caused by measurement error, for pupils with an ability close to the cutoff point (Eggen, 1999). Hypotheses are formulated at the lower and upper end of the indifference region:

$$H_0: \hat{\theta} < \theta_c - \partial, \tag{1.10}$$

$$H_1: \hat{\theta} > \theta_c + \partial, \tag{1.11}$$

where ∂ signifies half the size of the indifference region. Type I and Type II errors are acceptable when:

$$P(reject\ H_0 | H_0\ is\ true) \leq \alpha, \tag{1.12}$$

$$P(fail\ to\ reject\ H_0 | H_1\ is\ true) \leq \beta, \tag{1.13}$$

with small constants for the rate of type I error α and the rate of type II error β. The SPRT (Wald, 1973) can be used to test this, with the ratio between the values of the likelihood function (see equation 1.8) under both hypotheses as the test statistic (Eggen, 1999):

$$LR(\boldsymbol{u}) = \frac{L(\theta_c + \partial | \boldsymbol{u})}{L(\theta_c - \partial | \boldsymbol{u})}. \tag{1.14}$$

The following rules are used to make a classification decision (Eggen, 1999):

Ability below θ_c $$LR(\boldsymbol{u}) \leq \beta/(1-\alpha), \tag{1.15}$$

Ability above θ_c $$LR(\boldsymbol{u}) \geq \frac{1-\beta}{\alpha}. \tag{1.16}$$

If the above rules do not lead to a classification decision, the pupil is classified as having an ability above the cut-off point when the log of the likelihood ratio is larger than the midpoint of the log of the interval $\beta/(1-\alpha) < LR(\boldsymbol{u}) < (1-\beta)/\alpha$. When the log-likelihood ratio is smaller than this midpoint, the pupil is classified as having an ability below the cut-off point.

This procedure can be generalized to cases with multiple classification categories and multiple test dimensions. To do so, the likelihood ratio in equation 1.14 is expanded so all dimensions and all items in the test are included (van Groen, 2014):

$$LR(\boldsymbol{u}) = \prod_{j=1}^{k} \frac{L(\theta_{c,j} + \partial | \boldsymbol{u_j})}{L(\theta_{c,j} - \partial | \boldsymbol{u_j})}, j = 1, \dots, k, \tag{1.17}$$

where $\theta_{c,j}$ is the classification cut-off point for dimension j, and $\boldsymbol{u_j}$ the vector of responses for the dimension j. It is assumed that all dimensions share the same value for ∂. The classification cut-off

points, which separate adjacent classification categories, have to be determined for each dimension covered by the test.

1.8.2 Sum of the probability of correct responses on all items

A pupil can also be classified using the sum of the probabilities of correct responses on all items in the item bank:

$$\sum P_i(\hat{\theta}), for\ i \in V_{all}, \tag{1.18}$$

where V_{all} is the set of items in the MST. The outcome of equation 1.18 can be compared to the pre-specified cut-off points in order to make a classification decision. These cut-off points are determined by defining the minimal required total sum score across all domains covered by the test, for each classification level. In practical terms, this sum score represents which share of the items on the test should be answered correctly for a pupil to be classified into a certain classification level.

1.8.3 Estimated ability classification method using the Rasch model

After a test has been completed, the ability of a pupil can be estimated using maximum likelihood estimation, as described in section 1.4. However, in the Centrale Eindtoets, multiple abilities are tested. In the case of a multivariate ability distribution, the ability of a pupil cannot easily be expressed in a single measure. This is because the different abilities are not measured on the same scale. As an alternative, the items for all domains in the test are put into one item bank. The item parameters for this item bank are estimated under the Rasch model. The item parameters under the Rasch model are then used to estimate the ability of the pupils. The resulting ability estimates are used for classification purposes by defining an ability interval for each category. Pupils are classified in the category corresponding with the ability interval that contains the estimated ability.

To obtain these ability intervals, it has to be determined which ability interval belongs to which classification level. To do so, a criterion has to be defined regarding the ability in each classification level. One example of such a criterion is a sum score. By using the sum score as a criterion, the classification cut-off points from the classification method described under 1.8.2 can be used. To do so, the minimal required sum score for each level are obtained from the method described under 1.8.2. Secondly, the ability estimates corresponding to those sum scores are obtained. These ability estimates serve as the cut-off points for each classification level.

2 Methodology

In this study, the multistage and linear versions of the Centrale Eindtoets were compared with respect to the precision of the ability estimates and classification accuracy. Furthermore, the effects of three classification methods and two module designs were investigated. Specifically, the following research questions were answered:

1. How do the linear and multistage version of the Centrale Eindtoets compare with respect to the precision of the ability estimates?
2. How do the linear and multistage version of the Centrale Eindtoets compare with respect to classification accuracy?
3. What is the influence of different classification methods on the classification accuracy of the linear and multistage version of the Centrale Eindtoets?
4. What is the influence of different module designs on the precision of the ability estimates and the classification accuracy of the linear and multistage version of the Centrale Eindtoets?

2.1 Research Design

In order to answer the four research questions above, several versions of the Centrale Eindtoets have to be administered to respondents. The respondents' responses are used to estimate their abilities

and make classification decisions with different classification methods. The resulting data is used for four separate analyses in order to answer the research questions of this study: (1) a comparison between the linear and multistage version of the Centrale Eindtoets with respect to the precision of the ability estimates, (2) a comparison between the linear and multistage version of the Centrale Eindtoets with respect to the classification accuracy, (3) the influence of different classification methods on the classification accuracy of the linear and multistate versions of the Centrale Eindtoets, and (4) the influence of different module designs on the precision of the ability estimates and the classification accuracy of the linear and multistage versions of the Centrale Eindtoets. To perform these steps, a traditional research design (i.e., Campbell & Stanley, 1963) is not applicable for three reasons.

Firstly, to provide results that are stable across reruns of this study, an amount of respondents comparable to the number of respondents on the current version of the Centrale Eindtoets (about 150,000) would be required. For example, these respondents might participate in a study that randomly assigns them to one of the proposed new variants of the Centrale Eindtoets. Such a study cannot easily be performed without implementing the new test variants in the actual test administration. Because of the uncertainty about the test quality of different configurations of the MST, it is unwise to implement these new test variants in a high stakes testing situation. Secondly, the current test administration system is not yet suitable for the MST that is the subject of this study, as the results of this and other studies have to provide more information about the optimal design of the system. Thirdly, a traditional research design does not lend itself well for an iterative process of testing a design and adapting the MST based on the results. Given the aforementioned reasons, a simulation study was performed.

This study is based on the following procedure. Firstly, using a sample of existing response data and the known item parameters from the Centrale Eindtoets 2015, the multivariate ability distribution for the different subjects of the Centrale Eindtoets 2015 was estimated. The ability distribution was then used as a starting point for the generation of ability parameters for the simulees in the simulated tests. Furthermore, using this ability distribution, item parameters were generated that fit the specifications of this test design. Secondly, the simulees´ responses to all items in the simulated tests were generated.

To determine the response of a simulee to each item, the probability that the simulee correctly responds to an item is calculated based on the OPLM. This probability is compared to a randomly generated value from a uniform distribution between zero and one. If this randomly generated value is higher than the probability that the simulee correctly responds to an item, the item is marked as answered correctly (1) for this simulee. If this randomly generated value is lower than the probability that the simulee correctly responds to an item, the item is marked as answered incorrectly (0) for this simulee.

Lastly, the precision of the ability estimates and the accuracy of the classification decisions was determined for both versions of the Centrale Eindtoets. For a more detailed description of the output measures of this simulation, the reader is referred to section 2.4.8. The procedure was repeated for three different classification methods and two different module designs, as explained in the data analysis section.

With the research design as described above, the four analyses that are part of this study will be performed. In the following sections, the methodology for all four analyses will be detailed as a whole. When there is a difference in methodology between the analyses, those differences will be made explicit.

2.2 Respondents

In the present study, there was no sampling performed in the traditional sense. Instead, this study used existing data from the response file from the Centrale Eindtoets 2015, modeled under OPLM, as input for the simulation study. To ensure ethical integrity, the responses in the file cannot be traced back to individual respondents. From this file, the distribution of the ability parameters of the respondents was obtained given the calibrated item parameters. This ability distribution will be used to generate the abilities of the simulees. The procedure used to generate simulees is discussed below.

2.2.1 Simulee generation

To obtain realistic values for the ability parameters in the simulation study, the response data from all pupils (N = 149,158) of the paper-and-pencil edition of the *Basis* and *Niveau* versions of the Centrale Eindtoets 2015 were selected. The response data consists of a score set for each pupil. Each set consists of a series of zeros and ones for each simulee, representing an incorrect and a correct answer to an item, respectively. From these response data, the multivariate normal distribution of the pupils was

obtained. Table 2.1 shows the standard deviations and the means for the ability distribution. The ability parameters of the simulees were drawn from this observed multivariate distribution. The simulee sample contains 100,000 simulees, obtained by simple random sampling. This sample size was chosen, because preliminary tests showed that this sample size allows one simulation run to be completed in one and a half hours, while providing stable simulation results.

2.3 Instrumentation and Procedure

The simulation test in this study was based on the proposal, as available at the start of the study for the design of the multistage version of the Centrale Eindtoets and the response data from the linear versions of the Centrale Eindtoets 2015 (CvTE, 2015). The item pool consisted of items for five subjects: reading, language skills, mathematics, vocabulary, and writing. Table 2.2 shows the proposal for 2018 for the amount of items per subject administered in each part of the simulation test.

2.3.1 MST design

The design of the multistage test is based on the current proposal for the 2018 version of the multistage the Centrale Eindtoets, consisting of three stages, as illustrated in Figure 2.1. Three subjects (i.e., reading, language skills, and mathematics) are tested in an adaptive way, with an initial stage in part one, after which pupils are routed to a second stage module depending on their performance. The three modules in stage two and three are targeted at the percentile scores in the population of 1-25 (module 1), 26-60 (module 2), and 61-100 (module 3), respectively. The students' performance in stage one and two determines to which stage three module they are routed. It should be noted that routing occurs per subject: for example, a simulee can be routed to the module corresponding with percentile score 1-25 for mathematics and the module corresponding with percentile score 25-60 for reading. The last two subjects (i.e., vocabulary and writing) are tested in a linear way, with each pupil receiving the same set of items regardless of their performance.

In the proposal for the 2018 version of the multistage the Centrale Eindtoets, there are different targets for p-values per module. In this context, the p-value for an item indicates the probability that, given the average ability of a (sub)population, the response to this item is correct. The p-value targets have been defined to ensure that the average difficulty of the items in this test adhere to a predefined standard.

For the subjects that are tested in an adaptive way, the target p-value for stage one is .70 for the total population of simulees. For stage two and three of these subjects, the target p-value is linked to the average ability of the target simulee group of each module. In these stages, the p-value for the target simulee group in each module is .60. For the subjects that are tested in a linear way, as well as the linear version of the test, all items have a target p-value of .70 for the total population of simulees.

2.3.2 Standard score classification

As stated in the problem statement, the Centrale Eindtoets offers advice on the most appropriate level of Dutch secondary education for the pupil. To do so, the Centrale Eindtoets classifies pupils into one of eight categories corresponding to five levels in Dutch secondary education and three intermediate categories, which are a mix of two adjacent levels of education. The levels are, from low to high ability: (1) vmbo-bb, (2) vmbo-bb/kb, (3) vmbo-kb, (4) vmbo-gt, (5) vmbo-gt/havo, (6) havo, (7) havo/vwo, (8) vwo.

To classify simulees into one of the eight categories, the pupils' sum scores on the items of the Centrale Eindtoets 2015 are calculated for each subject. Secondly, these sum scores are added to each other. Using these sum scores, a standard score is calculated with the following formula:

$$Standard\ score = sum\ score * A^* + B^*. \tag{2.1}$$

The formula uses two constants A^* and B^* to make sure that students with the same performance that take the test in year X and year X+1 will get the same standard score. The constants are determined annually using a statistical procedure called equating. In 2015, A^*=0.3338 and B*=482.23. The standard score ranges corresponding to each classification level are shown in Table 2.3. For this study, the standard score ranges of 2016 are used. The the Centrale Eindtoets 2015 made use of overlapping

Table 2.1

Means and Standard Deviations of the Multivariate Ability Distribution of the Centrale Eindtoets 2015

Subject	*M*	*SD*
Reading	0.4800	0.2670
Vocabulary	0.4740	0.3350
Writing	0.4600	0.2680
Language skills	0.4740	0.3020
Mathematics	0.3600	0.2750

Table 2.2

Proposed Amount of Items per Test Domain and Test Part for the Simulation Study

Subject	Part 1	Part 2	Part 3
Reading	15	15	15
Vocabulary	10	5	5
Writing	10	5	5
Language skills	16	17	17
Mathematics	25	30	30

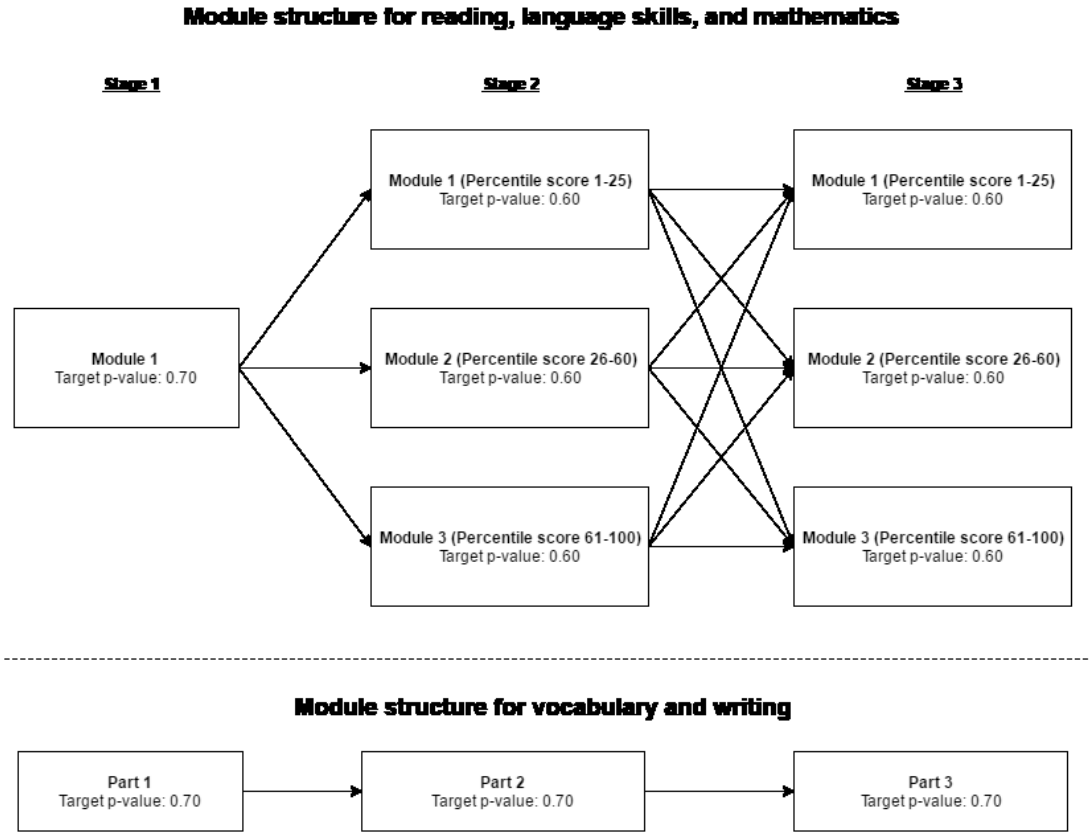

Figure 2.1. Module structure for the multistage variant of the test with target p-values

standard score ranges, which would make it impossible to classify all pupils into a distinct classification level.

Table 2.3

Classification Categories and the Corresponding Standard Score Ranges from the Centrale Eindtoets 2016

Classification Category	Lowest Standard Score	Highest Standard Score
Vmbo-bb	501	518
Vmbo-bb/kb	519	525
Vmbo-kb	526	528
Vmbo-gt	529	532
Vmbo-gt/havo	533	536
Havo	537	539
Havo/vwo	540	544
Vwo	545	550

2.4 Method of Data Analysis

The proposal for the multistage version of the Centrale Eindtoets 2018 as outlined above serves as the blueprint for the simulated tests. Two variants of the test were created by changing in which stage the items with the highest discrimination parameters are used. By doing so, it could be determined at what stage in this multistage test highly discriminatory items should be placed to obtain the most precise ability estimate. Furthermore, a linear variant of the test was created. This was done to facilitate a comparison between a multistage version of the Centrale Eindtoets and a linear version of the Centrale Eindtoets. In summary, there are three test variants:

(a) the multistage test variant with the highly discriminatory items in stage two and three,

(b) the multistage test variant with the highly discriminatory items in stage one, and

(c) the linear test variant.

To make the results across the three test variants comparable, all variants should share the same item bank per subject. In this case, an item bank is defined as a collection of test items with known item parameters. The most obvious way to create the item bank for this study would be to select the required amount of the items from the Centrale Eindtoets 2015. The items from the Centrale Eindtoets 2015, after all, form the basis of this study. For each test variant, the items in the item bank would then have to be divided in such a way that the target p-values, as described in section 2.3.1, are met. In practice, however, it was not possible to use this approach while conforming to the target p-values.

Another way to create the item bank for the simulated tests would be to generate item parameters in such a way that they conform to the target p-values. However, it would be undesirable that these items are unrelated to the existing the Centrale Eindtoets 2015 items. Using unrelated items makes it difficult to generalize the results of this study to the Centrale Eindtoets.

To solve this problem, a combination of these two item bank creation methods was used. In this procedure, the item bank was created by first preparing the item selection for the first test variant. This item bank formed the basis of the item selection for the second and third test variant. The procedure for the creation of the item bank is described in section 2.4.1. Furthermore, the procedure to make an item selection from this item bank for the second and third test variant is described in section 2.4.3.

For all three test variants, the item selection for the subjects that are tested in a linear way (i.e. writing and vocabulary) is the same. Therefore, the item selection procedure for these two subject is discussed separately in section 2.4.4.

2.4.1 Item bank creation

To relate the items in the simulated tests to the existing the Centrale Eindtoets 2015 items, the required amount of discrimination parameters were randomly sampled from the discrimination parameters of the existing the Centrale Eindtoets 2015 items. This sampling procedure was performed for each test domain that was tested in an adaptive way (i.e., reading, mathematics and language skills). Furthermore, the location parameters for the test items were randomly generated. In this process, the

average p-value for each module in stage two and three was taken into account. In each module, the average p-value for the target population (i.e., the subpopulation of the pupils that is intended to go to a specific module) of that particular module should be 0.65. Afterwards, the discrimination parameters and location parameters were paired to form the item parameters for the items. The resulting item bank was used as a base for the item selection in the first, second, and third test variant.

2.4.2 Item selection for the first test variant

To divide the selected items over the stages in such a way that the target p-values were met for the first test variant, a three step procedure was used. To explain this procedure, the following symbols are used:

$n_{stage\,x}$	number of items in stage *x,* where *x* = 1..3
N	total number of available items for one test domain

The procedure is illustrated in Figure 2.2. Note that more advanced, automated item selection methods are available for use with IRT. These items selection methods, also known as optimal test design, use mathematical optimization algorithms to select the items that best contribute to features such as measurement precision and content balance (Oakland, 1995, p. 100). However, optimal test design falls outside the scope of this study. The item selection procedure that was performed for the first test variant is as follows:

1. The generated pairs of item parameters were ordered by discrimination parameters, from low to high.
2. From these ordered items, items 1 until and including $n_{stage\,1}$ were selected for stage one.
3. From these ordered items, items $n_{stage\,1} + 1$ until and including $n_{stage\,2}$ were selected, and ordered by location parameters. These items were divided over the three modules of stage two. In this way, the items with the lowest location parameters end up in module one, and the items with the highest location parameters end up in the module three.
4. The remaining items $N - n_{stage\,3}$ until and including item *N* were ordered by location parameters, and assigned to the three modules of stage three, analogous to the previous step.

The item generation procedure outlined in this section was repeated several times to arrive at the optimal item selection for the first test variant. Repetition was required because the prerequisites for the item generation procedure resulted in a lower chance to attain the desired average p-value. Firstly, for reading, the procedure was repeated 1600 times. Secondly, for mathematics, the procedure was repeated 600 times. Thirdly, for language skills, the procedure was repeated 3500 times. The amount of repetitions that were required to reach the desired average p-values were dissimilar, because of the difference in the amount of available test items for each test domain. For example, in the domain mathematics there are 30 items available in each second stage module, while language skills contains only 17 items in each second stage module. The p-values of the item selections, which were attained in this way, for each subject, are presented in Table 2.4.

2.4.3 Item selection for the second and third test variant

For the second and third test variant, the item bank, as described in section 2.4.1, was used as a starting point for the item selection. To arrive at the item selection for the second and third variant, either a different order for the items in the item bank was used, or a subset of the items in the item bank was selected. For the second test variant, the item selection was obtained by reordering the items in the item bank. This procedure is illustrated in Figure 2.3. The procedure was performed as follows:

1. The item parameters from the item bank are ordered by discrimination parameters, from high to low.
2. From these reordered items, items 1 until and including $n_{stage\,1}$ were selected for stage one.
3. From the remaining items, items $n_{stage\,1} + 1$ until and including *N* were selected, and ordered by location parameters. For stage two, the odd items were selected, and for stage three, the even items were selected.

In this way, the items with the lowest location parameters ended up in module one of stage two and three, and the items with the highest location parameters ended up in module three of the stage two and

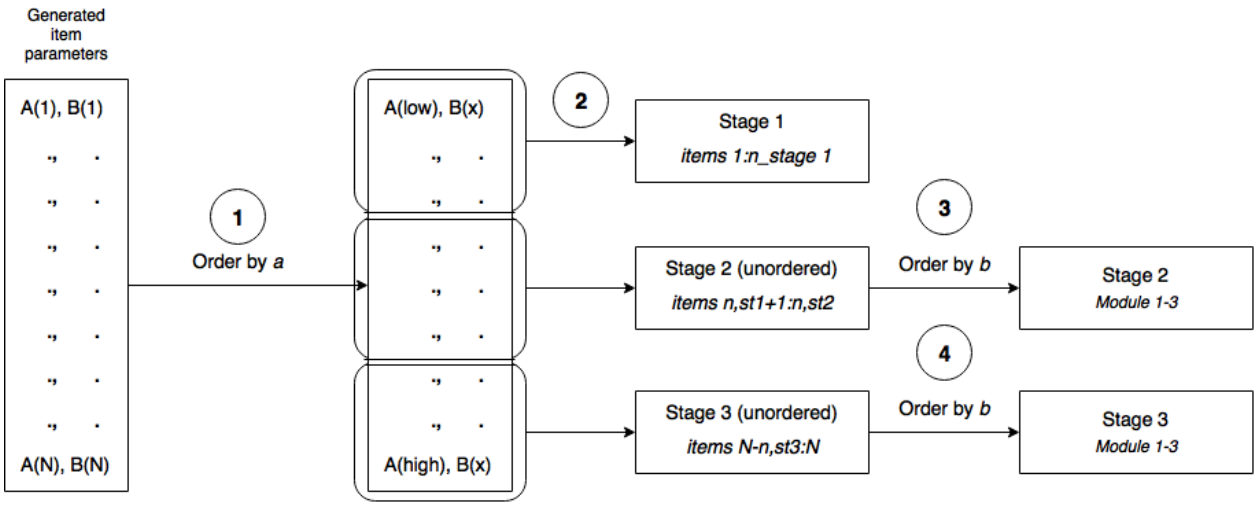

Figure 2.2. Overview of the steps taken in item bank creation

Table 2.4

The Average p-value per Module for the Test Variant with Highly Discriminatory Items in Stage Two and Three

Subject	Stage 1	Stage 2 Module 1	Stage 2 Module 2	Stage 2 Module 3	Stage 3 Module 1	Stage 3 Module 2	Stage 3 Module 3
Reading	0.6899	0.6070	0.5895	0.5970	0.5751	0.6216	0.6293
Mathematics	0.6818	0.6346	0.6096	0.5882	0.5848	0.5944	0.6301
Language Skills	0.7059	0.5803	0.6075	0.5947	0.6375	0.5914	0.6254

Note. The p-values in the stage two and three modules are for the intended target population of those modules.

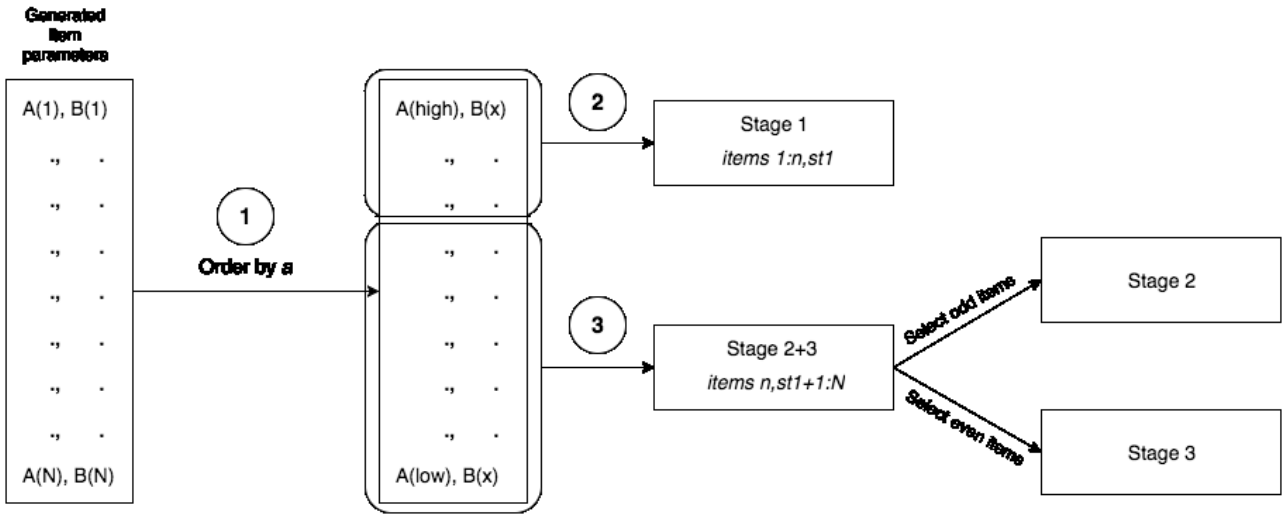

Figure 2.3. Overview of the item selection procedure for the second test variant

three. The p-values of the item selections, which were obtained in this way, for each subject, are presented in Table 2.5.

For the third test variant, a subset of the items in the item bank was selected. In stage one, the same items as in stage one of the item bank were selected. For stage two and three, ideally a random sample from the items in stage two and three of the item bank would have been obtained. However, a sample of items obtained in this way resulted in p-values that were too low. Therefore, another method was used to select the items for stage two and three from the item bank. The sample of the required amount of items was instead taken from the first two modules of stage two and three from the item bank. This procedure was repeated 100 times for each subject in order to end up with items with average p-values that come closest to the desired p-values. The p-values of the item selections, which were attained in this way, for each subject, are presented in Table 2.6.

2.4.4 Item selection for vocabulary and writing

In all variants of the test, there were two subjects (i.e., vocabulary and writing) that were tested in a linear way. As such, the item selection for these two subjects was shared across test variants. In order to select the items for these two subjects, a stratified proportional random sample of the required amount of items was taken from all items that are available for the subject. The strata were (a) items from the Basis version of the Centrale Eindtoets and (b) items from the Niveau version of the Centrale Eindtoets. For each subject, fifteen items from the Basis version and five items from the Niveau version were selected. By using these proportions, items with an average p-value close to the desired value of $p_{total\ population} = 0.70$ were obtained. The p-values that were attained in this manner are presented in Table 2.7.

2.4.5 Routing procedure

Routing in the multistage version occurs after the first stage and after the second stage. In the current test design there are four routing points, as shown in Figure 2.4. In this figure, the letter R followed by a number represents a routing point. In the simulated tests, routing is performed by calculating the unweighted sum score of the items that have been answered up to that point, and selecting the path that provides the most information, given this unweighted sum score. As there are several paths a simulee can take through the MST, not every simulee responds to the same items. As such, sum scores are not comparable across simulees. Furthermore, the item information formula, as presented in section 1.3, takes the ability of the simulee, and not the sum score, as an argument. Therefore, a list with the following information per routing point had to be generated:

(a) the routing point from which a routing decision is made,
(b) all possible unweighted sum scores for this routing point,
(c) ability estimates corresponding with those unweighted sum scores, and
(d) the optimal paths given the ability estimates in (c).

This list contains each point in the MST from which a routing decision is made. To generate this list, every routing point for each domain was determined. Afterwards, the following procedure was performed for each routing point:

1. The item parameters of the items up until the routing point were collected.
2. Using these item parameters, for each possible unweighted sum score, the corresponding ability was estimated.
3. The ability estimates were used to determine which path through the test provides the most information for each possible unweighted sum score. When determining the best path after stage one, the routing decision after stage two was also considered. This is because the purpose of routing after stage one is not to select the optimal stage two module, but to find the path through the test that provides the most accurate ability estimate for the simulee at the end of the test. Therefore, the path through stage two and three with the most information for the ability estimate after stage one is chosen. When determining the best path after stage two, the path is selected that leads to the stage three module that provides the most information based on the ability estimate after stage one and two.

This procedure results in a list of unweighted sum scores and optimal path combinations, for each routing point, per subject. Because the items that are presented until each routing point differ across the multistage test variants, two versions of the list had to be generated. The first version of the list was

Table 2.5

The Average p-value per Module for the Test Variant with Highly Discriminatory Items in Stage One

Subject	Stage 1	Stage 2 Module 1	Stage 2 Module 2	Stage 2 Module 3	Stage 3 Module 1	Stage 3 Module 2	Stage 3 Module 3
Reading	0.7440	0.5983	0.5958	0.6081	0.5782	0.5821	0.6067
Mathematics	0.6662	0.6363	0.6127	0.5974	0.6197	0.6130	0.5994
Language Skills	0.64027	0.6348	0.6169	0.6312	0.6341	0.6233	0.6329

Note. The p-values in the stage two and three modules are for the intended target population of those modules.

Table 2.6

The Average p-value per Stage for the Linear Variant of the Test

Subject	Stage 1	Stage 2	Stage 3
Reading	0.6895	0.6986	0.7050
Mathematics	0.6818	0.7090	0.6966
Language Skills	0.7058	0.6969	0.7036

Table 2.7

The Average p-Value per Stage for the Total Population of Simulees, for the Subjects That Are Tested in a Linear Way

Subject	Stage 1	Stage 2	Stage 3
Vocabulary	0.6884	0.7220	0.7027
Writing	0.7086	0.7022	0.7176

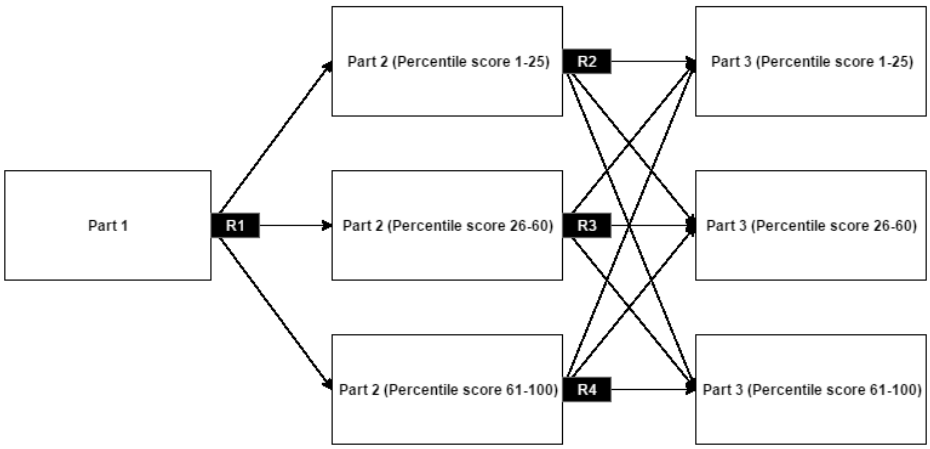

Figure 2.4. Overview of each point in the MST where are routing decision is made. The routing points are denoted by R, followed by a number.

used for the test variant with high discrimination parameters in stage one, and the second version was used for the test variant with high discrimination parameters in stage two and three.

2.4.6 Classification methods

For the third analysis in the present study, in which the influence of different classification methods on the classification accuracy is investigated, three methods to make this classification decision were used: the sequential probability ratio test (SPRT), the simulee's ability estimated after the test under the Rasch model, and the summed probabilities of correct responses on all items in the MST given the simulee's estimated ability. These methods are described in more detail in theoretical framework, under section 1.8. For the other analyses in the present study, only true classification is used, as described in section 2.4.7.

Each classification method needs its own set of classification cut-off points to perform the classification into one of the eight categories. The procedures to calculate these cut-off points are described. For the first method, the SPRT, the cut-off points had to be determined for all five subjects in the simulated tests. Furthermore, values for α, β, and δ had to be determined. After comparing several values for α, β, and δ on the resulting amount of forced decisions and the resulting PCD, it was decided to use both α=β=0.05 and α = β = 0.1 for classification with SPRT for all test domains. Regarding the indifference region, $\delta = 0.09 \cdot sd_{ability\ scale}$ was used for the test domains reading, mathematics, language skills, and writing, while $\delta = 0.015$ was used for the test domain vocabulary. For information on how the settings of the SPRT were determined, the reader is referred to appendix A.

For the second method, the summed probabilities of correct responses on all items in the MST method, the ability cut-off points of the SPRT are used as a starting point. For each ability cut-off point, the corresponding expected score on all items in the MST is calculated for each test domain. Lastly, these sum scores per test domain are summed:

$$\sum_{sum\ scores} = \sum_{reading} + \sum_{mathematics} + \sum_{language\ skills} + \sum_{writing} + \sum_{vocabulary}$$

This process results in the cut-off points for this method.

For the third classification method, which uses the simulee's ability estimated after the test with the Rasch model, the items in the MST first have to be modeled under the Rasch model. In order to do so, a response set for 100.000 simulees was generated using the observed multivariate ability distribution of the Centrale Eindtoets 2015 and the item parameters of the simulated test.

Using this response set, the item parameters of the items from the MST are estimated with Rasch. With the same response set, and the Rasch item parameters, the distribution of the abilities of the population is calculated. In practical terms, modelling the items from the MST under Rasch means that one is no longer looking at the ability of a pupil on each individual test domain. Instead, one is looking at the ability of the pupil on the whole MST. Thirdly, the item parameters and the mean and standard deviation of the ability distribution are used to calculate the ability for each sum score. Finally, the abilities corresponding with the second classification method's sum score cut-off points are selected as cut-off points for this method.

2.4.7 True classification

To determine whether the correct classification decisions were made in the simulated tests, a benchmark for the classifications is needed. In the present study, this benchmark was based on the classification method that is used in the Centrale Eindtoets test administration. This classification method, known as the standard score classification method, is described in section 2.3.2. The standard score classification method is used in the first, second and fourth analysis in the present study. In the third analysis, it serves as an upper limit for the classification methods as described in section 2.4.5.

2.4.8 Simulation results

The simulation will provide two results: a) the root mean square error (RMSE) of the pupil's abilities (a measure of the precision of the ability estimates that the simulation produces, see equation 1.9), and b) the proportion of correct decisions (PCD) for the classifications (a measure for the accuracy of the classification decisions). A correct classification is made when the classification decision from a classification method is equal to the classification decision that is made by the true classification method.

A variant of the PCD was calculated, based on overlapping educational levels. Table 2.3 shows all possible classification levels in the Centrale Eindtoets. However, there is overlap between some of the classification levels in this table. For example, an advice for Havo is both in the sixth and seventh

classification level. Therefore, it could be argued that a pupil who is classified as Havo, with a true classification of Havo/vwo, is in fact not incorrectly classified. With this rationale in mind, the second variant of the PCD only counts classifications as a misclassification, when the classification level that a pupil was assigned to does not contain the pupil's true school advice. Table 2.8 presents the classification levels with overlap.

Table 2.8

Classification Levels with Overlap

Classification Level	Has overlap with
Vmbo-bb	Vmbo-bb/kb
Vmbo-bb/kb	Vmbo-bb and Vmbo-kb
Vmbo-kb	Vmbo-bb/kb
Vmbo-gt	Vmbo-gt/havo
Vmbo-gt/havo	Vmbo-gt and Havo
Havo	Vmbo-gt/havo and Havo/vwo
Havo/vwo	Havo and Vwo
Vwo	Havo/vwo

3 Results

In this study, the linear and multistage version of the Centrale Eindtoets were compared. Specifically, four test variants were compared:

- a variant in which simulees respond to all items of the multistage version,
- a variant of the multistage version with highly discriminatory items in the second and third stage (MST variant 1),
- a variant of the multistage version with highly discriminatory items in the first stage (MST variant 2),
- a linear test variant.

With these four test variants, four analyses were performed. Firstly, a comparison was made between the multistage test variants and the linear test variant on precision of the ability estimates. Secondly, the differences between the multistage test variants and the linear test variant in terms of classification accuracy were examined. Furthermore, the effects of four different classification methods on classification accuracy of the linear and multistage variants were studied. Lastly, the influence of the different composition of the two multistage test variants on the precision of the ability estimates and the classification accuracy were compared. The results of these four analyses are discussed in the sections 3.1, 3.2, 3.3 and 3.4. The discussion of the results is presented in section 3.5.

3.1 Precision of the Ability Estimates

The first question in this study was how the linear and multistage version of the Centrale Eindtoets compare on precision of the ability estimates. In order to answer this question, the RMSE was calculated for each test domain in each test variant. The results of this comparison are presented in Table 3.1. In this comparison, the test variant with all items from the multistage test is the benchmark test. The RMSEs for writing and vocabulary are identical across all test variants, because all variants shared the same test items for these two domains. For the domains reading, mathematics, and language skills, the multistage test variants produced the best RMSE values. In other words, the best measurement precision is achieved with the multistage test variants.

3.2 Classification Accuracy

The second question in this study was how the linear and multistage version of the Centrale Eindtoets compare on classification accuracy. In order to answer this question, the PCD was calculated for each

Table 3.1

RMSEs for the Four Test Variants, per Domain

	RMSE			
Subject	All Items	MST Variant 1	MST Variant 2	Linear
Reading (45 items)	0.0835	0.0997	0.0954	0.1134
Mathematics (85 items)	0.0644	0.0730	0.0711	0.0905
Language Skills (50 items)	0.0797	0.0944	0.0894	0.1060
Writing (20 items)	0.2197	0.2197	0.2197	0.2197
Vocabulary (20 items)	0.2048	0.2048	0.2048	0.2048

test variant using the standard score classification method. The result of this comparison is presented in Table 3.2.

The test variant with all items from the multistage test is the benchmark test in this comparison. When comparing the other three test variants, the multistage test variants have a higher PCD than the linear test variant. In other words, the multistage test variants produced the best classification accuracy.

As the classification categories in the Centrale Eindtoets overlap, it could be argued that the PCD, as presented in Table 3.2, does not represent the actual amount of correct decisions. Instead, one could argue that a classification decision is only a misclassification when the level of secondary education that a pupil should be assigned to is not contained in the classification category for this pupil. For example, when a pupil should be assigned to the Havo classification category, but is classified as Havo/Vwo, it could be argued that the classification decision is still correct. The reader is referred to section 2.4.8 for a description of how the classification categories based on overlapping educational levels are composed. The results of the comparison between the linear and multistage version of the Centrale Eindtoets, without the overlapping classification categories, is shown in Table 3.3. In this table, the same overall trend as in Table 3.2 can be observed. Again, the multistage test variants produced the best classification accuracy.

3.3 Influence of the Classification Method

The third question in this study pertains to the effects of four different classification methods on classification accuracy of the linear and multistage version of the Centrale Eindtoets. In order to answer this question, the PCD was calculated for each test variant using all classification methods: (a) the standard score classification method, (b) the sum of the estimated probability on all items classification method, (c) the Rasch classification method, and (d) the SPRT. For the SPRT, two different settings were used: (a) $\alpha = \beta = 0.1$, and (b) $\alpha = \beta = 0.05$. In both cases, the δ for each test domain was 0.09, multiplied by the standard deviation of the ability distribution of the test domain. This results in a value for δ that is on the same scale as the ability distribution of the test domain.

An exception was made for vocabulary, as this setting for δ would have resulted in overlapping indifference regions. Therefore, the δ for vocabulary was set to 0.015. For the exact configuration of the indifference regions, the reader is referred to Appendix A. The results of this comparison is presented in the second section of Table 3.4. The full classification tables, which contain the PCD per classification category for each test variant, can be found in Appendix B.

In order to answer the third question in this study, it is more intuitive to look at the proportion of incorrect decisions (PID), which is the inverse of the PCD. For an ideal test, one wants zero incorrect classification decisions, or a PID of 0. In this comparison, the test variant with all items of the multistage test is the benchmark. Regardless of the test variant, the sum of the estimated probability on all items classification method consistently had the lowest PID. However, when looking at the PCD per classification category, this finding no longer holds, as shown in Appendix B. In this table, it is shown that, in each classification category, a different classification method results in the highest PCD. In sum, the highest overall classification accuracy was obtained with the sum of the estimated probability on all items method. However, when looking at the highest classification accuracy per classification category, no classification method consistently produced the best results. As in section 3.2, the PID for this research question was also calculated without the overlapping classification categories. The results of

Table 3.2

PCD for Each Test Variant, Using the Standard Score Classification Method

All Items	MST Variant 1	MST Variant 2	Linear
0.7428	0.7109	0.7210	0.6855

Table 3.3

PCD Without Overlapping Classification Categories, for Each Test Variant, Using the Standard Score Classification Method

All Items	MST Variant 1	MST Variant 2	Linear
0.9606	0.9527	0.9550	0.9452

the comparison between the four classification methods, without the overlapping classification categories, is shown in the third section of Table 3.4. In this table, the same overall trend as in the second section of Table 3.4 can be observed. Again, the highest overall classification accuracy was obtained with the sum of the estimated probability on all items method.

3.4 Influence of the MST Design

The last question that was answered is this study pertains to the influence of two different module designs on the precision of the ability estimates and the classification accuracy of the multistage version of the Centrale Eindtoets. In order to answer this question, the PCD was calculated for the two multistage test variants, using the standard score classification method, and the RMSE was calculated for each test domain in the two multistage test variants.

The comparison in terms of the precision of the ability estimates is presented in Table 3.5. The RMSEs for writing and vocabulary are identical across the two test variants, because all test variants shared the same test items for these two domains. For the domains reading, mathematics, and language skills, the RMSE values for the second multistage variant are lower than for the first multistage variant. In other words, better measurement precision is achieved in the multistage test variant with the highly discriminatory items in the first stage for the domains reading, mathematics, and language skills.

The comparison between the two multistage test variants in terms of classification accuracy is presented in Table 3.6. When looking at the classification accuracy, the second multistage variant has a higher PCD than the first multistage variant. In other words, higher classification accuracy is achieved in the multistage test variant with the highly discriminatory items in the first stage.

As in section 3.2, the PCD for this research question was also calculated without the overlapping classification categories. The results of the comparison between the two multistage test variant, without the overlapping classification categories, are shown in Table 3.7. When the PCD without overlapping classification categories is considered, the highest classification accuracy was again obtained with the second multistage test variant.

3.5 Discussion of the Results

The main goal of this study was to investigate the difference between the multistage and linear version of the Centrale Eindtoets in terms of measurement precision and classification accuracy. Secondly, the effect of different classification methods on classification accuracy was studied. Lastly, the influence of different module design on the measurement precision and classification accuracy of the multistage the Centrale Eindtoets was examined. The results from this study show that a multistage version of the Centrale Eindtoets outperforms the linear version of the Centrale Eindtoets on both measurement precision and classification accuracy. Furthermore, the sum of the estimated probability on all items classification method consistently provides the highest classification accuracy, regardless of the test variant. Finally, the second variant of the multistage test – with highly discriminatory items in the first stage – outperforms the first variant of the multistage test, both in terms of measurement precision and classification accuracy.

Table 3.4

PCD for Each Test Variant, Using All Classification Methods

	PCD			
Classification Method	All Items	MST Variant 1	MST Variant 2	Linear
Standard Score	0.7428	0.7109	0.7210	0.6855
Sum of the Estimated Probability on All Items	0.7266	0.6937	0.7014	0.6630
Rasch	0.7238	0.6862	0.6973	0.6502
SPRT ($\alpha = \beta = 0.1$)	0.7158	0.6805	0.6946	0.6528
SPRT ($\alpha = \beta = 0.05$)	0.7158	0.6530	0.6808	0.5707
	Proportion Incorrect Decisions With Overlapping Classification Categories			
Classification Method	All Items	MST Variant 1	MST Variant 2	Linear
Standard Score	0.2178	0.2418	0.234	0.2597
Sum of the Estimated Probability on All Items	0.2338	0.2577	0.2531	0.2803
Rasch	0.2362	0.2643	0.2543	0.2911
SPRT ($\alpha = \beta = 0.1$)	0.2421	0.2673	0.2576	0.283
SPRT ($\alpha = \beta = 0.05$)	0.2421	0.2789	0.2633	0.3212
	Proportion Incorrect Decisions Without Overlapping Classification Categories			
Classification Method	All Items	MST Variant 1	MST Variant 2	Linear
Standard Score	0.0394	0.0473	0.045	0.0548
Sum of the Estimated Probability on All Items	0.0396	0.0486	0.0455	0.0567
Rasch	0.04	0.0495	0.0484	0.0587
SPRT ($\alpha = \beta = 0.1$)	0.0421	0.0522	0.0478	0.0642
SPRT ($\alpha = \beta = 0.05$)	0.0421	0.0681	0.0559	0.1081

Table 3.5

RMSEs for the Two Multistage Test Variants, per Domain

	RMSE	
Domain	MST Variant 1	MST Variant 2
Reading (45 items)	0.0997	0.0954
Mathematics (85 items)	0.0730	0.0711
Language Skills (50 items)	0.0944	0.0894
Writing (20 items)	0.2197	0.2197
Vocabulary (20 items)	0.2048	0.2048

Table 3.6

PCD for the Two Multistage Test Variants, with the Standard Score Classification Method

MST Variant 1	MST Variant 2
0.7109	0.7210

Table 3.7

PCD Without Overlapping Classification Categories, for the Two Multistage Test Variants, with the Standard Score Classification Method

MST Variant 1	MST Variant 2
0.9527	0.9550

4 Conclusion

At the end of the primary school, pupils take a test that provides them with an independent advice on the level of secondary education that is most suitable for their ability. An example of such a test is the Centrale Eindtoets. The the Centrale Eindtoets classifies pupils into one of eight partly overlapping levels that correspond with the Dutch levels of secondary education. This classification is dependent on the pupils' performance on the test. Furthermore, the Centrale Eindtoets provides pupils with a score representing their ability. To meet these requirements, the Centrale Eindtoets should achieve high classification accuracy and give precise ability estimates. As the Centrale Eindtoets is currently administered in a linear format, and thus every pupil has to respond to the same items, pupils are likely to respond to items that are either too easy or too difficult. These items will contribute less than optimal to a precise measurement of their ability. This reduced measurement precision increases the chance of misclassifications, especially for pupils with an ability near a classification cut-off point.

Currently, Stichting Cito is developing an adaptive version of the Centrale Eindtoets. One of the advantages of this adaptive the Centrale Eindtoets is an increase in measurement precision. In this version of the Centrale Eindtoets, pupils first receive a block of items to obtain an initial estimate of their ability. Based on this initial ability estimate, pupils are routed to one of three blocks of items that best suits their ability. After this second block of items, the pupil's ability estimates are updated, and the final block of items most suitable for their ability is presented. This type of testing is called multistage testing. Although the advantage of multistage testing over linear testing is described in literature, choose to the exact amount of influence the test design has on the measurement precision and the classification accuracy of the Centrale Eindtoets was unknown. Therefore, the aim of the study was to compare the multistage and linear version of the Centrale Eindtoets on measurement precision and classification accuracy. The influence of different configurations for the multistage test on measurement precision and classification accuracy was also investigated. Lastly, the effect of three classification methods on classification accuracy was compared.

In order to test the different version of the Centrale Eindtoets with a representative sample size, the different versions must be put into the actual test administration. However, this procedure is unethical, as there could be a difference in measurement precision and classification accuracy between the different versions. As a result, some pupils might be put at a disadvantage when they are not assigned to the best version. Therefore, a simulation study was used to make a comparison between the different version of the Centrale Eindtoets. The ability parameters in this study were based on the ability distribution of the five test domains of the Centrale Eindtoets 2015: reading, mathematics, language skills, writing and vocabulary. The items parameters in this study were generated to adhere to the specifications for the adaptive the Centrale Eindtoets 2018, which will be a multistage test. Two variants of the multistage test were developed. In the first version, the highly discriminatory items were placed in the first stage of the multistage test. In the second version, the highly discriminatory items were place in the second and third stage of the multistage test. Secondly, a linear variant of the test was developed based on a selection of items from the multistage test. This linear variant enabled the comparison between the multistage and linear version of the Centrale Eindtoets. Lastly, a test variant in which pupils respond to all items in the multistage test was developed. This variant provides information on the maximum amount of improvement that can be made by using the items from the multistage test, when compared to the linear test. To determine the precision of the ability estimates in the test, the RMSE per test domain was calculated for each test variant. Furthermore, the classification accuracy was measured with the proportion of correct classification decisions. To make these classification decisions, three

classification methods were used: (a) the sequential probability ratio test, (b) the estimated ability classification method using the Rasch model, and (c) the sum of probabilities on all items in the test method.

The results from this study show that the multistage versions of the Centrale Eindtoets outperforms the linear version of the Centrale Eindtoets on both measurement precision and classification accuracy. For example, when looking at measurement precision for the mathematics domain across the test variants, the RMSE of the linear variant is 0.0905, while the RMSE of the first and second MST variant is 0.0730 and 0.0711, respectively. This finding is in line with research stating that multistage tests have greater measurement accuracy than linear tests at the same test length (e.g. Yan et al., 2014). In terms of classification accuracy, the PCD for the linear the Centrale Eindtoets is 0.6855, compared to 0.7109 and 0.7210 for the first and second MST variant. Again, this finding is in line with research, which states that opting for multistage testing over linear testing leads to an increase in the amount of information available for measurement, which in turn decreases the error associated with classification decisions (Weissman, 2014).

Furthermore, the sum of the estimated probability on all items classification method consistently provides the highest classification accuracy, regardless of the test variant. For example, when comparing the different classification methods for the first variant of the MST, the sum method has a PCD of 0.6937, compared to a PCD of 0.6862, 0.6805 and 0.6530 for the Rasch method, the SPRT with $\alpha = \beta = 0.1$, and the SPRT with $\alpha = \beta = 0.05$, respectively.

The fact that the sum method outperforms the Rasch method can be explained by the fact that, for the Rasch classification method, the test items were modeled under Rasch. By contrast, under Rasch the discrimination parameter cannot be specified. As such, modeling the test items under Rasch causes a loss of information, which explains why the sum method outperforms the Rasch method.

Next to that, the fact that the SPRT produces the least favorable PCD out of all classification methods used in this study can be explained by the specific classification requirements for the Centrale Eindtoets. Firstly, pupils are classified into one of eight categories, which results in classification cut-off points that are located close together. For classification with the SPRT this is less than ideal, as it limits the maximum size of the indifference region. This limitation can in turn lead to a high amount of forced decisions, which negatively affects the classification accuracy. Secondly, as the Centrale Eindtoets contains multiple test domains, classification with the SPRT requires a separate set of cut-off points for each test domain. This makes it much harder to configure the cut-off points for the highest possible classification accuracy, when compared to classification methods that only require one set of cut-off points for all test domains. Thirdly, as the SPRT for multiple domains originally used a multidimensional IRT model, its classification accuracy is limited here by the use of multiple unidimensional models.

The difference in PCD between the two settings of the SPRT can be explained by the fact that the decreasing the value for α increases the amount of forced decisions from 82.29% to 85.70% for this particular case. As a result, the SPRT with $\alpha = \beta = 0.1$ provides a higher classification accuracy than the SPRT with $\alpha = \beta = 0.05$.

Finally, the second variant of the multistage test – with highly discriminatory items in the first stage – outperforms the first variant of the multistage test, both in terms of measurement precision and classification accuracy. For example, when comparing the RMSE for the mathematics domain across the MST test variants, the RMSE of the first variant is 0.0730, while the RMSE of the second variant is 0.0711. This confirms the importance of the characteristics of the first stage for the measurement precision of the whole test, as stated by Kim and Plake (1993). In terms of classification accuracy, the PCD for the first MST variant is 0.7109, compared to 0.7210 for the second MST variant.

4.1 Limitations

Based on the results from the study, one can conclude that the adaptive the Centrale Eindtoets in 2018 will indeed be an improvement, when compared with a linear the Centrale Eindtoets. However, this does not mean that the adaptive the Centrale Eindtoets is indeed better than the current linear the Centrale Eindtoets. The present study is based on simplified case, in which not all specifications of the final adaptive the Centrale Eindtoets are taken into account. On the other hand, the present study was performed with a representative amount of simulees, which inspires confidence in the results presented in the study. There are some limitations to this study.

Firstly, in the current test design, there are eight classification categories. As shown in the results section, having a high amount of classification categories negatively affects classification accuracy. When the amount of classification categories is reduced to, for example, five levels, as is the case with the classification categories without overlapping categories, classification accuracy goes up. Therefore, the high amount of classification categories in present test design is a limitation in the study.

As a result of the high amount of classification categories, the cut-off points separating the classification categories are very close together. In case of classification using the SPRT method, this causes over 75% of the classification decisions to be forced. The exact percentage of forced classification decisions for each test variant is presented in Table 4.1. When forcing a classification decision using the SPRT method, the method does not adhere to the predefined rates of type I and type II error that are seen as acceptable. In practical terms, adhering to the predefined rates of type I and type II error would mean that the response data from the MST can be used to provide the majority of the pupils with an advice on the level of secondary school they should attend, with reasonable certainty. As a result, the amount of misclassifications will be higher than what is deemed acceptable. This is also a possible explanation for the fact that the SPRT performs the worst out of all of classification methods. Again, this is another indicator that the high amount of classification categories is a limitation of the present test design.

Secondly, instead of using existing, actual test items, parameters for the items in the item bank of the present test design were generated. This was done to construct items with P-values that are suitable for the test design used in the present study. However, in constructing the new items, not all requirements that will be present in the test design of the Centrale Eindtoets 2018 were considered (see next paragraph). If the P-values in the present study are not the same as in the final adaptive the Centrale Eindtoets because of these in a more strict requirements, the classification accuracy in the final adaptive the Centrale Eindtoets might differ from the present study. Furthermore, the actual test items that formed the basis of the generated items in this study have not been designed with an MST in mind. Therefore, classification accuracy might be improved by using items from the test version of the Centrale Eindtoets 2018, which have been designed with an MST in mind. Therefore, working with virtual test items might be a limitation of this study.

Thirdly, one of the goals in designing the adaptive version of the Centrale Eindtoets is to maximize the measurement precision of the test. When looking at the kinds of adaptive testing that are available, CAT offers the highest level of measurement precision, because CAT tests are the most adaptable to the ability of the pupil (Jodoin, Zenisky, & Hambleton, 2006). In this regard, it is a limitation of this study, and in the design of the adaptive version of the Centrale Eindtoets, that MST was chosen over CAT. On the other hand, MSTs offers some advantages over CATs. For example, in MST tests it is possible for pupils to go back to an item in order to change the answer, given the item exists in the current module (Jodoin et al., 2006). By contrast, this behaviour would not be possible in CAT, because changing the answer of a previous question might change the selection of the next questions, thereby invalidating some answers that have already been given. In short, opting for MST over CAT means a compromise in measurement precision in return for a more practical test taking procedure. Future research might investigate how big this compromise is.

Lastly, maximum likelihood estimation was used in the present study to estimate the ability of the simulees at the end of the simulated tests. However, it is known that the ability estimates as estimated by maximum likelihood estimation contain bias to the order n^{-1}. This bias can be reduced to the order n^{-2} by opting for weighted maximum likelihood estimation (Warm, 1989). Another disadvantage of maximum likelihood estimation is the fact that this method is unable to estimate abilities when all items are answered correct or incorrect. When these extreme ability estimations are used in the calculation of the RMSE, the resulting value will be overstated. In the present study, this problem was solved by omitting simulees with a zero or maximum score for a test domain, when calculating the RMSE. This results in an RMSE value that is more representative of the measurement precision capabilities of the test version under study. To overcome this limitation, an ability estimation method that is capable of estimating abilities at the extreme ends of the range should be employed. Suggestions for such an ability estimation method are the weighted likelihood method.

Table 4.1

Percentage of Forced Classification Decisions per SPRT Variant, for each of the Test Variants

	PCD			
Classification Method	All Items	MST Variant 1	MST Variant 2	Linear
SPRT ($\alpha = \beta = 0.1$)	76.87%	82.29%	81.98%	86.24%
SPRT ($\alpha = \beta = 0.05$)	83.10%	85.70%	85.27%	90.08%

4.2 Directions for future research

The results of the present study gives rise to new questions for future research. Firstly, the MST design as employed in the present study will not be used for the adaptive version of the Centrale Eindtoets from 2020 onwards. The present MST design contains one module in the first stage, followed by two stages with three modules each. However, the adaptive the Centrale Eindtoets from 2020 onwards will contain either (a) three modules in stage two, and five modules in stage three, or (b) five modules in stage two, and five modules in stage three. With this in mind, the influence of a test design, with more modules, on the measurement precision and classification accuracy could be investigated.

Secondly, one of the reasons for switching to digital testing is the fact that new, more authentic item formats, which can be scored automatically, can be used. However, those new item formats are not part of the present study. Therefore, the influence of this new item formats could be investigated.

Thirdly, the present study considers only one module in first stage. One of the reasons for switching to an adaptive the Centrale Eindtoets is the fact that this way of testing is more suitable for pupils who attend special education then the future one-level linear the Centrale Eindtoets. However, having only one module in the first stage might not be optimal for this target population. This is because the items in the first stage might be too difficult for this population. Therefore, the effects of having two modules in the first stage on the precision of the ability estimates of this population can form the basis of a new research topic.

Fourthly, the test domains vocabulary and writing have a RMSE value that is meaningfully higher than for a test domains reading, mathematics, and language skills. As such, these two test domains might have a negative impact on the precision of the ability estimates in this study, which could negatively affect classification accuracy. Furthermore, it can be investigated whether the current number of items for vocabulary and writing is sufficient. A higher number of items might improve the classification accuracy and measurement precision

Fifthly, the current the Centrale Eindtoets optionally contains the test domain environmental studies. However, this test domain is not considered in the present study because this test domain does not contribute go off to future research to the standard score. Therefore, it is of interest to investigate what the optimal test design is for an adaptive version of the Centrale Eindtoets that includes the test domain environmental studies.

Sixthly, in the present study, most classification methods classify pupils on the basis of a single score. When converting the pupils scores on the individual test domains into a single score, some information is lost. Therefore, future research could investigate alternative classification methods that do not have this problem.

Seventhly, the effects of the routing configuration that was chosen in the current study could be further investigated. It is of interest to look into the effects of the routing configuration on the utilization of the paths between the stages, as well as its effect on measurement precision and classification accuracy. For example, in this study, it was decided to route pupils based on certain target p-values, without considering if this would result in the optimal routing in terms of measurement precision and classification accuracy. Furthermore, recall that routing decisions are based on the unweighted sum score of the items that have been answered up to that point by a pupil, as described in section 2.4.5. The results of this study show that placing the highly discriminatory items in the first stage of the test yields higher the measurement precision and classification accuracy, compared to placement of the highly discriminatory items in the second and third stage of the test. This indicates that opting for a routing configuration in which routing decisions are based on the weighted sum score might be more optimal.

Lastly, the present study makes use of unidimensional IRT. When using unidimensional IRT, ability estimates for different test domains are independent. Conversely, when using multidimensional IRT , updating the ability estimate for one test domain influences the ability estimates the other test domains (Segall, 1996). This process could result in more accurate ability estimates. Therefore, the influence of modelling the items in the present study under multidimensional IRT could be investigated.

5 Reference List

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized Adaptive Testing With Multiple-Form Structures. *Applied Psychological Measurement*, *28*(3), 147–164. http://doi.org/10.1177/0146621604263652

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: CRC Press.

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation*, *18*(14), 1–7. http://doi.org/10.4324/9780203874776.ch27

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

CvTE. (2015). Verantwoording centrale eindtoets po [Justification central final test po]. Retrieved from https://www.hetcvte.nl/document/verantwoording_centrale_eindtoets

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249–261. http://doi.org/10.1177/01466219922031365

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, *65*(3), 337–362. http://doi.org/10.1007/BF02296150

EP-Nuffic. (2015). Education system The Netherlands. Retrieved from https://www.epnuffic.nl/en/publications/find-a-publication/education-system-the-netherlands.pdf

Fischer, H. G. (1995). Derivations of the Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York, NY: Springer.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, *23*(2), 165–170. http://doi.org/10.1016/0167-7152(94)00109-L

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*. http://doi.org/10.1097/01.mlr.0000245426.10853.30

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

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

Huff, K. L., & Sireci, S. G. (2005). Validity Issues in Computer-Based Testing. *Educational Measurement: Issues and Practice*, *20*(3), 16–25. http://doi.org/10.1111/j.1745-3992.2001.tb00066.x

Jehangir, K. (2005). *Evaluation of relations between scales in an IRT framework*. University of Twente.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the Psychometric Properties of Several Computer-Based Test Designs for Credentialing Exams With Multiple Purposes. *Applied Measurement in Education*, *19*(3), 203–220. http://doi.org/10.1207/s15324818ame1903_3

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in r. *Journal of Statistical Software*, *20*(10), 1–24. Retrieved from http://www.jstatsoft.org/v20/a10/paper\npapers2://publication/uuid/1D92ECF7-56E7-4CF8-91FB-BA008986BAB2

Kim, H., & Plake, B. S. (1993). Monte Carlo Simulation Comparison of Two-Stage Testing and Computerized Adaptive Testing. Paper presented at the annual meeting of the National Council

on Measurement in Education, Atlanta, G.A.

Linacre, J. M. (2000). Computer-Adaptive Testing : A Methodology Whose Time Has Come. *Development of Computerized Middle School Achievement Test*, (69).

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education*, *19*(3), 189–202. http://doi.org/10.1207/s15324818ame1903_2

Luecht, R., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–259. http://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Oakland, T., & Hambleton, R. K. (1995). *International perspectives on academic assessment*. Boston, MA: Kluwer Academic.

Rijksoverheid. (2016). Verplichte eindtoets basisonderwijs [Mandatory final test primary education]. Retrieved February 13, 2016, from https://www.rijksoverheid.nl/onderwerpen/toelating-middelbare-school/inhoud/verplichte-eindtoets-basisonderwijs

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. http://doi.org/10.1007/BF02294343

van Groen, M. M. (2014). *Adaptive testing for making unidimensional and multidimensional classification decisions*.

Veldkamp, B. P. (2014). Item pool design and maintenance for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 39–54). New York, NY: CRC Press.

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York, NY: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: One parameter logistic model*. Arnhem: Cito.

Wald, A. (1973). *Sequential analysis*. New York, NY: Courier Corporation.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. http://doi.org/10.1007/BF02294627

Weiss, D. J. (1985). Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology*, *53*(6), 774–789. http://doi.org/10.1037//0022-006X.53.6.774

Weiss, D. J., & Betz, N. E. (1974). Simulation studies of two-stage ability testing. Minneapolis: University of Minnesota, Department of Psychologym Psychometric Methods Program.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. http://doi.org/10.3354/cr030079

Yan, D., von Davier, A. A., & Lewis, C. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 1–20). New York, NY: CRC Press.

Appendices

Appendix A: SPRT Settings

For the first method, the SPRT, the cut-off points had to be determined for all five subjects in the simulated tests. To determine the cut-off points, response data from the Centrale Eindtoets 2015, containing the ability and the classification of each pupil, were used as a starting point. Firstly, the medians of the abilities belonging to each classification category were determined for every domain. Secondly, the midpoint of each two adjacent medians was determined, as illustrated in Figure A.1. To calculate the cut-off point between two categories, the median of the abilities in the first category – M1 in the figure – and the median of the abilities in the second category – M2 in the figure – is determined. Afterwards, the midpoint between these medians is selected as the cut-off point between two categories for the domain.

Next to the cut-off points, values for α, β, and δ have to be determined. Recall that in the SPRT, one works with the ratio between two likelihood functions as a test statistic. In this test statistic, one has to set an indifference region, δ, that is appropriate for the cut-off points under review. Setting an indifference region that is too narrow increases the chance of misclassifications, while an indifference region that is too broad increases the chance of forced decisions. Forced decisions can in turn lead to misclassifications. Additionally, when δ is too high, the indifference regions between the cut-off points will overlap, which causes unexpected results.

First, the value for δ was determined. δ should be related to the ability scale of each test domain. In order to achieve this, several methods exist. One of these methods is to determine a base value for δ, and to multiply this value by the standard deviation of the ability scale. The standard deviations for the ability scales of the five test domains are shown in Table A.1. In the case of this particular multistage test, a value of δ_{base}=0.09 resulted in indifference regions that are not too broad for most test domains. For test domains reading, mathematics, language skills, and writing, this resulted in indifference regions without overlap. However, for the test domain vocabulary, this resulted in a small overlap between indifference regions. To solve this, a fixed value of $\delta_{vocabulary}$=0.015 was chosen for vocabulary.

The process described above resulted in the values of δ as presented in Figure A.2. After having determined the maximum value for δ_{base} and $\delta_{vocabulary}$, two additional, more strict values for δ_{base}, 0.06 and 0.03, were also considered.

To arrive at the final values for α and β, two requirements were considered. First of all, for simplicity sake, it was decided to select α symmetrical value for α and β. Secondly, in literature, the values 0.05 and 0.1 are commonly used. With these requirements in mind, both values were selected for α and β. The process of determining α, β, and δ resulted in 9 possible combinations for the final settings of the SPRT. In order to select the final value for α, β, and δ, SPRT classification was performed with all possible combinations. For this comparison the multistage test variant with the highly discriminatory items in stage two and three was used. The results were compared on PCD and the amount of forced decisions. The result of this comparison can be found in Table A.2. As can be seen in table, making δ smaller causes an increase in the amount of forced decisions, while it decreases the PCD. For example, at δ=0.09 and α=β=0.1, the amount of forced decisions is 82288, and the PCD is 0.6805. By contrast, at δ=0.03 and α=β=0.1, the amount of forced decisions is 95365, and the PCD is 0.3006. Furthermore, it can be seen that decreasing α=β from 0.1 to 0.05 has the same effect, but to a small extent. For example, at δ=0.09 and α=β=0.1, the amount of forced decisions is 82288, and the PCD is 0.6805. By contrast, at δ=0.09 and α=β=0.05, the amount of forced decisions is 85697, and the PCD is 0.6530. For the final version of the SPRT, it was decided to pick a δ_{base} of 0.09, as this yielded the highest PCD, and lowest amount of forced decisions. Furthermore, the SPRT was run with both α=β=0.1 and α=β=0.05, to see what the effect of the two different settings was on the PCD and the amount of forced decisions for the different variants of the test.

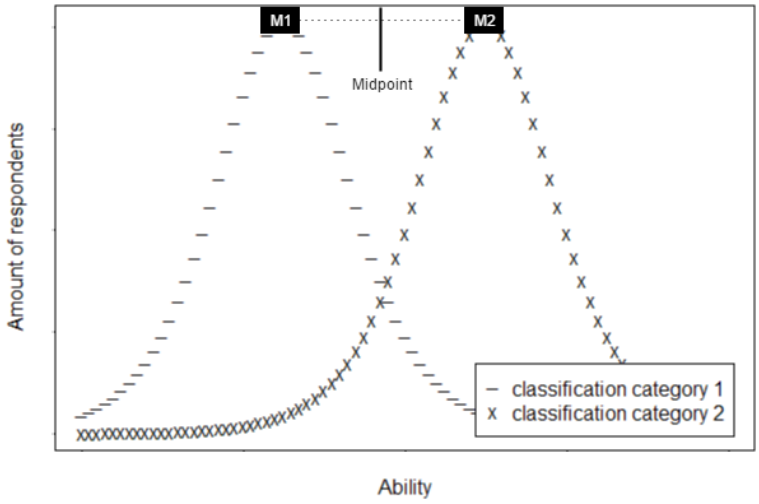

Figure A.1. Illustration of the procedure to determine the cut-off point between to classification categories. A *M* followed by a number denotes the median of the corresponding classification level.

Table A.1

PCD and the Amount of Forced Decisions for Different Values of δ_{base}, α and β

δ_{base}	α[a]	PCD	Forced Decisions[b]
0.09	0.1	0.6805	82288
0.09	0.05	0.6530	85697
0.06	0.1	0.6309	86918
0.06	0.05	0.4979	90879
0.03	0.1	0.3006	95365
0.03	0.05	0.2245	98085
0.09	0.1	0.6805	82288
0.09	0.05	0.6530	85697
0.06	0.1	0.6309	86918

Note. $\delta_{vocabulary}$ is 0.015 for every entry. The SPRT procedure was performed with the first test variant, in which the highly discriminatory items were placed in stages two and three.
[a]α and β are symmetrical.
[b]Forced decisions out of 100,000 classification decisions.

Table A.2

SDs of the Ability Scales for the Five Test Domains in the Simulation Study

Test domain (Abbreviation)	*SD*
Reading (LEZ)	0.2670
Mathematics (REK)	0.3350
Language skills (TAV)	0.2680
Writing (SCH)	0.3020
Vocabulary (WST)	0.2750

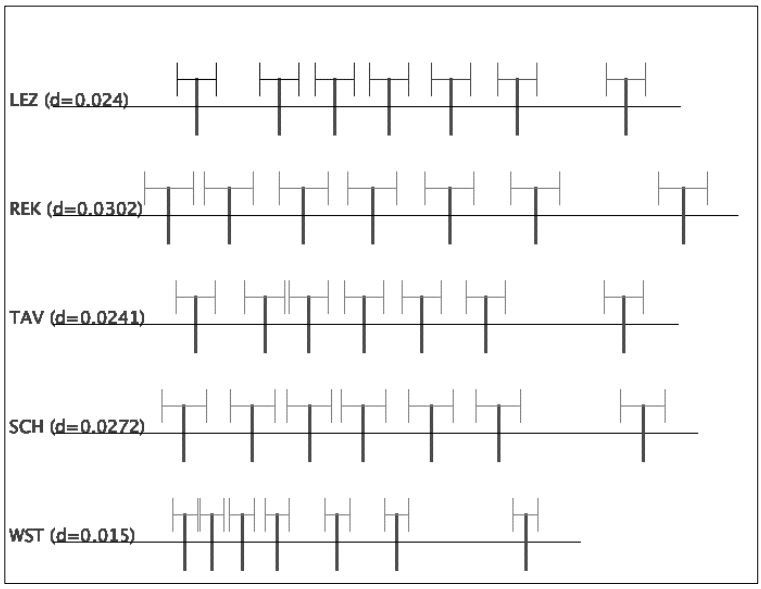

Figure A.2. Overview of the indifference regions around the classification cutoff points in all test domains, for the SPRT classification method

Appendix B: Proportional Classification Decisions Tables per Test Variant for Three Classification Methods

Table B.1

Proportional Classification Decisions per classification category for the Standard Score classification method.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
				All Items				
Cat. 1	**0.0986**	0.0097	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0109	**0.0856**	0.0139	0.0009	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0000	0.0156	**0.0342**	0.0164	0.0003	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0014	0.0182	**0.0685**	0.0206	0.0003	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0003	0.0222	**0.0828**	0.0213	0.0008	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0003	0.0224	**0.0629**	0.0220	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0007	0.0223	**0.1435**	0.0198
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0171	**0.1666**
				MST Variant 1				
Cat. 1	**0.0974**	0.0109	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0128	**0.0815**	0.0152	0.0017	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0000	0.0174	**0.0308**	0.0176	0.0007	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0026	0.0197	**0.0640**	0.0222	0.0005	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0010	0.0251	**0.0768**	0.0231	0.0014	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0005	0.0263	**0.0573**	0.0235	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0017	0.0254	**0.1390**	0.0203
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0196	**0.1642**
				MST Variant 2				
Cat. 1	**0.0980**	0.0103	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0122	**0.0830**	0.0146	0.0014	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0000	0.0170	**0.0313**	0.0176	0.0005	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0020	0.0195	**0.0649**	0.0221	0.0005	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0005	0.0242	**0.0790**	0.0224	0.0012	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0005	0.0241	**0.0598**	0.0232	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0013	0.0245	**0.1402**	0.0203
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0190	**0.1648**
				Linear				
Cat. 1	**0.0967**	0.0117	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0146	**0.0787**	0.0155	0.0024	0.0001	0.0000	0.0000	0.0000
Cat. 3	0.0000	0.0190	**0.0287**	0.0177	0.0010	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0039	0.0208	**0.0595**	0.0236	0.0011	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0016	0.0263	**0.0729**	0.0241	0.0025	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0011	0.0279	**0.0528**	0.0258	0.0001
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0024	0.0263	**0.1337**	0.0240
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0210	**0.1627**

Note. The correct decisions are bolded

Table B.2

Proportional Classification Decisions per classification category for the Sum of the Estimated Probability on All Items method.

	Cat. 1	**Cat. 2**	**Cat. 3**	**Cat. 4**	**Cat. 5**	**Cat. 6**	**Cat. 7**	**Cat. 8**
				All Items				
Cat. 1	**0.1074**	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0383	**0.0576**	0.0150	0.0004	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0002	0.0141	**0.0396**	0.0124	0.0002	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0011	0.0236	**0.0641**	0.0199	0.0003	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0005	0.0221	**0.0780**	0.0263	0.0004	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0002	0.0173	**0.0743**	0.0158	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0004	0.0290	**0.1407**	0.0162
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0188	**0.1649**
				MST Variant 1				
Cat. 1	**0.1065**	0.0018	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0391	**0.0543**	0.0165	0.0013	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0006	0.0154	**0.0359**	0.0138	0.0007	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0024	0.0249	**0.0585**	0.0225	0.0008	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0014	0.0243	**0.0720**	0.0287	0.0009	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0006	0.0209	**0.0683**	0.0177	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0011	0.0320	**0.1357**	0.0176
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0212	**0.1626**
				MST Variant 2				
Cat. 1	**0.1067**	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0393	**0.0545**	0.0167	0.0008	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0003	0.0154	**0.0363**	0.0139	0.0005	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0017	0.0246	**0.0600**	0.0221	0.0007	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0009	0.0239	**0.0733**	0.0286	0.0007	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0005	0.0194	**0.0706**	0.0171	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0009	0.0314	**0.1367**	0.0175
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0203	**0.1635**
				Linear				
Cat. 1	**0.1061**	0.0023	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0408	**0.0511**	0.0175	0.0018	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0010	0.0173	**0.0321**	0.0150	0.0011	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0036	0.0257	**0.0542**	0.0239	0.0016	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0021	0.0259	**0.0669**	0.0308	0.0016	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0010	0.0226	**0.0626**	0.0213	0.0001
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0019	0.0333	**0.1290**	0.0221
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0226	**0.1611**

Note. The correct decisions are bolded.

Table B.3

Proportional Classification Decisions per classification category for Rasch classification method.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
				All Items				
Cat. 1	**0.1047**	0.0037	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0241	**0.0612**	0.0250	0.0010	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0000	0.0078	**0.0422**	0.0160	0.0005	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0005	0.0202	**0.0637**	0.0242	0.0004	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0003	0.0193	**0.0799**	0.0274	0.0004	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0001	0.0182	**0.0731**	0.0163	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0005	0.0301	**0.1482**	0.0076
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0327	**0.1511**
				MST Variant 1				
Cat. 1	**0.1013**	0.0069	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0213	**0.0583**	0.0293	0.0023	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0002	0.0084	**0.0385**	0.0183	0.0011	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0011	0.0213	**0.0598**	0.0259	0.0009	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0010	0.0234	**0.0750**	0.0270	0.0009	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0005	0.0247	**0.0644**	0.0180	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0018	0.0343	**0.1420**	0.0083
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370	**0.1467**
				MST Variant 2				
Cat. 1	**0.1047**	0.0036	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0300	**0.0567**	0.0230	0.0015	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0002	0.0114	**0.0389**	0.0151	0.0008	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0012	0.0247	**0.0575**	0.0247	0.0009	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0010	0.0228	**0.0736**	0.0287	0.0012	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0005	0.0206	**0.0635**	0.0229	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0011	0.0269	**0.1444**	0.0139
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0257	**0.1581**
				Linear				
Cat. 1	**0.0974**	0.0105	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0178	**0.0549**	0.0349	0.0035	0.0002	0.0000	0.0000	0.0000
Cat. 3	0.0002	0.0078	**0.0377**	0.0187	0.0022	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0013	0.0232	**0.0524**	0.0306	0.0016	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0020	0.0220	**0.0724**	0.0290	0.0018	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0009	0.0268	**0.0595**	0.0204	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0027	0.0377	**0.1386**	0.0074
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0463	**0.1373**

Note. The correct decisions are bolded.

Table B.4

Proportional Classification Decisions per classification category for the SPRT classification method, with settings $\alpha = \beta = 0.1$.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
				All Items				
Cat. 1	**0.1055**	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0283	**0.0609**	0.0211	0.0009	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0001	0.0106	**0.0409**	0.0145	0.0004	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0006	0.0229	**0.0633**	0.0217	0.0005	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0006	0.0229	**0.0767**	0.0266	0.0006	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0003	0.0200	**0.0710**	0.0164	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0007	0.0317	**0.1431**	0.0109
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0292	**0.1545**
				MST Variant 1				
Cat. 1	**0.1048**	0.0035	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0319	**0.0558**	0.0218	0.0016	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0003	0.0127	**0.0376**	0.0150	0.0009	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0017	0.0258	**0.0573**	0.0229	0.0013	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0016	0.0263	**0.0679**	0.0299	0.0017	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0008	0.0223	**0.0634**	0.0211	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0014	0.0315	**0.1388**	0.0146
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0287	**0.1550**
				MST Variant 2				
Cat. 1	**0.1050**	0.0033	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0291	**0.0589**	0.0219	0.0013	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0001	0.0121	**0.0386**	0.0150	0.0007	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0012	0.0247	**0.0595**	0.0227	0.0010	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0010	0.0247	**0.0719**	0.0284	0.0013	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0006	0.0212	**0.0663**	0.0196	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0011	0.0320	**0.1399**	0.0134
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0293	**0.1545**
				Linear				
Cat. 1	**0.1038**	0.0045	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0310	**0.0604**	0.0189	0.0009	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0004	0.0175	**0.0379**	0.0102	0.0005	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0035	0.0368	**0.0533**	0.0148	0.0006	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0044	0.0379	**0.0625**	0.0217	0.0008	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0026	0.0314	**0.0602**	0.0134	0.0001
Cat. 7	0.0000	0.0000	0.0000	0.0001	0.0032	0.0462	**0.1223**	0.0146
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0310	**0.1525**

Note. The correct decisions are bolded.

Table B.5

Proportional Classification Decisions per classification category for the SPRT classification method, with settings $\alpha = \beta = 0.05$.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
				All Items				
Cat. 1	**0.1055**	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0283	**0.0609**	0.0211	0.0009	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0001	0.0106	**0.0409**	0.0145	0.0004	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0006	0.0229	**0.0633**	0.0217	0.0005	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0006	0.0229	**0.0767**	0.0266	0.0006	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0003	0.0200	**0.0710**	0.0164	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0007	0.0317	**0.1431**	0.0109
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0292	**0.1545**
				MST Variant 1				
Cat. 1	**0.1048**	0.0035	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0319	**0.0648**	0.0140	0.0004	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0003	0.0221	**0.0362**	0.0077	0.0003	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0042	0.0424	**0.0511**	0.0109	0.0004	0.0000	0.0000
Cat. 5	0.0000	0.0001	0.0049	0.0439	**0.0609**	0.0170	0.0006	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0031	0.0346	**0.0578**	0.0121	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0001	0.0036	0.0456	**0.1224**	0.0146
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0287	**0.1550**
				MST Variant 2				
Cat. 1	**0.1050**	0.0033	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0291	**0.0635**	0.0182	0.0004	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0001	0.0158	**0.0415**	0.0089	0.0002	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0019	0.0369	**0.0566**	0.0132	0.0005	0.0000	0.0000
Cat. 5	0.0000	0.0000	0.0026	0.0382	**0.0655**	0.0202	0.0008	0.0000
Cat. 6	0.0000	0.0000	0.0000	0.0015	0.0286	**0.0627**	0.0148	0.0000
Cat. 7	0.0000	0.0000	0.0000	0.0000	0.0021	0.0394	**0.1315**	0.0134
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0293	**0.1545**
				Linear				
Cat. 1	**0.1038**	0.0045	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cat. 2	0.0310	**0.0725**	0.0076	0.0002	0.0000	0.0000	0.0000	0.0000
Cat. 3	0.0004	0.0333	**0.0292**	0.0035	0.0001	0.0000	0.0000	0.0000
Cat. 4	0.0000	0.0111	0.0548	**0.0379**	0.0051	0.0001	0.0000	0.0000
Cat. 5	0.0000	0.0003	0.0139	0.0574	**0.0479**	0.0076	0.0001	0.0000
Cat. 6	0.0000	0.0000	0.0003	0.0098	0.0505	**0.0434**	0.0037	0.0001
Cat. 7	0.0000	0.0000	0.0000	0.0005	0.0117	0.0760	**0.0835**	0.0146
Cat. 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014	0.0299	**0.1525**

Note. The correct decisions are bolded.