Computerized Adaptive Testing in Industrial and Organizational Psychology

Guido Makransky

Promotiecommissie

Promotor
Prof. dr. Cees A. W. Glas

Assistent Promotor
Prof. dr. Svend Kreiner

Overige leden
Prof. dr. T.J.H.M. Eggen
Prof. dr. R.R. Meijer
Prof. dr. P.J.C. Sleegers
Dr. P.M. ten Klooster

Computerized Adaptive Testing in Industrial and Organizational Psychology
Guido Makransky
Ph.D. thesis
University of Twente
The Netherlands
30 March, 2012
ISBN: 978-90-365-3316-4

COMPUTERIZED ADAPTIVE TESTING IN INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, 30 March, 2012 at 14:45

by
Guido Makransky
born on 23 June, 1978
in Coban, Guatemala

This dissertation is approved by the following promotores:

Promotor: Prof. dr. Cees A.W. Glas
Assistant Promotor: Prof. dr. Svend Kreiner

Contents

Chapter 1

Introduction

Psychological tests are commonly used for making employment decisions because research has shown that they are good predictors of performance on the job (e.g., Barrick and Mount, 1991; Schmidt and Hunter, 1998). Although, the majority of employment tests are currently administered by computer, most of these instruments still use a traditional fixed testing format in which all test takers are administered the same items. This method of administering tests does not take advantage of the potential a computer has as a test administration device. In industrial/organizational psychology there is a need to have precise and flexible assessments in an environment where testing time is a valuable commodity. Computers can support this need by administering and scoring items intelligently with computerized adaptive testing (CAT). The fundamental aim of this dissertation is to investigate methodological CAT solutions to the practical challenges facing industrial/organizational psychological testing, in order to improve the quality and flexibility of the assessment tools used in this field.

1.1 Computerized Adaptive Testing

Instead of giving each examinee the same fixed test, a computerized adaptive test (CAT) is designed to adapt to the ability level of the individual examinee. This tailored item selection can result in increased reliability with only a limited number of properly selected items. CAT algorithms successively select items so as to maximize the precision of the test based on information about the examinee's level from previous items. For example, if an examinee performs well on an item, he/she would then be presented with a more difficult item. Or, if he/she performs poorly, he/she would be presented with an easier item. After each response the examinee's trait estimate is updated and the CAT algorithm selects the subsequent item to have optimal properties at the new estimate. The implementation of CAT is attractive because research indicates that CATs are more precise (Rudner, 1998; van der Linden & Glas, 2010), take a shorter period of time to complete (Rudner, 1998; Wainer, 2000), and can be more motivating for the test taker (Daville, 1993) compared with traditional tests.

1.2 Item Response Theory

The challenge with CAT is to compare respondents on the same scale even though they have responded to different items. This can be achieved with item response theory (IRT). IRT is a family of mathematical models that describe, in probabilistic terms, the relationship between a person's response to a test item and his or her standing on the construct being measured by the test (Wainer, Bradlow, & Wang, 2007). Consequently, IRT provides a modeling-based approach which makes it possible to predict a person's trait level based on the observed responses to a particular set of items on a test. In IRT, the responses to items are modeled as a function of one or more person ability parameters and item parameters. The item parameters in the models described in this dissertation are item difficulty parameters and item discrimination parameters. The difficulty parameters define the overall salience of the response categories; and the discrimination parameters define the strength of the association of the item responses with the latent person parameters. The latent person parameters locate the person on a latent scale where high values are related to high item scores. The fundamental concept of IRT is that individuals and items are characterized on the same metric in terms of their location on the latent scale.

IRT models can be applied to both dichotomous (two answer categories, e.g., correct/false) and polytomous data (more than two answer categories, e.g., 5-point Likert scale). There are different types of IRT models and more detailed information can be found in Embretson and Reise (2000); and Hambleton, Swaminathan, and Rogers (1991). One distinction is that IRT models can be unidimensional or multidimensional. In a unidimensional model, the assumption is made that all items in the test measure the same unidimensional trait. Multidimensional item response theory models (MIRT) also extend to cases where a test is constructed to measure multidimensional constructs (e.g., Reckese, 2009). MIRT models are built on the premise that a respondent possesses a vector of latent person characteristics that describe the person's level on the different traits, and the items are described by a vector of item characteristics that describe the location and sensitivity of the items in the test to multiple traits.

This dissertation uses several IRT models including dichotomous and polytomous models; as well as unidimensional and multidimensional models. Details of these models are discussed in the chapters where the specific models are applied.

1.3 Outline

This dissertation explores and develops methods for dealing with feasibility issues related to developing CAT solutions to practical challenges facing assessment in industrial/organizational psychology. One of these challenges is that a CAT requires a large item bank that is accurately calibrated. However, it is difficult to collect a large amount of data in the development phase of an organizational test. This is the case because companies that purchase an organizational test are usually unwilling to invest time and resources in letting their employees take the test unless they can use the results. Chapter 2 illustrates an automatic online calibration design that can be used in adaptive testing. This study explores three possible automatic online calibration strategies, with the intent of calibrating an item bank online in an operational setting. That is, the item bank is calibrated in a situation where test takers are processed and the scores they obtain can be used. The method makes it more attractive for respondents to participate during calibration, and increases the speed with which a CAT item bank can be calibrated.

Another challenge in the development of a CAT is the possibility that measurement invariance (MI) is violated across contexts. Several assumptions need to be empirically assessed to confirm that the test does not systematically discriminate against members of a particular group. Chapter 3 demonstrates a straightforward method for conducting a test of MI and illustrates a method for modeling differential item functioning (DIF) by assigning group-specific item parameters in the framework of IRT. The chapter exemplifies three applications of the method for an international organizational assessment context. These examples pertain to context effects due to the test administration method, motivation effects due to the stakes of the test, and language effects. The method leads to a more flexible and practical way of dealing with MI across groups.

A general challenge facing the testing industry is the issue of test security. This issue is specifically relevant in organizational testing where the majority of tests are conducted via the Internet, without a the presence of a human proctor. This method of test administration is called unproctored Internet testing (UIT). When the test is a high stakes ability or achievement test, the International Guidelines on Computer-Based and Internet-Delivered Testing recommend to follow up the results with a confirmation test in a controlled setting in order to limit cheating. A confirmation test is a short computerized test given under supervision to verify the result obtained in the UIT. Chapter 4 illustrates a method for verifying the results of an unproctored Internet test by using adaptive confirmation testing. This is conducted with an extension of the stochastic curtailed truncated sequential probability ratio test (SCTSPRT).

Most psychological tests used in employment testing assess multidimensional constructs with several correlated sub-scales. Companies that use these tests are interested in obtaining precise scores on each of the sub-scales. However, there is also an expectation that the tests should not be too time consuming. Chapters 5 and 6 investigate the possibility of increasing the precision of typical employment tests by scoring items with multidimensional item response theory (MIRT) and explore the possibility of shortening the test length of these measures by efficiently administering and scoring items with multidimensional computerized adaptive testing (MCAT). These methods can increase the precision of sub-scale scores by effectively administering and scoring items based on the correlations between the sub-scales. Chapter 5 explores the possibility of using MCAT for administering and scoring a cognitive ability test consisting of dichotomously scored items. Chapter 6 explores the potential of administering and scoring items with MCAT for the NEO PI-R (Cost & McCrea, 1992) a widely used personality test consisting of polytomously scored items.

The chapters in this dissertation are self-contained, hence they can be read separately. Therefore, some overlap could not be avoided.

Chapter 2

An Automatic Online Calibration Design in Adaptive Testing

Abstract

An accurately calibrated item bank is essential for a valid computerized adaptive test. However, in some settings, such as organizational testing, there is limited access to test takers for calibration. As a result of the limited access to possible test takers, collecting data to accurately calibrate an item bank is usually difficult. In such a setting, the item bank can be calibrated online in an operational setting. This study explores three possible automatic online calibration strategies, with the intent of calibrating items accurately while estimating ability precisely and fairly. That is, the item bank is calibrated in a situation where test takers are processed and the scores they obtain have consequences. A simulation study was used to identify the optimal calibration strategy. The outcome measure was the mean absolute error of the ability estimates of the test takers participating in the calibration phase. Manipulated variables were the calibration strategy, the size of the calibration sample, the size of the item bank, and the item response model.

Key Words: *computerized adaptive testing, item bank, item response theory, online calibration*

This chapter has been published as:
Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology, 11*.

2.1 Introduction

The past 15 years have seen a steady increase in the use of online testing applications in a variety of testing settings. Computers can be used to increase statistical accuracy of test scores using computerized adaptive testing (CAT; van der Linden & Glas, 2000a). The implementation of CAT is attractive because research indicates that CATs can yield ability estimates that are more precise (Rudner, 1998, van der Linden & Glas, 2000a), can be more motivating (Daville, 1993), easier to improve (Linacre, 2000; Wainer, 2000), and take a shorter period of time to complete (Rudner, 1998; Wainer, 2000) compared with traditional tests. Although CATs have been widely implemented within large scale educational testing programs, the use of CATs in other settings such as in organizational testing has been limited because of several practical challenges.

One of the major obstacles to cost-effective implementation of CAT is the amount of resources needed for item calibration, because of the large item banks needed in CAT. Large testing programs have been able to overcome this with the availability of extensive resources. Nevertheless, there has been broad interest in investigating procedures for optimizing the calibration process (e.g. Berger 1991; 1992; 1994; Berger, King & Wong, 2000; Jones & Nediak 2000; Lima Passos & Berger, 2004; Stocking 1990). Unfortunately, this research is based on the assumption that a large number of test takers is available in the development phase of a test. However, this is not the case in many applied settings.

In reality, the lack of available test takers is one of the greatest challenges in the development phases of a test in an organizational setting. This is the case because the companies that purchase an organizational test are usually unwilling to invest time and resources in letting their employees take a test unless they can use the results. To circumvent this problem, test developers usually access test takers from a context other than the one in which the test is to be used, that is, they access a low-stakes sample. The use of a low-stakes calibration sample comes with several limitations. First, there is evidence that large motivational differences exist between test takers in low stakes calibration samples and the intended population of test takers (Wise & DeMars, 2006). These motivational differences introduce bias in the estimation of item parameters in the calibration phase, which will result in biased test scores. Further, the use of a separate sample usually means extra resources in terms of time and money in test development.

The resources required for item calibration would be reduced if a test could be calibrated and implemented for the intended population as quickly and fairly as possible. This would make it attractive for possible customers to be involved in the

calibration process because they could use the results. Therefore, it is worthwhile to identify designs that make it possible to simultaneously calibrate items and estimate ability, while treating test takers fairly. The present study differs from previous studies in that this is an investigation of the problem of calibrating a set of items where there is no previously available information, with the practical constraint of maintaining fairness in test scoring. This problem is common for test development companies that are interested in developing a new CAT when there is no previously available version of the instrument.

The purposes of this paper are to discuss calibration strategies that will make it more practical and cost effective to develop and implement CATs in small testing programs, and to report on a simulation study that was conducted to choose an optimal strategy. More specifically, the paper investigates three different calibration strategies for calibrating an item bank from scratch, with the primary objectives of calibrating items in a fair and effective manner, while providing accurate ability estimates throughout the calibration design.

2.2 The Model

The present study was carried out in the framework of item response theory (IRT). The fundamental concept of IRT is that each test item is characterized by one or more parameters and each test-taker is characterized by ability parameters, in this study by a single ability parameter. The probability that a given test-taker answers a given item correctly is given by a function of both the item's and the test taker's parameters. Conditional on those parameters, the response on one item is independent of the responses to other items. One IRT model used in this study, is the two-parameter logistic, or 2-PL model,

$$P_i(\theta) = \frac{1}{1+\exp(-a_i(\theta - b_i))} \tag{2.1}$$

(Birnbaum, 1968). Here $P_i(\theta)$ is the probability of a correct response for item i, θ is the test taker's ability, and a_i and b_i are item parameters. Further, a_i is called the discrimination and b_i the difficulty parameter. A specific form of this model that is also used in this study is the one-parameter logistic or1-PL model (often referred to as the Rasch model; Rasch, 1960). In the 1-PL model the assumption is made that all items have the same discrimination parameter. The 1-PL and 2-PL models are viable alternatives to the 3-PL model because the guessing parameter in the 3-PL model can be difficult to estimate in small sample sizes, as those used in this study. An

additional reason for not using the 3-PL model in this study is that a CAT algorithm is used to administer items from an early stage in the calibration strategies described in this study. Therefore, the chances of guessing are not as high because the ability of the respondent is matched with the difficulty of the item.

Calibration pertains to the estimation of the item parameters a_i and b_i from response data, say data from a calibration sample. In the operational phase of CAT, the item parameters are considered to be known and the focus becomes the estimation of θ. In IRT θ can be estimated using several different strategies. The weighted maximum likelihood estimator derived in Warm (1989) was used to estimate ability in this study. This method is attractive because of its negligible bias (van der Linden & Glas, 2000b).

What differentiates CAT from traditional tests is that items are selected optimally by an item selection algorithm that finds the next available item from the item bank that provides the most information about the test taker. A selection function that is often used in item selection for CAT is Fisher's information function. For an introduction regarding Fisher's information and alternative criteria for item selection, refer to Wainer (2000) or van der Linden and Glas (2000a). For dichotomously scored items, the information function has the following form:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{2.2}$$

where $P_i(\theta)$ is the response function for item i, $P'(\theta)$ its first derivative with respect to θ, and $Q_i(\theta) = 1 - P_i(\theta)$. In CAT, the item is selected that has the maximum information in the item pool at $\theta = \theta^*$, where θ^* is the current ability estimate for the test taker (van der Linden & Glas, 2000b). Maximization of information minimizes the estimation error of θ.

2.3 Calibration Strategies

This study investigated the online calibration of an item bank where there was no available information about item parameters at the beginning of the testing process. Therefore the most equitable way to select items during the initial phase of testing was to administer items randomly. Although random item administration does not guarantee tests with equal difficulty levels, it does ensure that there are no systematic differences in difficulty which would result in unfairness. Then, once sufficient data become available, optimal item selection can be carried out with Fisher's information function. A viable calibration strategy would be able to progress from random to

optimal item selection in a fair and effective manner. The next section describes three plausible calibration strategies that were evaluated in this study.

2.3.1 Two-Phase Strategy

In this strategy, labeled P2, items are administered randomly up to a given number of test takers. For the remaining test takers the items are calibrated and administered optimally in the form of a CAT. In the random phase, tests are scored with the assumption that all items have a difficulty parameter equal to 0 (that is, $b_i = 0$), and in the optimal phase tests are scored based on the item parameters obtained in the random phase. The reason for the scoring rule in the random phase is to obtain scores that are on the same scale as in the optimal phase. This scoring rule is analogous to the scoring rule used in classical test theory, where a proportion-correct score is computed assuming that all items have the same weight. Here this score is simply converted to a score on the θ scale. The clear transition from one phase to the next means that stakeholders can be informed about the current precision of the test, and policy decisions about how the test should be used can be clearly defined based on the level of precision. The transition is made when the average number of item administrations is above some predefined value T. The optimal transition point T from the random to the optimal phase is one of the topics in this study.

2.3.2 Multi-Phase Strategy

An alternative strategy labeled M consists of more than two phases. As in the previous strategy, the items are calibrated at the end of each phase. Table 2.1 illustrates an example with the five phases that the design follows. In Phase 0, all item selection is random and ability is estimated with the assumption that $b_i = 0$. As in the previous strategy, also here the transition is made when the average number of item administrations is above some predefined value T. In the next phase, labeled Phase 1, the first three parts of the test are random, and the final part is CAT using the item parameter estimates from data collected in the previous phase. A transition takes place when the average number of administrations over items has doubled. In general, a transition takes place when this average exceeds (Phase + 1) * T. This continues until the final phase, where all of the items are administered optimally and the item bank is calibrated.

The motivation for the strategy is as follows. In phase 0, the amount of uncertainty regarding the item parameters and the person parameters is too high to allow for optimal item selection. In fact, this high uncertainty might introduce bias because the uncertainty estimate in item parameters and ability could compound the error in

the ability estimate. Therefore, items are administered randomly. After the random part, ability is estimated using the item parameters obtained in the previous phase, and this estimate serves as an initial estimate for the adaptive part. In later phases, it is assumed that the parameters are estimated with sufficient precision to support optimal item selection. The inclusion of an adaptive part at the end makes the test more effective in terms of scoring ability and in terms of calibrating items. As with the P2 strategy there is a clear transition point between phases in this strategy.

Table 2.1

M strategy design

Part 1	Part 2	Part 3	Part 4
Random	Random	Random	Random
Random	Random	Random	CAT
Random	Random	CAT	CAT
Random	CAT	CAT	CAT
CAT	CAT	CAT	CAT

2.3.3 Continuous Updating Strategy

Labeled C, this strategy is analogous to the previous two strategies in that items are administered randomly and tests are scored with the assumption that $b_i = 0$ in the first phase. An item becomes eligible for CAT if the number of administrations of the item is above a transition point labeled T. The proportion of optimally administered items in a test is proportional to the number of eligible items in the item bank. Therefore, during this phase the first part of the test is random, and the final part is CAT where items are calibrated after each exposure and tests are scored based on the parameters computed after the latest administration of the items. In the final phase all item selection is optimal and the items are calibrated after each exposure, therefore, the precision of the θ estimates is continuously improved.

The three calibration strategies were chosen because they represent a sample of possible designs on a continuum ranging from one extreme where items are calibrated at a single point in time, to the other extreme where items are calibrated constantly after each exposure, once the items become eligible for CAT. The P2 strategy is similar to a typical random item administration calibration design where the items are calibrated at a single point, with the difference that ability scores are reported up to that point. Therefore, this strategy is the easiest to implement and can be considered a control strategy for comparison purposes. The P2 and M strategies have the advantage that test takers within the same phase are given the same probability

of success. This can help define policy decisions about how the test should be used, based on the level of precision in the test. The C strategy has the advantage that changes in the calibration sample can be quickly detected because calibration occurs continuously. This would make it easier for test developers to detect mistakes in the items, and would make it possible to get a rough measure of the characteristics of the items in the test at an early stage of the calibration process. In addition, it is easier to detect fluctuations in the item parameters which may be caused by item exposure with the C strategy.

2.4 Research Questions

The main research question in this study was: Which of the three calibration strategies is the most effective for calibrating a new item bank effectively, while estimating ability precisely? In order to assess this in more detail the three strategies described above were compared based on a number of criteria: the global and conditional precision of ability estimates in a large calibration sample; the precision of the strategies at different points in the calibration process, and for different size calibration samples; the uniformity of item exposure; and their application under the assumptions of the 1-PL and 2-PL models. A secondary research question was: Could accounting for the uncertainty in the parameters in the calibration phase of a test improve the precision of the ability estimates? Both questions were evaluated with simulation studies.

2.5 Simulation Studies

To investigate which of the considered calibration strategies leads to the lowest overall mean absolute error (MAE) in the estimation of ability, simulation studies were conducted. Simulation studies make it possible to determine the true ability level of the test taker; next the calibration design can be reproduced in order to investigate the precision of the test result for each test taker. This cannot be done with operational data, because in practice it is impossible to assess the actual accuracy of a test since it is not possible to know the real ability level of a test taker. The simulation studies were programmed in Digital Visual Fortran 6.0 standard for Windows. The simulations were designed to measure the impact of each of the three strategies across a variety of conditions by varying the following variables:

1. The transition point *T* from one phase to the next. These points were varied as *T* = 10, 25, 50, 100, 200 item administrations.
2. The calibration sample sizes, which were varied as N = 250, 500, 1,000, 2,000, 3,000, 4,000.
3. The IRT model, varied as the 1-PL model and the 2-PL model.
4. The size of the item bank, varied as *K* = 100, 200, 400 items.
5. Accounting for uncertainty in the parameter estimates.

Upper and lower baselines were also simulated to compare the precision of the simulation strategies to external criteria. MAE for an optimal test administered with a completely calibrated item bank, labeled O, was set as a lower baseline. This was simulated by calibrating items using strategy P2 with a transition point of 4,000. The precision of a test administered randomly with all items having difficulty parameters of 0.0 was set as an upper baseline. This procedure is labeled R. The length of the test was also varied as: 20, 30 and 40 items in certain conditions, however, only the results of the test with 20 items are reported. The test length and item bank sizes selected for this simulation are typical for an organizational testing or certification program that uses a test battery with several unidimensional CAT's.

2.6 Method

The three calibration strategies were compared by assessing the accuracy of the ability estimate while in the calibration phase of the test. Once the number of test takers becomes large and the item bank is accurately calibrated, it is expected that different calibration designs result in similar precision, so then the calibration design is no longer of interest. Therefore, it was important to differentiate the calibration sample from the post-calibration sample of test takers. A calibration sample of 4,000 test takers was set in this study.

The test takers' θ parameters were drawn from a standard normal distribution. An item bank was simulated by drawing item difficulty parameters from a standard normal distribution, and item discrimination parameters from a lognormal distribution with an expectation of 1. After each phase, items were calibrated under either the 1-PL or 2-PL model using the method of marginal maximum likelihood estimation (Bock & Aitkin, 1981). Optimal item selection was implemented using maximal expected information. The item parameters were the current estimates at that point in the design of the strategy. MAE was computed as the mean absolute difference between the true ability drawn from the N (0,1) distribution and the ability estimated

by the weighted maximum likelihood procedure. The MAE for each strategy was then calculated by averaging across all test takers to give an estimate of the global precision of the strategy.

In addition to global precision, it was also of interest to investigate the precision with which a certain test taker's score was estimated. This conditional precision was measured at specific points on the ability continuum (θ = -2, -1, 0, 1, 2), to give an estimate of the precision with which a test taker with a specific θ could be expected to be assessed within each condition. Therefore, after each phase 4,000 test takers were simulated at each of the five ability values, and the MAE was computed for each of the five ability values.

2.7 Results

2.7.1 Global Precision and Optimal Transition Points

Before the research questions could be investigated, it was necessary to identify the optimal point at which item selection should transition from one phase to the next in each of the three calibration strategies. Five conditions were investigated (T = 10, 25, 50, 100, 200) for the 1-PL and 2-PL models. The results are shown in Table 2.2.

Table 2.2

Comparison of the MAE for different transition points within each calibration strategy

Model	Item bank	Strategy		MAE				
				T = 10	T = 25	T = 50	T = 100	T = 200
1-PL	K = 100	R	0.418					
		P2		0.489	0.420	0.404	**0.392**	0.394
		M		0.453	0.395	**0.389**	0.396	0.402
		C		0.381	**0.379**	0.380	0.381	0.392
		O	0.376					
	K = 200	R	0.418					
		P2		0.577	0.435	0.409	**0.393**	0.396
		M		0.417	0.397	**0.392**	0.408	0.420
		C		0.390	**0.382**	0.393	0.398	0.404
		O	0.381					
	K = 400	R	0.418					
		P2		0.533	0.439	0.406	**0.401**	0.414
		M		0.430	0.410	**0.405**	0.414	0.420
		C		0.400	**0.396**	0.397	0.397	0.413
		O	0.384					
2-PL	K = 100	R	0.405					
		P2		0.475	0.388	0.353	**0.352**	0.361
		M		0.362	0.366	**0.349**	0.368	0.380
		C		**0.342**	0.345	0.353	0.355	0.366
		O	0.342					
	K = 200	R	0.405					
		P2		0.460	0.352	0.351	**0.349**	0.373
		M		0.366	**0.340**	0.346	0.381	0.394
		C		0.335	**0.324**	0.329	0.352	0.369
		O	0.323					
	K = 400	R	0.405					
		P2		0.450	0.368	**0.356**	0.362	0.416
		M		0.344	0.354	0.375	0.401	0.406
		C		0.339	**0.330**	0.348	0.366	0.414
		O	0.311					

Note: Best results for each strategy within each condition are printed in bold.

The table gives the MAE obtained for the three calibration strategies as well as a completely random (R) and completely calibrated test (O), for a calibration sample size of 4,000, with item bank sizes of 100, 200 and 400 (K = 100, 200, 400), using the 1-PL and 2-PL models. A comparison of the MAE for the three strategies indicated that the C strategy consistently resulted in the best ability estimates across all conditions.

The results for the 1-PL model were consistent across the item bank sizes, and indicated that a transition point of 100 (T = 100) had the lowest MAE for the P2 strategy, T = 50 for the M strategy, and T = 25 for the C strategy. Therefore, the most effective transition point became lower as the number of calibration points for the strategy increased (from P2 to C). Note that for T = 10, the MAE of the P2 and M strategies was often above the MAE of the upper baseline (strategy R). This occurred because, in that case, the item parameters were calculated based on 10 observations only. Therefore these estimates of the item parameters were very poor and performed worse than the baseline estimate of $b_i = 0$.

The results for the 2-PL model were similar to those for the 1-PL; but they were not as consistent. Specifically, a faster transition seemed to be optimal for the M strategy with larger item bank sizes. This finding seems to be a consequence of the M strategy taking a long time to transition through the five phases in the design with large item banks.

The general pattern in these findings is consistent with the hypothesis that a balance between efficiency and accuracy in terms of switching from one phase to the next is important. A quick transition resulted in a premature progression through the phases in each strategy, because item parameter estimates still had much error. Therefore, the use of an optimal item selection algorithm to administer items, assuming that the item parameters were accurate, resulted in inaccurate ability estimates. On the other hand, the slower progression through the phases resulted in loss of efficiency because the calibration procedure did not react quickly enough in switching to the next phase, even though item parameter estimates had stabilized. Since the results were similar across the different item bank sizes, and between the two models, transition points of T = 100, T = 50, T = 25 were used respectively, for the P2, M, and C calibration strategies in subsequent analyses for both the 1-PL and 2-PL models, in order to have comparable results across settings.

2.7.2 Local Comparison of the Calibration Strategies

In addition to global precision, the conditional precision of the three strategies for specific points on the ability scale was investigated. A comparison of these and random item administration with $b_i = 0$ (R) as a baseline is presented in Figure 2.1.

Figure 2.1 illustrates the conditional precision of the three strategies with the 1-PL model on the left side and the 2-PL model on the right side, for item bank sizes of 100, 200, and 400 items. The horizontal axis represents the ability level of the test takers at five points on the θ scale (-2, -1, 0, 1, 2), and the vertical axis represents the MAE across the first 4,000 test takers within each design. For the 1-PL model, the graph shows that the C strategy measured ability more precisely than the other strategies at extreme ability scores, while all three strategies performed fairly equally at $\theta = 0$. The use of random item administration with item parameter estimates of $b_i = 0$ performed well at $\theta = 0$; however, this method performed much poorer at extreme ability levels. For the 2-PL model, the three strategies performed quite similarly with a smaller item bank, but the C strategy performed better than the other two as the item bank size became larger. All three strategies also performed better than random item administration for the 2-PL model, with the largest differences occurring at extreme ability values.

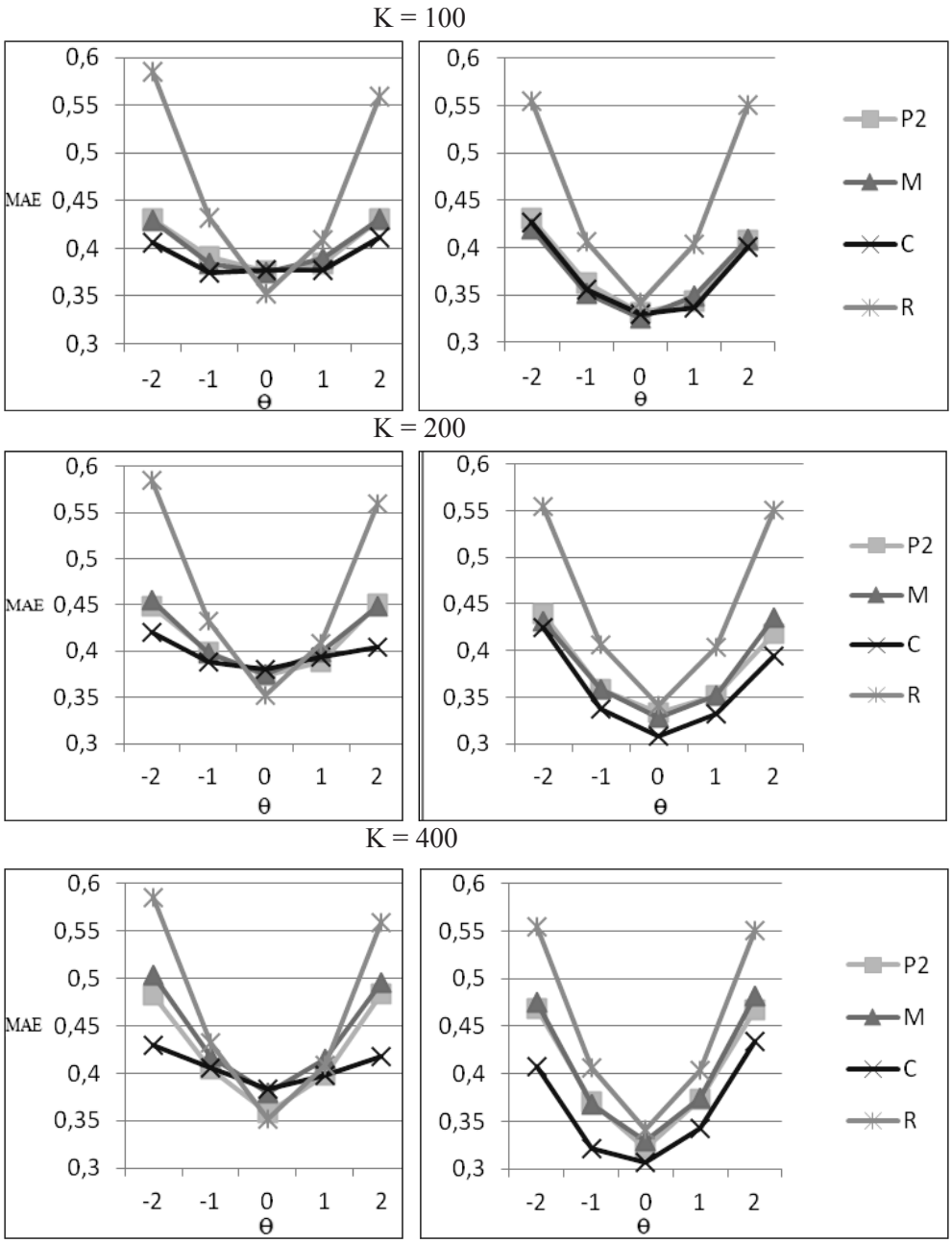

Figure 2.1: Conditional precision: MAE at specific points on the θ continuum for strategies P2, M, C and random item administration (R) for the 1-PL model on the left, and the 2-PL model on the right, presented for an item bank size of 100 on the top, followed by an item bank size of 200, and finally an item bank size of 400 on the bottom.

2.7.3 A Comparison of the Strategies at Different Points in the Calibration Process

The next research question investigated was the precision of the strategies for settings with a limited number of test takers. In this section the examples are limited to an item bank size of 100, because the general results across the different item bank sizes led to similar conclusions. Figure 2.2a and 2.2b display the specific precision for each strategy at a particular point in the calibration process. In other words, these figures present the results for how accurately the particular test estimates ability for the n^{th} test taker in the calibration design. This provides information about the point at which a test can be confidently used in a high-stakes situation. The horizontal axis represents the n^{th} test taker in the calibration design, and the vertical axis shows the MAE for the three strategies, as well as random item administration (R), and a fully calibrated test (O).

The results indicate that strategy C performed nearly as well as a fully calibrated test after as few as 500 test takers for the 1-PL model; it took strategy M 1,000 test takers to reach a similar level of precision. Strategy P2 never reached the same precision as a fully calibrated test, which implies that the P2 strategy needs to be supplemented with additional calibration points later in the design in order to reach the same level of accuracy. The results for the 2-PL model were similar to the 1-PL model, with the exception that the C strategy took a longer time to reach precision estimates comparable to a completely calibrated test.

These results consider the accuracy of a given test taker at a particular point in the calibration process. Figures 2.2c and 2.2d present the cumulative precision of each strategy, which is the average precision with which a test taker is assessed in the calibration phase of the test, for different size calibration samples. The figure plots the average MAE of the sample on the vertical axis, based on the number of test takers in the calibration sample on the horizontal axis. The results were similar for the 1-PL and 2-PL models, in that the C strategy performed considerably better than the other two strategies and random item administration. The difference was evident after the number of test takers in the calibration sample reached 500 for the 1-PL model, and after as few as 250 for the 2-PL model. The M and P2 strategies resulted in ability estimates that were considerably better than random item administration; however the calibration sample had to be at least 1,000 before a significant difference was evident. The difference between the precision of the three strategies decreased as the calibration sample became larger, suggesting that the benefits of using the C strategy are highest when there is a limited number test takers.

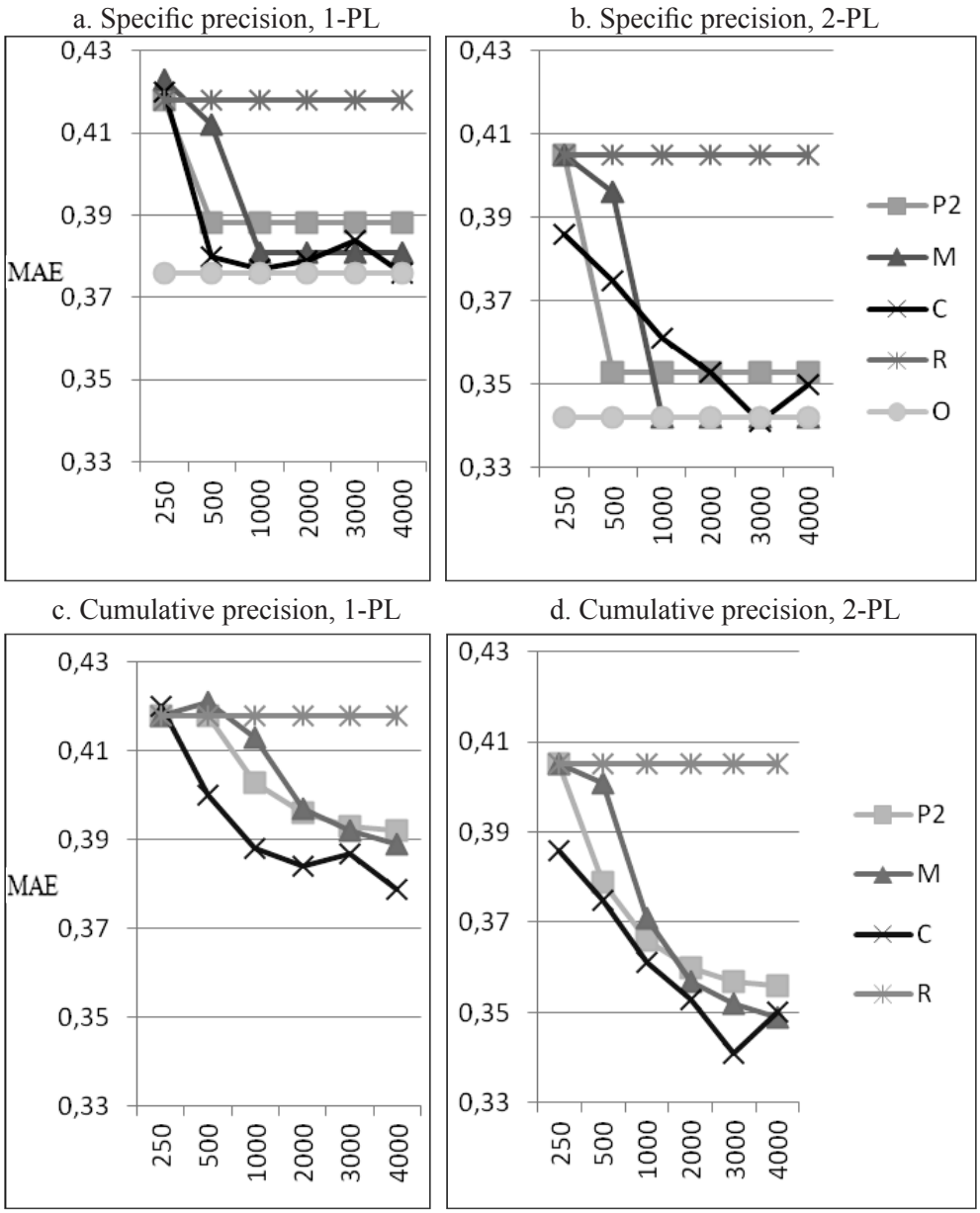

Figure 2.2: Top: Specific precision of each strategy, random item administration (R), and a completely calibrated test (O) for test taker number 250, 500, 1000, 2000, 3000 and 4000 for the 1-PL model on the left, and the 2-PL model on the right.

Bottom: Cumulative precision of each strategy and random item administration (R), for 250, 500, 1000, 2000, 3000 and 4000 test takers for the 1-PL model on the left, and the 2-PL model on the right.

2.7.4 Item Exposure

The calibration strategies have been compared in terms of how accurately ability is assessed in the calibration process. However, the calibration strategy should also calibrate the entire item bank. Therefore, it was important to investigate the frequency with which items were administered using the calibration strategies in the two models. Table 2.3 displays the number of times items were administered in the three calibration strategies, for item bank sizes of $K = 100$ and $K = 400$, in a calibration sample of 4,000 test takers. The results for the 1-PL model are presented in the upper portion, and the 2-PL model in the lower portion of the table.

Table 2.3

Number of times items were administered for each strategy within each model

Model	Item bank	Strategy	Number of administrations						
			<100	100-199	200-399	400-599	600-799	800-999	>1000
1-PL	K = 100	P2	0	2	13	21	31	18	15
		M	0	0	6	28	31	23	12
		C	0	0	0	32	39	6	23
	K = 400	P2	0	286	103	6	1	0	4
		M	0	316	64	8	0	0	12
		C	0	262	142	0	0	0	0
2-PL	K = 100	P2	12	30	11	8	4	5	30
		M	0	45	7	8	5	6	29
		C	39	6	6	10	6	5	28
	K = 400	P2	136	198	22	11	15	6	12
		M	1	355	19	7	8	8	10
		C	320	17	18	8	6	4	27

Table 2.3 shows a fairly uniform administration of items for all three calibration strategies for the 1-PL model. Item administration for the 2-PL model was highly uneven for the P2 and C strategies, but fairly balanced for the M strategy. In the C strategy, 39%, and 80% of the items were administered fewer than 100 times, for item banks consisting of 100 and 400 items respectively.

2.7.5 Accounting for Uncertainty in the Parameter Estimates

In IRT item parameters are usually estimated, and then these estimates are treated as true parameters in subsequent analyses. Most of the literature on IRT takes this

assumption for granted. However van der Linden and Glas (2000b) discovered that the impact of estimation error can have dramatic consequences on ability estimation.

In the current study there is known uncertainty in the parameters, because ability is estimated with items that are in the calibration process. Therefore it is important to investigate the consequences of taking uncertainty into account in the model. Uncertainty can be taken into account by using a distribution of the parameter instead of a point estimate in the estimation equation. The distribution is simply the likelihood distribution associated with the parameter estimate, which represents the current level of confidence related to each parameter. Here four different conditions were assessed.

1. All item parameters were treated as true parameters.
2. Uncertainty in theta was taken into account in the model, but uncertainty in the item parameters was ignored.
3. Uncertainty in the item parameters was taken into account in the model, but uncertainty in theta was ignored.
4. Uncertainty in all parameters was taken into account in the model.

Table 2.4

Global precision: MAE for four methods of calculating theta based on taking uncertainty in parameters into account

Model	Strategy	All parameters treated as true parameters	Uncertainty in theta is taken into account	Uncertainty in the item parameters is taken into account	Uncertainty in all parameters is taken into account
1-PL	P2	.392	.394	.388	.390
	M	.389	.400	.392	.394
	C	.379	.392	.381	.388
2-PL	P2	.352	.366	.353	.367
	M	.349	.356	.354	.354
	C	.345	.350	.344	.359

Table 2.4 presents the MAE for theta calculated under each of the four conditions for the 1-PL and 2-PL models for a calibration sample of 4,000. In general, taking uncertainty into account in all parameters decreased precision in theta estimation. The results also indicate that the use of a point estimate is better than a distribution in estimating theta. Taking uncertainty in the item parameters into account did not decrease precision greatly and slightly increases precision in the P2 strategy with the 1-PL model, and in the C strategy with the 2-PL model.

2.8 Discussion

The purposes of this paper was to investigate three different calibration strategies for calibrating an item bank from scratch, with the primary objectives of calibrating items in a fair and effective manner, while providing accurate ability estimates throughout the calibration design. The benefits of the three strategies were tested in terms of several possible conditions.

The C strategy consistently outperformed the other two strategies across all test lengths, and all item bank sizes. An example is that ability was estimated nearly as well as in a fully calibrated test after as few as 500 test takers in a test consisting of 20 items and an item bank consisting of 100 items for the 1-PL model. A weakness of this strategy was the non-uniform administration of items with the 2-PL model, which lead to the calibration of a few items at the expense of others. The M strategy might be preferred in settings where the 2-PL is used, because this strategy resulted in a more uniform administration of items with both models. However, a larger number of test takers were required before the precision in ability estimation increased, which made this strategy ineffective with large item bank sizes. The P2 strategy generally resulted in a lower level of precision compared to the other two, because items were calibrated only at one point. An alternative method would be to use the P2 strategy with follow-up calibrations instead of simply calibrating one time. The use of random item selection with $b_i = 0$ for all parameters at the beginning of each strategy, led to good ability estimates for test takers with ability estimates near the mean; however, this method was inaccurate at estimating test takers with extreme ability values due to a consistent shrinkage toward the mean.

In a context where stakeholders need to know the level of precision in the test in order to make procedural decisions about how the test should be used, it might be important that test takers within the same phase are given the same probability of success. Here the P2 or the M strategy would be preferred over the C strategy because the precision in the C strategy is continuously improved.

The C and P2 strategies resulted in a non-uniform administration of items for the 2-PL model, because the item selection algorithm in the 2-PL model quickly resorted to selecting the items with high discrimination parameters at the expense of the other items. This resulted because the discrimination parameter has a multiplicative effect on the information for the items for the 2-PL model, which leads to the selection of items with greater information at specific points on the ability scale, over items that provide information across a broader area. This can be efficient when there is little error in ability and item parameter estimates; however, it is not optimal at the beginning of a test when there is a lot of insecurity concerning a test taker's

ability, and is undesirable when there is error in the item parameters. The use of the 2-PL model for these strategies could be a disadvantage because items can receive a small discrimination parameter by chance due to inconsistent answering in a small test taker population. Therefore, good items might never get the opportunity to be accurately calibrated and used in the test with the 2-PL model, which would result in a waste of resources for the test development organization. The optimal selection of items in the development phases of a test with the 2-PL model could also be an advantage, however, in settings where there is an abundant number of items, and it does not matter if some items are never used, because the algorithm in the 2-PL model concentrates on calibrating the items that are likely to be the best and most frequently used in the test.

A study by van der Linden and Glas (2000b) found dramatic impact of capitalization on estimation errors on ability estimation using the 2-PL model with a fully calibrated test. They highlighted four solutions for controlling the capitalization of error in ability estimation: cross validation, controlling the composition of the item pool, imposing constraints, and using the 1-PL model. The final two are possibilities for the current context. Imposing an exposure constraint would lead to a more uniform administration of items; however, the constraints would also limit the efficiency of the item selection algorithm. In the context of the 1-PL model all three calibration strategies resulted in improved ability estimates, in addition to a uniform calibration of items. The results suggest that the 1-PL model could be used in selecting items for the calibration phase of the test, and then once items have been accurately calibrated, the selection algorithm could switch to the 2-PL model.

The study also investigated the consequences of accounting for the uncertainty in the parameter estimates with the 1-PL model. Accounting for this uncertainty lead to lower precision in most contexts, and a slight increase in others. The mixed results and the extra calculation time needed to account for the uncertainty in the parameters suggests that a point estimate would be preferred in most settings, even though there is possible error in the parameter estimates in the calibration phases of a test.

The results of the study provide viable calibration design options for test development organizations that find it difficult to attract test takers in the development phases of a test. In these settings, these calibration strategies offer more cost effective and practical methods for developing large item banks, which makes it more attractive for smaller test development organizations to take advantage of the benefits of CAT. All three methods have the advantage over traditional booklet calibration designs in that they offer the possibility to assess test takers' ability throughout the calibration of the test. This makes it more attractive for test users and companies that purchase tests to become involved in the development phases of the test because the results

can be used. It is important for practitioners to be aware of the ethical and legal consequences of administering scores while the test is in the calibration phase. Therefore, it is vital to have clear guidelines about how the results should be used at different points in the calibration process.

The cost of developing a CAT compared to its benefits will always be compared to other test designs. It is considerably more expensive to develop a CAT compared to a linear test. However, the long term benefits of a CAT may outweigh the initial costs, because items can be used longer, since they are exposed less frequently in this format. A cost-benefit analysis based on the expected item exposure, and the benefits of CAT for the specific testing program, can be conducted before a decision to develop a CAT is made.

Future research could investigate the consequences of using the 2-PL model with item exposure constraints to investigate if it can lead to a uniform calibration of items while simultaneously estimating ability accurately. In this study, the assumption was made that items fit the model that was used; future research could also estimate the consequences of bad items by varying the degree to which the items fit the model. In addition accounting for uncertainty in parameters did not increase accuracy greatly in this study; however Bayesian methods could be explored in future studies to investigate if these models can lead to better ability estimates when accounting for uncertainty. In addition these models can be used to incorporate pre-existing hypotheses about item parameters. Finally, methods for filtering and assessing fit in items during the calibration process could be considered.

Chapter 3

Assessing and Modeling Measurement Invariance in Computerized Adaptive Testing

Abstract

Computerized adaptive tests (CAT) are commonly used because compared to fixed tests they are more accurate and take a shorter period of time to complete. In the process of developing a CAT several assumptions need to be empirically checked to confirm the measurement invariance of the particular instrument. The consequence of ignoring these assumptions is the possibility of systematic discrimination against members of a particular group. This can lead to possible legal and economic consequences if the test is used for making decisions in an organizational context. This article demonstrates a straightforward method for conducting a test of measurement invariance (MI) and illustrates a method for accounting for the lack of MI by assigning group-specific item parameters in the framework of IRT. The article exemplifies three applications of the method for an international organizational assessment context. These examples pertain to context effects due to the test administration method, motivation effects due to the stakes of the test, and language effects.

Key Words: *computerized adaptive testing (CAT), item response theory (IRT), measurement invariance (MI), differential item functioning (DIF), international assessment*

This chapter has been submitted for publication as:
Makransky, G., & Glas, C. A. W. (submitted). *Assessing and modeling measurement invariance in computerized adaptive testing.*

3.1 Introduction

Many personnel decisions within organizations are made based on the assessment of a given attribute under different conditions. Examples of these different conditions include stability of measurement through time (Golembiewski, Billingsley, & Yeager, 1975), across different populations (Riordan & Vandenberg, 1994), or using different mediums of test administration (Taris, Bok, & Meijer, 1998). The degree to which measurements conducted under these different conditions yield measures of the same attribute is known as measurement invariance (MI; Horn & McArdle, 1992). Vandenberg and Lance (2000) highlight the fact that inaccurate inferences are often made if the assumptions of MI are not assessed across these conditions. Furthermore, they stress the need for more focus on MI in organizational assessment.

After the work of Vandenberg and Lance (2000), the topic of MI has enjoyed increased attention among researchers and practitioners in organizational assessment. Although a great deal of this work has been conducted in the framework of confirmatory factor analysis (e.g., Vandenberg, 2002; Meade, Michels, & Lautenschlager, 2007), item response theory (IRT) approaches are more desirable when the equivalence of a single scale or specific scale items is of interest (Meade & Lautenschlager, 2004). Although the assessment of MI is necessary in many organizational assessment contexts, the importance of MI is even greater when test respondents are compared based on their responses to different items, such as the case in computerized adaptive testing (CAT; Zwick, 2010).

There are currently numerous operational CAT applications in a wide variety of fields including organizational assessment. Test development organizations choose to develop CAT's because they are more precise (Rudner, 1998, van der Linden & Glas, 2010), can be more motivating (Daville, 1993), and take a shorter period of time to complete (Rudner, 1998; Wainer, 2000) compared to fixed tests. Furthermore, there is now commercial software available for developing and administering a CAT (e.g., Weiss, 2008), and the number of researchers and practitioners with CAT knowledge has grown significantly over the last decade, as evidenced by the development of organizations such as the International Association for Computerized Adaptive Testing (IACAT).

Although CATs are more accessible and the advantages are evident, there are also several issues that complicate their implementation. One example is the complexity of developing the large item bank that is essential in CAT. In the process of developing an item bank several assumptions need to be empirically checked to confirm the MI of the particular instrument. For instance, there are often motivation differences between the calibration and the intended test populations (e.g., Makransky

& Glas, 2010). Furthermore, comparability studies should be made when a CAT is developed from an existing test (e.g., Green, Bock, Humphreys, Linn, & Reckase, 1984; Shaeffer, Steffen, Smith, Mills, and Durso, 1995; Wang, Jiao, Young, Brooks, & Olson, 2007). The consequence of ignoring these assumptions is the possibility of systematic discrimination against members of a particular group, which can have negative legal and economic ramifications for the company using the test.

MI can be checked by assessing if there are items where individuals at the same trait level but from different subgroups have unequal probabilities of responding correctly. This is known as assessment of differential item functioning (DIF). The presence of items with DIF results in bias against all of the members of a group in a fixed test, because all respondents are administered the same items. In a CAT, respondents are administered different items, therefore, the existence of items that exhibit DIF can produce bias within a group as well as between groups. The additional effect within groups occurs because not all respondents are administered the DIF items. So some are disadvantaged and others are not. Furthermore, fewer items are typically administered in a CAT. An item that exhibits DIF can consequently have a large effect on the test result. An item that exhibits DIF can also have major repercussions in a CAT because the sequence of items administered to the examinees depends in part on their responses to that item. Therefore, DIF can become a more critical issue for a CAT compared with a fixed test (Zwick, 2010).

Most of the current literature on DIF has focused on developing sophisticated statistical methods for detecting or "flagging" items that function differently across groups (Zumbo, 2007). Less literature has focused on practical methods for dealing with DIF when it exists. As a consequence, most literature assumes that items that produce DIF between groups should be identified and eliminated from the test. This approach is also evident in the international test commission (ITC) guidelines. According to the ITC test adaptation guideline D.9: "Test developers/publishers should provide statistical evidence about the equivalence of items in all intended populations". In his interpretation of the guideline, Hambleton (2005, p. 29) states that when performance is not equivalent, a sound reason must be available or the item should be deleted from the test.

Eliminating items that exhibit DIF can have two disadvantages. The first is that the items could in fact measure important components of the construct in both groups, but do so in a different way. The elimination of these items can leave gaps in the measurement of the construct that can make it difficult to maintain the validity of the test. The second disadvantage is that it can be costly to eliminate items. From the test developer's perspective this can result in a reluctance to eliminate items. Including

such items without taking DIF into account can lead to invariance in the results which undermines the validity of the decisions that are made with the test.

An alternative approach could be considered when the number of items that exhibit DIF is relatively small. This approach is to investigate if the items that exhibit DIF actually measure the same construct in both groups even if they do so in a different way. In IRT, such differences can be modeled by group-specific item parameters (e.g., van Groen, ten Klooster, Taal, van de Laar, & Glas, 2010; Weisscher, Glas, Vermeulen, & De Haan, 2010). This approach is only defensible if it can be explicitly shown that the responses to the items given in the two groups pertain to the same latent variable. In other words, the construct that is being measured must remain the same in both groups. This can be shown by investigating if the same IRT model holds for the entire set of response data (Glas, 1999).

The present article has two main objectives: First, we illustrate a straightforward method that can be used for investigating MI and for modeling DIF with group-specific item parameters. Second, we exemplify the method by investigating the possibility of modeling DIF with group-specific item parameters for a computerized adaptive cognitive ability test developed for organizational testing in Denmark and Sweden.

The remainder of the article follows the following format. First, we will describe a method used to investigate MI and introduce the possibility of modeling DIF with group-specific item parameters (virtual items). Next, we will illustrate three applications of the methodology that are typically necessary when developing a CAT in an international organizational context. These include context effects due to the test administration method, motivation effects due to the stakes of the test, and language effects. Finally, we discuss practical issues, and look ahead at possible future applications.

3.2 Modeling DIF with group-specific item parameters

The present study was carried out in the framework of IRT. The fundamental concept of IRT is that each test item is characterized by one or more parameters and each test-taker is characterized by ability parameters, in this study by a single ability parameter. The probability that a given test-taker answers a given item correctly is given by a function of both the item's and the test taker's parameters. Conditional on those parameters, the response on one item is independent of the responses to

other items. The IRT model used in this study, is the two-parameter logistic, or 2-PL model,

$$P_i(\theta) = \frac{1}{1+\exp(-a_i(\theta-b_i))} \tag{3.1}$$

(Birnbaum, 1968). Here $P_i(\theta)$ is the probability of a correct response for item i, θ is the test taker's ability, and a_i and b_i are item parameters. Further, a_i is called the discrimination and b_i the difficulty parameter.

The first step in modeling DIF is to use model-fit statistics to identify the items with DIF. There are many methods available for conducting such an analysis (e.g., Holland & Thayer, 1988; Swaminathan & Rogers, 1990; Zwick & Thayer, 2002). In the present study, we use two statistics which are asymptotically equivalent, but highlight different aspects of the data: the Wald statistic and the Lagrange Multiplier (LM) statistic. Both test the assumption that item parameters are the same in the two or more subgroups of the sample. The Wald statistic directly contrasts the parameter estimates in the subgroups and supports making graphics where misfitting items can be easily identified. The LM statistic is a general tool for the evaluation of fit to IRT models, and in addition to the evaluation of DIF, it can also be used for evaluation of other assumptions of IRT, such as the form of the response curves and local independence (Glas, 1998, 1999). The LM test was used since the IRT model should not only fit between groups but also within groups. The general framework of the statistic is as follows. The sample of respondents is divided into subsamples from subpopulations labeled g = 1,....,G. These might be either focal and reference groups used to investigate DIF; or score-level groups used to evaluate model fit within subpopulations; or even a combination of the two. The statistic is based on the difference between average observed scores on every item i in the subsamples, that is,

$$S_{ig} = \frac{1}{N_g}\sum_{(n|g)}^{N_g} b_{ni}\, X_{ni} \tag{3.2}$$

where summation is over the N_g respondents in the subgroup. X_{ni} is either an observed response (here 0 or 1) or a dummy (say, 9) if the response is unobserved. Further, b_{ni} is equal to 1 if the response is observed or 0 otherwise. Using this coding, the statistic can both be used for a fixed test where all respondents respond to all items, and a CAT, where respondents are administered only a subset of items from the whole item pool. These values are compared to their posterior expectations $E(S_{ig})$. The differences are squared and weighted by their covariance matrix (for more details see Glas, 1998, 1999). It can be shown that the hypothesis tested is equivalent to testing the hypothesis that the parameters of the items are equal for the subgroups. The LM statistic has an asymptotic chi-squared distribution with *G-1* degrees of freedom. The statistic can be accompanied by the effect sizes

$$d_{ig} = \max_g \left| S_{ig} - E(S_{ig}) \right| \quad (3.3)$$

which show the seriousness of the model violation in the metric of the observed score scale. So the effect sizes d_{ig} are on a scale ranging from 0 to the maximum score m_i (in this study, m_i = 1). As a rule of thumb, effect sizes $d_{ig} > 0.10$ can be considered indicative of more than minor model violations (van Groen et al., 2010, p. 1257). Since the power of the LM test increases with sample size, the effect sizes are usually considered more important than the significance probabilities.

Once items that exhibit DIF have been identified, the next step is to model the DIF in such a way that the measures obtained in the subgroups are still comparable. This can be done by dividing each DIF item into several virtual items, one for each subgroup. Each virtual item is then given group-specific item parameters. Therefore, it is assumed that the same construct is measured in all groups, but the item parameters are different for some groups. If it can be shown that the items without DIF and the items with the group-specific item parameters fit a concurrent IRT model, the conclusion that all items relate to the same underlying attribute is supported (Glas, 1999; Glas & Verhelst, 1995). Again, this can be investigated by computing LM statistics targeted at the form of the item response curves and at the assumption of local independence. The latter assumption implies that item responses are independent given a person's value on the latent variable. If this does not hold it means that, other, unaccounted, variables influenced response behavior and unidimensionality is violated. Thus, if a concurrent estimate of the parameters of all the groups and all the items is obtained, the estimated person parameters of all groups are on the same scale and can be meaningfully compared (Weisscher et al. 2010, p. 545). This means that all items can be used to estimate the value of the person parameters and can contribute to the precision of the estimates.

3.3 Method

3.3.1 Instruments

The Master Competence Analysis (MCA) is a cognitive ability test that measures competence in logical, analytical reasoning. The MCA is used by human resources departments within multinational companies for making employment decisions. The test is made up of dichotomously scored free response and multiple choice items.

Three different versions of the test will be used in the applications presented in the following pages.

1) The computer based fixed version of the MCA (MCA-CBT) consisting of 18 items. This test was used for high-stakes selection and recruiting decisions.
2) A randomly administered test with an item bank of over 300 items. A total of 201 items that fit the 2-PL model are included in the analyses in this article. The test was used in a low-stakes sample with the intention of calibrating new items for the MCA-CAT item bank.
3) The CAT-MCA with an item bank of 201 items. The 81 items with the highest exposure rates are included in the analyses in this article. Test respondents were administered a maximum of 24 items adaptively from the item bank. This test was used for high-stakes selection and recruiting decisions.

3.3.2 Statistical Analyses

The item parameters for the different versions of the MCA were estimated by marginal maximum likelihood (MML; Bock & Atkin, 1981). DIF and fit within subpopulations were examined using the Wald and LM statistics. These were calculated using the free MIRT software package (Glas, 2010). Items were flagged for DIF in case of a significant Wald test or for effect sizes $d_{ig} > 0.10$ in the LM test. This was done in an iterative procedure until the item bank contained no common DIF items.

3.4 Applications

3.4.1 Study 1: Investigating MI across context effects between CBT and CAT formats

The sample of respondents for this study consisted of 2318 job candidates who had taken either the CBT (N = 1517) or CAT (N = 801) version of the MCA. The item analysis consisted of 18 items for the CBT and 81 items for the CAT. A total of 15 of the items were common items across the two tests. The two samples were similar and consisted of job candidates with varied educational backgrounds, working in a variety of job categories.

When a CAT is developed based on an existing fixed test, professional standards require that test equivalence be demonstrated empirically (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999). Direct comparability of scores is also necessary

to make valid comparisons of job candidates. However, research results have demonstrated lack of MI across CBT and CAT formats. For instance, Shaeffer and his colleagues (1995) conducted a comparability study between the CBT and the CAT version of three GRE measures. They found MI for the analytic measure, but the numeric and verbal measures were comparable across test formats. Wang and his colleagues (2007) conducted a meta-analysis of comparability studies for K-12 reading and math achievement for CBT and CAT. Their findings suggest that the adaptive algorithm can have an effect on the test results. Here we investigate context effects between CBT and CAT in organizational cognitive ability testing.

The first step in the analysis was to evaluate the model fit within the two populations independently. In general the model fit the data. Five of the 81 items in the CAT, and none of the 18 items in the CBT had effect sizes above $d_{ig} > 0.10$ on the LM test. The next step was to compute the Wald test for the parameter estimates in the subgroups. Figure 3.1 displays the MML estimates of discrimination and difficulty parameters for the two samples independently. The comparability of the two estimates is based on norming the two ability scales by setting the means and standard deviations equal to zero and one, respectively. Subsequent analyses showed that the difference between the two ability distributions was similar. The Wald test identified four items with significantly different item parameters (item numbers: 4, 5, 11,15).

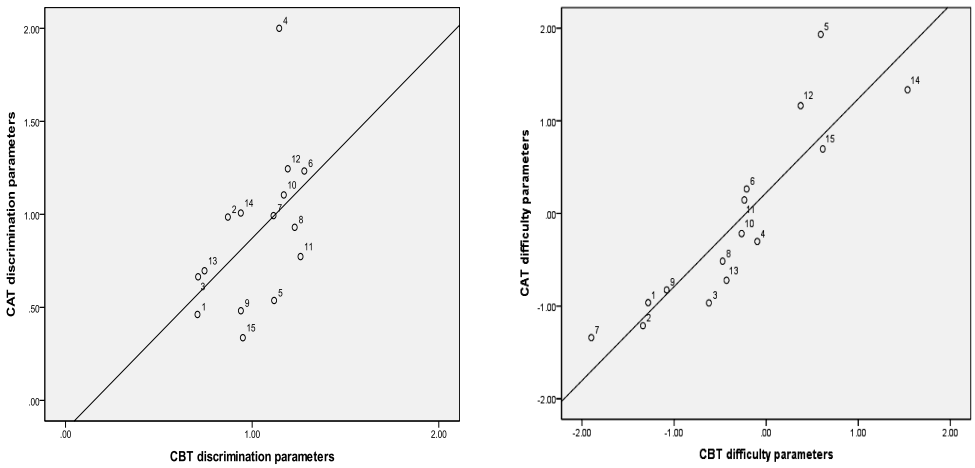

Figure 3.1: Item parameters estimated independently for the CBT and the CAT versions of the MCA.

The next step in the analysis was to conduct a common item equating analysis for the CBT and CAT versions of the MCA. The anchor items for this analysis were the 15 common items that were included in both tests. The scale was identified by

setting the mean and standard deviation for the sample that took the CBT to zero and one, respectively. The MML estimate of the mean for the sample that took the CAT was -0.17 (SE = 0.06). In this analysis the LM statistic for the 15 common items were computed to investigate DIF further. Table 3.1 presents the results of the first iteration of this analysis. The columns in Table 3.1 represent the item number, the LM test statistic, the significance probability, the observed and expected proportions of correct responses for the CBT and CAT samples respectively, and the signed effect sizes. The signed effect sizes show the direction of DIF. For instance, item 14 had context DIF favoring the CBT group, and item 15 had context DIF favoring the CAT group.

Table 3.1

LM tests for DIF between CBT and CAT samples for the 15 common items (df=1)

			CBT sample		CAT sample		Effect
Item	LM	Prob	Obs	Exp	Obs	Exp	Size
1	0.00	1.00	0.69	0.69	0.76	0.76	0.00
2	1.54	0.21	0.74	0.72	0.76	0.77	0.02
3	4.31	0.04	0.63	0.58	0.64	0.65	0.05
4	0.23	0.63	0.29	0.32	0.52	0.52	-0.03
5	1.87	0.17	0.25	0.34	0.38	0.38	-0.09
6	3.68	0.06	0.54	0.52	0.53	0.54	0.02
7	11.04	0.00	0.64	0.72	0.82	0.82	-0.09
8	2.05	0.15	0.52	0.50	0.58	0.58	0.02
9	0.07	0.79	0.63	0.63	0.72	0.72	0.01
10	2.44	0.12	0.52	0.49	0.55	0.55	0.02
11	8.90	0.00	0.47	0.51	0.54	0.52	-0.03
12	25.65	0.00	0.27	0.35	0.42	0.40	-0.08
13	6.52	0.01	0.61	0.55	0.60	0.61	0.06
14	18.54	0.00	0.42	0.30	0.22	0.23	0.12
15	2.38	0.12	0.43	0.56	0.37	0.37	-0.12

The items with effect sizes over $d_{ig} > 0.10$ on the LM test were split into virtual items with group-specific item parameters in an iterative procedure. In the first iteration presented in Table 3.1, two items (14 and 15) were assigned group-specific item parameters. In the subsequent steps, the number of items assigned group-specific parameters were two, and two. The results of the combined analyses for the Wald and the LM tests were that seven items (item numbers: 4, 5, 7, 11, 12, 14, 15) were assigned group-specific item parameters.

Concurrent analyses using the LM test were conducted to test whether the combined model including the seven items with group-specific item parameters fit the data. Tables 3.2 and 3.3 illustrate that the model does fit the data. Table 3.2 illustrates that the LM test of DIF for the remaining eight common items displayed no significant DIF across contexts.

Table 3.2

LM tests for DIF between CBT and CAT samples for the remaining eight common items (df=1)

			CBT sample		CAT sample		Effect
Item	LM	Prob	Obs	Exp	Obs	Exp	Size
1	0.53	0.47	0.76	0.75	0.69	0.70	0.00
2	0.02	0.88	0.76	0.76	0.74	0.74	0.00
3	1.75	0.19	0.64	0.65	0.63	0.60	-0.01
6	0.02	0.88	0.53	0.53	0.54	0.54	0.00
8	0.15	0.69	0.58	0.58	0.52	0.52	0.00
9	0.59	0.44	0.72	0.71	0.63	0.65	0.00
10	0.12	0.73	0.55	0.54	0.52	0.52	0.00
13	2.90	0.09	0.60	0.60	0.61	0.57	-0.01

Table 3.3 presents an example for the results of the LM test of fit for the form of the item response curves in the CBT population. Since concurrent estimates obtained using both groups were used, the tests target non-uniform DIF. The table is similar to the previous tables with the exception of the observed and expected values. In this table these values represent three score-level groups instead of focal and reference groups. The results in Table 3.3 indicate that none of the 15 common items had LM effect sizes that were greater than $d_{ig} > 0.10$. Similar results were found in the CAT population. Therefore, we conclude that the items all measure the same construct and the fit of the resulting model is statistically acceptable.

Table 3.3

LM tests for model fit for the 15 common items in the CBT population (df=2)

			Score group 1		Score group 2		Score group 3		Effect
Item	LM	Prob	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Size
1	0.01	1.00	0.62	0.62	0.75	0.75	0.86	0.85	0.00
2	2.54	0.28	0.59	0.60	0.78	0.76	0.86	0.87	0.02
3	12.03	0.00	0.43	0.48	0.66	0.63	0.76	0.77	0.03
4	0.02	0.99	0.31	0.31	0.53	0.53	0.73	0.74	0.00
5	4.54	0.10	0.17	0.19	0.40	0.36	0.58	0.59	0.02
6	0.68	0.71	0.31	0.30	0.54	0.54	0.77	0.77	0.01
7	2.30	0.32	0.66	0.67	0.85	0.83	0.92	0.93	0.01
8	9.33	0.01	0.39	0.35	0.57	0.60	0.80	0.80	0.02
9	0.73	0.69	0.53	0.52	0.71	0.70	0.84	0.85	0.00
10	2.68	0.26	0.34	0.32	0.56	0.56	0.77	0.78	0.01
11	9.31	0.01	0.32	0.30	0.54	0.55	0.77	0.78	0.01
12	7.57	0.02	0.22	0.20	0.39	0.40	0.65	0.65	0.01
13	1.57	0.46	0.45	0.46	0.61	0.62	0.75	0.76	0.01
14	5.85	0.05	0.10	0.10	0.21	0.19	0.33	0.35	0.02
15	1.19	0.55	0.20	0.21	0.37	0.35	0.53	0.54	0.01

Accounting for DIF in the model had an effect on the mean scores for the CAT respondents. The MML estimate of the mean for the CAT sample became -0.03 (SE = 0.07) after splitting the DIF items into virtual items with group-specific item parameters. The mean and SD for the CBT sample was set to zero and one, respectively. Therefore, the two samples had very similar mean scores once the items exhibiting DIF had been modeled with group-specific item parameters.

The consequences of using the equating design was investigated further in terms of the purpose of the MCA. In an applied setting the MCA is used for selecting qualified job candidates in terms of their level of cognitive ability. A cut-off point was set at the ability score of zero in this study. This would be a cut-off point that could be used in an applied setting where a company is interested in selecting from the top pool of candidates based on their cognitive ability. Given this cut-off point the candidates that would be selected for consideration is outlined in Table 3.4. For the sample of CAT candidates, 311 (39%) would be considered for the job in both analyses. However, 64 additional candidates would have scores that qualified them for the job after accounting for DIF across test contexts. This means that ignoring DIF has a negative effect on 8% of the sample of candidates that took the CAT. In the CBT, 739 (49%) would be considered for the job in both analyses. Accounting

for context DIF across tests would change the decision for 0.8% of this sample. Two candidates would be positively affected, and 10 candidates would be negatively affected.

Table 3.4

Selection decisions with a cut-off point of zero for the CAT and CBT versions of the MCA in a hypothetical selection procedure

	MCA-CAT			MCA-CBT		
		DIF is ignored			DIF is ignored	
	Selection decision	No	Yes	Selection decision	No	Yes
Model with group-specific item parameters	No	426	0	No	766	10
	Yes	64	311	Yes	2	739

3.4.2 Study 2: Investigating MI across the stakes of the test

An accurately calibrated item bank is essential for a valid computerized adaptive test. When a fixed test with established item parameters is available, it is possible to use these items in an equating study to establish the item parameters for use in the CAT item bank. However, in some settings, such as organizational testing, there is limited access to test takers for calibration. As a result, collecting data to accurately calibrate an item bank is usually difficult. This study investigated the effects of using a randomized item administration design in a low-stakes sample to establish the item parameters of the new CAT item bank which should function in a high-stakes situation.

The item bank was administered to a sample of over 9000 respondents in a low-stakes setting. The respondents were recruited through Google add words commercials where they were offered a free cognitive ability test result for participating. The data were filtered by eliminating respondents who did not take the test seriously. This was assessed based on several data filtering techniques one of which is the elimination of short answering times according to the model described by Wise and DeMars (2006). The final sample consisted of 4768 test respondents.

Each test respondent was administered between 10 and 20 randomly selected items. An effort was made to keep the test short in order to ensure that the low-stakes sample maintained a relatively high level of motivation. The item trial was used to assess item parameters that could be used in the CAT version of the test. The item

parameters from a sample of 1362 job candidates who had taken the CBT version of the MCA in Danish was used as an anchor. The 18 items from the CBT were included in random calibration test in order to form an anchor for the calibration design.
The analysis procedure in this study was analogous to Study 1, therefore we only present the main results of this study. The first step in the analysis was to evaluate the model fit within the two populations independently. The model fit the data, because none of the 201 items in the random test or the 18 items in the CBT had effect sizes above $d_{ig} > 0.10$ on the LM test. The next step was to compute the Wald test for the parameter estimates in the subgroups. Figure 3.2 displays the MML estimates of discrimination and difficulty parameters for the two samples independently. The two ability scales were normed by setting the means and standard deviations equal to zero and one respectively. The regression line is shifted in this graph because the ability distributions for the two samples are not the same. Accounting for this shift, the Wald test identified nine items with significantly different item parameters (item numbers: 1, 2, 3, 4, 11, 12, 13, 15,18).

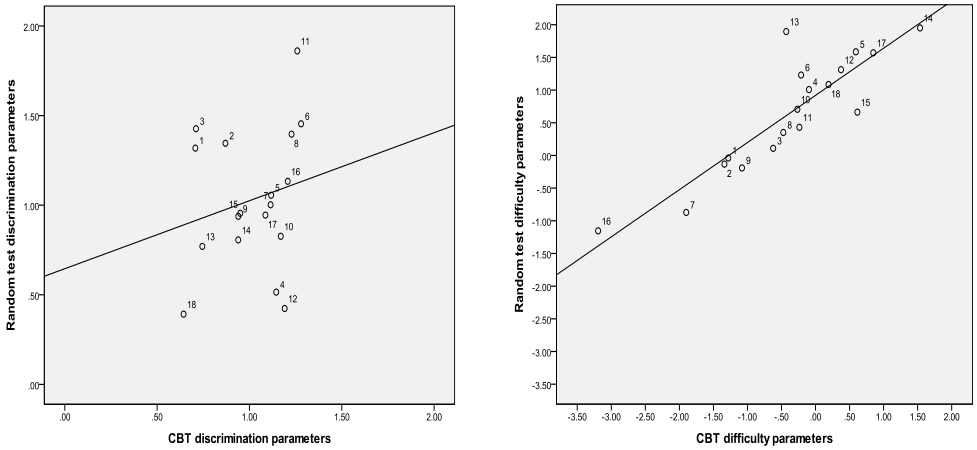

Figure 3.2: Item parameters estimated independently for the CBT and the random calibration versions of MCA.

A common item equating analysis for the CBT and the randomly administered calibration versions of the MCA was conducted. The scale was set to have a mean of zero and a standard deviation of one for the high-stakes sample that took the CBT. The MML estimated mean for the low-stakes calibration sample was -0.95 (SE = 0.02). This means that the low-sakes calibration sample scored significantly lower than the high-stakes operational sample.

Then, a DIF analysis using the LM statistic for the 18 common items was conducted. The results of the combined analyses with the Wald and LM tests were that nine items (item numbers: 1, 2, 3, 4, 11, 12, 13, 15,18) were assigned group-specific item parameters.

The LM test was also conducted to test whether the model including the nine items with group-specific item parameters fit the data. None of the remaining nine common items displayed significant DIF across the stakes of the test. Furthermore, the LM test of fit for the form of the item response curves indicated that none of the 18 common items had LM effect sizes that were greater than $d_{ig} > 0.10$. Therefore, we conclude that the items all measure the same construct and the fit of the resulting model is statistically acceptable.

Taking the DIF into account in the model had an effect on the mean scores for the low-stakes calibration test respondents. The MML estimated mean for the CAT sample became -0.80 (SE = 0.02) after splitting the DIF items into virtual items with group-specific item parameters. This was an increase in the average scores of 0.15 compared to the analysis where DIF was ignored.

The consequences of using the equating design was investigated further based on the purpose of the study. Although the low-stakes calibration respondents were not job candidates, part of the agreement for participating in the study was to obtain a realistic cognitive ability estimate. Given this purpose, we compare the results of the two analyses. Figure 3.3 presents a scatter plot where the scores from the analysis where DIF is ignored is on the vertical axis, and the difference between the two ability estimates is on the horizontal axis. The difference was calculated by subtracting each respondents ability score where DIF was modeled with virtual items from their score where the DIF was ignored. The figure clearly illustrates that the respondents would get a more favorable score after modeling the DIF in the items. It is also apparent that the effect is conditional on ability. Specifically, the difference between the scores was larger for the respondents who had low ability estimates.

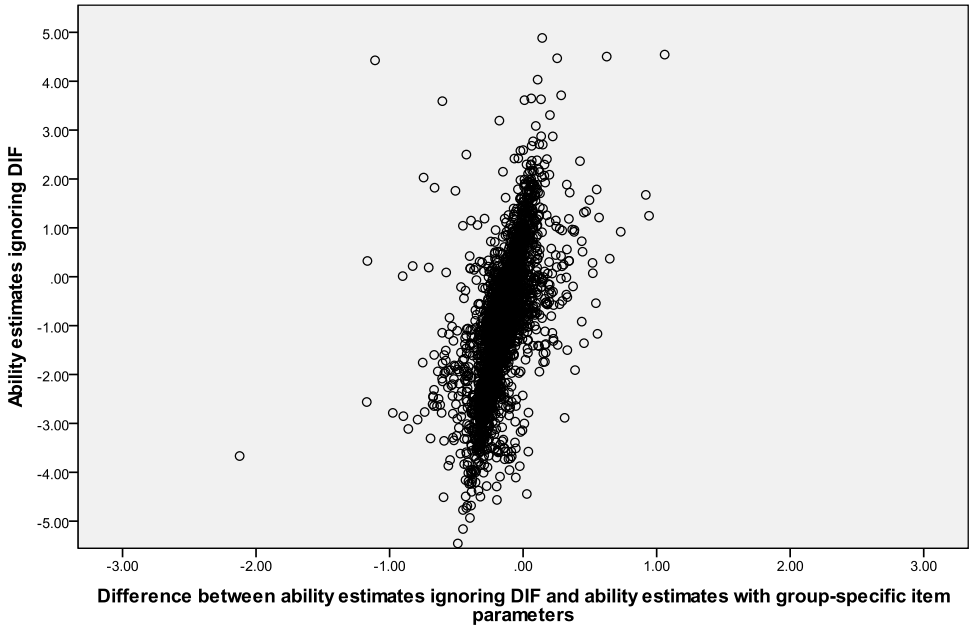

Figure 3.3: Scatter plot representing the difference in the ability estimates for the analysis where DIF was ignored on the vertical axis, and the difference between this analysis and the analysis with group-specific item parameters on the horizontal axis.

In conclusion this study demonstrated that there are motivation differences between low and high-stakes samples that can be modeled by calibrating the low and high-stakes samples in a common item calibration design. In addition, there are some items that exhibit DIF across the stakes of the test. Modeling DIF resulted in mean score changes of 0.15. The changes were more significant for low ability respondents.

3.4.3 Study 3: Investigating MI across languages

Ensuring linguistic and cultural equivalence across various language versions of assessment instruments has become an increasingly crucial challenge. Factors such as globalization and the increasing mobility of the world's workforce, the emergence of complex multicultural societies such as the European Union, and the continuous increase in the number of countries participating in international comparative assessments are examples of trends that have enhanced the importance of this challenge. Evidence also suggests that the need for multi-language versions of achievement, aptitude, and personality tests, and surveys will continue to grow (e.g.,

Ercikan, 2002; Hambleton, 2002, 2005; Hambleton & de Jong, 2003; Choi et. al, 2009; Matsumoto & van de Vijver, 2010). Therefore, methods are needed to easily access and model language differences in tests that are used across cultures, in order to provide an easy and flexible means of modeling these differences.

The applicability of the method for modeling DIF described in this study was investigated across the Danish and Swedish language versions of the MCA-CAT. The 79 items with the highest exposure rates from the CAT item bank were used. Data from 801 test respondents from the Danish (517) and Swedish (284) language versions of the MCA-CAT were used for this study. A test respondent is typically administered the language version that s/he feels most comfortable with. Therefore, comparisons are often made across language versions of the instrument.

The analysis procedure in this study was analogous to the previous studies, therefore we only present the main results of this study. The first step in the analysis was to evaluate the model fit within the two language versions of the test independently. Three of the 79 items in the Danish and two items in the Swedish version of the test had effect sizes above $d_{ig} > 0.10$ on the LM test. However, in general the results suggest that the model has acceptable fit.

A common item equating analysis was conducted by setting the mean to zero and a standard deviation to one for the Danish language version sample. The MML estimated mean for the Swedish language version of the MCA was 0.18 (SE = 0.11). DIF was investigated across languages with the Wald and LM tests as in the previous studies. A total of 10 items were assigned language-specific item parameters based on the results of the Wald and LM tests.

The LM test was also conducted to test whether the model including the 10 items with language specific item parameters fit the data from the MCA-CAT. The 69 items that were not split into virtual items displayed no DIF over the effect size value of $d_{ig} > 0.10$ between the Swedish and Danish versions of the test. The LM test of fit for the form of the item response curves identified five items with a LM effect size that was greater than $d_{ig} > 0.10$ in the Swedish, and four items in the Danish version of the test. None of these nine items had been given language-specific item parameters. Furthermore, the items fit the 2-PL model when the language versions were combined. Therefore, it is likely that these items did not fit the model because the sample sizes in the individual language versions of the test became small. Therefore, we conclude that the fit of the resulting model is statistically acceptable.

The MML estimated mean for the Swedish language version of the MCA became 0.11 (SE = 0.11) after modeling DIF with language-specific item parameters. The consequences of using virtual items with language-specific item parameters was investigated further with a hypothetical selection procedure. A cut-off point was set

at the ability score of zero as in Study 1. Given this cut-off point the candidates that would be selected for consideration is outlined in Table 3.5. Table 3.5 presents two contingency tables (one for the sample of Swedish, and one for the sample of Danish MCA-CAT respondents). The table outlines the number of respondents who would be selected as valid job candidates. For the Swedish sample, 151 (53%) candidates would be considered for the job in both analyses. However, the decision would change for 12 candidates (4%). Three of these candidates would benefit from the inclusion of language-specific item parameters. By contrast nine candidates would no longer be considered because the results show that their score was artificially inflated by responding to items that had DIF that favored the Swedish language version of the test. For the sample that took the Danish version of the test, a total of 239 (46%) would be considered for selection in both analyses. Furthermore, nine candidates would benefit from the inclusion of language specific item parameters, and six would not.

Table 3.5:

Selection decisions with a cut-off point of zero for the Swedish and Danish language versions of the MCA-CAT in a hypothetical selection procedure

	Swedish			Danish		
		DIF is ignored			DIF is ignored	
	Selection decision	No	Yes	Selection decision	No	Yes
Model with group-specific item parameters	No	121	9	No	263	6
	Yes	3	151	Yes	9	239

The results indicate that modeling DIF with group-specific item parameters did not have a considerable impact on the MCA-CAT. This may have been a consequence of the similarity between the Danish and Swedish languages. Additionally, considerable effort was made to ensure that the ITC guidelines for translating and adapting tests (2010) were followed when adapting the MCAT to the different languages (for more information about this process refer to Makransky & Kirkeby, 2007). Therefore, the DIF effect sizes that were observed were not large. Greater effects would be expected when language versions of the test differ greatly.

3.5 Discussion

Lack of measurement invariance (MI) can result in systematic discrimination against members of a group thereby undermining the validity of the decisions that are made with a test. This can lead to possible legal and considerable economic consequences if the test is used for making decisions in an organizational context. Nonetheless, checking for MI is typically not one of the first analyses performed when establishing the validity of a test (Vandenberg & Lance, 2000). This article presented a straightforward method for conducting a test of MI and modeling differential item functioning (DIF) by assigning group-specific item parameters in the framework of IRT. Instead of eliminating items that exhibit DIF, this method provides a flexible way of scoring tests with items which measure the same construct across groups even though they do so in a different way. This method is particularly useful in the development of a computerized adaptive test (CAT) because there is a need to test several assumptions related to MI in this process, and because DIF can have a greater impact on the results compared to a fixed test (Zwick, 2010).

The article exemplified three applications of the method for the master competence analysis (MCA), a cognitive ability test used in international organizational assessment. These examples pertained to context effects due to the test administration method, motivation effects due to the stakes of the test, and language effects. In general, the study demonstrated that MI was violated across test administration contexts, the stakes of the test, and across languages. The results showed that simply ignoring the potential presence of DIF would result in decisions that could affect a significant proportion of test respondents. Furthermore, the studies provided evidence that modeling DIF with group-specific item parameters is a viable methodology for making comparisons across contexts without the need to eliminate items from a test. That is, evidence was provided that showed that the same IRT model fit the data across groups with a combination of common items and virtual items with group-specific item parameters, thereby providing evidence of MI.

Specifically, the results of Study 1 indicated that there were context effects that violated MI across the CBT and CAT versions of the MCA. Seven of the 15 common items exhibited DIF across formats. Modeling DIF with group-specific item parameters had a significant effect on the conclusions derived from a comparison of the CAT and CBT samples. A common item equating analysis ignoring the possibility of DIF resulted in the conclusion that the sample of CAT respondents had a significantly lower score (Mean = -0.17, SE = 0.06), compared to the CBT sample (Mean = 0). Modeling DIF with group-specific item parameters lead to the conclusion that the CAT sample (Mean = -0.03, SE = 0.07) did not score significantly lower

than the CBT sample (Mean = 0). Similarly, modeling DIF with group-specific item parameters had an impact on the decisions that were made in a hypothetical selection procedure. Selection decisions changed for 8% of the CAT sample and 0.8% of the CBT sample.

Study 2 also identified the lack of MI across the stakes of the test. Nine of the 18 common items exhibited DIF across the low-stakes volunteer and the high-stakes operational samples. A common item equating analysis ignoring the possibility of DIF resulted in the conclusion that the sample of low-stakes respondents had a mean score that was 0.95 (SE = 0.02) lower than the high-stakes sample. Modeling DIF with group-specific item parameters leads to the conclusion that the low-stakes sample only had a mean score that was 0.80 (SE = 0.02) lower than the high-stakes sample. These findings are vital because an incorrect assessment of the difference between the groups can have a significant impact on the future assessment decisions that are made with the operational CAT, since items in a CAT are often calibrated in a low-stakes sample (e.g., Makransky & Glas, 2010).

Finally, Study 3 identified 10 of 79 items with DIF across the Danish and Swedish language versions of the MCA. A common item equating analysis ignoring the possibility of DIF resulted in the conclusion that the Swedish sample (mean = 0.18, SE = 0.11) had a mean score that was not significantly higher than the Danish sample (mean = 0). The conclusion remained the same after modeling DIF with group-specific item parameters (Swedish mean = 0.11; SE = 0.11). Nonetheless, the consequence of modeling DIF with group-specific item parameters did have an impact on the decisions that were made in a hypothetical selection situation. Selection decisions changed for 4% of the Swedish and 3% of the Danish sample. Although the consequence of modeling DIF across language versions of the MCA did not result in large effects, it is likely that the impact of this method would be larger when the cultural and language differences are greater. An example is a preliminary analysis of the MCA-CBT across Danish and Polish language versions (not described in this study). The results of this analysis indicate that the effect of DIF across these language versions is considerably greater than that identified between the Danish and Swedish versions.

3.5.1 Implications for Practice

Although MI has enjoyed increased attention among researchers and practitioners, it is still common practice for test developers and organizational researchers to make assumptions of MI across different testing conditions. Our results suggest that such assumptions are not warranted in the development of a CAT. We recommend that test administrators conduct MI analyses to ensure that the properties are the same for the

different formats and test contexts before organizational decisions are made on the basis of the scores. Since the impact of MI can systematically discriminate against test respondents within a particular condition or group, the assessment of MI should be conducted alongside other commonly applied procedures such as calibrating item parameters, and checking the reliability of the test.

The method described in this article for modeling DIF with group-specific item parameters is easy to implement because it can be applied with a free software program (MIRT, Glas, 2010) that is available on the Internet. Furthermore, it is easy to develop and maintain an item bank consisting of virtual items with group-specific item parameters with standard item banking software (e.g., Weiss, 2008). This method of modeling DIF provides more flexibility than the traditional method of eliminating items when they display DIF across groups. There are many examples of applied settings where such flexibility is needed. For instance, many test development companies produce tests that are available across dozens of languages, and the presence of DIF is inevitable. In such a setting it would be quite effective if the differences between the language versions of the test could be modeled instead of having to eliminate the items that do not function similarly across all languages.

3.5.2 Future Research

The present article examined three applications of the method. Future research could investigate the methodology for different operational comparisons where violations of MI may be greater. Some examples are the transformation of a fixed speeded test to CAT, and a setting where the language differences are larger. Similarly, the examples assessed in this article investigated comparisons across two levels. There are several applied settings where there is a need to assess MI across a large number of groups or conditions. One example is the need to assess MI across several language and cultural groups when a test is available in many cultures. This is a typical challenge for international test providers, and in international comparative assessments. Future research should investigate the feasibility of this method in such conditions.

Chapter 4

Unproctored Internet Test Verification: Using Adaptive Confirmation Testing

Abstract

Unproctored Internet testing (UIT) is commonly used in employment test administration. When the test is high stakes, the International Guidelines on Computer-Based and Internet-Delivered Testing recommend to follow up the results with a confirmation test in a controlled setting. This article proposes and compares two methods for detecting whether a test taker's original UIT responses are consistent with the responses from a follow-up confirmation test. The first method is a fixed length adaptive confirmation test using the likelihood ratio (LR) test to evaluate cheating and the second method is a variable length adaptive confirmation test using an extension of the stochastic curtailed truncated sequential probability ratio test (SCTSPRT) to evaluate cheating. Simulation studies indicated that the adaptive confirmation test using the SCTSPRT was almost four times shorter while maintaining the same detection power. The study also demonstrated that cheating can have a detrimental effect on the validity of a selection procedure and illustrated that the use of a confirmation test can remedy the negative effect of cheating on validity.

Key Words: *unproctored Internet testing, computer adaptive testing, item response theory*

This chapter has been published as:
Makransky, G., & Glas, C. A. W. (2011). Unproctored Internet test verification: Using adaptive confirmation testing. *Organizational Research Methods, 14,* 608-630.

4.1 Introduction

The increase in computer availability and the widespread use of the Internet have led to an increased application of unproctored tests that can be administered at any place and time via the World Wide Web. Unproctored Internet testing (UIT) is Internet-based testing of a candidate without a traditional human proctor. Although definitive usage data are lacking, it is probable that UIT accounts for the majority of individual employment test administrations that currently take place in the U.S. private sector (Pearlman, 2009). Fallaw, Solomonson, and McClelland (2009) found that more than two thirds of the employers who conduct testing for selection are already engaging in UIT.

The flexibility of Web-based test administration is attractive because it limits the resources necessary for administering tests, meaning that test proctors need not be hired, trained, or sent to testing locations; testing equipment does not have to be purchased, distributed, or maintained; and job candidates do not have to travel to testing locations. UIT also allows continuous access to assessments and limits the time it takes to process candidates.

The primary disadvantage of UIT stems from the many forms of cheating that are possible such as assistance from others who have knowledge of the items before the test, assistance from others during the test, or the substitution of test takers (Tippins, 2009). There is evidence that cheating is widespread in unproctored as well as proctored settings. Automatic Data Processing Inc. (2008) reports data indicating that 45% of job applicants falsify work histories. These applicants would presumably be willing to cheat if given the opportunity. Cizek (1999) and Whitley (1998) found that approximately half of all college students report cheating on an exam at least once during their college education. There is also abundant literature motivated by the concern that applicants may fake their responses in organizational testing in an attempt to obtain employment (e.g., Anderson, Warner, & Spector, 1984; LaHuis & Copeland, 2009; Ones, Viswesvaran, & Reiss, 1996). It is likely that cheating in UIT is particularly widespread because proctoring is the primary means by which cheating is curtailed (Arthur, Glaze, Villado, & Taylor, 2009). In fact, Chapman and Webster (2003) describe cheating as one of the most common reasons that human resources professionals are hesitant to implement UIT.

Some possible solutions for limiting the detrimental effects of cheating in a high stakes unproctored Internet ability test are to use a speeded test (e.g., Arthur et al., 2009; Nye, Do, Drasgow, & Fine, 2008) or to have a two-step testing process where a confirmation test is used to verify the results of the first test (e.g., Beaty, Dawson, Fallaw, & Kantrowitz, 2009; Burke, van Someren, & Tatham, 2006).

Nye et al. (2008) did not detect any differences due to cheating when comparing proctored and unproctored test results for a perceptual speed test in a selection setting. Similarly, Arthur et al. (2009) concluded that a speeded test format appears to reduce the prevalence of cheating. However, more research is needed before a conclusion about the validity of using a speeded test format without a verification test for UIT can be made.

Currently, the International Guidelines on Computer-Based and Internet-Delivered Testing (2005), developed by the International Testing Commission, explicitly stipulate the need for verification testing. Guideline 45.3 states:

> For moderate and high stakes assessment (e.g., job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e., at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores. Procedures should be used to check whether the test-taker's original responses are consistent with the responses from the confirmation test. Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (e.g., not seek assistance, not collude with others, etc.). This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept.

The most commonly used method for conducting a confirmation test is the test–retest approach. With this approach, an unproctored test is used as a prescreen, followed by a proctored administration of the same or a parallel version of the test in a proctored setting. Variations are to use randomized item selection (e.g., Burke et al., 2006) or computerized adaptive testing (CAT).

There are many statistical methods available for detecting forms of cheating such as collusion in proctored settings, including Angoff's b index (Angoff, 1974), error similarity analysis (Belleza & Belleza, 1989), and the Z index (van der Linden & Sotaridona, 2002). There is also research investigating how to identify respondents who distort their responses in personality scales when they are used for personnel selection (e.g., Drasgow, Levine, & Zickar, 1996; LaHuis & Copeland, 2009). However, studies describing the statistical methods used in UIT confirmation testing or literature comparing the efficiency of different methods are still lacking. This is surprising, given the widespread use of UIT and the well-documented criterion validity evidence of ability tests used for personnel selection (e.g., Schmidt & Hunter, 1998).

The problem with current confirmation test approaches is that they largely undermine the purpose and attractiveness of UIT in the first place, which is to have a quick, seamless, automated, and candidate-friendly application-selection process that will reduce cost by limiting on-site or proctored testing. This is the case because most confirmation tests are unnecessarily long due to suboptimal designs. The lack of short confirmation tests means that many possible UIT benefactors discard or partially discard using UIT (Kaminski & Hemingway, 2009). Therefore, research is needed to investigate whether efficient follow-up testing is possible and to find effective methods of detecting suspicious test responses.

This study is designed to fill the gap in the literature by identifying a psychometric method for developing a confirmation test designed to verify the veracity of unproctored ability test results as efficiently (in terms of the number of items administered) as possible. Thus, a confirmation test, as it is investigated in this article, is used exclusively to confirm or reject the result from the unproctored test and is not intended to be used as a supplement to the unproctored test result.

An effective framework for developing such a confirmation test is to use CAT combined with item response theory (IRT). CAT has the advantage over traditional tests that the test can be tailored to the individual respondent, ensuring a precise result as quickly as possible. The use of CAT has been found to increase the efficiency of classification decisions in classification testing (Eggen & Straetmans, 2000; Rudner, 2002) by decreasing the number of items necessary for making an accurate decision without compromising precision. In classification testing, the goal of the examination is to classify respondents into a limited number of categories based on cutoff points selected on an ability scale. Confirmation testing can be considered a special form of classification testing, where the goal of the test is to classify each respondent by either accepting or rejecting the hypothesis that they have the same ability level in the controlled proctored condition as they did in the unproctored setting. The fundamental difference between confirmation and classification testing is that a unique cutoff point is used for each respondent in a confirmation test, whereas a classification test uses a general cutoff point for the whole sample. This difference is essential and implies that the methods used in classification testing cannot be applied directly to confirmation testing and that extensions of these methods are required.

Consequently, the design of the algorithm for an adaptive confirmation test could take the form of a classification CAT where the result from each respondent's unproctored test is used as a unique classification cutoff point in the confirmation test. This presents a challenge because the unproctored test result is an ability estimate and not a fixed point. In the following sections, we explore and test the consequences of using adaptive confirmation test methods for verifying unproctored test results,

and propose a means of generalizing the methods used in classification testing to the present context, with the intention of identifying a way of increasing the efficiency of confirmation tests compared to those currently being used.

The article is organized as follows: The IRT model is presented in the next section. This is followed by an introduction of a statistical test that can be used in a fixed length confirmation test. The subsequent section covers two sequential procedures for conducting a classification test: the truncated sequential probability ratio test (TSPRT) and the stochastically curtailed TSPRT (SCTSPRT) and presents a method for generalizing these procedures to the current context. This is followed by an outline of the research questions in the study, and a description of the simulation studies designed to answer these questions. The final section provides an overview of the results including benefits and limitations of using the proposed methods for confirmation testing, as well as future research possibilities.

4.2 The IRT Model

The proposed methods are defined in the framework of IRT. The fundamental concept of IRT is that each test item is characterized by one or more parameters and each test taker is characterized by ability parameters, in this study by a single ability parameter. The probability that a given test taker answers a given item correctly is given by a function of the test taker's ability level θ and the item parameters. Conditional on those parameters, the response on one item is assumed independent of the responses to other items. The IRT model used is the 2-Parameter Logistic, or 2-PL model, where the probability of a correct response is given by

$$P_i(\theta) = \frac{1}{1+\exp(-a_i(\theta-b_i))} \tag{4.1}$$

(Birnbaum, 1968). Here, $P_i(\theta)$ is the probability of a correct response for item i, θ is the test taker's ability, and a_i and b_i are item parameters. Furthermore, a_i is called the discrimination and b_i the difficulty parameter. It is common to scale ability levels to have a mean of zero and a standard deviation of one. The item difficulty parameter represents the point on the ability scale at which the probability of answering the item correctly equals .5. Higher discrimination parameters indicate that the probability of answering an item correctly rises dramatically with small changes in ability in the region of the item difficulty.

Alternative IRT models that might be considered are the 1- and 3-Parameter Logistic models (1-PL and 3-PL models). In the first model, also known as the Rasch

model (Rasch, 1960), the *a* parameters of all items are considered equal. This model was not pursued here, because it is often too restrictive to obtain acceptable model fit. In the 3-PL model (Birnbaum, 1968), a guessing parameter is added as a third item parameter to model guessing. This model was not pursued in this article because many authors report identification problems with the model (see, for instance, Luecht, 2006 or Partchev, 2009). Luecht (2006, p. 578) remarks that attempting to estimate stable item parameters in small samples can be extremely challenging and that the estimation problems include non-convergence of the numerical estimation solutions from ill-conditioned likelihood functions and empirical underidentification of the model parameters leading to large error covariances of the parameter estimates. In addition, we focus on adaptive testing where the items are tailored to the ability level of the respondents, θ is usually close to b_i, and the probability of a correct response converges to .5. Therefore, there is little information in the data to support the estimation of a guessing parameter on one hand and, on the other hand, a guessing parameter adds little to the precision of the estimate of θ.

What differentiates CAT from a classical linear test is that an item selection function is used to optimally select the next item in the test. A selection function that is often used is Fisher's information function. For dichotomously scored items, the information function has the following form:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{4.2}$$

where $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ, and $Q_i(\theta) = 1\text{-}P_i(\theta)$. The next item is selected that has the maximum information in the item pool at $\theta = \theta^*$, where θ^* is the current estimate of the ability of the test taker. CATs can also be used to classify examinees into a limited number of categories, by evaluating a hypothesis that θ is above or below a cut-off point on the ability scale. These algorithms usually select the next item with maximal information at the cut-off point rather than at the current ability estimate.

The efficiency (in terms of the number of items administered) of CAT in classification testing can be improved by adjusting the decision method and the termination criteria, as well as the item selection method that is used. Different item selection methods are discussed later in this article. In the sequential procedures used in classification testing, the response pattern is analyzed after the administration of each item to determine whether there is sufficient information to make a classification decision without administering unnecessary extra items. However, the test length requirements for a confirmation test are more stringent than those for a classification test. Therefore, it is not certain that a sequential procedure will increase the

efficiency of the confirmation test without severely decreasing the decision accuracy. The following sections describe both fixed length and sequential procedures for developing a confirmation test used for validating UIT.

4.3 Fixed Length Confirmation Test

A hypothetical job candidate will be used to illustrate the confirmation testing procedures used in this article; we will call this candidate C. Upon applying for a job, C is typically administered an initial unproctored online test that he can take at his convenience. If his result meets the criteria for selection, he is then offered a follow-up confirmation test under supervised conditions. A fixed length confirmation test consists of a fixed number of items. After the administration of all of the items in the test, a statistical test is used to evaluate the hypothesis that C's ability level is significantly lower than his unproctored test score. If C's confirmation test score is not significantly lower than his unproctored test score, then his unproctored test result is accepted and this score is used in making a selection decision. Note that the score from the confirmation test is not used to supplement C's unproctored score; it is simply used to confirm or reject the unproctored score. The example of C will be used again later in this article. We will now explain the fixed length confirmation test in more technical detail.

The LR test can be used to identify aberrant response patterns in a fixed length test. The LR test uses the ratio of the maximum likelihood under two different hypotheses. In the current context, the null hypothesis is that one ability parameter is sufficient to describe the response pattern in the two tests. This hypothesis represents the assumption that the respondent has the same level of ability in both tests. The alternative hypothesis is that two ability parameters are necessary to describe the response pattern: one for the unproctored test and one for the confirmation test. This represents the situation where the respondent's score has changed from the unproctored to the confirmation test.

Investigating the log-likelihood of each model compares the two models under the two hypotheses. The log-likelihood function of the ability parameter y, given a response pattern x on a test of N items is given by

$$LogL(\theta; x) = \log[\prod_{i-1}^{N} P_{\mathrm{i}}(\theta)^{x_i}(1 - P_{\mathrm{i}}(\theta))^{1-x_i}], \tag{4.3}$$

where $P_i(\theta)$ is the item response function from the 2-PL model given in Equation 4.1, $x_i = 1$ for a correct response, and $x_i = 0$ for an incorrect response to item i. The

LR statistic for the test of the null hypothesis that the unproctored test is valid is given by

$$LR = \frac{L(\theta; x_1, x_2)}{L(\theta_1; x_1)L(\theta_2; x_2)}, \tag{4.4}$$

where x_1 and θ_1 are the response pattern and estimated ability from the unproctored test, and x_2 and θ_2 are the response pattern and estimated ability from the confirmation test. $L(\theta_1; x_1)$ and $L(\theta_2; x_2)$ are the likelihood of the unproctored and confirmation test, respectively, and $L(\theta; x_1, x_2)$ is the likelihood of one single θ computed using the responses of the unproctored and confirmation test concurrently. The *LR* statistic has a chi-squared distribution with degrees of freedom equal to the number of parameters in the unrestricted model minus the number of parameters in the restricted model. This difference is one in the current context.

In confirmation testing the one-sided test is used because the only matter of interest is if the estimate for the unproctored test is significantly higher than the estimate for the confirmation test, because this would suggest that there is a possibility that the respondent has cheated.

The one-sided version of the null-hypothesis of the *LR* test is

$$H_0 : \theta_1 \leq \theta_2. \tag{4.5}$$

The decision to reject (Decision = r) the null hypothesis is made if the *LR* test is significant and the unproctored test score (θ_1) is higher than the confirmation test score (θ_2). Otherwise, the null hypothesis that the result from the unproctored test is valid, is accepted (Decision = a).

4.4 Sequential Confirmation Tests

An organization that is interested in hiring C (the hypothetical job candidate introduced earlier) may want to reduce the resources necessary for test administration by reducing assessment time. In this situation, a sequential confirmation test could be used. In a sequential confirmation, test items are administered to C one at a time. After his response to each item, the decision algorithm evaluates if there is enough information available to accept or reject the null hypothesis that C's unproctored test result is valid. If a decision can be made to accept or reject C's unproctored test score at this point, then the test is terminated. If there is not enough information available to make a decision, then an additional item is administered. Additional items are

administered until a decision is made or until the maximum number of items is reached, in which case a statistical test similar to the fixed length confirmation test is used to determine the decision. Two methods for conducting a sequential confirmation test are explored in this article: the TSPRT and the SCTSPRT.

The TSPRT is a method that is commonly used for classifying respondents in computerized classification testing (Eggen, 1999; Eggen & Straetmans, 2000; Reckase, 1983). The TSPRT offers substantially shorter tests than a conventional full-length fixed-form test, while maintaining a similar level of classification accuracy (Eggen & Straetmans, 2000; Rudner, 2002). The SCTSPRT is an extension of the TSPRT that can be used to shorten testing even further, without substantially compromising error rates (Finkelman, 2003, 2008).

The TSPRT and the SCTSPRT cannot be applied directly to the context of confirmation testing, because the hypothesis tested in this context is based on the change between two testing sessions, rather than the comparison of an estimate to one or more fixed cutoff points, which is the basis for the hypotheses tested in classification testing. Therefore, confirmation testing is fundamentally different from other applications where the TSPRT and the SCTSPRT are used. The following sections outline the TSPRT and the SCTSPRT as they are used in computerized classification testing. This is followed by an extension of these tests to confirmation testing.

4.4.1 TSPRT

The sequential probability ratio test (SPRT) is a sequential procedure used to make classification decisions in computerized classification testing. In applied conditions, a maximum for the number of items to be administered is usually set (e.g., Eggen & Straetmans, 2000; Jiao, Wang, & Lau, 2004). This is a truncation of the SPRT labeled TSPRT.

When implementing the TSPRT, first a cutoff point (θ_o) on the θ scale is selected, followed by a small region on each side of the point. The combination of the two regions of size $\delta > 0$ is referred to as the indifference zone. The indifference zone serves the purpose of identifying two points on the ability scale that can subsequently be used to calculate the LR. The statistical hypotheses are formulated as

$$H_0: \theta_2 \geq \theta_o + \delta = \theta_+ \text{ against } H_A: \theta_2 \leq \theta_o - \delta = \theta_-. \tag{4.6}$$

After the administration of the k-th item ($1 \leq k < N$, where N is the maximum number of items in the test), the following statistic is used to make a decision to accept or reject H_0:

$$LR_k = L_k(\theta_+;\boldsymbol{x})/\, L_k(\theta_-;\boldsymbol{x}) = \log L_k(\theta_+;\boldsymbol{x}) - \log L_k(\theta_-;\boldsymbol{x}), \tag{4.7}$$

where $L_k(\theta_+;\boldsymbol{x})$ and $L_k(\theta_-;\boldsymbol{x})$ are the likelihoods, θ_+ is the model under the null hypothesis, and θ_- is the model under alternative hypothesis, both given the response pattern observed up to the *k*-th item. The log-likelihood ratio of the values θ_+ and θ_- after *k* observations represents the relative strength of evidence for selecting θ_+ and θ_- as the correct decision. High values of the ratio indicate that the examinee's ability is above the cut-off point, and low values indicate that their ability is below the cut-off point. Note that no estimation of θ is involved. Decision error rates are specified as

$$P\,(\text{accept } H_0 | H_0 \text{ is true}) \geq 1 - \alpha, \text{ and } P\,(\text{accept } H_A | H_A \text{ is true}) \geq 1 - \beta, \tag{4.8}$$

where α and β are small constants that can range from .01 and .4. Further, we define $A = \alpha/(1-\beta)$, and $B = (1-\alpha)/\beta$. The TSPRT uses the following procedure after each item has been administered:

$$\text{Continue sampling if: } \log A < LR_k < \log B, \text{ and } k < N; \tag{4.9}$$
$$\text{Accept } H_0 | \text{ Decision = a, if: } LR_k \geq \log B \tag{4.10}$$
$$\text{Reject } H_0 | \text{ Decision = r, if: } LR_k \leq \log A. \tag{4.11}$$

If the maximum number of items in the test is reached (that is, if $k = N$), an additional decision rule is implemented:

$$\text{Accept } H_0 | \text{ Decision = a, if: } LR_k \geq \log C \tag{4.12}$$
$$\text{Reject } H_0 | \text{ Decision = r, if: } LR_k < \log C \tag{4.13}$$

Here, C is a constant satisfying $A \leq C \leq B$; its selection is an important factor in the error rates of the TSPRT. Spray and Reckase (1996) proposed using the relation $\log C = (\log A + \log B)/2$, which would result in $\log C = 0$, whenever $A = 1/B$.

4.4.2 The SCTSPRT

An alternative to the TSPRT used to shorten testing without substantially compromising error rates is the SCTSPRT (Finkelman, 2003, 2008). Often, the TSPRT is inefficient because an extra item is presented even though the classification decision for the respondent can no longer be changed. The SCTSPRT stops testing once the decision cannot be altered on the basis of the remaining items in the test. This is called curtailment. The SCTSPRT also stops testing in cases in which the

probability of changing the classification decision is smaller than a predefined value (y).

To evaluate the probability that the decision will change, it is first necessary to define what the current decision is, based on the k responses so far. D_k^* represents the tentative decision at stage k; where $k < N$. Finkelman (2008) suggests using the following definition: $D_k^* = a$, if $LR_k \geq \log C$ and $D_k^* = r$ otherwise. The probability of changing decisions depends on the probabilities of the response pattern of future items starting at stage $k + 1$, which in turn depends on θ. Therefore, a value for θ must be selected for evaluating the probability that the decision will change between stage k and N, that is, evaluating whether $D_k^* \neq D_N$, where D_N is the decision at stage $k = N$. A conservative option is to use the conditional power approach by Jennison and Turbull (2000). The conditional power approach implies setting $\theta = \theta_-$ when $D_k^* = a$, and $\theta = \theta_+$ when $D_k^* = r$ (Finkelman, 2008).

The stochastic curtailed TSPRT is defined as the following sequential rule.

If $k < N$, set $k = N$ and D = a if

$$\{\log LR_k \geq \log B\} \text{ or } \{\log LR_k > \log C \text{ and } P_{\theta-}(D_N = a \mid LR_k) \geq y\}; \quad (4.14)$$

and set $k = N$ and D = r if

$$\{\log LR_k \leq \log A\} \text{ or } \{\log LR_k < \log C \text{ and } P_{\theta+}(D_N = r \mid LR_k) \geq y'\}; \quad (4.15)$$

$$\text{If } k = N, \text{ set D = a if and only if } \log LR_k \geq C, \quad (4.16)$$

where y and y' are thresholds along the probability scale which take values between .5 and 1. In particular, they determine how high the probability of retaining the current decision must be for the test to be stopped early. That is, the statement $P_{\theta-}(D_N = a \mid LR_k) \geq y$ in Equation 4.14 specifies that the probability of event $D_T = a$, given $\log LR_k$, must be greater than or equal to y under θ_- for the criterion to be satisfied. Similarly the statement $P_{\theta+}(D_N = r \mid LR_k) \geq y'$ in Equation 4.15 specifies that the probability of the event $D_T = r$, given $\log LR_k$, must be greater than or equal to y' under θ_+ for the criterion to be satisfied.

Setting $y = y' = 1$ results in a test that has the same error rates as the TSPRT with the possibility of reducing the number of items. The stochastic curtailed SPRT is straightforward when all items that will be administered are known, and k is close to N. Finkelman (2008) uses an approximation of the probabilities $P_{\theta-}(D_T = a \mid LR_k)$, and $P_{\theta+}(D_N = r \mid LR_k)$ in Equations 4.14 and 4.15 when the remaining items in the test are unknown. Applying Finkelman's approximation in the current context is not straightforward, so we approximate these probabilities by simulation as a computational method. A simulation of the remaining item responses in the test at

point k can be applied to calculate the probability of retaining the current decision at point $k = N$. The simulation uses the conditional power approach described above for selecting θ at point k. The remaining item responses from point k to $k = N$ can then be generated by replicating a number of CATs given θ_+ or θ_- 100 times. The probability of retaining the decision is computed by adding up the number of times the decision remains unchanged in the 100 replications.

4.4.3. Extending the SPRT Framework to Computerized Confirmation Testing

The SPRT framework cannot be applied directly to confirmation testing because the cutoff point is not a fixed point but varies over respondents based on their unproctored test result. Therefore, an extension of the SPRT framework is needed where the cutoff point is defined at the level of the individual based on the unproctored test result. Defining the cutoff point based on the unproctored test result presents a further challenge because the unproctored test result is an estimate, which contains error, and is not a fixed variable. A solution is to develop a test taking into account the sampling distribution of the estimate. Using a cutoff point from the sampling distribution, the probability of exceedance or significance probability can be determined.

When using the SPRT framework for a confirmation test, the selection of the cutoff point determines the ratio of Type I and Type II error rates in the test. Decreasing the Type I error rate automatically means an increase in Type II error; therefore, a prioritization of the two is important. Incorrectly rejecting a respondent who has completed the unproctored test without cheating (a Type I error) can have serious ethical and possible legal consequences. Burke et al. (2006) provide a good overview of considerations that are important in handling such respondents, which can be demanding in terms of required resources for those administering the test. Alternatively, the cost of committing a Type II error includes hiring the incorrect applicant, which can result in negative organizational consequences and negative consequences for the applicants who were not selected. In the current context, raising questions about someone's integrity usually outweighs the negative consequences of a Type II error. Therefore, the development of a confirmation test method that controls for the Type I error rate is prioritized in this article.

The cutoff point for the confirmation test can be established using the lower limit of the confidence interval for the ability estimate in the unproctored test. This confidence level would hold for the Type I error rate in the confirmation test, if the confirmation test had perfect precision. In practice, a more conservative value is needed because the error is compounded due to a combination of the error in the two

tests. One method for establishing the Type I error rate for the confirmation testing procedure is to take the error of both tests into account with the following equation:

$$\theta_0 = \theta_1 - (SE(\theta_1) + g), \tag{4.17}$$

where θ_1 and $SE(\theta_1)$ are the ability estimate and standard error obtained from each respondent's unproctored test, respectively, and g is a positive constant representing a combination of the anticipated error in the confirmation test and the desired Type I error rate of the test.

The SPRT requires a predefined cutoff point before the confirmation test begins; therefore, the exact value of g that corresponds to a specific Type I error rate cannot be obtained because it will depend on the items that are administered in the confirmation test. However, a value of g can be chosen in such a way that the Type I error rate aggregated over respondents is close to a predefined overall nominal Type I error rate. The choice of g only takes the characteristics of the confirmation test into consideration; the error in the unproctored test is already included in equation 4.17. Therefore, one value of g would result in a stable Type I error rate, regardless of the precision of the unproctored test. The value of g that corresponds to a Type I error rate can be obtained by conducting an a priori simulation of the confirmation test.

In summary, the purpose of the proposed method is to set a cutoff point at some position below the θ estimate in the unproctored test, to take into account the error in the unproctored test and anticipated error in the confirmation test with the intention of controlling the Type I error rate in the confirmation test.

4.5 Research Questions

There are four main research questions in this study: Which item selection method performs best in fixed length and sequential confirmation testing procedures? What are the benefits of using a sequential confirmation test, rather than using a fixed length confirmation test based on efficiency in terms of the number of items administered and the power to identify cheaters? What are the benefits of using the stochastic curtailed version of the TSPRT compared to using the TSPRT? Does a confirmation test decrease the effect of cheating on the validity of a selection procedure?

4.6 Simulation Studies

The research questions outlined above were investigated by means of simulation studies. The first simulation study was designed to compare different methods for conducting a fixed length confirmation test. The second study was used to define the cutoff points that correspond to nominal Type I error rates for a sequential confirmation test and to test whether the empirical Type I error rates matched the nominal Type I error rates. The third study explored different options for conducting a TSPRT for sequential adaptive computerized confirmation testing. This study also compared the results for the fixed length adaptive confirmation test with those of the TSPRT to investigate if the TSPRT improved the efficiency of the fixed length adaptive confirmation test. The fourth study investigated the possibility of shortening the confirmation test even further using the SCTSPRT procedure. Finally, a selection procedure was simulated to test the utility of the computerized adaptive confirmation test for a personnel selection procedure.

The simulation studies were programmed in Digital Visual FORTAN 6.0 for Windows. The ability estimation methods used in the fixed length confirmation test can also be calculated by using standard IRT software packages such as Parscale (Muraki & Bock, 1996), or ConQuest (Wu, Adams, Wilson, & Haldane, 2007). It was assumed that during the unproctored test a respondent who had cheated had another θ value than during the confirmation test. This was simulated by drawing θ from a standard normal distribution for the confirmation test. The θ value in the unproctored test was set to $\theta = \theta + \Delta$, where Δ represented the effect of cheating. Three groups of respondents were simulated. The first was non-cheaters: here the effect of cheating was zero ($\Delta = 0$). The next group represented a moderate cheating effect of one standard deviation unit ($\Delta = 1$). This means that all of the respondents in this group had a true ability level that was one standard deviation unit higher in the unproctored test compared to the confirmation test. The final group represented a large cheating effect of two standard deviation units ($\Delta = 2$). In this group all of the respondents had a true ability level that was two standard deviation units higher in the unproctored test compared to the confirmation test. The values of Δ were selected based on previous literature. Burke, Someren, and Tatham, (2006) report values of $\Delta = 2$ in testing the efficiency of a confirmation test used for high stakes organizational ability testing. Van Krimpen-Stoop and Meijer (2000) use values of $\Delta = 1$, and $\Delta = 2$ in investigating non-fitting response behavior in the second half of a CAT. These values are large; however, in this context they are appropriate because the consequences of rejecting an unproctored test score are serious, and would only be considered when the effect of cheating is large. In addition, the type of cheating expected in an

unproctored test, such as the substitution of test takers, would likely lead to greater effect size differences than the cheating that is widespread in proctored settings.

An item bank was simulated by drawing item difficulty parameters from a standard normal distribution and item discrimination parameters from a lognormal distribution with an expectation of 1. The item responses were generated according to the 2-PL model in Equation 1. This was repeated for each respondent, both for the unproctored test and for the confirmation test. The unproctored test had a fixed length where the first three items were selected with maximum information for $\theta = 0$, and the remaining items were administered by selecting the item with maximal information at the current θ estimate. The estimation of θ was done with either maximum likelihood estimation (MLE) or weighted likelihood estimation (WLE; Warm, 1989). Each condition was repeated for 10,000 respondents.

Items in the confirmation test were administered according to one of three item selection procedures: random (RAN), maximum information at the cut-off point (MIC), and maximum information at the current θ estimate (MIE). The first procedure randomly selects the next item from the available item bank, excluding items used before. This selection method is similar to a non-adaptive confirmation test (e.g., Burke, van Someren, & Tatham, 2006). The other two procedures select the next item maximizing the item information in Equation 2. In the MIC item selection algorithm, the next item is selected to maximize the information for the cut-off point. In MIE, the next item selected is the item for which the information at the current ability estimate is maximal. The MIE item selection algorithm was initiated after three items had been administered in the confirmation test in order to ensure that a fairly stable ability estimate had been established. The item with maximum information at the cut-off point was administered up to this point. MIC has been found to obtain comparable power with fewer items than MIE in computerized classification testing (Spray & Reckase, 1994; Eggen, 1999), however MIE item selection is favored when the purpose of the test is to estimate the ability of an examinee (Thissen & Mislevy, 2000).

4.6.1 Study 1: Fixed length confirmation test using the LR test

The first simulation study was designed to answer two research questions: Which of the three item selection methods described above is the most efficient? Can ability estimation using WLE increase classification accuracy over the more commonly used MLE? The second question was motivated by research that has found that ability estimation accuracy can be increased by using alternative estimation algorithms compared to MLE (e.g., Warm, 1989). The questions were investigated for adaptive unproctored tests consisting of 10, 30 and 50 items, and fixed length confirmation

tests consisting of 5, 10 and 20 items for item bank sizes of 100, 200 and 400 items. The outcome variable was the proportion of correct classifications.

The results of the first simulation study did not vary by size of the item bank so only the results for an item bank consisting of 200 items are presented in Table 4.1. The table reports the percentage of respondents in each condition who had their unproctored test score rejected.

The number of accurate classifications was higher when using WLE compared to MLE for cheaters ($\Delta = 1$, and $\Delta = 2$), while both methods maintained similar Type I error rates for confirmation test lengths of 10 and 20 items.

The item selection algorithm also had an effect on the efficiency of the confirmation test. Although, there was not a great difference between MIC and MIE item selection procedures, both performed better than random item selection (RAN) across most conditions. In fact, MIC and MIE item selection methods performed similar to random item selection with half the number of items in the confirmation test (e.g., 10 instead of 20), when WLE was the estimation method. An unproctored test length of 10 items resulted in high Type I error rates in the confirmation test due to a lack of precision in the unproctored test.

The results of the first simulation study suggest that the most effective method for constructing a fixed length confirmation test is to use WLE for ability estimation and MIC or MIE for item selection. With these methods a confirmation test length of 10 items was necessary before the confirmation test obtained acceptable power. Additionally, the confirmation test performed best when the unproctored test was at least 30 items.

Table 4.1

Percentage of aberrant response classifications for a fixed length confirmation test using the LR test with a significance level of 5%

		MLE								
		Length of Unproctored Test								
Test		10 Items			30 Items			50 Items		
Length	Δ	RAN	MIC	MIE	RAN	MIC	MIE	RAN	MIC	MIE
5	0	2.28	0.88	1.09	1.60	0.81	0.50	1.37	0.32	0.70
	1	9.88	1.04	4.57	11.07	8.65	7.47	11.09	4 .14	8.23
	2	19.21	1.32	9.36	33.46	6.08	13.19	35.83	23.71	20.20
10	0	6.67	10.45	9.71	3.28	5.02	5.45	2.93	4.23	4.09
	1	22.55	32.03	29.19	26.96	38.84	37.97	27.07	37.36	40.73
	2	35.72	32.83	36.52	62.05	58.00	60.07	65.22	65.58	67.13
20	0	13.06	16.49	15.72	5.48	7.86	7.75	4.30	5.73	5.42
	1	39.98	56.60	54.02	45.80	64.03	64.76	47.45	65.05	67.30
	2	59.88	65.07	65.55	81.57	88.09	87.63	88.57	92.34	93.44

		WLE								
		Length of Unproctored Test								
Test		10 Items			30 Items			50 Items		
Length	Δ	RAN	MIC	MIE	RAN	MIC	MIE	RAN	MIC	MIE
5	0	3.99	6.43	6.63	3.18	4.80	3.84	3.04	4.12	4.06
	1	18.46	27.70	25.04	16.91	24.55	28.14	17.01	28.03	27.05
	2	43.42	57.70	61.20	45.11	66.59	64.17	45.79	57.48	66.33
10	0	6.30	10.13	8.94	4.15	4.79	5.57	3.97	3.84	3.92
	1	29.89	44.97	43.73	27.99	42.39	45.21	28.38	43.07	45.43
	2	66.04	81.37	81.28	68.30	87.25	89.02	70.82	88.65	89.84
20	0	9.73	13.72	15.13	5.17	6.77	6.81	4.25	4.95	5.89
	1	44.84	59.23	60.50	47.33	64.79	64.69	45.82	64.89	65.26
	2	82.32	92.30	94.14	87.47	96.52	97.79	89.51	98.02	98.90

Note: MLE: maximum likelihood estimation; WLE: weighted likelihood estimation; Δ: cheating effect in standard deviation units; RAN: random item selection; MIC: optimal item selection at the cut-off point; MIE: optimal item selection at the current ability estimate.

4.6.2 Study 2: Defining the cut-off points that correspond to nominal Type I error rates for a sequential confirmation test

The second simulation study was used to define the cut-off points that correspond to nominal Type I error rates for a sequential confirmation test; and to test whether the empirical Type I error rates matched the nominal Type I error rates. The decision algorithm for the sequential confirmation test procedures requires a cut-off point from Equation 17 before testing begins. The algorithm uses the cut-off point to test the probability that the respondent's ability based on the confirmation test response pattern is above or below that point. Equation 17 includes θ and the standard error obtained from the unproctored test, as well as *g,* which is a constant depending on a combination of the anticipated error in the confirmation test and the desired Type I error rate for the combined testing procedure. The value of *g* is obtained by simulating the confirmation test a number of times for a representative sample of respondents. Therefore, a simulation of the values of *g* that correspond to particular Type I error rates based on the characteristics of the confirmation test, was necessary before a sequential confirmation test could be performed.

The choice of *g* only takes the characteristics of the confirmation test into consideration; therefore, one value of *g* should result in a stable Type I error rate regardless of the precision of the unproctored test. Precision in the unproctored test was varied in the current example by varying the length of the unproctored test. Table 4.2 compares the empirical and nominal Type I error rates across several conditions, to test the empirical consequences of using *g* in Equation 17 to set a cut-off point for a sequential confirmation test. The optimal values of *g* that correspond to Type I error rates of .10, .05, and .01 are reported. In order to estimate the Type I error rates under optimal conditions, a conservative critical value of $\alpha = \beta = .01$, and an indifference zone of .1 were set for this simulation so all respondents were administered all ($N = 10$ or $N = 20$) items in the confirmation test.

The table provides optimal values of *g*, for setting the Type I error rate in the sequential confirmation test, these are used in the remainder of the article. The difference between empirical and nominal Type I error rates in Table 4.2 appear to be due to random error and do not provide evidence for systematic differences. Therefore, this method is acceptable for selecting a single value for *g* that maintains a consistent Type I error rate independent of the precision of the unproctored test.

Table 4.2

Theoretical and empirical type I error rates for an adaptive confirmation test using the TSPRT for different values of the constant g

Nominal Type I Error	Item Selection Method	Confirmation Test Length = 10				Confirmation Test Length = 20			
			Unproctored Test Length				Unproctored Test Length		
		g	10	30	50	*g*	10	30	50
.10	RAN	0.80	0.088	0.102	0.102	0.55	0.950	1.010	0.980
	MIC	0.50	0.091	0.097	0.110	0.35	0.097	0.089	0.102
	MIE	0.50	0.093	0.098	0.104	0.35	0.099	0.100	0.103
.05	RAN	1.20	0.053	0.056	0.051	0.85	0.520	0.490	0.520
	MIC	0.75	0.052	0.050	0.045	0.55	0.054	0.049	0.048
	MIE	0.75	0.055	0.055	0.051	0.55	0.055	0.049	0.048
.01	RAN	3.00	0.023	0.022	0.017	1.50	0.008	0.007	0.009
	MIC	1.30	0.010	0.004	0.007	0.95	0.016	0.075	0.072
	MIE	1.30	0.013	0.008	0.010	0.95	0.015	0.082	0.080

Note: *g:* constant representing a combination of the anticipated error in the confirmation test and the desired Type I error rate for the combined testing procedure; RAN: random item selection; MIC: optimal item selection at the cut-off point; MIE: optimal item selection at the current ability estimate.

4.6.3 Study 3: Sequential confirmation test using the TSPRT

There were two main objectives for conducting the third simulation study. The first was to establish an effective method for using a TSPRT for sequential confirmation testing. The second was to evaluate if a sequential procedure could reduce the number of items needed in the confirmation test while maintaining the same level of classification accuracy. A maximum number of items administered in the test was set at $N = 10$. In addition to the number of correct classifications, the average number of items needed to make a classification was used as a criterion to measure the increased efficiency of the procedure compared to the fixed length test. Varying the following variables assessed the most effective method of conducting a confirmation test using the TSPRT:

a) Item selection procedure: RAN, MIC, MIE.
b) Unproctored test length: 10, 30, 50.
c) Cheating effect in the unsupervised test: $\Delta = 0, 1, 2$.
d) Critical value for the SPRT test: $\alpha = \beta = .01, .05, .1, .2, .4$.
e) Indifference zone: $\delta = .1$ to $.3$ in increments of $.05$.

The results of the third simulation study are presented in Table 4.3 for an item bank size of 200 items; similar findings were obtained with different size item banks. To save space only the results for an indifference zone of $\delta = .25$ are presented in Table

4.3. $\delta = .25$ is larger than the indifference zone typically used in classification testing; however, it produced the best results in the current context. This is a consequence of a short maximum test length, because small values of δ resulted in nearly all tests progressing until the maximum number of items had been reached. A slight decrease in the percentage of correct decisions can occur when increasing δ, however the decrease in precision was not evident until $\delta = .3$ in this study. Therefore, a value of .25 will be used in the remainder of this article. Similarly, the results for critical values of $\alpha = \beta = .01$ and .05 were too conservative for the current context because the test was never terminated before the maximum number of items had been reached. Therefore, less conservative critical values of $\alpha = \beta = .1, .2, .4$ are reported in Table 4.3. The table reports the percentage of respondents who had their unproctored test result rejected (C), and the average number of items used in the confirmation test (K) for each condition.

The results of the TSPRT indicate that the MIC and MIE item selection procedures performed similarly in terms of classification accuracy. The MIC procedure used an average of .1 fewer items to obtain a similar level of accuracy. Both MIC and MIE outperformed random item selection in terms of power and efficiency across all conditions. These item selection methods increased the percentage of cheaters identified by approximately 20% over random item selection, while maintaining a similar Type I error rate. Changing the critical value from $\alpha = \beta = .1$ to .2 led to a similar number of correct classifications with fewer items. The use of the critical value $\alpha = \beta = .4$ led to inflated Type I error rates and decreased the power in the test. Therefore, a critical value of $\alpha = \beta = .2$ obtained the best results, and will be used in the remaining simulations.

Table 4.3

Percentage of aberrant response classifications and average number of items for a sequential adaptive confirmation test using the TSPRT

		Unproctored Test Length = 10						Unproctored Test Length = 30						Unproctored Test Length = 50					
α		RAN		MIC		MIE		RAN		MIC		MIE		RAN		MIC		MIE	
B	Δ	C	K	C	K	C	K	C	K	C	K	C	K	C	K	C	K	C	K
.1	0	5	9.7	5	9.2	5	9.3	5	9.8	5	9.5	5	9.7	6	9.9	4	9.4	5	9.6
	1	22	10.0	34	9.8	34	9.8	27	10.0	44	9.9	43	10.0	28	10.0	47	9.9	48	10.0
	2	53	10.0	76	9.6	74	9.8	67	10.0	90	9.6	89	9.8	72	10.0	93	9.6	93	10.0
.2	0	5	8.5	5	7.1	5	6.9	5	8.7	4	7.5	5	7.5	5	8.8	5	7.9	6	7.9
	1	22	9.4	34	8.6	34	8.6	26	9.7	44	9.1	45	9.3	29	9.8	48	9.2	48	9.3
	2	57	9.7	75	8.4	76	8.5	69	9.7	91	7.8	90	7.9	72	9.7	93	7.8	93	7.7
.4	0	6	3.9	9	2.5	9	2.4	8	4.2	11	2.6	10	2.8	7	4.2	10	2.5	12	3.1
	1	24	4.8	38	3.1	38	3.1	29	5.0	46	3.1	47	3.3	32	5.2	49	3.1	49	3.3
	2	55	5.0	72	2.9	73	2.8	67	5.1	84	3.0	83	3.1	70	5.0	87	3.0	87	2.9

Note: Type I error = 5%; α = β: values that determine the required magnitude that the LR test must surpass in order to stop the confirmation test early; Δ: cheating effect in standard deviation units; C: percent of aberrant response classifications; K: average test length; RAN: random item selection; MIC: optimal item selection at the cut-off point; MIE: optimal item selection at the current ability estimate.

The results from the sequential and fixed length tests can be compared for a confirmation test consisting of 10 items and an unproctored test of 30 items because the Type I error rate was approximately 5% for both methods. The sequential adaptive confirmation test using the TSPRT resulted in a greater number of correct classifications for the MIC and MIE item selection methods compared to the fixed length tests. The sequential method was also more efficient and saved an average of 1.9 items with the MIC item selection method, 1.8 with the MIE method, and .6 with random item selection method when averaging across the three cheating effects.

Using a single value for g in Equation 17 has the disadvantage that the Type I error rate applies to the entire distribution and is not defined conditionally at each point on the ability scale. Therefore, an analysis of classification accuracy across ability levels was done to assess whether the probability of being classified as a possible cheater depended on the respondent's ability. The conditional precision was simulated in exactly the same way as in the previous simulation, however, 10,000 response patterns were simulated for each of the following θ values: -2, -1, 0, 1, 2. Table 4.4 presents the conditional precision of the confirmation test using the TSPRT.

Table 4.4

Conditional precision of the sequential adaptive confirmation test using the TSPRT with an unproctored test of 30 Items

	$\theta = -2$		$\theta = -1$		$\theta = 0$		$\theta = 1$		$\theta = 2$	
Δ	C	K	C	K	C	K	C	K	C	K
	RAN									
0	19	9.8	8	9.3	3	8.5	2	8.1	3	8.9
1	36	9.9	30	9.8	25	9.6	23	9.7	23	9.9
2	72	10.0	73	9.7	72	9.6	64	9.8	52	10.0
	MIC									
0	6	7.8	5	7.7	5	7.4	4	7.6	5	8.0
1	46	9.3	46	9.1	46	9.1	41	9.1	36	9.1
2	93	7.9	92	8.0	91	8.1	87	8.5	78	8.9
	MIE									
0	7	8.7	4	7.5	5	7.8	5	7.7	5	7.9
1	45	9.6	47	9.4	45	9.2	42	9.3	40	9.2
2	92	7.8	92	8.2	91	7.9	88	8.4	80	8.9

Note: Type I error = 5%; Δ: cheating effect in standard deviation units; C: percent of aberrant response classifications; K: average test length; RAN: random item selection; MIC: optimal item selection at the cut-off point; MIE: optimal item selection at the current ability estimate.

For random item selection the empirical Type I error rate (displayed in the columns labeled C) ranged from 2% to 19% across ability levels. The Type I error rate ranged from 4% to 6%, and from 4% to 7% for MIC and MIE item selection methods respectively. The deviation from the nominal Type I error rate of 5% was greatest for the lowest ability level of $\theta = -2$. The power of the MIC and MIE methods at detecting cheaters was also relatively consistent for ability levels $\theta = -2, -1, 0$, and only became lower at ability levels $\theta = 1$, and 2. Therefore, it is justifiable to use a single value of g in Equation 17 with MIC and MIE item selection. For random item selection, the percentage of aberrant response classifications was far less stable across values of θ. For instance, for $\theta = -2$, the percentage was 19, while it was 3 for $\theta = 2$. So, this method leads to local bias and should not be used with random item selection.

4.6.4 Study 4: Sequential confirmation test using the stochastically curtailed TSPRT

The main purpose of the fourth simulation study was to check if the stochastically curtailed TSPRT could be used to shorten the TSPRT in computerized adaptive confirmation testing. A stochastically curtailed TSPRT was developed based on the most effective method identified in Study 3. The stochastically curtailed TSPRT halts testing in cases in which the probability of a change of the classification decision is smaller than the predefined value y from Equations 14 and 15. Therefore, y was varied from .75 to 1 in intervals of .05.

The results of the fourth simulation study are presented in Table 4.5. To save space results for only a selection of the values of y are presented. As expected Table 4.5 indicates that the accuracy of the confirmation test decreased as the value of y decreased. The Type I error rate was maintained as long as $y \geq .95$, however, the empirical Type I error rate was above the nominal Type I error rate, for $y \leq .9$. Therefore, values of $y \geq .95$ would be suggested for designing a confirmation test with a pre-defined Type I error rate. The detection accuracy was only .02% lower for $y = .95$ compared to $y = 1$ when averaging across MIC and MIE for all conditions of Δ. However, there was a decrease in the number of items necessary for making a decision from 5.96 to 5.4 items for MIC, and from 6.07 to 5.5 items for MIE. These results suggest that the SCTSPRT with $y = .95$ is the most efficient (in terms of the number of items used), yet accurate (in terms of correct classifications), method for conducting a computerized confirmation test. The accuracy of the MIC and MIE item selection methods was virtually identical; however the MIC used an average of .1 fewer items when averaging across all conditions for $y \geq .95$. The Type I error rate was consistent across unproctored test lengths; however the power of the test increased as the length of the unproctored test increased.

A comparison of the two sequential confirmation test methods, the TSPRT and the SCTSPRT showed that both tests retained similar precision; however the SCTSPRT led to classifications with fewer items than the TSPRT. The SCTSPRT led to an average decrease of 2.2 and 2.1 items for the MIC and MIE item selection methods, respectively, when y was set to 1; and of 2.8 and 2.7 when y was set to .95. This represents a reduction of over 25% of the items in the confirmation test.

Table 4.5

Percentage of aberrant response classifications and average number of items for a sequential adaptive confirmation test using the SCTSPRT

		Length of the Unproctored Test											
		10 Items				30 Items				50 Items			
		MIC		MIE		MIC		MIE		MIC		MIE	
y	Δ	C	K	C	K	C	K	C	K	C	K	C	K
1.00	0	4	4.0	5	5.0	5	5.9	4	5.3	4	5.3	5	5.6
	1	32	6.3	34	6.6	45	7.1	48	7.3	45	7.0	46	7.5
	2	77	6.3	75	6.2	91	5.6	91	5.7	93	6.1	93	5.4
.95	0	5	4.5	5	4.2	5	4.2	5	5.0	5	5.1	5	5.2
	1	34	6.1	34	6.2	43	6.4	41	6.2	48	6.3	48	6.7
	2	78	5.6	77	5.9	91	4.9	90	5.3	93	5.5	93	4.8
.90	0	6	4.1	7	4.3	6	4.4	6	4.7	6	4.8	6	4.6
	1	34	5.4	34	5.7	44	6.1	43	6.3	50	6.0	47	6.2
	2	73	5.2	75	5.1	88	4.3	88	4.7	92	5.2	92	5.2
.80	0	7	3.9	7	3.9	8	3.9	8	4.1	7	4.2	7	4.3
	1	36	4.9	36	5.0	43	4.9	45	4.8	47	5.4	48	5.1
	2	75	4.6	74	4.7	87	4.3	87	4.3	90	4.5	91	4.4

Note: Type I error = 5%; y: determines how high the probability of retaining the current decision must be for the test to be stopped early; Δ: cheating effect in standard deviation units; C: percent of aberrant response classifications; K: average test length; MIC: optimal item selection at the cut-off point; MIE: optimal item selection at the current ability estimate.

4.6.5 Study 5: A simulated personnel selection procedure

So far we have discussed the accuracy of a confirmation test in terms of classification accuracy. However, we were also interested in the effect of cheating and the returns obtained from using the sequential adaptive confirmation test for a selection procedure where those candidates who obtain the highest score on a test are hired for a job. A sample of 10,000 job candidates was generated with a normal distribution with a mean of 0 and a standard deviation of 1 representing their true ability. Cheating was defined in the same way as in the previous simulations with the cheating effect (Δ) = 0, 1, and 2 standard deviation units in the confirmation test. The proportion of cheaters in the sample was varied as 0%, 10%, and 20%. Candidates were assigned randomly to a cheating or honest condition. Based on research findings presented in Cizek (1999) assignment to the cheating condition was also based on an assumed correlation between ability and the propensity to cheat of -0.3. The selection ratio for the example presented below was 50%, although other ratios were also explored.

The benefits ratio of the test is measured as the ratio of correct selections to incorrect selections. Table 4.6 presents the benefits ratio based on the cross-classification of true and reported scores for the unproctored test, without a confirmation test. As a more traditional validity measure, Table 4.6 also presents Pearson's correlation coefficient between the true ability and the unproctored test score for all 10,000 job candidates.

Table 4.6

Selection accuracy for an unproctored adaptive test of 30 items: selection of the top 5,000 candidates based on the unproctored test score in a sample of 10,000 applicants

Cheating %	Δ	Selections		Benefit Ratio	Correlation Between True Ability and Unproctored Test Score
		False	Correct		
0%	0	493	4507	9.14	0.95
10%	1	735	4265	5.80	0.90
10%	2	738	4262	5.78	0.80
20%	1	974	4026	4.13	0.87
20%	2	1053	3947	3.75	0.72

Note: Δ: cheating effect in standard deviation units.

The benefits ratio of the unproctored test alone, when there was no cheating was 4507:493; or 9.14 correct selections for each incorrect selection. The effect of cheating decreased the benefits ratio of the test to 5.8 with 10% cheaters, and 4.13 with 20% cheaters, when the cheating effect was moderate. The decrease was even larger when the effect of cheating was large; 5.78 with 10% cheaters, and 3.75 with 20% cheaters. The correlation between the true ability and the unproctored test score was 0.95, when there was no cheating in the sample of job candidates. The correlation was reduced to 0.90 and 0.80 respectively for moderate and large cheating effects when 10% of the candidates in the sample were cheaters. The correlation became 0.87 and 0.72 respectively for moderate and large cheating effects when 20% of the candidates in the sample were cheaters.

Table 4.7 presents a continuation of the selection procedure presented above where the candidates who were selected based on their unproctored test result were administered an adaptive confirmation test using the SCTSPRT as a follow-up. The table illustrates a cross-classification of the candidates' true ability and the decision based on the confirmation test. A benefits ratio of the combination of the unproctored test and the confirmation test is presented, based on the candidates whose unproctored test result was accepted. The table also presents Pearson's correlation

coefficient between the true ability and the unproctored test score, after eliminating the candidates who had their unproctored test score rejected in the confirmation test. The last two columns of Table 4.7 present a cross-classification of candidates' true condition (honest/cheater) and their confirmation test result.

Table 4.7

Selection accuracy for an unproctored test of 30 items with a follow-up SCTSPRT confirmation test

Cheating %	Δ	Decision from Confirmation Test	Selections		Benefit Ratio	Correlation Between True Ability and UIT Score*	Candidates True Condition	
			False	Correct			Cheater	Honest
0%	0	a	432	4274	9.90	0.95		4498
		r	61	233				286
10%	1	a	458	3773	8.24	0.93	476	3755
		r	277	492			478	291
10%	2	a	245	3608	14.73	0.93	75	3778
		r	493	654			921	226
20%	1	a	535	3360	6.28	0.91	980	2915
		r	439	666			866	239
20%	2	a	147	2896	19.44	0.93	150	2893
		r	906	1051			1801	156

Note: * Correlation between true ability and unproctored test score after eliminating the candidates who had their unproctored test score rejected in the confirmation test; Δ: cheating effect in standard deviation units; a: accept UIT score; r: reject UIT score.

The addition of the confirmation test in the selection procedure only slightly increased the benefits ratio of the selection procedure from 9.14 to 9.90 (4274:432) when there was no cheating. The increase occurred because a larger percentage of falsely selected candidates (14%) had their unproctored test result rejected compared to the group of respondents that was correctly selected (5%). This occurred because the unproctored test score was inflated based on random error for the falsely selected candidates. This finding illustrates that the use of a confirmation test also helps detect candidates who are falsely selected even when the source of the inflated unproctored test result is not due to cheating. The benefit of using the confirmation test increased as the percentage of cheaters in the applicant pool increased. When the effect of cheating was moderate the increase in benefits ratio went from 5.80 to 8.24 with 10%

cheaters, and from 4.13 to 6.28 with 20% cheaters. The benefits ratio increased from 5.78 to 14.73 with 10% cheaters and from 3.75 to 19.44 with 20% cheaters when the effect of cheating was high. This occurred because the proportion of respondents who had cheated became larger in the sample of selected candidates. Therefore, the benefit of using the confirmation test was larger because the percentage of cheaters who were identified by the confirmation test increased. The benefit of using a confirmation test was also larger when the ratio of the candidates who were selected became smaller because the proportion of cheaters in the group of selected candidates increased.

A comparison of the correlation between the true ability and the unproctored test score in Table 4.6 and Table 4.7 gives an indication of the impact of the confirmation test in terms of a correlation coefficient. The correlation remained unchanged at 0.95 when there was no cheating. The inclusion of a confirmation test increased the correlation coefficient from 0.90 to 0.93 and from 0.80 to 0.93, for moderate and large cheating effects respectively, when 10% of the candidates had cheated. The correlation between the true ability and the unproctored test score increased from 0.87 to 0.91 and from 0.72 to 0.93, for moderate and large cheating effects respectively, when 20% of the candidates had cheated on the unproctored test.

4.7 Discussion

With the increased need for flexible selection procedures, we foresee an increased use of UIT in the future. This study showed that cheating can have a detrimental effect on the validity of a selection procedure, and the possibility of using a confirmation test can decrease this effect. In addition, there is evidence that a significant number of test candidates would cheat in an unproctored high stakes test; therefore, confirmation testing is important for validating UIT results. The article investigated four research questions related to the problem of verifying unproctored Internet test results by means of a short confirmation test. The research questions are repeated below with conclusions and practical considerations based on the results of the study.

Which item selection method performs best in fixed length and sequential confirmation testing procedures? The results illustrated that adaptive selection methods were more powerful than random item selection in terms of accurately classifying respondents as honest or cheaters; and were more efficient in terms of the number of items required to make a decision in the sequential procedures. A comparison of the two adaptive methods lead to the conclusion that maximum information at the cut-off point (MIC) was slightly more efficient than maximum information at the current θ estimate (MIE) for the sequential procedures. This is

consistent with previous findings from adaptive classification test literature (e.g., Spray & Reckase, 1996; Eggen, 1999). However, this difference was small and may not have practical significance in applied settings.

What are the benefits of using a sequential confirmation test, rather than using a fixed length confirmation test based on; efficiency in terms of the number of items administered, and the power to identify cheaters? Regarding this research question, we first discuss the benefits in terms of efficiency. The primary advantage of using the sequential methods was that they required fewer items to obtain a similar level of power. The SCTSPRT method required one half of the items compared to the fixed length procedure using the WLE for ability estimation; and approximately one fourth compared to using the MLE for ability estimation. One of the criticisms of UIT has been that it lacks a quick method for verifying unproctored test results. This improvement will make it more attractive to take advantage of the benefits of UIT because it provides a quick testing process that can reduce cost by limiting on-site or proctored testing. An additional advantage was that the sequential methods provide the option of controlling the Type I error rate in the test. This allows the test users to set the error rate according to their particular needs. Although the sequential procedures were more efficient than the fixed length confirmation test methods, there may be applications where a fixed number of items is desired. When a fixed length confirmation test was used the results showed that ability estimation using WLE increased classification accuracy compared to the more commonly used MLE.

The power of the sequential confirmation test methods introduced in this article is directly related to the characteristics of the test, and the degree of cheating. The adaptive SCTSPRT method resulted in over 90% accurate detection of cheaters when the unproctored test length was at least 30 items, and the respondent's cheating effect was large (two standard deviation units). The power of the test was lower for short unproctored test lengths, and for moderate cheating effects (one standard deviation unit). It is expected that an unproctored test length of 30 items is acceptable for most applied settings because the flexibility of UIT allows the respondent to take the test at their convenience. The practical consequences of the lower power for moderate cheating effects will vary based on the intended use of the test for the particular testing organization. The item bank in the confirmation test can be expanded if the power of the test is not sufficient in certain situations. Other alternatives, such as a long supervised test which counts as the result of record could also be considered. The SCTSPRT method proposed in this study provides the option of increasing power by increasing the Type I error rate in the test; however, this results in an increase in the number of false positives, which can lead to a number of negative ethical and possible legal consequences.

What are the benefits of using the stochastic curtailed version of the TSPRT compared to using the TSPRT? The stochastically curtailed version of the TSPRT led to a reduction in test length of over 25% or over 2 items compared to the TSPRT. Therefore, the effect of using the stochastically curtailed version of the TSPRT for confirmation testing is even larger than what has been reported for computerized classification testing (e.g., Finkelman, 2008).

Does a confirmation test decrease the effect of cheating on the validity of a selection procedure? The simulation study demonstrated that cheating could have a detrimental effect on the validity of a selection procedure, and illustrated that the use of a confirmation test could hinder the negative effect of cheating on validity. The advantage of using a confirmation test increased when the effect of cheating increased, and as the number of cheaters in the applicant pool increased.

4.7.1 Limitations

The confirmation test methods proposed in this article are limited to the context of knowledge and ability tests, and are only applicable to situations in which the unproctored test score is used as the operational score. A further limitation is that the SCTSPRT is a CAT/IRT-based confirmation test approach that demands a large item bank, which requires considerable resources to develop. Many developments in CAT research have been focused on decreasing the resources required for CAT item bank development. Advances such as automatic calibration procedures (e.g., Kingsbury, 2009; Makransky & Glas, 2010), and automated item generation (e.g., Glas & van der Linden, 2003; Glas, van der Linden, & Geerlings, 2010) provide options for decreasing the resources required for developing an adaptive test, however, the cost is still higher than the price of developing a traditional test. A related issue is that an adaptive approach requires computer access at the supervised testing location, and relies on fast and reliable Internet access in order to allow communication between the local computer and the decision algorithm. These technological limitations are no longer as prominent in most testing locations, but they may still be prominent in others.

An additional limitation to using the SCTSPRT for confirmation testing is that this method is designed to test the hypothesis that the result in the first test is valid. Therefore, it does not add to the reliability of the ability estimate. Another disadvantage to using the SCTSPRT is that it is an adaptive procedure, which means that respondents are administered a different number of items depending on their responses. This can result in perceptions of injustice, specifically for respondents who have their UIT score rejected after the administration of few items. However, in practice CAT procedures prove easy to explain to respondents. Some test

administrators may also want to know the number of items that will be administered so that they can plan the assessment in relation to other activities such as interviews. Therefore, it is conceivable that these practitioners may prefer a fixed length test even though it is less efficient.

Finally, the SCTSPRT method for confirmation testing proposed in this article takes into account the Type I error in the test but not the Type II error. This limits the generalizability of the method to contexts where the importance of Type I error outweighs the importance of Type II error. The generalization of the method to control the Type II error would be straightforward, and would consist of setting a less conservative cut-off point with the consequence of increasing the number of false positives. A related issue is the modest power of this method for detecting moderate cheating effects. Although the SCTSPRT method represents an improvement compared to existing methods, the lack of power for detecting moderate cheating effects continues to be a challenge for confirmation testing.

A general challenge for practitioners using a confirmation test is the issue of how to deal with UIT scores that are rejected by the confirmation test. The International Guidelines on Computer-Based and Internet Delivered Testing (2005) provide a good outline of the procedures that should be in place before conducting a confirmation test. Burke, van Someren, & Tatham (2006) also offer a detailed description of a procedure that can be followed when a UIT score is rejected. One suggestion is to provide the possibility of taking a full-length supervised test to respondents who have their unproctored score rejected.

4.7.2 Future research

We anticipate several future research directions related to confirmation testing, the application of the SCTSPRT and UIT in general. The first is to add to the results of the current study by incorporating practical constraints such as content balancing and item exposure control. Also, methods for controlling the Type I and Type II error rates simultaneously might be investigated. The current article generalized research from classification testing to the context of confirmation testing. Future research could investigate other approaches for verifying UIT results in order to explore if there are alternative methods that can increase the power to detect cheaters while maintaining test efficiency. Another area of research is related to candidate's perceptions of confirmation tests and UIT in general. Different methods for processing UIT scores that are rejected could be investigated.

A further line of future research is related to methods of proctoring online tests. Biometric authentication systems are currently used together with online proctoring through the use of webcams, or a confirmation test (e.g., Foster, 2009). Online

proctoring can limit the flexibility of the test and can be expensive. Therefore, future research and technological advances could investigate ways to make these methods cheaper and more widely available.

Finally, the adaptive confirmation test using the SCTSPRT was applied specifically to the verification of UIT in this article. However, there are many assessment applications where the goal of the examination is to evaluate if there has been a change on the latent trait. Some examples could be to assess change: after a training program, after the administering a drug in clinical trials, or after administering a test at two different time points. In these applications the fundamental objective is not to re-assess the position of the individual on the latent trait, but alternatively to measure if a change on the position of the latent trait has taken place. Future research could investigate the possibility of using the SCTSPRT in these contexts.

Chapter 5

Optimizing Precision of Cognitive Ability Measurement in Organizational Assessment with Multidimensional Computerized Adaptive Testing

Abstract

Cognitive ability tests are widely used in organizations around the world because they have high predictive validity in selection contexts. Although these tests typically measure several sub-domains, testing is usually carried out for a single sub-domain at a time. This can be ineffective when the sub-domains assessed are highly correlated. This article illustrates the applicability of using multidimensional computerized adaptive testing (MCAT) for a cognitive ability test used in organizational contexts. MCAT can increase the validity of sub-domain ability scores by effectively administering and scoring items based on the relations between the sub-domains. This can be particularly useful in settings where there is a need to provide detailed feedback regarding sub-domain scores, but there is limited time for assessment. Results indicate that MCAT leads to improved test precision, shorter tests, and increased selection utility compared to more traditional testing methods.

Key Words: *multidimensional computerized adaptive testing, item response theory, international assessment, personnel selection, cognitive ability measurement*

This chapter has been submitted for publication as:
Makransky, G., & Glas, C. A. W. (submitted). *Optimizing precision of cognitive ability measurement in organizational assessment with multidimensional computerized adaptive testing.*

5.1 Introduction

Organizations around the world frequently use cognitive ability tests to make selection and promotion decisions because these tests have high predictive validity in organizational contexts (e.g. Schmidt & Hunter, 1998; Salgado, Anderson, Moscoso, & Bertua De Fruyt, 2003). The majority of these tests assess multidimensional constructs with several correlated sub-domains, such as: numeric, verbal and spatial ability. Although research has concluded that a general cognitive ability measure is valid for predicting performance in both educational and work domains (e.g. Schmidt & Hunter, 1998; Kuncel, Hezlett & Ones, 2004), there is often a desire to report sub-domain scores in the feedback process. A great deal of variation exists concerning the amount of feedback that is expected from cognitive ability tests in different countries. An example is that feedback reports are typically not provided in China (Burke, Tong, & Liu, 2011). Conversely, a detailed feedback report that highlights strengths and weaknesses in sub-domains is customary in northern Europe (this was vividly discussed during the development of the ISO 10667 standard for assessment service delivery, 2011).

There is a trend of moving from reporting general scores to reporting sub-domain scores in order to identify strengths and weaknesses in job candidates (Botempo, 2011). More detailed reports are preferred because managers and HR professionals want detailed information about potential or current employees in order to improve organizational decisions. In addition providing test takers with a more detailed feedback report may improve the test taker's perception of, and reaction to the testing procedure, which has been identified as an important variable in terms of managers' prioritizations of which selection procedures to use (Köning, Klehe, Berchtold & Kleinmann, 2010).

Lubinski (2004, p. 105) also concludes that both general and specific abilities are essential for understanding why certain learning and work environments are found attractive as well as aversive. According to Lubinski (2004, p. 105) general intellectual ability is critical for predicting migration up and down niches that differ in complexity (Wilk, Desmarais, & Sackett 1995; Wilk & Sackett, 1996), whereas specific abilities refine predictions about content or the nature of learning and work wherein cognitive abilities are expressed (Gottfredson, 2003).

Although the sub-domains in cognitive ability tests are correlated, testing is usually carried out for a single sub-domain at a time. This approach can sometimes lead to inaccurate sub-domain ability scores when the number of test items related to a sub-domain is small. This is often the case in organizational contexts where there is a trend to use short test forms because companies want to reduce testing time in

order to reduce costs. But even in these cases, there are often pressures from users of the test to report the inaccurate results (Reckese, 2009). Therefore, it is important to consider efficient measurement methods that decrease the amount of time needed for testing, without compromising measurement precision.

The assessment of a score on a single unidimensional sub-domain at a time is typical of traditional scoring methods such as: classical test theory and unidimensional item response theory (IRT). According to van der Linden and Hambleton (1997, p. 221) unidimensional models are not always appropriate for real tests. When the sub-domains measured by a test composite are correlated, responses to items measuring one sub-domain can provide clues about the test taker's ability on other sub-domains in the battery. Knowledge of the magnitude of the correlation between the sub-domains in the population of interest, in addition to the individuals' performance levels, can add a unique source of information that can provide more precise estimates of ability. This cross-information is ignored by conventional classical test theory and IRT scoring methods but can be effectively modeled by multidimensional item response theory (MIRT: e.g., Reckese, 2009) models.

MIRT models account for the multidimensional nature of the complex constructs that are often assessed in psychological tests by using a measurement model that is built on the premise that a test taker possesses a vector of person characteristics that describe the different abilities that he or she brings to the test, and the items are described by a vector of item characteristics that describe the difficulty and sensitivity of the items in the test to multiple constructs. MIRT opens up the opportunity to include theoretical model assumptions directly into the measurement process. An additional advantage of MIRT is that a test may be shorter while maintaining the same level of precision (Segall, 1996).

One of the instruments used to limit testing time without compromising precision is the administration of tests via the computer using computerized adaptive testing (CAT). A CAT selects items that match the test taker's ability so as to maximize the precision of the test based on information about the test taker from previous items. CATs reduce the number of items necessary in a test by approximately 50% (Weiss, & Kingsbury, 1984). Recent literature has investigated the feasibility of combining MIRT and CAT called multidimensional computerized adaptive testing (MCAT: e.g. Segall, 1996, 2000, 2010; Wang and Cheng 2004). With MCAT items are selected and scored based on the location of the test taker and the items on the multidimensional construct instead of each unidimensional construct in isolation. The use of this information can lead to more intelligent item selection and improved ability estimates.

There have been a few studies that have assessed the viability of MCAT with positive results. Luecht (1996) and Segall (1996) compared MCAT and CAT results in realistic test contexts and found savings in test lengths of 25% to 40%. Wang and Cheng (2004) compared the results of MCAT and CAT by simulating several conditions. The MCAT method improved test precision most when there were many sub-domains, and when these sub-domains were highly correlated. Several studies have evaluated MCAT in the field of medical science with positive results (Gardner, Kelleher, & Pajer, 2002; Heley, Pengsheng, Ludlow, & Fragala-Pinkham, 2006; Petersen, Groenvold, Aaronson, Fayers, Sprangers & Bjorner, 2006). These studies have used item parameters from real tests, but the generalization of the findings is limited because the item banks are common in the medical science field but differ from item banks typically found in psychological and educational assessments. MCAT is a methodological development that provides an intriguing opportunity for the field of organizational measurement where there is a need for short instruments that are highly precise. However, to our knowledge there is no research investigating the potential of MCAT for a cognitive ability test used in an organizational context.

The present article has two objectives: First, we investigate if MCAT can improve the precision of ability sub-domain scores for a cognitive ability test used in organizational testing compared to more traditional test administration and scoring methods. Second, we illustrate how the results can impact different HR procedures where cognitive ability tests are used.

The remainder of the study is organized as follows. First, we will describe how IRT and CAT can be used to estimate cognitive ability scores. Second, we will provide a background to the method we use to estimate scores using MIRT and MCAT. Then, we will investigate the precision and utility of the method by designing simulation studies based on real data from a cognitive ability test used for personnel selection and development in several European countries. Finally, we will discuss the utility and practical implications of using the MCAT method in an organizational context.

5.2 Unidimensional IRT and CAT

The proposed methods are defined in the framework of IRT. The fundamental concept of unidimensional IRT is that each test item is characterized by one or more parameters, and each test taker is characterized by a single ability parameter. The probability that a given test taker answers a given item correctly is given by a function of the test taker's ability level and the item parameters. Conditional on those parameters, the response on one item is assumed independent of the responses to other

items. There are different types of IRT models and more detailed information can be found in Embretson and Reise (2000); and Hambleton, Swaminathan, and Rogers (1991). The two parameter logistic (2-PL) response model has been frequently used for cognitive ability tests, where the probability of a correct response is given by the logistic function

$$P(x) = \frac{\exp(x)}{1+\exp(x)}, \tag{5.1}$$

with,

$$x = a_i(\theta - b_i), \tag{5.2}$$

In equation (5.2), θ is the test taker's ability, and a_i and b_i are parameters of item i. Furthermore, a_i is called the discrimination and b_i the difficulty parameter. It is common to scale ability levels to have a mean of zero and a standard deviation of one. The item difficulty parameter represents the point on the ability scale at which the probability of answering the item correctly equals .5. Higher discrimination parameters indicate that the probability of answering an item correctly rises dramatically with small changes in ability in the region of the item difficulty.

IRT provides several advantages over classical test theory for designing tests, test assembly, test scaling and calibration, construction of test item banks, investigation of test item bias and other common procedures in the test development process (for an overview see Hambleton et al., 1991). One of the advantages is that IRT provides more flexibility because items do not have to be used in the context of the other items in the test, but can be selected adaptively in a computer adaptive test (CAT; van der Linden & Glas, 2010). A CAT adapts to the test taker's ability level by selecting the next item dynamically so that it provides the most information about the test taker's ability; thereby improving the precision of the ability estimate.

5.3 Multidimensional IRT and MCAT

In a multidimensional IRT model several sub-domains may be estimated simultaneously. In this way a response to an item can give information about several sub-domains, either directly through the dependencies of item response and sub-domain score, or indirectly through the correlations between the sub-domain scores. Many MIRT models exist (for an overview see Reckese, 2009). In this article we have chosen to use the multidimensional extension of the 2-PL model, because it is a straightforward generalization of the unidimensional 2-PL model. The MIRT extension of the 2-PL model extends to cases where test takers can be characterized

by their standing on multiple traits. In the MIRT 2-PL model the x from Equation 5.2 has the following form:

$$x = \sum_m^M a_{im}\theta_m - b_i. \tag{5.3}$$

In this model the θ_m (m = 1,…,M) becomes an m dimensional vector of person coordinates with M indicating the number of dimensions in the coordinate space. The discrimination parameter from Equation 5.2 has a specific value for each dimension a_{im}. In other words the model has only one intercept, but has a specific discrimination parameter a_{im} for each sub-domain.

In most applied tests, items are designed to measure a single sub-domain. Therefore, we investigate a model where each item only contributes information directly to the sub-domain that it is intended to measure; and only contributes indirectly to the other sub-domains through the correlations of the sub-domains. Consequently, an item only has one non-zero discrimination parameter. This is known as a between item MIRT model. An alternative model is a within-item MIRT model that has several non-zero discrimination parameters and uses the loading on each domain to score the test directly. The between-item MIRT 2-PL model is given by:

$$x = a_i\theta_{m(i)} - b_i. \tag{5.4}$$

Here, $\theta_{m(i)}$ is the test taker's standing on the dimension that item i is intended to measure.

Four criteria are required to administer items adaptively with MIRT: a rule for starting the test, a selection criterion that defines which item should be selected and presented in the next step, a stopping criterion that defines when the CAT procedure should stop, and a method for calculating a score. These are described in more detail in the simulation studies section of this article.

5.4 Research Questions

There are three main research questions in this study: 1) Is the precision of the ability estimates for the cognitive ability sub-domains improved when items are administered by MCAT compared CAT and a random item administration for a cognitive ability test used in organizational testing? 2) Does the MCAT method require fewer items to achieve the same level of precision compared to CAT and random item administration for a cognitive ability test used in organizational testing?

3) Does the MCAT method result in biased estimates for test takers with unusual ability profiles (high ability on one sub-domain and low ability on other domains) compared to CAT and random item administration?

5.5 Methods

5.5.1 Questionnaire

The Adjustable Competence Evaluation (ACE) is a cognitive ability test used for screening, selection, and development activities in Austria, Denmark, Finland, Germany, Norway, Poland, Sweden, and Switzerland. The test measures competence in logical, analytical reasoning, which relates to the so-called g-factor (Spearman, 1904). The item bank includes 206 dichotomous (scored correct/incorrect) items measuring three cognitive domains: numeric (55 items), spatial (49 items), and verbal (102 items) ability. The test feedback report presents the general result and the selected sub-domain scores in isolation. The test is a power test based on CAT technology. The ACE test originally consisted of a single long version. A second shorter version of the test was developed because customers requested a version that could be administered in half of the time. The high precision version of ACE consists of 24 items (eight measuring each sub-domain) and takes approximately 45 minutes to complete. The standard precision version of ACE consists of 12 items (four measuring each sub-domain) and takes approximately 22 minutes to complete (for more information about ACE see Makransky & Kirkeby, 2010a).

5.5.2 Sample

The sample consisted of 1350 test respondents (762 men and 588 women) between the ages of 20 and 76. The test results were gathered from international companies in Denmark (N = 415), Germany (N = 200), Poland (N = 420), and Sweden (N = 315). The sample consisted of respondents with varied educational backgrounds, working in a variety of job categories. The test responses were collected in connection with personnel selection, recruitment, and individual development procedures.

5.5.3 Simulation studies

The research questions outlined above were investigated by means of simulation studies. A sample of 10,000 simulated test takers' numeric, spatial, and verbal ability

parameters were drawn from a normal distribution based on the characteristics of the sample described above. The latent correlations between the sub-domains (see Table 5.1) and the item parameters were computed with the free MIRT software package (Glas, 2010).

Table 5.1

Latent correlations between the sub-domains in the ACE sample

	Numeric Ability	Spatial Ability
Spatial Ability	0.72	
Verbal Ability	0.81	0.62

The simulation studies were programmed in Digital Visual FORTRAN 6.0 for Windows. Multidimensional computerized adaptive testing (MCAT) was compared to two other testing methods: computer adaptive testing (CAT), and random item administration (RAN). CAT and RAN item selection and scoring were carried out independently for one sub-domain at a time. For the MCAT method all three sub-domains were assessed simultaneously. Items were selected randomly from the item bank for the RAN method. The first three items were selected randomly for the MCAT and CAT methods. After three items, the item selection algorithm used Segall's (1996, 2000, 2010) Bayesian approach. For the MCAT method, a preliminary ability estimate on the three sub-domains was estimated based on the items that had been administered up to that point using the multidimensional IRT model. Then the item selection method chose the item that was expected to contribute most to the precision of the ability estimates, based on each item's multidimensional information function. The unidimensional model is a special case of multidimensional model, in the sense that the covariance between the sub-domain is ignored, that is, the covariances are set equal to zero. Therefore, in this case, the selection criterion corresponds to selecting the item that is expected to contribute most to the precision of the ability estimate in a unidimensional CAT. The test was stopped when a fixed number of items had been administered.

Ability estimates for each test taker were calculated using the expected a posteriori (EAP; Bock & Mislevy, 1982) for all three methods. EAP ability estimation is the mean of the posterior distribution of θ, which is a product of the prior distribution and the likelihood function. In the MIRT model, the prior distribution is based on the assumed distribution of respondents' ability scores. This is obtained based on the latent correlations between the sub-domains and the distribution of ability scores

in the calibration sample. The likelihood function is the probability of a response pattern given the ability level θ and the item parameters.

5.6 Results

Research question 1 was explored by administering the standard and high precision versions of ACE using the MCAT, CAT, and RAN methods. Pearson's correlation coefficient, and mean absolute error (MAE; average absolute difference) were calculated between the true ability scores drawn from the normal distribution and the observed ability scores obtained from using each testing method. Pearson's correlation coefficient and MAE are outlined in Table 5.2. The high precision ACE test is outlined in the top portion of the table. This version consists of 24 items, 8 measuring each sub-domain. The standard precision ACE test is outlined in the bottom portion of the table. This version consists of 12 items, 4 measuring each sub-domain.

Table 5.2

Correlation and MAE between true and estimated score for each scoring method

		Correlation			MAE		
		MCAT	CAT	RAN	MCAT	CAT	RAN
ACE High Precision	Numeric	0.92	0.90	0.82	0.32	0.36	0.46
	Spatial	0.86	0.86	0.73	0.41	0.41	0.54
	Verbal	0.89	0.88	0.74	0.37	0.38	0.54
	Average	0.89	0.88	0.76	0.37	0.38	0.51
ACE Standard Precision	Numeric	0.89	0.83	0.71	0.36	0.45	0.55
	Spatial	0.83	0.81	0.61	0.44	0.48	0.63
	Verbal	0.87	0.80	0.61	0.39	0.48	0.63
	Average	0.86	0.81	0.64	0.40	0.47	0.60

The results indicate that the MCAT method had the highest precision, and the RAN method had the lowest precision. In the high precision test the differences between the MCAT and CAT method were not large (average correlation difference = 0.01); however the MCAT method increased the correlation between the observed and the true ability scores by an average of 0.13 per sub-domain as compared to the RAN method. The magnitude of the differences was more pronounced in the shorter test where the MCAT method increased the correlation between the observed and true ability scores by an average of 0.05 per sub-domain compared to the CAT method,

and by 0.22 compared to the RAN method. When comparing the precision of the high versus the standard precision versions of ACE, it is clear that the correlation between the observed and true scores only decreased by an average of 0.03 per sub-domain when the MCAT method was used. This is not a big difference, given that the standard precision version is only half as long. The decrease in the correlations was an average of 0.07 for the CAT method and 0.12 for the RAN method. In general, the results indicate that the MCAT and CAT methods were superior to the RAN method for both test lengths. Additionally, the benefits of using MCAT compared to CAT were clear for the standard precision version of ACE; however, the differences were not large for the high precision version of ACE.

Research question 2 was tested by calculating the MAE of the standard and high precision versions of ACE using the MCAT method. Then, given these values, the number of items required to obtain the same precision was calculated with the CAT and RAN methods. Table 5.3 highlights the number of items required to obtain a similar level of precision using the CAT and RAN methods compared to the MCAT method. The high precision test is outlined in the top portion of the table, and the standard precision test is outlined in the bottom portion of the table.

Table 5.3

Number of items necessary to obtain a similar level of precision

		MCAT	CAT	RAN
ACE High Precision	Numeric	8	11	22
	Spatial	8	8	20
	Verbal	8	9	25
	Total	24	28	67
ACE Standard Precision	Numeric	4	8	17
	Spatial	4	5	15
	Verbal	4	8	20
	Total	12	21	52

For the high precision ACE, 28 items were required with the CAT method, and 67 items were required with the RAN method compared to 24 items with the MCAT method. This means that the CAT method required 17% more items, and the RAN method almost three times more items compared to the MCAT method. The difference was even larger for the standard precision ACE. Here the CAT method required 21 items and the RAN method 52 items compared to 12 items with the MCAT method.

This is an increase of 75% with the CAT method, and over 400% with the RAN method compared to the MCAT version of the test.

Additional precision is obtained in the MCAT method because information about the correlations between the sub-domains is used in order to calculate specific sub-domain scores. Therefore, it is not unexpected that the method produces more accurate results when averaging across all test takers. However, it is important to investigate if the MCAT method introduces bias for test takers with unusual ability profiles. An unusual ability profile is one where a test taker has a high ability level in one sub-domain and low ability levels in one or more of the other sub-domains.

Research question 3 was examined by simulating two unusual ability profiles 10,000 times, to measure the bias in the three testing methods. Unusual profile one is a test taker who has a numeric ability level that is one standard deviation (SD) unit above the mean, but has spatial and verbal ability levels that are one SD unit below the mean. Unusual profile two is a test taker who has a very high numeric ability level (two SD units above the mean), an average spatial ability level (at the mean), and a very low verbal ability level (two SD units below the mean). Figure 5.1 outlines the true ability score, and the average score obtained from using each of the three testing methods for the two test takers. The top panel of the figure outlines the results for unusual profile one and the bottom panel summarizes the results for unusual profile two. The left part of the figure outlines the results for the standard precision ACE, and the right side highlights the results for the high precision ACE.

When comparing the methods, the results indicate that the RAN method was consistently more biased than the MCAT and CAT methods across all conditions. For the high precision test the MCAT and CAT methods performed almost identically. For the standard precision test the MCAT method performed better than the CAT method in terms of numeric ability, but worse than the CAT method in terms of verbal ability and spatial ability for unusual profile one. In general it does not seem like the MCAT method introduces additional bias compared to the CAT and RAN methods for unusual ability profiles. It should be noted that the ability estimates for all three methods were biased toward the mean. This resulted because the expected a posteriori (EAP) estimation method was used. The EAP estimation method is known to produce estimates that are biased toward the mean (Warm, 1989; van der Linden & Pashley, 2010).

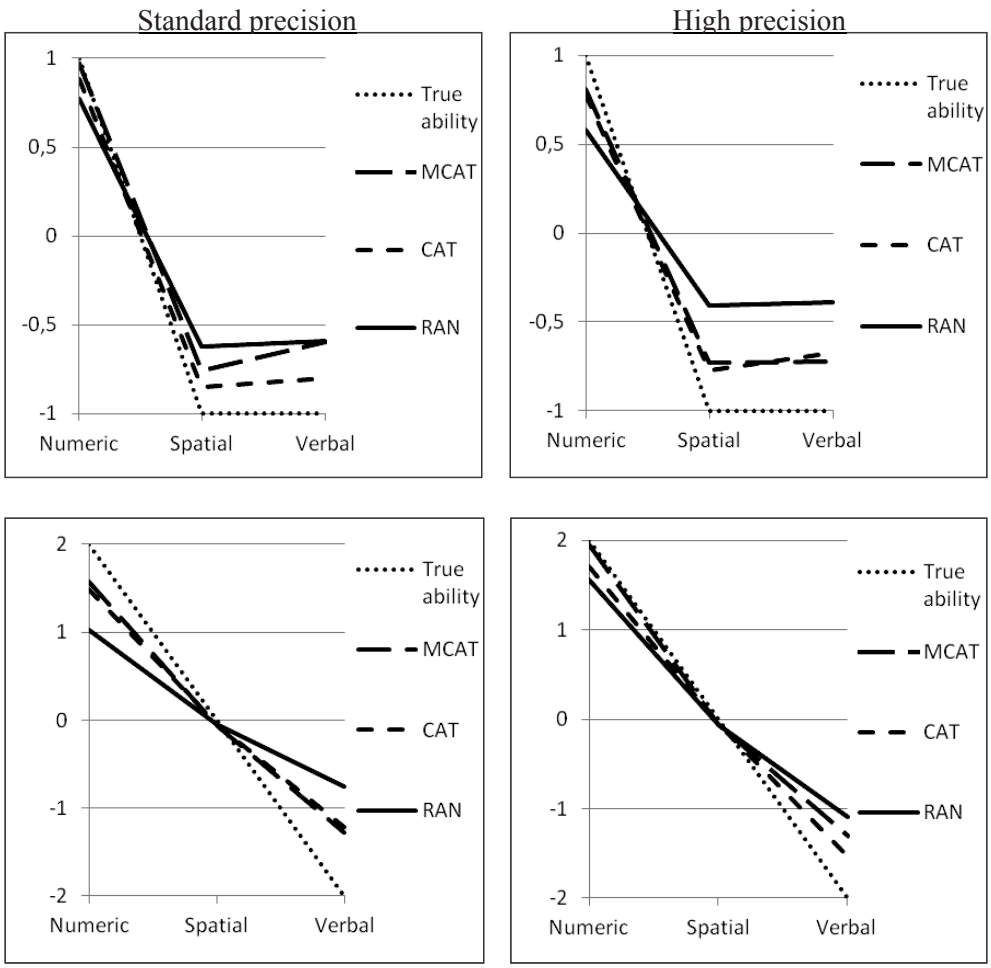

Figure 5.1. True ability, and the average score obtained from using MCAT, CAT and RAN methods, for unusual profile one with ability scores 1, -1, -1 (top panel), and unusual profile two with ability scores 2, 0, -2 (bottom panel) for numeric, spatial and verbal ability, respectively, in the standard (left panel), and high (right panel) precision versions of ACE.

Given the added precision of the MCAT method, it was interesting to exemplify the increased precision in terms of added utility (the monetary value of productivity gains) when using the MCAT method for a multiple cutoff selection procedure. This was done by conducting a simulated selection procedure. In the simulated selection procedure, the three sub-domains were considered equally important for the job. Therefore, a proportion of candidates were selected who all met the same minimum criteria on all three sub-domains. The test results were the only criteria used for selecting candidates. Job performance was calculated by creating a variable that

correlated r = 0.51 with general cognitive ability (the average of the true scores for the three sub-domains). A correlation of 0.51 was based on findings from a meta-analysis of the correlation between general cognitive ability and job performance across a wide variety of jobs (Hunter, 1980; Hunter & Hunter, 1984, Schmidt & Hunter, 2004). Similar findings have also been identified in a wide variety of European based jobs (Salgado et al., 2003). The utility of the selection procedure was assessed by using a utility model

$$\delta = \bar{O}_x SD_y \ , \tag{5.5}$$

similar to the model described by Schmidt and Hunter (1998). Here δ is the change in average utility per hire per year. $\overline{O}_x$ is the average output or performance for the selected candidates, and SD_y is the standard deviation of performance. The standard deviation of performance has been found to be at a minimum 40% of the mean salary of the job (Schmidt & Hunter, 1983; Schmidt, Hunter, McKenzie, & Muldrow, 1979; Schmidt, Mack, & Hunter, 1984).

The sample of simulated job candidates was replicated three times with different selection ratios. Therefore, the minimum cut-offs were set so that the best 10%, 20%, and 50% of the sample would be selected for the job based on their test results. Given the same number of test items, Figure 2 outlines the amount of money saved in a selection procedure by using the CAT and MCAT methods compared to the RAN method. These savings are calculated for a typical position where a cognitive ability test would be used, with a yearly salary of $65,000.

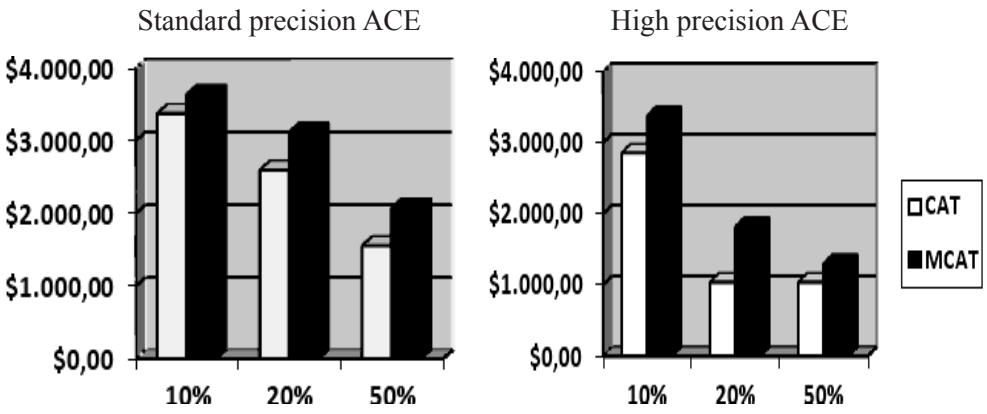

Figure 5.2. Utility (the value of productivity gains) of the MCAT and CAT methods compared to the RAN method as a baseline for a job with a yearly salary of $65,000.

The figure illustrates that the savings of using the MCAT and CAT methods compared to the RAN method are greatest for the standard precision test. This is the case because the difference in precision between the MCAT and CAT methods compared to the RAN method were greatest in the standard precision test. The savings were highest when the selection ratio was small. The savings of the MCAT method compared to the CAT method were fairly stable across all conditions, and were around $500 per employee per year. So, if a company hires 10 employees. The savings of using the MCAT method compared to CAT method would be approximately $25,000 assuming that these employees stay at the job for an average of 5 years. The savings of using the MCAT compared to RAN method in this example would range from $65,000 to $182,000 depending on the version of the test that is used and the selection ratio.

5.7 Discussion

There is a trend in personnel selection practice to use short test forms because companies want to reduce testing time in order to reduce costs. This can become a problem in settings where there is a need to provide feedback regarding the separate sub-domains in the test, since a limited number of items may result in inaccurate sub-domain scores. Therefore, it is important to consider efficient measurement methods that decrease the amount of time needed for testing, without compromising measurement precision. In this article we illustrate the benefits of

using multidimensional computerized adaptive testing (MCAT) for a cognitive ability test used in organizational contexts.

The results of the study suggest that the benefits of using the MCAT method are particularly evident in short tests. Specifically, the results showed that the MCAT method increased the correlation between the observed and true score by an average of 0.05 per sub-domain compared to computer adaptive testing (CAT), and by an average of 0.22 compared to random item administration (RAN) with the standard precision ACE test. The benefits of the MCAT method compared to the CAT method were minor (average correlation difference = 0.01) for the longer test; however, the MCAT method increased the correlation between the observed and true score by an average of 0.13 compared to the RAN method. The increased precision can also be illustrated in terms of the number of items necessary in order to obtain the same level of precision. The RAN method was almost three, and over four times longer, than MCAT method for the standard and high precision ACE tests, respectively. In addition, the MCAT method was 75% shorter with the standard precision ACE test, and 17% shorter with the high precision ACE test, compared to the CAT method. Previous research has reported savings in test lengths between 25 and 40% for MCAT compared to CAT methods (Luecht, 1996; Segall, 1996). Therefore, the findings in this study suggest that the benefits of MCAT may be greater in organizational contexts. The MCAT and CAT methods performed comparably in terms of estimating ability scores for test takers with unusual ability profiles, however, both methods performed better than the RAN method.

The utility of using each testing method was assessed by simulating a multiple cutoff personnel selection procedure. For a single employee with a salary of $65,000 a year, the savings could range from $1,300 to $3,640 per year by using the MCAT method compared to the RAN method for selection. The savings were approximately $500 when the MCAT method was used rather than CAT. Added utility is one of the possible benefits of obtaining improved precision in sub-domain scores for HR screening and selection procedures. An additional benefit is the increased quality of the feedback that is provided to a job applicant. An increase in the quality of the feedback the applicant receives is likely to improve their perception of, and reaction to the testing procedure. This is important in the field of organizational testing because research indicates that managers may prioritize applicant reactions even higher than variables such as predictive validity in deciding on what selection procedures to use (Köning, Klehe, Berchtold & Kleinmann, 2010). This is the case because positive applicant reactions can result in a better impression of the organization.

Important development and promotion decisions are also made based on the domain-specific ability scores, and the quality of these decisions could be affected

by the accuracy of their estimates. Even among people with comparable general cognitive ability, those with sharp specific ability differences have distinct preferences for processing and working with different mediums: these in turn characterize contrasting learning and work environments. Newer talent search programs take advantage of these individual differences (Lubinski, 2004, p. 106).

Beyond the advantages of higher precision and shorter test length, MCAT provides the advantage that one can map the theoretical model directly into the measurement model. This opens up an attractive new step to more theoretically-based testing that is likely to enhance the validity of test score interpretations within psychological and education assessments (Frey & Seitz, 2009). The use of MCAT could increase test developers and psychometricians' awareness of the content of test items in terms of the bigger multidimensional construct, and not just the specific sub-domain. This could lead psychologists to begin to focus more attention on several important psychological questions that are essentially disregarded in current psychological research. Borsboom (2006, p. 431) highlights questions such as: What are the psychological processes that the test items evoke? How do these processes culminate in behaviors, like marking the correct box on an IQ-item? How do such behaviors relate to individual differences? What is the structure of individual differences themselves? What is the relation between such structures and the test scores?

5.7.1 Limitations

A possible limitation in this study is that it uses the adjustable competence evaluation (ACE) item bank to investigate the benefits of MCAT for a cognitive ability test meant for personnel selection. Since item selection is the driving force behind the increased precision in MCAT and CAT, it is possible that the results do not generalize across cognitive ability tests that have large differences in item bank characteristics. However, previous studies with larger (e.g. Segall, 1996, 2000, 2010; Wang & Cheng 2004), and smaller (e.g. Gardner, Kelleher, & Pajer, 2002; Heley, Pengsheng, Ludlow, & Fragala-Pinkham, 2006; Petersen, Groenvold, Aaronson, Fayers, Sprangers & Bjorner, 2006) item banks have also concluded that MCAT increases the precision of the test compared to CAT and other testing methods. Consequently, similar results could be expected across tests.

Possibly the greatest barrier to the widespread implementation of MIRT and MCAT is the lack of commercially available software to make these methods easily accessible to organizations that are convinced of their value. It is not until recently that commercial software has become available for administering a CAT (e.g., Weiss, 2008). The availability of accessible software has lead to countless CAT applications in a wide variety of fields. The requirements to apply MCAT methods are similar to

those of a CAT. Therefore, given the large amount of research interest and advantages of MCAT, it is likely that commercial software will soon be available to implement these methods.

5.7.2 Future research

In the current article we have focused on a cognitive ability test. However, most constructs measured in organizational testing are multidimensional, with sub-domains that are highly correlated. Personality tests are an obvious example of a test type where the advantaged of MIRT and MCAT should be assessed. In a personality test it is the scores on the individual sub-domains that are of primary interest. In addition, these sub-domains are highly correlated. Therefore, an attractive research area is to investigate the benefits of using MIRT and MCAT for non-cognitive organizational tests.

In this article we have focused on between-item multidimensionality. There are existing tests where a single item loads on several sub-domains, and within-item MIRT could be considered. Future research could investigate the benefits and possible limitations of using a within-item MIRT model for scoring some of these measures.

Although there has been research available on MIRT and MCAT for over 15 years, to our knowledge there are still no large scale applications of these methods. This trend may be similar to the trend in CAT. The first research on CAT was conducted in the late 1970's and early 1980's; however, it was not until the last decade that there was broad implementation of the method in applied settings. The application of MIRT and MCAT to real tests will likely highlight new challenges that require new research directions.

Chapter 6

Improving Personality Facet Scores with Multidimensional Computerized Adaptive Testing: An Illustration with the NEO PI-R

Abstract

Narrowly defined personality facet scores are commonly reported and used for making decisions in clinical and organizational settings. Although these facets are typically related, scoring is usually carried out for a single facet at a time. This method can be ineffective and time-consuming when personality tests contain many highly correlated facets. This article investigates the possibility of increasing the precision of the NEO PI-R facet scores by scoring items with multidimensional item response theory (MIRT) and by efficiently administering and scoring items with multidimensional computerized adaptive testing (MCAT). The increase in the precision of personality facet scores is obtained from exploiting the correlations between the facets. Results indicate that the NEO PI-R could be substantially shorter without attenuating precision when MCAT methodology is used. Furthermore, the study shows that the MCAT methodology is particularly appropriate for constructs that have many highly correlated facets.

Key Words: *multidimensional computerized adaptive testing (MCAT), multidimensional item response theory (MIRT), personality assessment, NEO PI-R*

This chapter has been accepted for publication as:
Makransky, G., Mortensen, E. L., & Glas, C. A. W. (in press). Improving personality facet scores with multidimensional computerized adaptive testing: An illustration with the NEO PI-R. *Assessment*.

6.1 Introduction

In personality research there is debate regarding the use of fine-grained versus broad personality variables, known as the bandwidth-fidelity controversy (e.g., Ones & Viswesvaran 1996; Paunonen & Ashton, 2001; Dudley et al. 2006). Results of several studies have concluded that narrow personality traits can substantially increase the quality of prediction achieved by broad personality factors (e.g., Ashton, 1998, Paunonen et. al. 1999; Stewart, 1999; Paunonen & Ashton, 2001; Dudley et al. 2006). These findings suggest that the use of multiple facets of personality provides important advantages over the use of multidimensional aggregates of these facets. These advantages pertain not only to the empirical accuracy in predicting outcomes, but also to the psychological meaningfulness in explaining behavior (Paunonen et. al. 1999). A more detailed picture of personality obtained through the facet scores also provides the test respondent with a richer source of information in the feedback process.

By contrast, Ones and Viswesvaran (1996) advocate the use of broader personality traits instead of narrow facets. They point out that the reliabilities of the narrow personality facets are typically lower than the broad measures, which decreases their predictive validity. They suggest that for specific personality trait measures to reach adequate reliability, scales would have to be lengthened by three to six times. This seems unrealistic in most test settings because assessment time is a precious commodity, which means that there is often a limit to the amount of time respondents are expected to spend on a test.

When personality tests contain many facets, it is time-consuming and inefficient to accurately assess each one individually. Consequently, most personality tests are either long, or are designed to assess broad general constructs, thereby ignoring the individual facet scores that make up these general constructs. An example is the widely used NEO-PI which has two versions: the NEO PI-R, and the NEO FFI (Cost and McCrea, 1992). The NEO PI-R consists of 240 items. With this version score reports are available for the 30 facets within the Big Five model of personality. The NEO FFI consists of 60 items; however, scores are only reported for the five general factors.

The ideal personality test would be an instrument that reports reliable facet scores with few items, thereby limiting assessment time. Currently, most personality test scoring procedures ignore the fact that the facets included in a test are highly correlated. In addition, items are administered without consideration for the information that has been collected from previous items. More efficiency could be achieved if the known correlations between the facets on a personality test and the

previous item responses could be used during a test. If information has been collected which suggests that a given person is dutiful, organized, and achievement-striving, then it is likely that the person also has a high level of self-discipline. Therefore, an appropriate item may be one that provides high information in the high score range on this facet. If the respondent answers as expected, then more confidence can be placed in the hypothesis that the respondent is self-disciplined. Thus, the information reflected in correlations among related facets can be used to derive efficient scoring procedures.

Multidimensional item response theory models (MIRT: e.g., Reckese, 2009), which use information about the correlations between facets to efficiently score test results provides a framework for administering items adaptively based on the characteristics of the items and information about the test respondent from previous items. This is known as multidimensional computerized adaptive testing (MCAT: e.g., Segall, 1996) and is similar to the general idea of computerized adaptive testing (CAT: e.g., van der Linden & Glas, 2010) but within a multidimensional framework. MCAT has been found to increase the precision of trait scores compared to CAT and more traditional scoring methods for tests used in ability testing (Segall, 1996, 2000, 2010; Wang & Cheng 2004), certification testing (Luecht, 1996), and medical testing (Gardner, Kelleher, & Pajer, 2002; Heley, Pengsheng, Ludlow, & Fragala-Pinkham, 2006; Petersen, Groenvold, Aaronson, Fayers, Sprangers & Bjorner, 2006). MIRT and MCAT are methodological developments which provide an intriguing opportunity for the field of personality testing where there is a need to assess a large number of correlated personality facets with a high level of precision.

The objective of this article is to investigate the possibility of increasing the precision of the NEO PI-R facet scores by efficiently administering and scoring items with MIRT methodology. Furthermore, we will investigate if it is possible to make the NEO PI-R shorter without attenuating reliability with MCAT. The remainder of the article is organized as follows: First, we will outline the research questions posed in this study. Second, we will describe how item response theory (IRT) can be used to estimate personality facet scores. Next, we will provide a background to the method we use to estimate scores using MIRT and MCAT. Then, we will investigate the precision of the method by simulation studies based on real data from the NEO PI-R. Finally, we will discuss the practical implications of the results for personality testing.

6.2 Research Questions

There are four main research questions in this study: 1) Is the precision of the trait estimates for facet scores in the NEO PI-R improved when using MIRT compared to unidimensional IRT scoring methods? 2) Is it possible to make the NEO PI-R shorter without attenuating precision when using the MCAT method? 3) What is the loss of accuracy with even shorter test lengths when the CAT and MCAT methods are used? 4) How similar are the results of a linear test scored with either IRT or MIRT, and a test administered and scored using CAT or MCAT methodology in a real testing situation?

6.3 Item Response Theory

IRT has become the dominant psychometric framework for the construction, analysis, and administration of large-scale aptitude, achievement, and ability tests (Embretson & Reise, 2000). Advances in IRT research have been applied to improving the interpretability and validity of widely used personality tests (e.g., Reise & Henson, 2000; Waller, Thompson, & Wenk, 2000; Egberink, Meijer & Veldkamp, 2010). In IRT, the responses to items are modeled as a function of one or more person ability parameters and item parameters. The item parameters are item difficulty parameters and item discrimination parameters. The difficulty parameters define the overall salience of the response categories and the discrimination parameters define the strength of the association of the item responses with the latent person parameters. The latent person parameters locate the person on a latent scale where high values are related to high item scores. The fundamental concept of IRT is that individuals and items are characterized on the same metric in terms of their location on the latent scale. IRT models can be applied to both dichotomous (two answer categories, e.g., true/false) and polytomous data (more than two answer categories, e.g., 5-point Likert scale). There are different types of IRT models and more detailed information can be found in Embretson and Reise (2000), and Hambleton, Swaminathan, and Rogers (1991). In this article, we use the Generalized Partial Credit model (GPCM: Muraki, 1992). In the GPCM, items are conceptualized as a series of ordered thresholds where examinees receive partial credit for successfully passing each threshold.

6.4 Multidimensional Item Response Theory and MCAT

Up to now most of the research that has applied IRT to personality data has used unidimensional IRT models. Unidimensional IRT models typically rely on the assessment of a score on a single unidimensional domain at a time. This does not take into account the multidimensional nature of the complex constructs which are assessed in most personality tests. Furthermore, when the facets measured by a test are correlated, responses to items measuring one facet can provide information about the examinee's standing on other facets in the test (van der Linden & Hambleton, 1997). That is, knowledge of the magnitude of the correlation between the facets in the population of interest, in addition to the individuals' performance levels, can add a unique source of information that can provide more precise trait level estimates.

MIRT models account for the multidimensional nature of the complex constructs based on the premise that a respondent possesses a vector of latent person characteristics that describe the person's level of personality on the different traits. In most personality tests, items are designed to measure a single facet. Therefore, we investigate a model where each item only contributes information directly to the facet that it is intended to measure; and only contributes indirectly to the other facets through the correlations of the facets. In this case, the item has only one non-zero discrimination parameter associated with the dimension on which the item response depends. This is known as a between-items MIRT model or a simple-structure MIRT model. An alternative model is a within-items MIRT model or a complex-structure MIRT model in which each item measures several facets. Therefore, each item has non-zero discrimination parameters for each dimension on which the item-response depends.

Many MIRT models exist (for an overview see Reckese, 2009). In this article we have chosen to use the multidimensional extension of the GPCM. In practice, results obtained using the GPCM can hardly be distinguished from results obtained using alternative models such as the graded response model and the sequential model (see, Verhelst, Glas & de Vries, 1997). The multidimensional GPCM is a straightforward generalization of the unidimensional GPCM. The multidimensional GPCM extends to cases where examinees can be characterized by their standing on multiple traits, which have a multivariate normal distribution.

The fundamental goal of MCAT is to locate the respondent's position in a multidimensional latent space with a high level of precision (or low level of error). This can be done by administering items adaptively by selecting the next item that is expected to contribute most to the precision of the trait estimates. Three criteria are required to administer items adaptively with MIRT: a selection criterion that defines

which item should be selected and presented in the next step, a stopping criterion that defines when the test should stop, and a method for estimating scores on the latent traits.

Segall (1996, 2000, 2010) developed a Bayesian method for selecting dichotomously scored items adaptively in a MCAT. This method was generalized to polytomously scored items by Wang and Chen (2004). The method estimates the running trait estimates on the facets based on the items that have been administered up to that point using the multidimensional IRT model. This information is used to choose the item that is expected to contribute most to the precision of the trait estimates, based on each item's multidimensional information function (that is, based on the predicted determinant of the information matrix). This model fits a so-called empirical Bayesian framework where the multivariate normal latent trait distribution is a prior. That is, this distribution models the correlations between the traits and these correlations are estimated along with the item parameters in the calibration phase, and are treated as known in the MCAT phase. The unidimensional model is a special case of multidimensional model, in the sense that the covariance between the traits is ignored, that is, the covariances are set equal to zero. Therefore, in this case, the selection criterion corresponds to selecting the item that is expected to contribute most to the precision of the trait estimate in a unidimensional CAT.

Although, there are many options available for stopping criteria, most operational adaptive tests use a fixed number of items as a stopping rule. There are also many options available for estimating the latent scores in adaptive testing. Previous research in the field of MCAT has favored the maximum a posteriori (MAP; Segall, 1996, 2000, 2010) method. The MAP estimate is the mode of the posterior distribution of latent trait, which is a product of the prior distribution and the likelihood function.

Recent research has successfully investigated the use of MIRT for modeling the relationships of examinees to a set of test items that measure multiple constructs (e.g., Yao & Boughton, 2007; de la Torre, 2008; Finch, 2010). In the framework of dichotomous items, MCAT has been found to increase the precision of trait scores compared to CAT and more traditional scoring methods for tests used in ability and achievement testing (e.g. Luecht 1996; Segall, 1996, 2000, 2010). Wang and Cheng (2004) adapted CAT to multidimensional scoring of polytomous (e.g., Likert scale) items. The present study examines the benefits of MCAT for a highly dimensional real polytomous test. Furthermore, this study investigates the benefits of MIRT and MCAT for a personality test that uses Likert scale items.

6.5 Method

6.5.1 Instruments

The Danish version of the NEO PI-R (Costa & McCrae, 1992) is a computer-based Big Five personality questionnaire. It consists of 240 items, distributed over five constructs (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). Each construct consists of 48 items, equally distributed over six facets: Neuroticism (N1 Anxiety, N2 Hostility, N3 Depression, N4 Self-Consciousness, N5 Impulsiveness, N6 Vulnerability to Stress), Extraversion (E1 Warmth, E2 Gregariousness, E3 Assertiveness, E4 Activity, E5 Excitement Seeking, E6 Positive Emotion), Openness (O1 Fantasy, O2 Aesthetics, O3 Feelings, O4 Actions, O5 Ideas, O6 Values), Agreeableness (A1 Trust, A2 Straightforwardness, A3 Altruism, A4 Compliance, A5 Modesty, A6 Tender mindedness), Conscientiousness (C1 Competence, C2 Order, C3 Dutifulness, C4 Achievement Striving, C5 Self-Discipline, C6 Deliberation). The items are scored on a 5-point Likert scale. The construct validity of the Danish version of the NEO PI-R has been established with exploratory factor analyses, and by comparing the test to other commonly used personality instruments (Skovdahl Hansen & Mortensen, 2003).

6.5.2 Sample

The NEO PI-R dataset consisted of 600 respondents (300 males, and 300 females) included in the Danish norm sample. Data were collected in the context of a large Danish twin study. The sample consisted of 1213 twins from ages 18 to 67. One twin was selected randomly from each pair to make up the norm sample. The latent correlations between the facets (see Table 6.1) and the item parameters were computed with the free MIRT software package (Glas, 2010). These values were used in all simulations presented below.

Table 6.1

Latent correlations between the facets in the NEO PI-R

	N1	N2	N3	N4	N5	Average correlation
N2	0.81					0.70
N3	0.91	0.85				
N4	0.87	0.72	0.87			
N5	0.40	0.59	0.41	0.34		
N6	0.85	0.79	0.87	0.82	0.38	
	E1	E2	E3	E4	E5	
E2	0.62					0.38
E3	0.30	0.32				
E4	0.40	0.35	0.46			
E5	0.18	0.33	0.38	0.26		
E6	0.59	0.45	0.33	0.48	0.23	
	O1	O2	O3	O4	O5	
O2	0.53					0.60
O3	0.68	0.63				
O4	0.75	0.55	0.55			
O5	0.72	0.64	0.75	0.69		
O6	0.49	0.47	0.67	0.36	0.48	
	A1	A2	A3	A4	A5	
A2	0.44					0.34
A3	0.45	0.42				
A4	0.30	0.42	0.37			
A5	0.19	0.47	0.29	0.36		
A6	0.23	0.25	0.34	0.22	0.31	
	C1	C2	C3	C4	C5	
C2	0.53					0.60
C3	0.68	0.63				
C4	0.75	0.55	0.55			
C5	0.72	0.64	0.75	0.69		
C6	0.49	0.47	0.67	0.36	0.48	

6.5.3 Analysis

The research questions outlined above were investigated by means of a Monte Carlo and a real data simulation study. The Monte Carlo simulation can provide information about the accuracy of the different scoring methods for a situation where the true trait

scores are known. The real data simulation was used because the model will not fit perfectly in a real testing situation. Therefore, the consequences of using CAT and MCAT methods were investigated in a simulation study where the actual responses from the Danish norm sample were used.

Four different scoring methods were compared: Unidimensional IRT scoring of a linearly administered test of all 240 items (Lin-IRT), multidimensional IRT scoring of a linearly administered test of all 240 items (Lin-MIRT), unidimensional computerized adaptive administration and scoring of 180, 120, 90, and 60 items (CAT), and multidimensional computerized adaptive administration and scoring of 180, 120, 90, and 60 items (MCAT).

Lin-IRT and CAT methods were carried out independently for one facet at a time. All six facets within each of the Big Five constructs were assessed simultaneously using the Lin-MIRT and MCAT methods. There were no constraints added to limit the item selection for the MCAT method. That is, the best item was selected regardless of its facet. An alternative approach is returned to in the discussion section. The test was stopped when a fixed number of items had been administered.

6.6 Results

6.6.1 Study 1: Recovery of true trait scores

Research questions 1, 2, and 3 pertaining to the precision of the test scores were investigated in the Monte Carlo simulation study. In this study 6,000 simulated respondents' true trait scores were randomly drawn from a normal distribution based on the scores from the NEO PI-R Danish norm sample. The root mean squared error (RMSE) was used to quantify the precision of the scoring methods. The RMSE is defined as the square root of the mean squared error (MSE). This error is the difference between the trait estimate and its true trait value. The RMSE incorporates both the variance of the estimator and its bias.

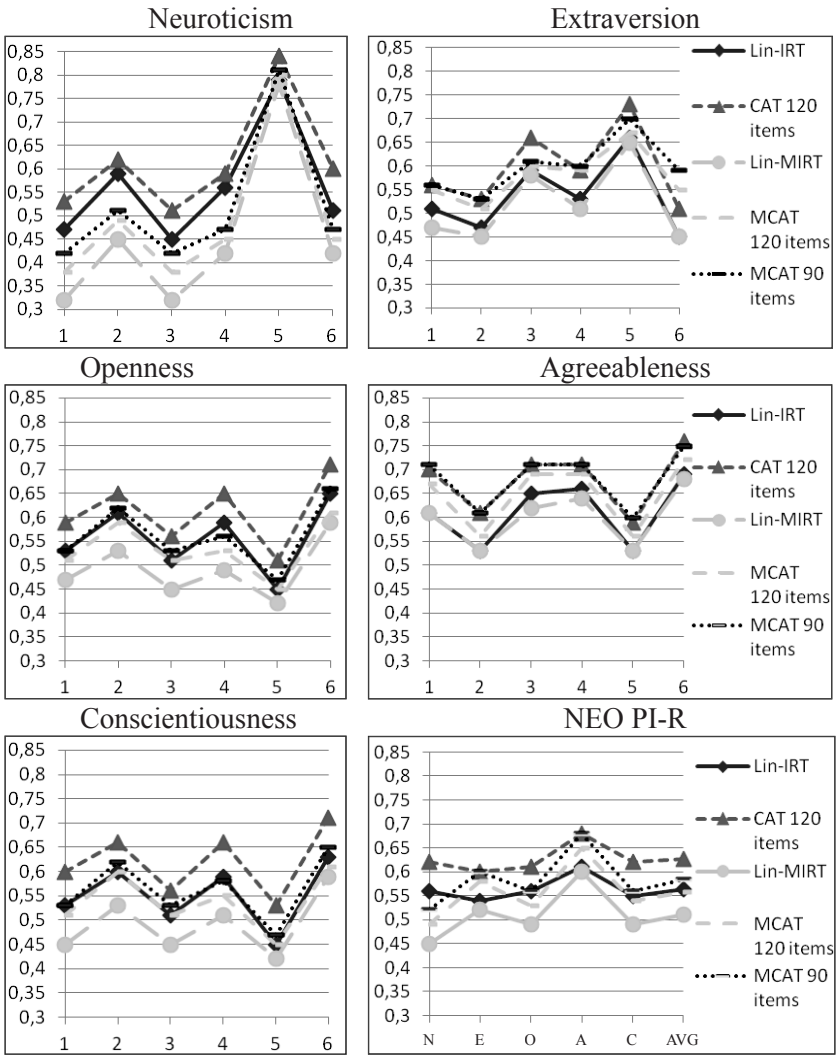

Figure 6.1: RMSE for the facets in the NEO PI-R.

Figure 6.1 presents the RMSE between the true trait scores and estimated trait scores obtained using each testing method. To ensure that Figure 6.1 is interpretable, only the results for the full length test (240 items) scored using Lin-IRT and Lin-MIRT, as well as the CAT with 120 items, and the MCAT with 90 and 120 items are presented. The RMSE is presented on the vertical axis. The first five panels present the six facets within each Big Five construct on the horizontal axis. The final panel presents the results averaged over facets for the Big Five constructs.

Clear results were found regarding research question one, pertaining to the precision of the trait estimates for facet scores when using MIRT compared to unidimensional IRT scoring methods. The results indicate that the Lin-MIRT scoring method consistently decreased the error of the facet scores compared to the Lin-IRT method (see Figure 6.1). This decrease in RMSE averaged over facets was 0.11 for Neuroticism, 0.07 for Openness, 0.06 for Conscientiousness, 0.02 for Extraversion, and 0.01 for Agreeableness. This means that the individual facet scores are more accurate when the correlation matrix between the facets is used to score the NEO PI-R. The difference in the improved accuracy was a consequence of the magnitude of the latent correlations between the facets in the different Big Five constructs. The average latent correlations between the facets in the Neuroticism, Openness and Conscientiousness constructs were 0.70, 0.60 and 0.60 respectively (see Table 6.1). By contrast, the average latent correlations for the Extraversion and Agreeableness constructs were 0.38 and 0.34 respectively.

Research question two, pertaining to the possibility of making the NEO PI-R shorter without attenuating precision when using the MCAT method was investigated by calculating the accuracy of the facet scores with different test lengths using the CAT and MCAT methods. Averaging across all facets, the final panel of Figure 6.1 illustrates that the NEO PI-R could be reduced to 90 items with the MCAT method, while only slightly decreasing the general accuracy of the test compared to the Lin-IRT scoring method. The MCAT method with 120 items resulted in a slight increase in the general accuracy compared to the full NEO PI-R with the Lin-IRT scoring method. The MCAT method also produced facet scores that were comparable in terms of accuracy to the CAT method with half as many items per facet. Again, the improvement in accuracy for the MCAT method depended on the average latent correlation between facets within each construct. The MCAT method resulted in facet scores that were on average at least as accurate as the Lin-IRT scores with the 60 item test for Neuroticism, the 90 item test for Openness, the 120 item test for Conscientiousness, the 150 item test for Extraversion, and the 210 item test for the Agreeableness construct.

The results in Figure 6.1 also show that the accuracy of the test decreases as the number of items that are administered is reduced. Therefore, the results related to research question three indicate that the full scale scores could not be completely recovered with the MCAT and CAT methods. The increase in the RMSE between the true and estimated scores was an average of 0.05 for the 120 item MCAT, and 0.07 for the 90 item MCAT compared to the Lin-MIRT method. Similarly, the increase in RMSE was 0.01 for the 180 item MCAT, and 0.12 for the 60 item MCAT method (not shown in Figure 6.1). The increase in error was larger for the CAT where the

RMSE between the true and estimated scores increased by an average of 0.03 for the 180 item CAT, 0.07 for the 120 item CAT, 0.11 for the 90 item CAT, and 0.18 for the 60 item CAT compared to the full Lin-IRT version of the NEO PI-R.

6.6.2 Study 2: The consequences of using the different scoring methods with real test responses

The second simulation study used the existing item response data collected from real examinees in the Danish NEO PI-R norm sample to investigate what the result would be if the NEO PI-R had been administered and scored by using each of the four testing methods described above in a real testing situation (research question four). The item exposure rates were calculated for the MCAT and CAT methods in order to assess the efficiency with which the NEO PI-R item pool was used. Finally, the study was used to examine if the conclusions from the first study remained accurate when real data rather than simulated data was used.

In general, the results showed that there was a high level of consistency across the different NEO PI-R factors. Therefore, Table 6.2 presents the correlations between the scoring methods averaged across all of the Big Five constructs in the NEO PI-R.

Table 6.2

Correlations between scoring methods averaged across all Big Five factors. The average correlations between true trait scores and estimated trait scores from the first simulation study are presented in the diagonal

		CAT					MCAT			
	Lin-IRT	180 items	120 items	90 items	60 items	Lin-MIRT	180 items	120 items	90 items	60 items
Lin-IRT	0.82	0.98	0.94	0.89	0.81	0.95	0.93	0.89	0.86	0.80
CAT 180		0.81	0.96	0.91	0.82	0.93	0.93	0.90	0.87	0.82
CAT 120			0.78	0.95	0.86	0.90	0.91	0.90	0.88	0.84
CAT 90				0.74	0.89	0.85	0.86	0.87	0.86	0.82
CAT 60					0.67	0.78	0.78	0.78	0.78	0.78
Lin-MIRT						0.86	0.98	0.94	0.92	0.87
MCAT 180							0.85	0.96	0.93	0.89
MCAT 120								0.82	0.96	0.92
MCAT 90									0.81	0.94
MCAT 60										0.77

The pattern of correlations are not unexpected. The Lin-IRT and Lin-MIRT methods had an average correlation of 0.95. In addition, tests of similar length correlated highly. For instance, the CAT and MCAT methods with 120 items had an average correlation of 0.90. Finally, test methods that used the same scoring methodology correlated highly (e.g., Lin-IRT or Lin-MIRT). The average correlations between the full NEO PI-R test using the Lin-IRT scoring method and shorter versions of the test using the CAT method were: 0.98 for the 180 item test, 0.94 for the 120 item test, 0.89 for the 90 item test, and 0.81 for the 60 item test. Similarly, the average correlations between the full NEO PI-R test using the Lin-MIRT scoring method and the shorter versions of the test using the MCAT method were: 0.98 for the 180 item test, 0.94 for the 120 item test, 0.92 for the 90 item test, and 0.87 for the 60 item test.

The pattern of correlations obtained in this study supports the findings obtained in the first study. That is, the testing methods that were identified as being the most accurate in the first study also had the highest correlations in the second study. Therefore, the results of the second study indicate that the conclusions based on the simulated data seem to hold when real data from actual NEO PI-R respondents are used.

The item exposure rates were also compared for the MCAT and CAT methods. This was done by calculating the proportion of respondents who were exposed to each item. Figure 6.2 groups the items based on the percent of exposures for

the MCAT and CAT with 120 items. Similar results were obtained with other test lengths. Although the figure illustrates that both methods resulted in a non-uniform use of items, the MCAT method had a slightly more effective use of the item pool compared to the CAT method. That is, 58 items in the MCAT, compared to 87 items in the CAT were not used at all. Similarly, 90 items were administered to over 90% of all test respondents in the MCAT, while this number was 100 in the CAT.

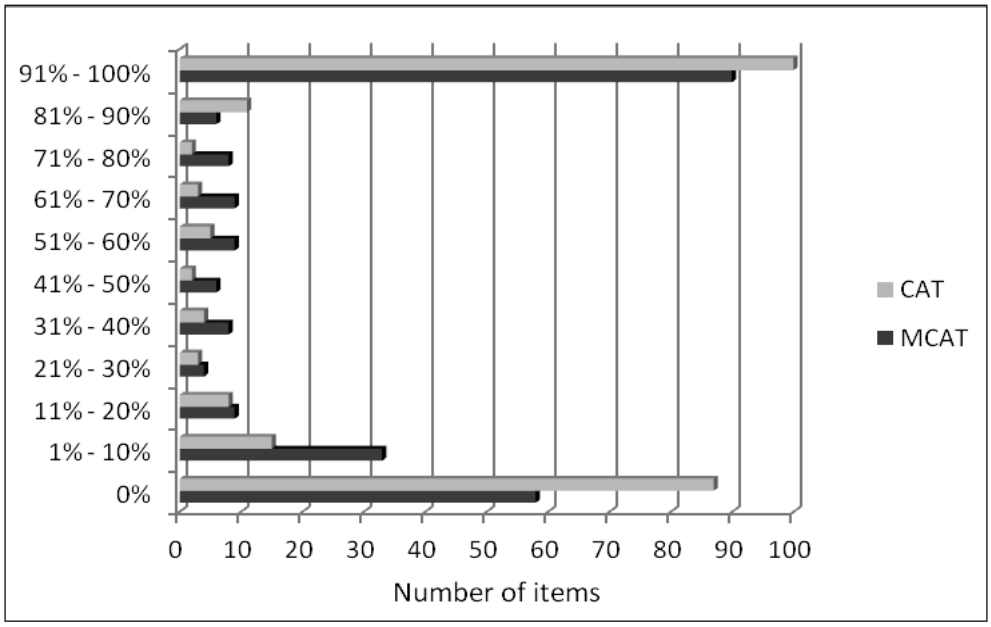

Figure 6.2: Item exposure rates for the MCAT and CAT methods with 120 items. The vertical axis shows the percent of test respondents who are exposed to the items, and the horizontal axis presents the number of items within each category.

The finding that the MCAT method resulted in a more uniform use of the item pool can be explained by the fact that more information is available for item selection in the MCAT method. This is the case because the scores from the other facets are taken into account when selecting the best item. Therefore, the item selection algorithm is more likely to select different items based on the different information that is available for each test respondent. By contrast, the unidimensional CAT starts by selecting the best item in each facet for all test respondents because there is no information available to differentiate the respondents. Even after the first item has been administered and there is information available to differentiate the test respondents, the CAT method tended to administer a similar set of items. This finding is consistent

with the results obtained by Reise and Henson (2000), and occurs because the NEO PI-R consists of polytomous items that have an information function that is spread out rather evenly over the latent trait range. Since some items have considerably higher discrimination parameters because they are highly related to the latent trait, these items are consistently selected over items that have a low discrimination parameter because they appear to be only weakly related to the latent trait.

6.7 Discussion

This article investigated the possibility of increasing the accuracy of personality facet scores in the NEO PI-R by scoring items based on the correlations between the facets in the test with multidimensional item response theory (MIRT). Furthermore, the article examined the possibility of making the NEO PI-R shorter without attenuating precision by selecting items adaptively based on the previous responses on the test and the correlation between the facets in the test, with multidimensional computerized adaptive testing (MCAT). These issues were investigated by setting up a Monte Carlo, and a real data simulation study based on data from 600 test respondents who had completed the Danish computer-based version of the NEO PI-R.

The results revealed that the MIRT and MCAT methods provide a promising alternative for administering and scoring personality tests, specifically when these include many highly correlated facets. The linear MIRT (Lin-MIRT) scoring method resulted in improved accuracy for all of the facets within the NEO PI-R compared to the unidimensional item response theory (Lin-IRT) scoring method. The results were particularly evident for the three NEO PI-R constructs that had the highest correlations between the facets: Neuroticism, Openness, and Conscientiousness. The results of the study also revealed that the NEO PI-R could be reduced to 120 items (by 50%) while slightly improving the general accuracy of the test, and further to 90 items (by 63%) with a slight loss in precision compared to the Lin-IRT method. The MCAT method also produced facet scores that were comparable in terms of accuracy to the CAT method with half as many items per facet. In addition, the real data simulation indicated that there was a relatively high correlation between the Lin-MIRT and MCAT methods and the other testing methods described in this study. For example, the Lin-MIRT testing method had an average correlation of 0.95 with the Lin-IRT method. Also the MCAT method with 120 items correlated by 0.94 with the full test scored using Lin-MIRT, and 0.89 with the full test scored using Lin-IRT. The MCAT method also resulted in a slightly better use of the item pool compared to the CAT method.

When making a decision about the appropriate length of a personality test, there is a trade-off between test accuracy and testing time. The benefits of saving time with a shorter test are difficult to quantify, because they vary greatly based on the test setting. Nevertheless, shorter tests can have several advantages for the test respondent and the organization or person administering the test. From the test respondent's perspective long tests require a high level of cognitive demand. It may be difficult to maintain a high level of motivation throughout the test which can result in careless mistakes. This may be particularly true for specific groups of respondents with clinical conditions. From the test organizations perspective additional testing time is often equivalent to increased costs. This could include the cost of administering the test, including the salary of test supervisors and the cost of maintaining an available testing location. Furthermore, test respondents are often assessed on other measures in additional to a personality test which can increase the cognitive demand of the assessment process. In some settings there is a limit to the amount of time that is available for testing, so a long personality test would mean that other important constructs are not assessed.

Before discussing the future perspectives of the MIRT and MCAT methods for personality testing, we will reflect on several practical and methodological issues and limitations involved in the current research. A consideration when deciding to score tests with MIRT is the potential impact on test structure. Is it possible that using information from all facets within a domain to inform the score on one facet will reduce the variability among the facets, which could narrow them to the domain and potentially limit criterion validity. This concern about MIRT might apply more strongly to MCAT in particular because in adaptive testing information is permitted to be lost in exchange for time. The issue is a greater concern for complex-structure MIRT models, because each item has several discrimination parameters. Consequently, an item that measures several facets simultaneously may be selected at the expense of an interstitial item that measures the facet of interest. The simple-structure MIRT model described in this article does not favor general items over interstitial items because only one discrimination parameter representing the strength of the item's relationship with the latent facet is considered. Selecting items adaptively using CAT and MCAT does, however, mean that specific information from the items that are not administered is permitted to be lost in exchange for time. Therefore, future research needs to assess the impact on criterion validity when administering and scoring personality facet scores with MCAT as well as CAT methods.

One of the advantages of a fixed test is that it can be designed optimally to ensure that the content of the construct is sufficiently covered. This is also possible with CAT and MCAT by establishing content constraints to ensure that the content of

the adaptive test meets content specifications. Veldkamp and van der Linden (2002) suggest a method using linear programming that can successfully incorporate a large number of constraints to ensure that the adaptive test lives up to a large number of practical criteria. In the context of personality testing, this would require a thorough classification of the constraints that should be considered in order to maintain the content validity of the test. The addition of content constraints would result in a decrease in the accuracy of the CAT and MCAT methods because the optimal items are no longer selected. In the present study, the MCAT method was performed without any constraints. A replication of the first simulation study was conducted to assess the impact of a simple constraint on the MCAT item selection algorithm. The constraint consisted of administering the same number of items within each facet. The result of adding the additional constraint was an average increase in RMSE of 0.02 per facet. Therefore, the result suggests that the addition of constraints does increase the error of the MCAT method; however the increase is not large enough to change the conclusions of the study. Further research could investigate the consequences of implementing more elaborate constraints.

An additional issue is the impact of the MIRT and MCAT models for test respondents who have a large variability in their facet scores within a particular construct. Preliminary results of a small number of replications (not described in this article) suggest that these respondent's facet scores are shrunk toward their mean standing on the construct. Although this is only an issue for a limited number of respondents, future research should further investigate the consequences of the MIRT and MCAT scoring methods for respondents who have atypical profiles.

A further concern with the implementation of MIRT models is the added complexity of the model compared to the more parsimonious IRT model. The added complexity means that more assumptions are necessary, which can result in the possibility of model misspecification. A general discussion of model misspecification is beyond the scope of this article. However, examples of such assumptions are the nature of the multidimensional structure and the additional response function assumptions. In the present study simple-structure was assumed. Therefore, the only additional parameters that were estimated were the correlations between the facets. These correlations are similar to the correlations reported in the existing NEO PI-R literature; therefore, it is expected that the estimates are quite robust. The complexity of the model also means that a number of choices must be made regarding the methods to be used in conducting a MCAT. In the present study a multivariate empirical normal prior was used, and the item selection algorithm used the predicted determinant of the information matrix. Other options such as Kullback-Leibler information (Mulder

& van der Linden, 2010) are available. Future research could investigate different options for conducting a MCAT for personality testing.

Another possible limitation in this study is the generalizability of the results to other samples and other personality tests. The sample consisted of a representative Danish norm sample; however, this sample may not be representative of all samples that take the NEO PI-R. Furthermore, the sample was relatively small so the item parameters could be biased. In addition, it is not certain that the results would generalize beyond the NEO PI-R. In order to investigate this issue further, we conducted a small number of replications (not described in this study) with EASI (Makransky & Kirkeby, 2010b) and found similar results. EASI is a personality typology test that is based on the Big Five model of personality. The benefits of MCAT would likely be greater than the results reported in this study with an item bank designed specifically for adaptive testing. In this study we used the existing NEO PI-R items that were designed for a fixed test. Therefore, many items were designed to differentiate between ”normal respondents”. The benefits of an adaptive test are maximized when the item bank is made up of items that differ in terms of where on the latent trait they provide the most information.

Possibly the greatest barrier to the widespread implementation of MIRT and MCAT is the lack of commercially available software to make these methods easily accessible to organizations that are convinced of their value. It is not until recently that commercial software has become available for administering a CAT (e.g., Weiss, 2008). The availability of accessible software has lead to countless CAT applications in a wide variety of fields. The requirements to apply MIRT and MCAT methods are similar to those of a CAT. Therefore, given the large amount of research interest and advantages of MIRT and MCAT, it is likely that commercial software will soon be available to implement these methods.

This brings us to the prerequisite that the MIRT and MCAT methods require computer administration. Although definitive usage data are lacking, it is becoming a standard to administer personality tests by computer. Hand held devices are beginning to be used in settings such as a doctor’s office where it has traditionally been difficult to administer tests and questionnaires by computer (Walter, 2010). Computers and the internet will provide an even bigger role in personality assessment in the future (Riese & Hanson, 2000). In fact, the administration of tests by computer has many advantages which are just beginning to be explored. One advantage that will likely increase the need for MIRT and MCAT is that a computer offers the option of administering different types of multi-media items. There are many existing tests that use complex interactive items such as video simulations. More realistic items increase the fidelity of the test, which could in turn increase the test respondent’s

motivation level. The greatest challenges in developing multi-media items include the difficulty of establishing a valid scoring system, and the higher cost associated with developing such items (Tuzinki et. al. 2011). Multi-media items challenge traditional psychometric models because it is difficult to ensure unidimensionality as the complexity of the items increases. This is the case because items often measure several important traits. In these cases MIRT and MCAT provide viable options for administering and scoring such test items.

In conclusion, the results of this article imply that MIRT and MCAT are methodological developments that provide an intriguing opportunity for the field of personality testing, in particular for personality tests that contain many facets that are highly correlated. There can be practical and statistical advantages to incorporating the theoretical model directly into the test design. This opens up an attractive new step to more theoretically-based testing that can enhance the validity of test score interpretations.

Summary

One of the fundamental goals of industrial/organizational psychology is to find the proper job candidate to match a particular job, thereby increasing the probability that they will succeed. In this process it is important to assess job candidates accurately in order to increase the quality of the selection decisions that are made. Tests are tools that are used to increase the efficiency of this assessment process. Efficiency is particularly important in industrial/organizational psychology because additional assessment time is highly related to added costs. The overarching goal of this dissertation was to increase the precision and efficiency of the measurement tools that are used to make selection decisions in industrial/organizational psychology, by introducing psychometric innovations in the framework of computerized adaptive testing (CAT).

Chapter 1 presented a general introduction to CAT and item response theory (IRT). Chapters 2 and 3 presented methodology that would increase the possibility of developing valid CAT's given the typical challenges within an organizational assessment context. In CAT, increased efficiency is obtained by administering items that match the test respondent's ability level. This can only be achieved if there is a sufficient number of accurately calibrated items that cover the entire ability range. This can be a challenge in organizational testing because there is limited access to test takers for calibration. Chapter 2 introduced a methodology that can be used to calibrate an item bank online in an operational setting. The item bank is calibrated in a situation where test takers are processed and the scores they obtain can be used. This study explored three possible automatic online calibration strategies, with the intent of calibrating items accurately while estimating ability precisely and fairly. These included: 1) A two-phase strategy in which items are calibrated at the end of each phase. 2) A multi-phase strategy in which items are administered randomly at the beginning of the test and adaptively at the end. The proportion of adaptive items increases after each phase in the calibration design. 3) A continuous updating strategy where item parameters are continuously updated. The results of the study provide viable calibration design options for test developers since these methods have the advantage that they offer the possibility of assessing test takers' ability throughout the calibration of the test, and because they can be used to calibrate items accurately with small calibration samples.

In the process of developing a CAT, several assumptions need to be empirically checked to confirm the measurement invariance (MI) of the particular instrument. The consequence of ignoring these assumptions is the possibility of systematic discrimination against members of a particular group. This can lead to possible

legal and economic consequences if the test is used for making decisions in an organizational context. Nonetheless, controlling for MI is typically not one of the first analyses performed when establishing the validity of a test (Vandenberg & Lance, 2000). Chapter 3 presented a straightforward method for conducting a test of MI and modeling differential item functioning (DIF) by assigning group-specific item parameters in the framework of IRT. Instead of eliminating items that exhibit DIF, this method provides a flexible way of investigating if the items measure the same construct across groups even though they do so in a different way. The article exemplified three applications of the method for the Master Competence Analysis (MCA), a cognitive ability test used in international organizational assessment. These examples pertained to context effects due to the test administration method, motivation effects due to the stakes of the test, and language effects. In general, the study demonstrated that MI was violated across test administration contexts, the stakes of the test, and across languages. The results showed that simply ignoring the potential presence of DIF would result in decisions that could affect a significant proportion of test respondents. Furthermore, the study provided evidence that modeling DIF with group-specific item parameters is a viable methodology for making comparisons across contexts without the need to eliminate items from a test.

Test efficiency can also be obtained by increasing the flexibility of the context in which a test is administered. The increase in computer availability and the widespread use of the Internet have led to an increased application of unproctored tests that can be administered at any place and time via the World Wide Web. Unproctored Internet testing (UIT) is Internet-based testing of a candidate without a traditional human proctor. UIT currently makes up the majority of individual employment test administrations. UIT is attractive because it limits the resources necessary for administering tests, and job candidates do not have to travel to testing locations. UIT also allows continuous access to assessments and limits the time it takes to process candidates. The primary disadvantage of UIT stems from the many forms of cheating that are possible. Chapter 4 proposed and compared two methods for detecting whether a test taker's original UIT responses are consistent with the responses from a follow-up confirmation test. The first method was a fixed length adaptive confirmation test using the likelihood ratio (LR) test to evaluate cheating and the second method was a variable length adaptive confirmation test using an extension of the stochastic curtailed truncated sequential probability ratio test (SCTSPRT) to evaluate cheating. Simulation studies indicated that the adaptive confirmation test using the SCTSPRT was almost four times shorter while maintaining the same power of detection. The study also demonstrated that cheating can have a detrimental effect on the validity of

a selection procedure, and illustrated that the use of a confirmation test can remedy the negative effect of cheating on validity.

The majority of tests used within industrial/organizational psychology assess multidimensional constructs with several correlated sub-domains. Scores on these tests are typically obtained for a single sub-domain at a time. This can be ineffective when the sub-domains assessed are highly correlated. Chapters 5 and 6 investigated methods for increasing the precision and the efficiency of a cognitive ability and a personality test by scoring items with multidimensional item response theory (MIRT) and by efficiently administering and scoring items with multidimensional computerized adaptive testing (MCAT). Chapter 5 illustrated the applicability of MCAT for the Adjustable Competence Evaluation (ACE); a computer adaptive cognitive ability test used in organizational selection. The MCAT method was compared to a unidimensional CAT method and random item administration. Results of this study indicated that the MCAT method leads to improved test precision, shorter tests, and increased selection utility compared to the other two methods. The results suggested that the benefits of using the MCAT method are particularly evident in short tests. The MCAT method was 75% shorter than the unidimensional CAT method, and over four times shorter than the random item administration method, while maintaining the same precision for the short version of ACE (12 items). For the long version of ACE (24 items), the MCAT method was 17% shorter than unidimensional CAT, and almost three times shorter than the random item administration method while maintaining the same precision. The conclusion of the study was that MCAT can be particularly useful in settings where there is a need to provide detailed feedback regarding sub-domain scores, while having limited time for assessment.

Narrowly defined personality facet scores are commonly reported and used for making decisions in clinical and organizational settings. Chapter 6 investigated the possibility of increasing the accuracy of personality facet scores in the NEO PI-R (a widely used personality test based on the Five-factor model of personality) by scoring items with MIRT. Furthermore, the article examined the possibility of making the NEO PI-R shorter without attenuating precision by selecting items adaptively based on the previous responses on the test and the correlation between the facets in the test, with MCAT. The results revealed that the MIRT and MCAT methods provide a promising alternative for administering and scoring personality tests, specifically when these include many highly correlated facets. The linear MIRT scoring method resulted in improved accuracy for all of the facets within the NEO PI-R compared to the unidimensional item response theory scoring method. The results were particularly evident for the three NEO PI-R constructs that had the highest correlations between the facets: Neuroticism, Openness, and Conscientiousness. The results of the study

also revealed that the NEO PI-R could be reduced to 120 items (by 50%) while slightly improving the general accuracy of the test, and further to 90 items (by 63%) with a slight loss in precision compared to the unidimensional IRT scoring method. The MCAT method also produced facet scores that were comparable in terms of accuracy to the unidimensional CAT method with half as many items per facet.

The chapters in this dissertation illustrate that CAT provides the possibility of increasing the precision and efficiency of assessment in industrial/organizational psychology. The various psychometric methods described increase the feasibility of developing CATs in organizational assessment, and improve the precision, efficiency and the flexibility of their administration in this field.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, C. D., Warner, J. L., & Spector, C. E. (1984). Inflation bias in self-assessment examination: Implications for valid employee selection. *Journal of Applied Psychology*, *69*, 574-580.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association. 69*. 44-49.

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored Internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 39-45.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19*, 289-303.

Automatic Data Processing Inc. (2008). *2008 ADP screening index*. Retrieved on December 21, 2009 from www.adp.com/media/press-releases/2008-news-releases/adp-annual-pre-employment-screening-index.aspx.

Barrick, M. R. & Mount M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist-practitioner model: How I-Os should respond to unproctored Internet testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 58-63.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16,* 151-155.

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, *15*, 283-306.

Berger, M. P. F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*, *57*, 521-538.

Berger, M. P. F. (1994). D-optimal designs for item response theory models. *Journal of Educational Statistics*, *19*, 43-56.

Berger, M. P. F., King, C. Y. J., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika, 65*, 377-390.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement 6,* 431-444.

Bontempo, B. D. (2011). *The future of computer-based testing is now…or at least real soon.* Paper presented at the ATP Innovations in Testing Conference, Phoenix, Arizona.

Borsboom, D. (2005). The attack of the psychometricians. *Psychometrika, 71,* 425-440.

Burke, E., Tong, A., & Liu, Y. (2011). *Partnering in meeting assessment needs in China: A case of east and west working together.* Paper presented at the ATP Innovations in Testing Conference, Phoenix, Arizona.

Burke, E., van Someren, G., & Tatham, N. (2006). *Verify range of ability tests*: *Technical Manual*. *SHL Group.*

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment, 11*, 113-120.

Choi B., Bjorner J. B., Ostergren, P. O., Clays, E., Houtman, I., Punnett, L., Rosengren, A., Bacquer, D. D., Ferrario, M., Bilau, M., & Karasek, R. (2009). Cross-Language Differential Item Functioning of the Job Content Questionnaire Among European Countries: The JACE Study. *International Journal of Behavioral Medicine*, *16*(2): 136–147.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

Costa, P. T. Jr., & McCrae, R. R. (1992*). NEO PI-R: Professional Manual.* Odessa, Florida: Psychological Assessment Resources, Inc.

Daville, C. (1993). Flow as a testing ideal. *Rasch Measurement Transactions 7:3*.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement, 32*, 355-370.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal detection of mismeasured individuals. *Applied Measurement in Education*, *9*, 47-64.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of Conscientiousness in the prediction of job performance:

Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40-57.

Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44*, 232-244.

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

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199-215.

Fallaw, S. S., Solomonson, A. L., & McClelland, L. (2009). *Current trends in assessment use: A multi-organizational survey*. Poster submitted for the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis based models. *Applied Psychological Measurement, 34*, 10-26.

Finkelman, M. (2003). *An Adaptation of Stochastic Curtailment to Truncate Wald's SPRT in Computerized Adaptive Testing* (CSE Report 606). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 31-34.

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

Gardner, W., Kelleher, K. J., & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care, 40*, 812-823.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273-294.

Glas, C. A. W. (2010). *MIRT: Multidimensional Item Response Theory*. (Computer Software). University of Twente. (http://www.utwente.nl/gw/omd/afdeling/Glas/).

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 249-263.

Glas, C. A. W., van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 303-331). New York: Springer.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In: G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Their foundations, recent developments and applications.* (pp.69-96). New York: Springer.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1975). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science*, *12*, 133–157.

Gottfredson, L. S. (2003). The challenge and promise of cognitive career assessment. *Journal of Career Assessment, 11,* 115–135.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347-360.

Haley, S. M., Pengsheng, N., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing in physical functioning using the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 87*, 1223-1229.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58-79). Washington, DC: National Academy Press.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K., Hambleton, P. F., Merenda, & C. D. Spieldberg (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & de Jong, J. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-240.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA, USA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117–144.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATE).* Washington, DC: U.S. Department of Labor, Employment Service.

Hunter, I. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

International Guidelines on Computer-Based and Internet Delivered Testing (2005). Retrieved on January 5, 2010 from www.psychtesting.org.uk/downloadfile.cfm?file_uuid=9BD783C8-1143-DFD0-7E98-798CD61E4F00&ext=pdf.

International Test Commission Guidelines for Translating and Adapting Tests (2010). Retrieved December 6, 2011, from http://www.intestcom.org/upload/sitefiles/40.pdf.

ISO International Standard for Assessment Service Delivery Procedures and Methods to Assess People in Work and Organizational Settings (2011). ISO 10667-2. Retrieved November 7, 2011 from http://www.iso.org/iso/search.htm?qt=iso+10 6672&sort=rel&type=simple&published=on.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall/CRC.

Jiao, H., Wang, S., & Lau, C. A. (2004). *An Investigation of Two Combination Procedures of SPRT for Three-category Classification Decisions in Computerized Classification Test*. Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

Jones, D. H., & Nediak, M. S. (2000). *A simulation study of optimal on-line calibration of testlets using real data* (RUTCOR research report). New Brunswick, NJ: Rutgers University, Faculty of Management and RUTCOR.

Kaminski, K. A., & Hemingway, M. A. (2009). To proctor or not to proctor? Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 24-26.

Kingsbury, G.G. (2009). *Adaptive Item Calibration: A Simple Process for Estimating Item Parameters within a Computerized Adaptive Test.* Paper presented at the 2009 GMAC conference on computerized adaptive testing, Minneapolis, Minnesota.

König, C. J., Klehe, U. C., Berchtold, M., & Kleinmann, M. (2010). Reasons for Being Selective When Choosing Personnel Selection Procedures. *International Journal of Selection and Assessment, 18*, 17-27.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology. 86*, 148–16.

La Huis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, *12*, 296-319.

Lima Passos, V., & Berger, M. P. F. (2004). Maximin calibration designs for the nominal response model: An empirical evaluation. *Applied Psychological Measurement*, *28*, 72-87.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come.* MESA Memorandum. No. 69.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence,' objectively determined and measured". *Journal of Personality and Social Psychology, 86,* 96-111.

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

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575-596). Mahwah, NJ: Lawrence Erlbaum.

Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology, 11*.

Makransky, G., & Kirkeby H. (2007). *MCA documentation manual*. Copenhagen, Denmark: Master International.

Makransky, G., & Kirkeby H. (2010a). *ACE documentation manual*. Copenhagen, Denmark: Master International.

Makransky, G., & Kirkeby, H. (2010b). *EASI documentation manual*. Copenhagen, Denmark: Master International.

Matsumoto, D., & van de Vijver, F. J. R. (2010). *Cross-cultural research methods in psychology.* New York, NY: Cambridge University Press.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable?: An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322-345.

Mulder J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback–Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77-101). New York: Springer.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3) [Computer software]. Chicago: Scientific Software.

Nye, C. D., Do, B., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16*, 112-120.

Ones, D. S. & Viswesvaran, C. (1996). 'Bandwidth-fidelity dilemma in personality measurement for personnel selection'. *Journal of Organizational Behavior, 17*, 609-626.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660-679.

Partchev, I. (2009). 3PL: A useful model with a mild estimation problem. *Measurement*, *7*, 94-96.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five Predictors of Academic Achievement. *Journal of Research in Personality, 35*, 78-90.

Paunonen, S. V., Rothstein, M. G., & Jackson, D. N. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior, 20*, 389-405.

Pearlman, K. (2009). Unproctored Internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 14-19.

Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M., & Bjorner, J. B. (2006). Multidimensional computerized adaptive testing of the EORTC

QLQ-C30: Basic developments and evolutions. *Quality of Life Research, 15*, 315-329.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

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

Reise S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, 7 (4), 347-364.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*, 643–671.

Rudner, L. (1998). *An applied study on computerized adaptive testing*. Rockland, MA: Swets & Zeitlinger.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*. 573-605.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adaptive GRE General Test* (Research Rep. No. 95-20). Princeton NJ: Educational Testing Service.

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68,* 407-415.

Schmidt, F. L., & Hunter J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274.

Schmidt, F. L., & Hunter J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162–173.

Schmidt, F. L., Hunter, I. E., McKenzie, R. C., & Muldrow, T. W. (1979). The impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64,* 609-626.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. Park Ranger for three modes of test use. *Journal of Applied Psychology, 69,* 490-497.

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

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-73). Boston: Kluwer Academic.

Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 57-76). New York: Springer.

Skovdahl Hansen, H., & Mortensen, E. L. (2003). *Documentation for the Danish version of the NEO PI-R and the NEO-FFI*. Copenhagen, Denmark: PsykologiErhverv.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computerized adaptive test*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, April 5-7, New Orleans, LA.

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

Stewart, G. L. (1999). Trait bandwidth and stages of job performance: Assessing differential effects for conscientiousness and its subtraits. *Journal of Applied Psychology, 84*, 959-968.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, *55*, 461-475.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Taris, T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology*, *132*, 301–316.

Thissen D., & Mislevy R. J. (2000). Test Algorithms. *In H. Wainer (Eds.) Computerized adaptive testing: A primer* (pp. 101-134): Hillsdale, NJ.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2,* 2-10.

Tuzinki, K., Shyamsunder, A., Sarmma C., & Hawkes, B. (2011). *3D Animation and Virtual Worlds: The Next Assessment Frontier.* Paper presented at the ATP Innovations in Testing Conference, Phoenix, Arizona.

van der Linden, W. J., & Glas, C. A. W. (2000a). *Computerized adaptive testing. Theory and practice,* Dordrecht: Kluwer Academic Publishers.

van der Linden, W. J., & Glas, C. A. W. (2000b). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35-53.

van der Linden, W, J., & Glas C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.

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

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). New York: Springer.

van der Linden, W. J., & Sotaridona, L. (2002). *A statistical test for detecting answer copying on multiple-choice tests*. University of Twente Research Report 02-04. Enschede, Netherlands.

van Groen, M. M., ten Klooster, P. M., Taal, E., van de Laar, M. A. F. J., & Glas, C. A. W. (2010). Application of the health assessment questionnaire disability index to various rheumatic diseases. *Quality of Life Research, 12*, 1255-1243.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston, MA: Kluwer-Nijhoff.

Veldkamp B, P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67,* 575-588.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In: W.J.van der Linden and R.K.Hambleton (Eds.). *Handbook of Modern Item Response Theory* (pp.123-138). New York: Springer.

Wainer, H. (2000). *Computerized adaptive testing. A primer.* Second edition. Hilsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5,* 125-146.

Walter, O. (2010). *Clinical and medical applications of computer adaptive testing*. Paper presented at the First International IACAT Conference on Computerized Adaptive Testing, Arnhem, The Netherlands.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219-238.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 450-480.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Weiss, D. J., (2008). *FastTEST Professional Testing System.* (Computer Software). Assessment Systems Corporation.

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

Weisscer, N., Glas, C. A., Vermeulen, M., & De Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: accounting for differential item functioning. *Journal of Clinical Epidemology, 63*, 543-549.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235-274.

Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology, 80,* 79–85.

Wilk, S. L., & Sackett, P. R. (1996). Longitudinal analysis of ability job complexity fit and job change. *Personnel Psychology, 49,* 937–967.

Wise, S. L., & DeMars, C. E. (2006). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest: General item response modeling software (Version 2.0) [Computer software]. Camberwell, Australia: ACER.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31,* 83-105.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zwick R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 331-354). New York, NY: Springer.

Zwick, R. & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenzel DIF analysis to computer-adaptive tests. *Applied Psychological Measurement, 26*, 57-76.

Samenvatting

Een van de belangrijke doelstellingen van de arbeids- en organisatiepsychologie (A&O psychologie) is het vinden van de juiste kandidaat voor een juiste baan. Daarbij is het belangrijk om kandidaten nauwkeurig te evalueren. Tests zijn belangrijke hulpmiddelen om de efficiëntie van dit evaluatieproces te vergroten. Efficiëntie is met name van belang in de A&O psychologie omdat de in het selectieproces geïnvesteerde tijd veel geld kost. Het algemene doel van dit proefschrift is om de nauwkeurigheid en efficiëntie van de meetinstrumenten die gebruikt worden om de selectie beslissingen te nemen in de A&O psychologie te vergroten door psychometrische innovaties zoals Computerized Adaptive Testing (CAT).

Hoofdstuk 1 presenteert een algemene inleiding in CAT en item response theorie (IRT). De hoofdstukken 2 en 3 presenteren methoden die de mogelijkheid van het ontwikkelen van een CAT ondersteunen gegeven de typische problemen binnen een organisatie assessment context. In CAT wordt een verhoogd rendement behaald door het selecteren van items uit een item pool die passen bij het niveau van de kandidaat. Dit kan alleen als er een voldoende nauwkeurig geijkte item pool beschikbaar is die het gehele vaardigheidsniveau bestrijkt. In de A&O psychologie kan een uitdaging zijn omdat de mogelijkheden voor het van te voren ijken (kalibreren) van de item pool beperkt zijn.

Hoofdstuk 2 introduceert een methodiek die kan worden gebruikt om een item pool online te kalibreren in een operationele omgeving. De item pool wordt dan gekalibreerd in een situatie waarin de test scores consequenties hebben voor de geteste personen. De studie onderzocht drie mogelijke online kalibratie strategieën, met de bedoeling de nauwkeurigheid van het kalibreren van de item pool te optimaliseren, onder de randvoorwaarde dat de test scores eerlijk en vergelijkbaar zijn. De methoden zijn: 1) Een tweefasige strategie waarbij de items worden gekalibreerd aan het iende van iedere fase. 2) Een meerfasige strategie waarbij iedere afname start willekeurig gekozen items en eindigt met een adaptieve item selectie. Het percentage adaptieve items neemt toe na elke fase van de kalibratie ontwerp. 3) Een strategie met voortdurende aanpassing waarbij de item parameters continu worden bijgewerkt. De resultaten van de studie leidden tot inzicht in haalbare kalibratie ontwerpmogelijkheden voor testontwikkelaars binnen de randvoorwaarden van hun specifieke context.

In de ontwikkeling van een CAT moeten verschillende veronderstellingen die bij item selectie en scoring worden gebruikt empirisch gecontroleerd worden. Met name de invariantie van de meting over subpopulaties (measurement invariance, MI)

van de betreffende item pool moet worden aangetoond. Het gevolg van het negeren van deze veronderstellingen is de mogelijkheid van systematische discriminatie van leden van een bepaalde groep. Dit kan leiden tot mogelijke juridische en economische gevolgen als de test wordt gebruikt voor het maken van beslissingen in een organisatorische context. Toch is het controleren van MI meestal niet een van de eerste analyses die uitgevoerd wordt bij de vaststelling van de validiteit van een test (Vandenberg & Lance, 2000). Hoofdstuk 3 presenteerde een eenvoudige methode voor het uitvoeren van een test van MI en modellering vraagonzuiverheid (differential item functioning, DIF).

In plaats van het elimineren van items die DIF vertonen, wordt voorgesteld om wordt voorgesteld om onzuivere items groep-specifieke item parameters.te geven. Deze methode biedt een flexibele manier hetzelfde construct te meten in verschillende groepen, ook al gebeurt dat op een andere manier. Het hoofdstuk geïllustreerd drie toepassingen van de methode voor de Master Competence Analysis (MCA), een cognitieve vaardigheden test die gebruikt wordt in internationale A&O assessments. De voorbeelden van de toepassing van de methoden voor het modelleren van DIF hebben betrekking op context effecten samenhangend met afnamemethode van de test, motivatie effecten als gevolg van de inzet van de test, en taal effecten. De resultaten toonden aan dat simpelweg negeren van de mogelijke aanwezigheid van DIF zou leiden tot onzuivere beslissingen voor een aanzienlijk deel van de test-respondenten. Voorts heeft de studie aangetoond dat modelleren DIF in groep-specifieke post parameters een goede methode is voor het garanderen van de vergelijkbaarheid van de test scores, zonder dat onderdelen van een test verwijderd hoeven te worden.

Test efficiëntie kan ook worden bevorderd door het verhogen van de flexibiliteit van de context waarin een test wordt afgenomen. De toename van computer beschikbaarheid en het wijdverbreide gebruik van het internet hebben geleid tot een grotere toepassing van niet-gesuperviseerde (unproctored) tests die kunnen worden afgenomen op elke plaats en tijd via het World Wide Web. Unproctored Internet testen (UIT) is via het internet testen van een kandidaat zonder een traditionele menselijke surveillant. Op dit ogenblik wordt de meerderheid van de individuele A&O tests afgenomen via UIT. UIT is aantrekkelijk omdat het de middelen die nodig zijn voor het beheer van tests beperkt, en omdat sollicitanten niet hoeven te reizen naar test locaties. UIT biedt ook continue toegang tot assessments en beperkt de tijd die het kost om kandidaten te verwerken. Het voornaamste nadeel van UIT zijn de vele vormen van bedrog die mogelijk zijn. In Hoofdstuk 4 worden twee methoden voor het detecteren fraude gepresenteerd. Het gaat hierbij om methodes om na te gaan of de item responsies in de UIT situatie in overeenstemming met de responsies op een follow-up test. De eerste methode is adaptieve follow-up test met een vaste lengte

in combinatie met een likelihood ratio test (LR) test en de tweede methode is een adaptieve follow-up test waarbij de lengte wordt bepaald door de stochastic curtailed truncated sequential probability ratio test (SCTSPRT). Simulatie studies gaven aan dat de adaptieve bevestiging test met de SCTSPRT was bijna vier maal korter was met behoud van dezelfde detectie kracht.

De meerderheid van de tests die worden gebruikt binnen de A&O psychologie beoordelen multidimensionale constructen met meerdere gecorreleerde sub-domeinen. Scores op deze test worden meestal berekend per sub-domein. Dit kan niet optimaal effectief zijn wanneer de sub-domeinen sterk gecorreleerd zijn. Hoofdstukken 5 en 6 onderzochten methoden voor het verhogen van de precisie en de efficiëntie van een cognitief vermogen test en een persoonlijkheidstest met het toepassen van multidimensionale item response theorie (MIRT) en door het afnemen van tests via multidimensionaal computer adaptief toetsen (MCAT). Hoofdstuk 5 geïllustreerd de toepasbaarheid van MCAT voor de Adjustable Competence Evaluation (ACE), een computer adaptieve cognitieve vaardigheden test gebruikt in A&O-selectie. De MCAT methode werd vergeleken met een eendimensionale CAT methode en willekeurig item selectie. De resultaten van dit onderzoek tonen aan dat de MCAT methode leidt tot verbeterde test nauwkeurigheid, kortere tests en verhoogde selectie utiliteit ten opzichte van de andere twee. De resultaten suggereerden dat de voordelen van het gebruik van de MCAT methode vooral duidelijk zijn in korte tests. De MCAT methode was 75% korter dan de unidimensionele CAT-methode, en meer dan vier keer korter dan de willekeurig item administratie-methode, met behoud van dezelfde precisie voor de korte versie van ACE (12 items). Voor de lange versie van ACE (24 items), was de MCAT methode 17% korter dan unidimensionele CAT, en bijna drie keer korter dan de willekeurig item selectie methode met behoud van dezelfde nauwkeurigheid. De conclusie van de studie was dat MCAT kan bijzonder nuttig zijn in situaties waar er behoefte is om gedetailleerde feedback te geven met betrekking tot sub-domein scores.

Smal gedefinieerde persoonlijkheidsfacet scores worden vaak gebruikt voor het nemen van beslissingen in klinische en organisatorische instellingen. Hoofdstuk 6 onderzocht de mogelijkheid van het verhogen van de nauwkeurigheid van de persoonlijkheid facet scores op de NEO PI-R (een veel gebruikte persoonlijkheidstest gebaseerd op de vijf-factor model van persoonlijkheid) door de test te scoren met MIRT. Verder is de mogelijkheid onderzocht voor het korten maken van de NEO PI-R zonder de precisie te verminderen, door het selecteren adaptief van items op basis van de eerdere antwoorden op de test en de correlatie tussen de facetten in de test met MCAT. Uit de resultaten bleek dat de MIRT en MCAT methoden een veelbelovend alternatief voor het beheer en het scoren van persoonlijkheidstests zijn,

in het bijzonder wanneer deze een groot aantal sterk gecorreleerd facetten bevat. De lineaire MIRT scoremethode resulteerde in een betere nauwkeurigheid alle facetten op de NEO PI-R in vergelijking met de eendimensionale IRT scoringsethoden. De resultaten waren vooral duidelijk voor de drie NEO PI-R constructies met de hoogste correlaties tussen de facetten: Neuroticisme, Openheid en Zorgvuldigheid. Uit de resultaten van het onderzoek is ook gebleken dat het NEO PI-R kan worden verlaagd tot 120 punten (50%), met enigszins betere algemene nauwkeurigheid van de test en na 90 items (met 63%) met een gering verlies in nauwkeurigheid opzichte van de eendimensionale IRT scoringsmethode. De MCAT methode produceerde ook facet scores die wat betreft nauwkeurigheid vergelijkbaar waren van de unidimensionele CAT methode met half zoveel items per facet.

De hoofdstukken in dit proefschrift tonen aan dat CAT de mogelijkheid biedt tot het verhogen van de precisie en efficiëntie van de assessments in de A&O psychologie. De verschillende beschreven psychometrische methoden verhogen de haalbaarheid van de ontwikkeling van CAT in A&O assessment en de haalbaarheid van het verbeteren van de precisie, efficiëntie en de flexibiliteit van toetsen op dit gebied.

Acknowledgements

In 2007 Master International A/S in cooperation with the Danish Ministry of Technology Science and Innovation agreed to sponsor my industrial Ph.D. at the University of Twente. The objective of the industrial Ph.D. program was to bridge the academic-practitioner gap in order to ensure that academic research has practical value, and to guarantee that Danish companies are on the cutting edge of innovation. The requirement is that the Ph.D. student must divide their time between work tasks and academic research. This complicated international cooperation between several parties required a large amount of logistic organization and flexibility from all of the parties involved, and from the people closest to me.

The presentation of this dissertation is an indication that an important milestone in my scientific journey has been reached, and is evidence that this cooperation has been successful. This milestone would not have been achieved without a great deal of flexibility and support shown by all of those around me. This is a great opportunity to look back on my achievement and to express my gratitude to all those people who guided, supported and stayed beside me during my Ph.D. study.

First of all, I would like to thank my colleagues at Master. My direct supervisor Louise Bergøe has been extremely helpful and understanding. She has helped organize my work tasks so that I could have the time and energy to work on my Ph.D. project. I would also like to thank the management including Thorkild Eriksen, Norbert Mörtl as well as Louise who have always supported me and have been willing to implement the outcomes of the Ph.D. project in the company's solutions. I have appreciated the opportunity to implement and watch the ideas from my project flourish into applied solutions. I would also like to thank my psychology colleagues Henriette Kirkeby, Erik Lyon, Mads Rung and Louise for showing interest in my project and reading through different drafts along the way, and Dorthe Petersen for helping me format the dissertation.

I would like to extend the greatest thanks to my Ph.D. supervisor Cees Glas. It has been a great privilege to be under the guidance of such a knowledgeable and inspiring mentor. His unbelievable ability to have an elegant solution to any scientific problem is inspiring. His enthusiasm and insight throughout the research, has inspired me to push myself and has helped me grow greatly throughout the project. I would also like to thank my Danish supervisor Svend Kreiner for his support and guidance on psychometric issues throughout the project. I also appreciate Bernard Veldkamp, Hanneke Geerlings, Rinke Klein Entik, Anke Weekers, Khurrem Jahangir, Iris Egberk, Britt Qiwei He as well as Birgit Olthof-Regeling and Lorette Bosch-Padberg

among others from the OMD department for helping and for making me feel at home in Twente.

Most importantly I would like to extend my heartfelt gratitude to my closest family. My wife Sidse Zimmermann for her constant love and support. Her caring motivation has carried me through the challenges that have regularly occurred throughout the project. I would also like to thank my son Elias. His joy and affection have added new meaning to my life which has helped me stay positive and push forward. My gratitude also goes out to my mother Nancy Makransky for always loving and believing in me, and my father Jerry Makransky for his support and interest in my project and for his comments and corrections on this dissertation.