COMPUTERIZED ADAPTIVE TESTING IN KINANTHROPOLOGY: MONTE CARLO SIMULATIONS USING THE PHYSICAL SELF DESCRIPTION QUESTIONNAIRE

A Dissertation Submitted to the Faculty of Physical Education and Sport

Charles University

by

Martin Komarc

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Supervised by
Jan Štochl

Prague, Czech Republic
March 2017

DECLARATION

I hereby declare that this dissertation exclusively represents my own original work, conducted under the supervision of Jan Štochl. All bibliographic sources information is cited in the references. This dissertation (or any of its part) was not used to obtain any other or similar degree.

Prague, 23.3.2017 ……………………………….

Martin Komarc

ACKNOWLEDGEMENTS

I would like to thank the people who provided me with their unrequited support during my Ph.D. doctoral studies and likewise for their support during preparation of the present thesis.

First and foremost, I must acknowledge my former supervisor Petr Blahuš, THE PERSON, who guided me in the correct direction at the very beginning of my career when I first set out on this journey. Petr also played an instrumental role in helping mold and further shape my critical reasoning skills. I will always remember all of what you did for me Petr – may you forever rest in peace!

I also would like to thank my doctoral thesis supervisor, Jan Štochl, for the opportunity to work on the interesting topic of Computerized Adaptive Testing and also for sharing his tremendous expertise with me. Besides many other things, Jan also gave me an opportunity to spend an inspiring time working and learning at the York Centre for Complex System Analysis, for which I am very grateful.

A BIG THANK YOU belongs to Lawrence M. Scheier from the LARS Research Institute, USA for his tireless support and encouragement so that I could finish this work. Larry your contribution to this dissertation, which took the form of motivating me and editing my English composition and grammar, was indeed crucial for its completion.

My further thanks go to Knut A. Hagtvet, whose excellent classes in generalizability theory and structural equation modeling significantly shaped my understanding of many fundamental psychometric concepts. Many thanks to Knut also for the possibility to arrange a research stay at the University of Oslo, Department of Psychology, which I thoroughly enjoyed.

Many other people deserve my acknowledgment for their contribution to my professional growth. Namely, I would like to thank Ondřej Pecha and Ladislav Čepička for their friendly and skillful guidance during my Ph.D. studies.

Finally I must express my deepest and warmest gratitude to my family and to my dear Ivana for their endless love, unfailing support, and continuous encouragement throughout my years of graduate study. Words cannot really convey the deep appreciation and heartfelt sincerity I have for your support. Thank you.

This study was supported by the following projects and grants: PRVOUK 39, PROGRES Q19, SVV 2016-260346, SVV 2017-260446, GAUK 962214 and GAUK 110217.

DEDICATION

I would like to dedicate this dissertation to my parents, Emília and Andrej, for their love, unlimited support and help throughout my life.

AGREEMENT

I agree that the Faculty of Physical Education and Sport library may lend or copy this dissertation upon request. The users should be indicated in the evidence list below.

Evidence list

First and last name	Faculty/department	Date	Signature

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis aims to introduce the use of computerized adaptive testing (CAT) – a novel and ever increasingly used method of a test administration – applied to the field of Kinanthropology. By adapting a test to an individual respondent's latent trait level, computerized adaptive testing offers numerous theoretical and methodological improvements that can significantly advance testing procedures.

In the first part of the thesis, the theoretical and conceptual basis of CAT, as well as a brief overview of its historical origins and basic general principles are presented. The discussion necessarily includes the description of Item Response Theory (IRT) to some extent, since IRT is almost exclusively used as the mathematical model in today's CAT applications. Practical application of CAT is then evaluated using Monte-Carlo simulations involving adaptive administration of the Physical Self-Description Questionnaire (PSDQ) (Marsh, Richards, Johnson, Roche, & Tremayne, 1994) – an instrument widely used to assess physical self-concept in the field of sport and exercise psychology. The Monte Carlo simulation of the PSDQ adaptive administration utilized a real item pool (N = 70) calibrated with a Graded Response Model (GRM, see Samejima, 1969, 1997). The responses to test items were generated based on item parameters and pre-specified true latent values of physical self-concept. The Monte Carlo simulation was designed to compare the number of administered items from PSDQ (test length) and accuracy of estimated latent levels of physical self-concept. The simulations also allowed for a comparison of different latent trait estimation methods, item selection procedures and other frequently used settings within current CAT research.

Results have shown that CAT can successfully be applied as a method of reducing test length when measuring physical self-concept using the PSDQ items. In particular, adaptive administration saved on average between 50% and 90% of the PSDQ items (depending on required measurement precision), while the obtained latent trait estimates were relatively unbiased and very similar to those based on the full version of the PSDQ. The maximum likelihood latent trait estimation, expected a posteriori estimation with uniform prior, and expected a posteriori estimation with standard normal prior distribution were similarly efficient with regard to the average number of administered items in PSDQ CAT. Similarly, comparison of Kullback-Leibler divergence-based and Fisher information-based item selection methods respectively, revealed almost identical results with regard to both the test length and bias of the latent trait estimates.

A possible limitation of this study is that the present findings are based exclusively on Monte-Carlo simulations. Real world applications of PSDQ CAT may however produce slight model and parameter deviations that do not completely conform to simulation findings. Notwithstanding, the promising findings from the simulation suggest a next step would entail evaluating the utility and precision of the PSDQ CAT administration in real testing conditions.

Key words: computerized adaptive testing, item response theory, self-concept, test administration

1 BRIEF INTRODUCTION TO MEASUREMENT (IN KINANTHROPOLOGY)

Mankind has always ventured to count and assign numbers to things. As part of organizing the world, we want to know "how much is out there and in what quantities do things exist?" Even counting how much fruit a tree bears, or ripened berries that fall to the ground involves developing an assignment scheme that utilizes collecting, counting, sorting, assigning and categorizing. It seems to be an integral part of our existence to assign numbers to observations according to some established set of rules; rules and procedures that are in today's world termed 'measurement' (Wood, 2006). The intent of measurement is to obtain information about particular characteristics, qualities or attributes of an object, and this process very much lies at the heart of every scientific inquiry. The processes and procedures that underlie measurement, and more formally testing generally involves assessing well-known attributes of objects – directly observable physical quantities such as time, weight, length as well as other non-physical attributes (e.g., how many numbers a person can memorize).

While our preoccupation with counting and measurement fulfills some aspect of our need to know about the observable world we inhabit, it is very often the case in the social and behavioral sciences that the attributes of interest we wish to measure are not directly observable. Many attributes, like a person's intelligence, test anxiety, well-being, motor abilities, are not observable but must be inferred. In essence, we can't touch or see these attributes, but rather infer them from observed patterns or sequences in behavior. These attributes are referred to as theoretical concepts (Bentler, 1978; Blahuš, 1985), given their abstract and ephemeral nature outside of the immediate and observable world. Given the unobservable nature of theoretical concepts researchers use specific, concrete and partial counterparts, so called empirical (observed) indicators, that are presumed to represent the abstract and generic theoretical concept of interest.

Unfortunately, by their very nature, empirical indicators are flawed and error prone. This is partly because they reflect the real world, which is "interpreted through our senses" and thus can never be known precisely (Popper, 2002). Observed indicators are also flawed given the uncertainty of measurement, which can never be perfectly precise. To provide a shared or consensual understanding of theoretical concepts they are linked to observable indicators by an operational definition (Bridgman, 1959); one that specifies variables defining the latent construct of interest. For example, researchers studying Kinanthropology might be interested in measuring "attitudes towards school physical education" with the goal of using knowledge of these attitudes to promote greater involvement by students in sports. As a result, a researcher might develop several true/false questionnaire items, that are presumed to reflect attitudes towards school physical education (e.g., "If for any reason a few subject areas have to be dropped from the school program, physical education should be one of the subjects dropped"). The skillfully chosen function of empirical indicators, questionnaire items in this case (e.g., sum of the total true responses), is then referred to as a 'test score' in the psychometric literature and is supposed to represent a quantifiable measure of the individual's "attitudes towards school physical education".

The process of concept formation, which according to Blahuš (1996) utilizes a form of so-called "weak associative measurement," raises several interesting questions. A researcher or a practitioner might wonder, for example, whether based on the administration of a set of questionnaire items it is reasonable to create a single general score that accurately assesses a person's "attitudes towards the school physical education". Additional questions that arise from this line of reasoning include: Are all the items equally good measures of the attitudes in question or are some items better than others? In the case of a single general score, how accurate is the resulting composite as a measure of attitudes? The last concern can also be expressed in terms of sufficiency, for instance, whether 20 items provide sufficient information to determine an individual's attitudes toward physical education. Furthermore, if 20 items are deemed insufficient, how many more items should be used? If large numbers of items must be used, we can pose the question whether two tests can be constructed as 'parallel forms', each form containing different items (McDonald, 1999)?

Interpreting the test scores (numbers produced by each of the research participants, students, or patients when they took a test) without answering the questions posed above may, according to Wood (2006), lead to incorrect conclusions regarding research hypothesis and/or practical recommendations (to clients/patients). These and similar questions are closely related to the two major problems of measurement and testing in behavioral and social sciences: reliability and validity of a test score. Validity "refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (Wainer, 2000, p. 16). Reliability, on the other hand, refers to the degree to which a test score, as a representation of the attribute or characteristic being assessed, is free from error (i.e. the accuracy of the measure).

The collection of techniques and statistical methods for evaluating the development and uses of a test is referred to as test theory in the literature (Embretson & Reise, 2000; McDonald, 1999; Zhu, 2006). The next section briefly mentions several of the key developments in the history of test theory, many of which still have practical implications in the behavioral and social sciences including the field of Kinanthropology.

2 HISTORICAL PATHS TO MODERN TEST THEORY

The history of measurement and testing in the behavioral and social sciences reflects several conceptual frameworks (classical test theory – CTT, item response theory – IRT) and empirical approaches (e.g., split-half reliability, internal consistency, factor analysis, …) used to formalize and rigorously test validity and reliability, two important psychometric benchmarks. Impetus for these approaches was mainly motivated by psychological research. Historically, psychologists tried to incorporate statistical methods that would assist them in solving their specific research questions (i.e., are two constructs related?). In the long run, these efforts combined with improved research designs provided impetus for the mathematical treatment of these problems on a more sophisticated basis (McDonald, 1999). Without the demand for statistical treatment of these important scientific questions, the development of statistical methods including correlation, linear regression analysis, and even factor analysis might certainly have been delayed (Blahuš, 2010).

In 1904 Charles Spearman, a student of William Wundt, published two seminal articles in the American Journal of Psychology, both of which provided a fundamental basis for the creation of psychometric theory. In the first of these articles titled "'General Intelligence' Objectively Determined and Measured," Spearman (1904a) demonstrated that cognitive performance is generated by a single, unitary quality – or what was then termed 'general intelligence'. Spearman proposed that a general factor of intelligence, which he labeled 'g', could be obtained from using a new statistical technique – factor analysis. Using factor analysis, a data summarization technique, Spearman showed that scores on all mental tests are positively correlated, and this positive association provided empirical evidence of a credible underlying "trait" of intelligence.

In a second article (The Proof and Measurement of Association between Two Things) Spearman (1904b) introduced the psychometric concept of reliability, providing a mathematical formula for estimating a test's precision or otherwise its accuracy in assessing a theoretical attribute (McDonald, 1999). Spearman argued that the observed test score is a composite of two independent components; the "true value" of the theoretical attribute/concept and a second component reflecting measurement error. By introducing both factor analysis and the concept of reliability, Spearman is generally considered as a father of CTT, a conceptual cornerstone of psychological testing and a theory that has stood the test of time through the 21st century.

The primary entity within CTT is a fixed test or some type of assessment protocol (e.g., questionnaire, test battery, etc.), usually consisting of several empirical indicators (e.g., survey or questionnaire items), which collectively provide a test score (e.g., total number of true responses that are correctly answered as "true" or false responses that are correctly answered as "false"). One of the most important features of empirical indicators in a test, what is called item difficulty is conceptualized within CTT as the probability, that a randomly selected examinee from the population of interest provides the keyed response (McDonald, 1999). For true/false (pass/fail) scored indicators/items, the relative frequency of true responses (passes) in a sufficiently large random sample from the population, is used as the estimate for an item difficulty.

Although CTT has been popular in test construction, particularly in the social and behavioral sciences, the theory contains several shortcomings (Gulliksen, 1950; Lord & Novick, 1968). One drawback of CTT is that test score reliability and item difficulties are population dependent. For example the relative frequency of true responses (passes) for a questionnaire item assessing frequency of hallucination (e.g., "I often experience hallucinations") would be much lower in the general population compared to a clinical sample of diagnosed schizophrenics (McDonald, 1999). Reliability of a test score, as another example, is higher in a heterogeneous population compared to homogenous population when using the same test (Thissen, 2000). This population/sample dependence that exist in CTT requires that new validity and reliability information is collected with each new population intended for a specific test's use (Wood, 2006). Emphasis of CTT on a test as a whole has shown to also be a drawback, since characteristics of the empirical indicators in a test (e.g., questionnaire items) are valid and interpretable only within the specified context for the particular test. Item difficulty, for example, cannot be considered outside of the particular test in which the items were administered – that is items are inseparable from the test (Verschoor, 2007). Moreover by using different scales for the item's and examinee's characteristics (e.g., probabilities for item difficulties vs. sum of the passes for examinees) respectively, CTT does not provide a means to make rigorous and methodologically sound conclusions about an individual's performance on the particular item. CTT also assumes that measurement error is distributed uniformly across the whole range of a test score, which is often unrealistic in practical applications of measurement in social and behavioral sciences (Embretson & Reise, 2000; Zhu, 2006).

3 GROUNDWORK FOR ITEM RESPONSE THEORY

The concerns outlined above with CTT sparked development of modern test theory, which according to Hambelton, van der Linden, and Wells (2010) consisted of a series of refinements in the underlying statistical proofs introduced by Lord's (1952, 1953) seminal publications. In these works, Lord introduced a theory to account for a test score which linked item responses to the underlying latent trait measured by the test. Work by the Danish mathematician George Rasch (1960) was also considered instrumental in the development of the modern test theory, and led to many advances in measurement theory and practice. It was, however, Lord and Novick's "Statistical theories of mental test scores" (1968), which is regarded by many as the real turning point in the transition from classic to modern test theory, the latter which is most commonly referred to as item response theory (IRT) today (Embretson & Reise, 2000; van der Linden & Glas, 2010; Wainer et al., 2000). Lord and Novick's (1968) book introduced, among other things, the work of Allan Birnbaum, who provided the statistical foundations for IRT based on his seminal work with the likelihood principle (Birnbaum, 1968).

Development of IRT was perhaps slowed by its computational complexity, which has been greatly facilitated by the increased computational capacity and speed of modern computers. The advent of powerful and relatively inexpensive computers introduced in the 1980s paved the way for IRT to be "the most dominant theory for test construction in all major testing organizations or agencies such as the Educational Testing Service (ETS) and American College Testing (ACT)" (Zhu, 2006, p. 53). The first systematic treatment of IRT in Kinanthropology is generally credited to Spray

(1987), who introduced its advantages and described its practical applications in the measurement of psychomotor behavior. Since this introduction, many successful applications in Kinanthropology have followed (see Wood & Zhu, 2006 for review). In the past few years, researchers in the Czech Republic have used IRT successfully to address several kinanthropological research questions (e.g., Čepička, 2004; Štochl, 2008, 2012); questions that would be difficult – if not impossible – to answer within the CTT framework.

Application of IRT offers several advantages over CTT. One advantage is that IRT employs a common logit scale for both test items characteristics (such as difficulty) and individual's level of the theoretical attribute being measured (often called ability or latent trait level in IRT). Therefore, researchers are able to conclude that when the latent trait level of an individual is higher than the difficulty of the particular item, the "person is more likely than not to provide a trait-indicating (positive, or true) response" (Nering & Ostini, 2010, p. 1). Another unique feature of IRT is that measurement error (the lack of precision in identifying a person's latent trait using the particular item) is conditionally dependent on a latent trait level of the examinee (Lord, 1952). This can be useful, for example, in mastery testing when a test developer wants to improve measurement precision for test takers at a certain latent trait level. Moreover, item characteristics in IRT are not affected by a particular sample used to obtain these characteristics (De Champlain, 2010), and likewise, individual latent trait estimates are not affected by particular items used to estimate them (Zhu, 2006). This item/latent trait invariance property in combination with the IRT's focus on the items rather than a test as a whole (Lord, 1953), enables a researcher to rank individuals on the same theoretical continuum (i.e., assessing some underlying trait or ability) even if they have been presented different set of items/indicators from a larger pool designed to measure the theoretical construct (latent trait) of interest. As Wainer (2000, p. 9) suggests, in all practical terms, this means the test developer does not have to "present all items to all individuals, only enough items to allow us to accurately situate an examinee on the latent continuum." Using this IRT approach, a tester can create a reliable test customized to each examinee. Customizing a test to examinee's trait level – or what is termed "adaptive testing" cannot be easily accomplished within the CTT framework but is a natural extension of using IRT.

4 ITEM RESPONSE THEORY (IRT)

4.1 Introduction

Item response theory (IRT), also known as latent trait theory or item characteristic curve theory (Hambelton et al., 2010), posits that the probability of a particular response to an item (or generally to any type of empirical indicator – such as questions in a survey questionnaire, tests of motor ability or measures of aptitude) is a mathematical function of the item properties (e.g., item difficulty) and an individual's level of the latent trait to be measured. IRT models can be categorized based on several different features (Thissen & Steinberg, 1986). One of the most common distinctions is whether they are uni- or multidimensional. In unidimensional models, responses to items are assumed to be accounted for by a single latent variable; that is, all items measure the same underlying theoretical construct – latent trait (Sijtsma & Molenaar, 2002). Items within a test may, however, capture several different, but possibly related latent traits. It is possible in such a case that different latent traits are measured by independent (non-overlapping) sets of items – a situation referred to as between-item multidimensionality. A common practice then is to apply unidimensional IRT models for each independent cluster of items separately. Within-item multidimensionality, on the other hand, occurs when more than one latent trait or ability underlie a response to a particular item within a test. Multidimensional IRT models (Mulder & van der Linden, 2010) are well suited to deal effectively with within-item multidimensionality. Given the focus of the empirical portion of this thesis only unidimensional models will be considered in subsequent passages.

Another frequently discussed classification distinguishes dichotomous and polytomous IRT models, respectively. Dichotomous IRT models were developed for test items with only two possible response outcomes – (binary-scored) items coded for example: correct/incorrect, true/false, yes/no, apply/not apply, etc. Increased use of polytomously scored items – items with more than two response alternatives (i.e., Likert-type items, multiple choice items when each category is scored separately) – led to the development of polytomous IRT models (see Nerning & Ostini, 2010; Ostini & Nerning, 2006). According to Ostini and Nerning (2006), advantages of modeling polytomous items is that "they are able to provide more information over wider range of the trait continuum than are dichotomous items." Nevertheless, dichotomous IRT models are still widely used and are considered a foundation for models used even to fit polytomously scored data. Although polytomous IRT models will be used later in this study, basic dichotomous IRT models are introduced briefly in the following section.

4.2 Unidimensional dichotomous IRT models

As already noted above IRT models yield the probability of a particular response to an item as a function of examinee's latent trait level and item properties, respectively. In the case of dichotomous items the simplest IRT model defines this probability as a logistic function:

$$P(\theta_j) = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)} \quad (1)$$

Here $P(\theta_j)$ indicates the probability of examinee j with latent trait θ responding to a keyed (correct, or trait indicating) category of item i with difficulty b. This model was first introduced by Danish mathematician George Rasch (1960) and is commonly referred to as dichotomous Rasch model or one-parameter logistic (1-PL) model in the literature (Embretson & Reise, 2000). The θ parameter of examinee j is theoretically unrestricted, and is quite similar to the well-known z-score, scaled to mean of 0 and standard deviation of 1 (it usually ranges from -3 to 3 in typical population)[1]. Larger values of θ indicate higher latent trait levels. Individuals with higher values of the latent trait are more likely to get the item correct, or generally to give a positive (keyed, trait indicating) response to a test item. The reference to one-parameter in the 1-PL model refers to the probability of a keyed response depending on one item parameter only – namely the item difficulty b. The difficulty parameter is expressed in the same metric as the examinee's latent trait parameter θ, and it is defined as the θ value at which there is a 50% probability to answer the item positively (correct or trait indicating answer). Values for b parameter also typically range from -3 to 3 where higher values indicate more difficult items. If an examinee's θ parameter is higher than the item difficulty, the examinee has more than 50% probability that he or she will respond to a keyed, latent

[1] It should be recalled, however, that the logit distribution is not identical to the standard normal distribution. There are 99.7% of observations within the 3 standard deviations around the mean in the standard normal distribution, whereas 90.5% of observations fall into the same interval around the mean in the logit distribution.

trait indicating response category for that item. The opposite is true when trait level of examinee is below the item difficulty.

Using Equation 1 it is possible to plot the relation between an item's difficulty level and the examinee's response, producing a typical S-shaped logistic function. In IRT this is termed the item characteristic curve (ICC), also known as trace line or item response function (Ostini & Nering, 2006; Zhu, 2006). In Figure 1 three ICC's for three items with different difficulty parameters (b_1 = -1; b_2 = 0; b_3 = 1) are graphically depicted.

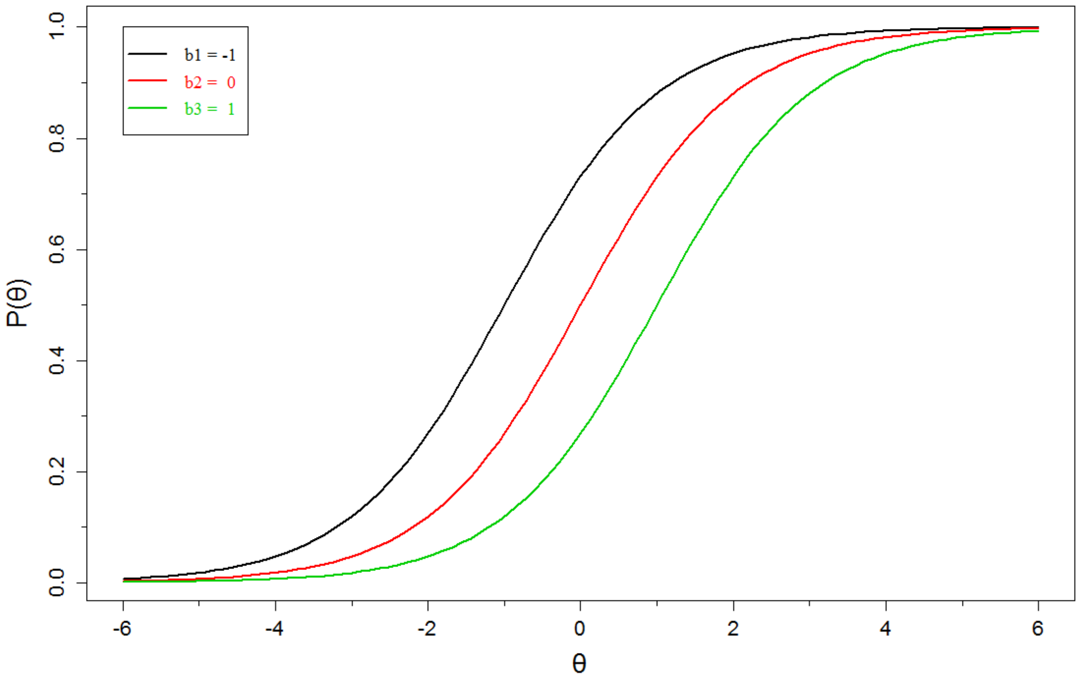


Figure 1 – Illustration of ICCs for 1-PL model

An important feature of the Rasch model is that all ICC's are parallel to each other. Practically speaking, this means that the model assumes the same relation between the item score and latent trait for all items (Štochl, 2008). A unique property of the Rasch model is so-called specific objectivity (Masters, 2010; McDonald, 1999). When a set of items corresponds to the model than "specific objectivity" means that comparison of examinees' latent trait levels is objective because it is independent of the items used for the comparison (Rasch 1960). Similarly, in the Rasch model it is possible to obtain a comparison of the item difficulty parameters, which is independent of the particular group of examinees used in the comparison (Masters, 2010). Another very appealing attribute of the Rasch model is that the examinee's observed sum score is a sufficient statistic for the latent trait estimation (McDonald, 1999). Since the observed sum score does not involve any item characteristics, the actual response pattern (e.g., which particular items were answered correctly) may be ignored in the latent trait estimation (de Ayala, 2009).

Even though the Rasch model has been demonstrated to have many attractive features, the parallel ICC's assumption may contribute to model misfit (lack of congruence between the model and the actually observed data) (Wainer & Mislevy, 2000). If this is the case it is possible to generalize the model by adding another parameter to the 1-PL model and thus allowing the items ICC's to have different steepness:

$$P(\theta_j) = \frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]} \quad (2)$$

In this model a denotes so-called discrimination parameter of item i, which defines the slope of the item's ICC. The inclusion of the a parameter converts the 1-PL IRT model into a two-parameter logistic (2-PL) model. By allowing items to have different discrimination 2-PL model broadens the applicability of IRT (Wainer & Mislevy, 2000). This is mainly caused by the fact that the parallel ICC assumption is unrealistic in many testing situations. According to Hambelton et al. (2010) the a parameter for psychometrically sound items typically ranges from 0.5 to 2.5.

Figure 2 portrays three items with the same b (b = 0) difficulty but different a discrimination parameters (a_1 = 0.5; a_2 = 1; a_3 = 2.5). The steeper the ICC, the better the item discriminates between examinees along the latent trait continuum. In other words, higher a parameter leads to a faster increase in the probability of the latent trait indicating item category response with increasing θ value (McDonald, 1985).

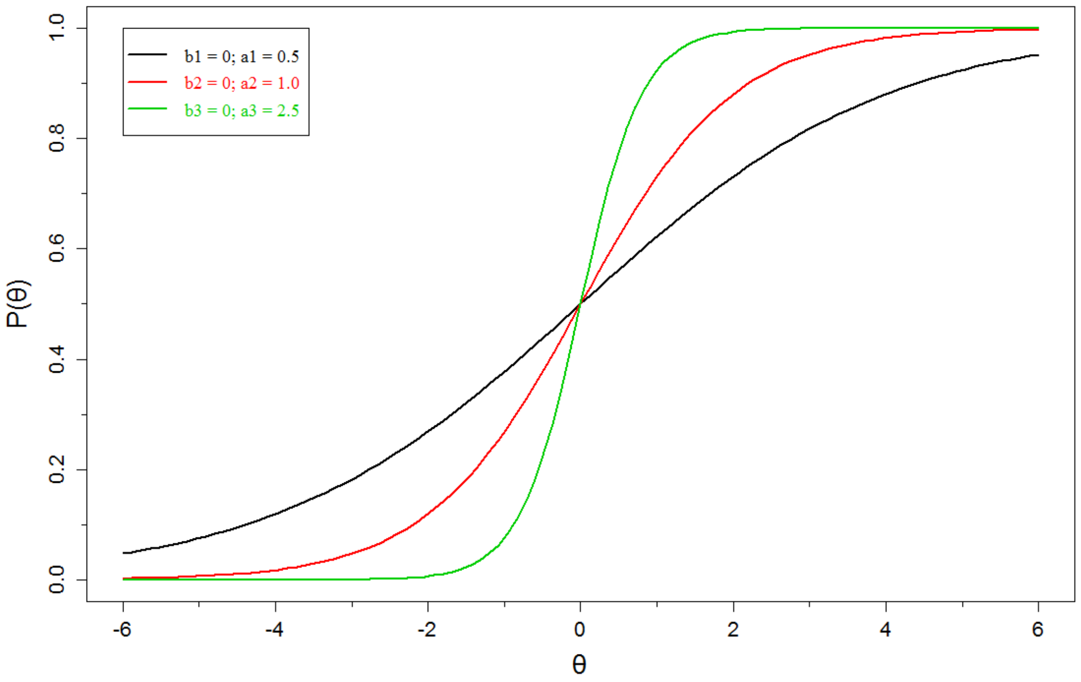


Figure 2 – Illustration of ICCs for 2-PL model

It is conceivable in various testing situations, for example in cognitive testing, that an examinee can get a difficult item correct even if his or her ability level is very low (i.e., their ability dictates they should not be able to answer the question correctly). This very feature can account for ‘guessing’ the correct answer. This case is especially relevant in educational testing where a list of item responses (of which only one is correct) is presented to a test-taker (so-called multiple choice items). Guessing behavior can also be modeled by adding another item parameter to the model in (2) – namely the guessing parameter *c*:

$$P(\theta_j) = 1 + (1 - c_i)\frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]} \quad (3)$$

The addition of a third parameter *c* results in a three-parameter logistic (3-PL) model (Birnbaum, 1968), where *c* dictates the lower asymptote of the item’s ICC. As a result the difficulty parameter of an item is now defined as the θ value where $\mathrm{P}(\theta) = (1 + c)/2$. The higher the *c* value, the higher the probability of a correct answer even for examinees with infinitely low ability level. Figure 3 presents ICC’s of three items based on the 3-PL model.

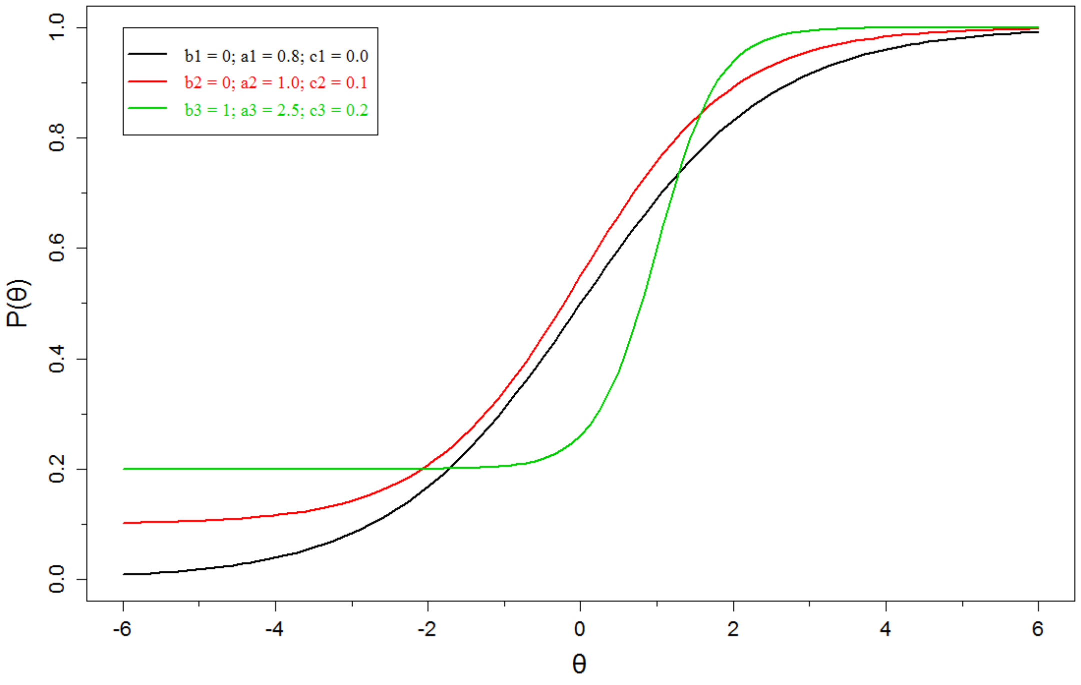


Figure 3 – Illustration of ICCs for 3-PL model

The three dichotomous IRT models so far discussed are still very popular in practical testing, however, they may not be applicable or appropriate in situations where polytomously-scored items are used to measure the latent trait of interest. For instance, if a researcher uses Likert-type items (i.e., a five-point scale ranging from strongly agree to strongly disagree), frequently used in survey questionnaires and attitudinal inventories, a model that can handle more than two response categories is required. This requirement led to the extensions of dichotomous IRT models and development of a broad class of IRT models capable of dealing with various polytomous item types.

4.3 Unidimensional polytomous IRT models

Polytomous IRT models can be used to model the interaction between examinee qualities and test items (how would a given examinee perform on a given item), where test items have several different response categories available. Thus, the main goal of the polytomous IRT model is to "describe the probability that an individual responds to a particular response category (e.g., strongly agree; agree; disagree; strongly disagree) given her or his level of ability and the item parameters" (Hambelton et al., 2010, p. 27). Statisticians have developed several variants of the polytomous IRT models, each suitable to use in a specific testing situation. For example a class of models was developed for items, which were presumed to have *a priori* ordered response categories (Thissen, Cai, & Bock, 2010). On the other hand, the nominal IRT model (Bock, 1972, 1997) was designed to analyze item responses with no predetermined ordering. The detailed description of all existing polytomous IRT models is beyond the scope of the present discussion. In keeping with the empirical focus of this thesis, only a restricted selection of models for polytomous items with ordered response categories will be introduced next (for example graded response model, rating scale model, partial credit model). Additional information that addresses other IRT models (e.g., nominal model, unfolding model, nonparametric models) can be found in several published works (Andrich & Luo, 1993; Nering & Ostini, 2010; Ostini & Nering, 2006; Ramsay, 1991; Roberts, Donoghue, & Laughlin, 2000; Sijtsma & Molenaar, 2002; Thissen et al., 2010; van der Linden & Glas, 2010; van der Linden & Hambelton, 1997).

One of the first polytomous IRT models developed, the graded response model (GRM), is an extension of the 2-PL model applied to polytomous items (Samejima, 1969, 1997, 2010). In order to derive the probability of responding to a given category of an item, the GRM uses a two-step procedure and is therefore considered an indirect model (Nering & Ostini, 2010; Ostini & Nering, 2006).

Consider a test item with K = 5 response categories ranging from 0 to 4 (k = 0,…, 4). In the first step the GRM models the probability $P_{ik}^{*}(\theta_j)$ of an examinee's response in the k^{th} **or higher** category of a test item i (i = 1, …, n):

$$P_{ik}^{*}(\theta_j) = \frac{\exp\left[a_i(\theta_j - b_{ik})\right]}{1 + \exp\left[a_i(\theta_j - b_{ik})\right]} \qquad (4)$$

where $P_{ik}^{*}(\theta_j)$ is often referred to as operating characteristic curve (OCC), or boundary characteristic function. The probability that an examinee's response falls in or above the lowest category is by definition $P_{i0}^{*}(\theta_j) = 1$, thus for an item with k response categories only K – 1 OCC's provide information about item functioning (Štochl, 2008). Consistent with the formulation of the dichotomous IRT model, a_i in Equation 4 represents the discrimination of item i and dictates the steepness of the OCC's. It should be obvious from Equation 4 that the discrimination parameter is the same for all item categories, which practically means that the OCC's are parallel within an item.

Samejima (2010) considered the model in Equation 4 as a homogenous case as opposed to the heterogeneous case of the GRM in which the discrimination parameters (and thus OCC's) are free to vary within an item. Parameter b_{ik} in the Equation 4 is associated with θ value at which there is 50% probability to respond in the ***k*^th^ or higher** category of item i. This parameter is often referred to as a threshold (or boundary), and it is analogous to the difficulty parameter in the dichotomous IRT models presented earlier. Again, since the probability of responding to the lowest or higher category is $P_{i0}^{*}(\theta_j) = 1$, there are only K – 1 meaningful thresholds per item and they dictate the positions of an item's OCC's along the trait continuum. An important feature of the GRM is that, by definition of the model, the threshold parameters are ordered from smallest to largest ($b_1 < b_2 < b_3 < \ldots$). Figure 4 illustrates OCC's for an ordered polytomous test item based on the GRM with the following parameters: a = 1.2, b_1 = -2.5, b_2 = -1, b_3 = 1, b_4 = 2.5.

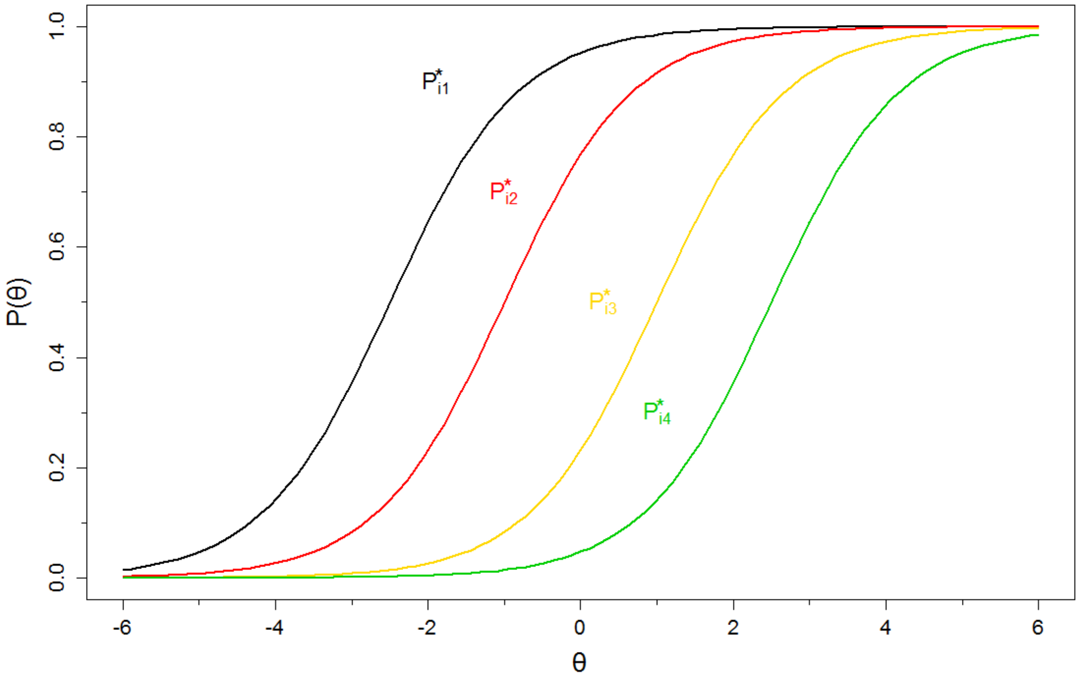


Figure 4 – Item operating characteristic curves for five-category item with a = 1.2, b_1 = -2.5, b_2 = -1, b_3 = 1, b_4 = 2.5, based on GRM

The OCC's, obtained through the Equation 4 in the first step are subsequently used within GRM to define the probability of responding to a particular category k of an item i:

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i(k+1)}(\theta) \quad (5)$$

Thus the probability of responding to a particular category for a five category item is computed as follows:

$$\begin{aligned}
P_{i0}(\theta) &= P^*_{i0}(\theta) - P^*_{i1}(\theta) = 1 - P^*_{i1}(\theta) \\
P_{i1}(\theta) &= P^*_{i1}(\theta) - P^*_{i2}(\theta) \\
P_{i2}(\theta) &= P^*_{i2}(\theta) - P^*_{i3}(\theta) \qquad (6) \\
P_{i3}(\theta) &= P^*_{i3}(\theta) - P^*_{i4}(\theta) \\
P_{i4}(\theta) &= P^*_{i4}(\theta) - 0 = P^*_{i4}(\theta)
\end{aligned}$$

If we plot P_{ik} along the whole range of the latent trait values we get what is called item category response functions (ICRFs), which are depicted for our example item in Figure 5.

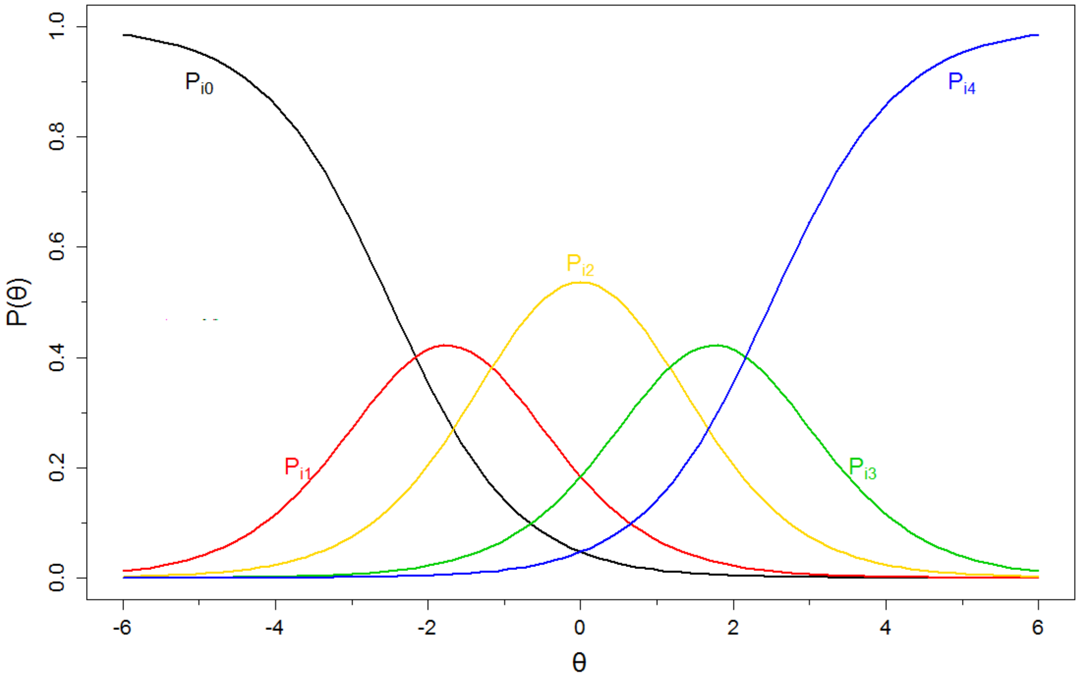


Figure 5 – Item category response function for five-category item with a = 1.2, b_1 = -2.5, b_2 = -1, b_3 = 1, b_4 = 2.5, based on GRM

Hambelton et al. (2010, p. 29) present some important characteristics of the GRM:

- for any given value of the latent trait the sum of the category response probabilities equals 1
- the first ICRF will always be monotonically decreasing while the last ICRF is always monotonically increasing
- the middle categories will always be unimodal with the peak located at the midpoint of the adjacent threshold categories.

In contrast to Samejima's GRM, which uses a two-step process to provide the probability of a particular response to an item, a class of models termed "direct" defines the item category response function for a test item i with K = 5 categories ranging from 0 to 4 (k = 0, 1…, m) as follows:

$$P_{ik}(\theta_j) = \frac{\exp \sum_{k=0}^{\mathrm{K}\leq k}(\theta_j - b_i^k)}{\sum_{m=0}^{m} \exp \sum_{k=0}^{\mathrm{K}\leq m}(\theta_j - b_i^k)} \quad (7)$$

where $P_{ik}(\theta_j)$ is the probability of an individual with a given latent trait level θ responding in the k^{th} category of item i; and b_i^k represents threshold parameter, that dichotomizes each pair of the adjacent item categories, so there are m-1 meaningful item thresholds. When computing the probability of a response, direct models in (7) compare only two adjacent item categories (1st vs 2nd), while indirect models (GRM) consider all item categories (GRM successively dichotomize the item categories into: 1st vs. all higher categories; 1st and 2nd vs. all higher categories; …). The manner in which the term b_i^k dichotomizes item categories within direct models involves local comparisons in contrast to so-called global comparisons used in GRM (Nering & Ostini, 2010; Ostini & Nering, 2006). One of the results of the successive dichotomization of adjacent item categories is that b_i^k's within an item are not necessarily ordered from smallest to largest in indirect models. The fact that item category thresholds are not restricted to be ordered in the IRT models defined in Equation 7 permits empirical testing whether ordered categories within an item function as intended (Andrich, 2010).

There are two possible representations of the category threshold parameters in Equation 7, which in turn define two distinct polytomous IRT models. By setting

$b_i^k = b_{ik}$ the partial credit model (PCM – Masters, 1982, 2010; Masters & Wright, 1984) states that every single item i has a separate location parameter b for each of the response categories k. A slightly different way of expressing the item category thresholds in Equation 7, namely $b_i^k = b_i + \tau_k$ defines the rating scale model (RSM – Andersen, 1973; Andrich, 1978, 2010), in which every item i has its own location parameter b, and τ_k is the common category threshold for all the items in a test. This means that the category thresholds in the RSM are equally spaced across all items within a test and they may be shifted on the latent continuum because of different item difficulties (Hambelton et al., 2010, p. 32).

The exponential in the numerator of the Equation 7 for both the PCM and RSM involves a sum of the differences between a given θ value and each of the category threshold parameters b_i^k up to the desired category k. The denominator is the sum of all possible numerators for a particular item and this is the reason why the PCM and RSM are also often referred to as divided-by-total models as opposed to the GRM, which is considered as a difference model (Ostini & Nering, 2006).

Since the successive dichotomization of adjacent item categories is meaningless in case of modeling the probability of a response in the lowest category $P_{i0}(\theta_j)$, it is convenient to set

$$\sum_{k=0}^{0}\left(\theta_j - b_i^k\right) = 0 \qquad (8)$$

and therefore the probability of responding to the third response category of a five-category item would be computed according to the Equation 7 as follows:

$$P_{i2}(\theta_j) = \frac{\exp\left[0 + \left(\theta_j - b_i^1\right) + \left(\theta_j - b_i^2\right)\right]}{\begin{array}{c}\exp 0 \\ + \exp\left[0 + \left(\theta_j - b_i^1\right)\right] \\ + \exp\left[0 + \left(\theta_j - b_i^1\right) + \left(\theta_j - b_i^2\right)\right] \\ + \exp\left[0 + \left(\theta_j - b_i^1\right) + \left(\theta_j - b_i^2\right) + \left(\theta_j - b_i^3\right)\right] \\ + \exp\left[0 + \left(\theta_j - b_i^1\right) + \left(\theta_j - b_i^2\right) + \left(\theta_j - b_i^3\right) + \left(\theta_j - b_i^4\right)\right]\end{array}} \qquad (9)$$

Figure 6 illustrates ICRF's of a 5-category item based on the PCM, where category thresholds are not ordered from smallest to largest (b_1 = -2, b_2 = 0.1, b_3 = -1.5, b_4 = 2.2). It could be seen from the Figure that the threshold parameter within the PCM (and it also holds for the RSM) represents a value on the latent continuum, where two successive ICRF's intersect (Hambelton et al., 2010). It is also possible to incorporate a discrimination parameter a_i for every item in Equation 7, which defines the generalized PCM (Muraki, 1992, 1993, 1997) and generalized RSM (Ostini & Nering, 2006).

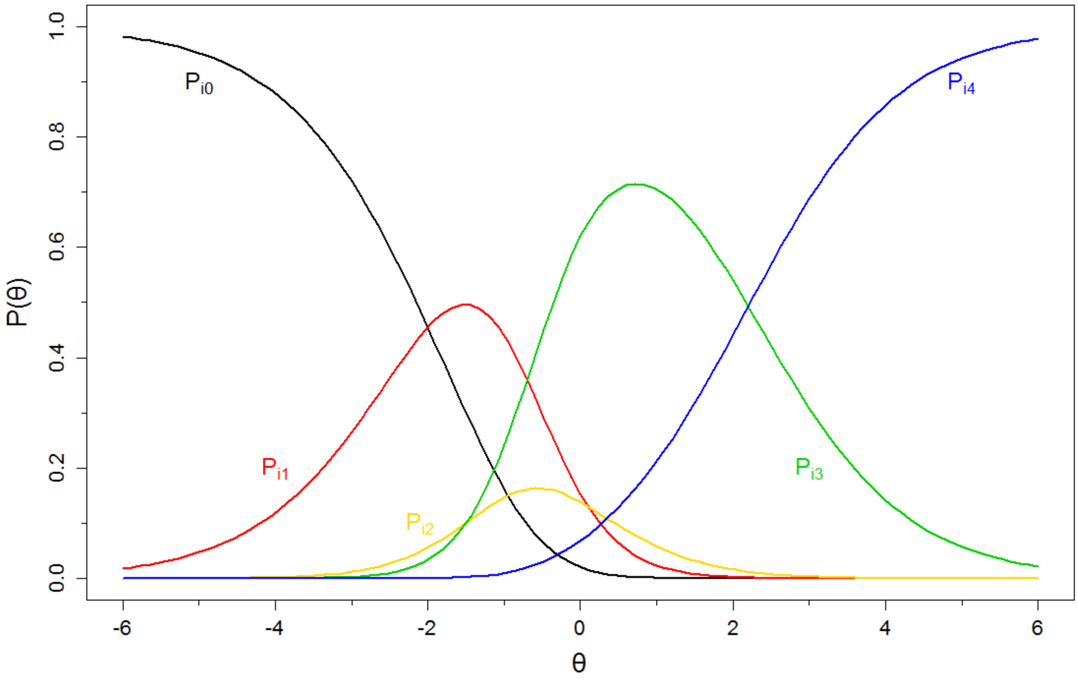


Figure 6 – Item category response function for five-category item with b_1 = -2, b_2 = 0.1, b_3 = -1.5, b_4 = 2.2, based on PCM (or RSM)

4.4 Assumptions required for unidimensional IRT models

There are two assumptions required to understand how the unidimensional IRT model works. These are important to gain insight how to interpret a given test score, estimating the model parameters, and deciding whether the model represents a reasonable approximation of the observed data.

The unidimensional IRT model assumes that there is a single underlying latent trait on which all the items within a test battery rely to some extent (Wainer & Mislevy, 2000). This supposition stems from Loevinger's (1947) concept of test homogeneity, which proposed that a collection of items should measure the same latent trait of interest. From a mathematical point of view the assumption of unidimensionality means that responses of examinees to a given set of items can be described by a single latent variable (Sijtsma & Molenaar, 2002). This assumption is never fulfilled perfectly in practical testing situations, however IRT models are robust to moderate violations of unidimensionality (Hambelton & Cook, 1983). The usual approach therefore is to consider a set of items as unidimensional as long as a dominant latent theoretical construct underlies responses to the items (Hambelton, Swaminathan, & Rogers, 1991; Zhu, 2006). Assessing dimensionality of a set of items is closely related to construct validity, that is, assessing "the extent to which a test designated to measure a specific theoretical trait or proficiency actually does so" (Steinberg, Thissen, & Wainer, 2000, p. 188). The dimensionality of a test instrument is traditionally evaluated using factor analysis (McDonald, 1985), a data summarization technique that provides a means to empirically confirm the number of dimensions needed to satisfactory describe the associations in the observed responses to the test items. Where there is evidence of strong multidimensionality it is advised to either delete small content areas from the test instrument, split a scale into two or create more subtests, or use one of the multidimensional IRT models (Wainer & Mislevy, 2000).

Until quite recently, a classical factor analysis (that is for example Thurstone's 1947 approach assuming that the indicators are continuously distributed and standardized to mean of 0 and variance 1) has been used to establish the dimensional structure of a test, even in cases when observed data were categorical in nature (e.g., binary/dichotomous and polytomous indicators of the theoretical construct). In the case of tests containing binary items this approach led to extraction of spurious factors related to difficulty of the items (Maydeu-Olivares, 2005; McDonald, 1967, 1999). This problem can be avoided by using categorical item factor analysis (see for example Bock & Gibbons, 2010) which utilizes tetrachoric or polychoric correlations, respectively. The most recent and comprehensive factor analytic approach assessing dimensionality, however, does not depend on the bivariate associations and uses all the information involved in the categorical responses (by using Full Information Maximum Likelihood estimator). This is the reason why it is referred to as full-information or IRT-based item factor analysis (Bock & Gibbons, 2010).

The fact that the full information item factor analysis is also referred to as an IRT-based approach is related to the second important assumption of the IRT models – the assumption of local independence. This assumption states that "for a given position on the latent dimension, responses to the items are statistically independent" (McDonald, 1967, p. 16). The assumption of local independence can be expressed as follows:

$$h(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} h(y_i|\theta) \quad (10)$$

where $\boldsymbol{y}$ is a vector of observed responses to items i (i = 1,…n) and $h(\boldsymbol{y}|\theta)$ is the conditional density function of the item responses at a given trait level θ. In case of dichotomous items this formula simplifies to:

$$P(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} P(y_i|\theta) \quad (11)$$

that is, the conditional probability of observing a vector of responses $\boldsymbol{y}$ at a given trait level θ, is a simple product of the model generated probabilities ($P(\theta)$ if y_i = 1 and 1 – $P(\theta)$ if y_i = 0) for each item i at that trait level (Segall, 2010). Statistical conditional independence stated in Equations 10 and 11 and that underlies the IRT models is a stronger requirement than the requirement of zero partial correlations in classical and item factor analysis. In this manner, the IRT model theoretically places more stringent demands by requiring an explanation of all the test item associations (bivariate and otherwise) which were not fully addressed statistically by classical or item factor analysis (McDonald, 1985). The assumption of local independence can be violated for example when items are administered naturally in clusters – e.g., a set of reading comprehension items about the same reading passage (Wainer & Mislevy, 2000). Other possible scenarios that may contribute to violations of the local independence requirement include learning through practice or functional dependence of items (Štochl, 2008).

4.5 Parameter estimation in IRT models

From the perspective of test theory, an important aim of psychological assessment is to obtain a quantifiable representation of an examinee’s latent trait. For instance, in the case of clinical psychology assessments, we want to know the examinee’s level of neuroticism or extroversion, both latent traits that we cannot observe directly but must infer from fallible empirical indicators. In order to be able to obtain such a representation in the context of IRT, one needs to know the characteristics of items used to measure the latent trait of interest, since items are considered the building block within the IRT model. One of the integral parts of IRT application is therefore estimation of item parameters (difficulty, discrimination) and the examinee's level of the latent trait being measured.

There are currently two broad classes of parameter estimation methods used within the context of IRT – maximum likelihood estimation methods and Bayesian estimation methods. In this section only a brief introduction to the logic behind some of the approaches to the latent trait (θ) and item parameters (difficulty, discrimination) estimation is outlined. An interested reader is referred to Baker (1965); Baker and Kim (2004); Bock (1972); Bock and Aitkin (1981); Bock and Lieberman (1970).

4.5.1 Latent trait (θ) estimation

It was already noted that the principle of local independence stated in (10) serves as a basis to estimate the model parameters in IRT. Let's assume that several dichotomous items i (i = 1, 2, …, n) based on the 2-PL model with parameters β_i (a_1, b_1, …, a_n, b_n) were administered to a sample of examinees and the task is to estimate the θ (latent trait) for each of them. If the item parameters β_i are known, it is possible to express the Equation 11 as a likelihood function:

$$L(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} P(y_i|\theta) \quad (12)$$

yielding a likelihood of observing a response vector $\boldsymbol{y}$ if θ were the true values (Wainer & Mislevy, 2000). Since $\boldsymbol{y}$ contains the observed (known) responses of an examinee, $L(\boldsymbol{y}|\theta)$ is a function of θ parameters only (Segall, 2010) and thus Equation 11 may be used to estimate θ. Figure 7 illustrates how the model generated probabilities (ICC's) are multiplied in order to get the likelihood function for two different patterns of responses to three dichotomous test items with parameters estimated according to 2-PL (i.e. difficulty and discrimination) listed in Table 1.

Table 1 – Fictional IRT item parameters and response patterns used in illustration of latent trait estimation on Figure 7

Item	*Discrimination*	*Difficulty*	Pattern 1 y_i	Pattern 2 y_i
1	0.6	-1	0	1
2	1.5	0	1	1
3	2.5	1	1	0

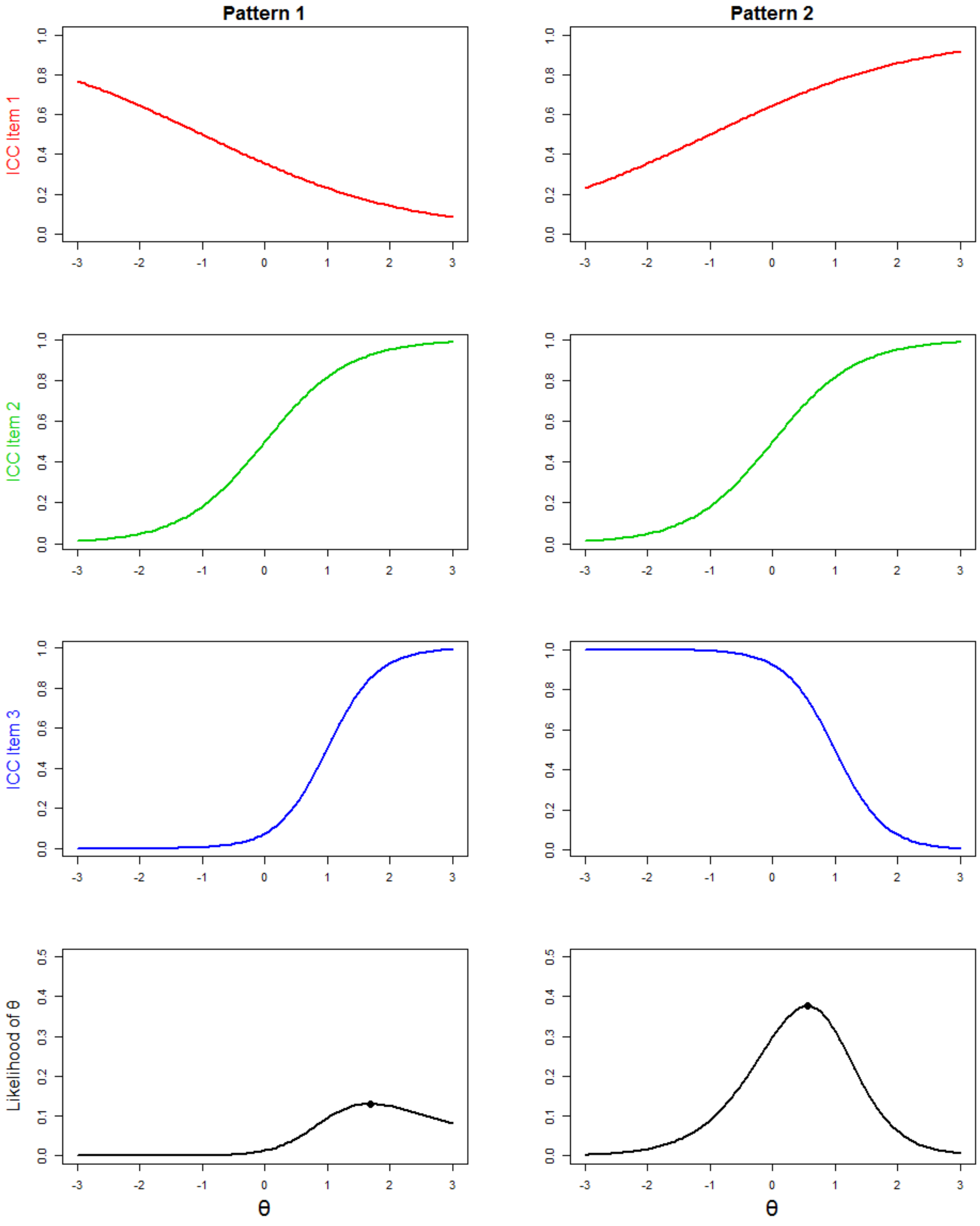


Figure 7 – Likelihood estimation of latent trait (θ) for two different response patterns

The maximum likelihood (ML) point estimate of θ for both response patterns would be simply the mode of the likelihood (Segall, 2010) – that is the highest point on the likelihood function portrayed in Figure 7. Thus one way to estimate θ is to calculate the value of the likelihood "at each of many points across the range of θ and simply note where it is the highest" (Wainer & Mislevy, 2000, p. 72). Usually however the θ is estimated iteratively using the Newton-Raphson successive approximations while maximizing the natural logarithm of the likelihood expressed in Equation 12 (details can be found in Baker, 1985 or Baker and Kim, 2004). A latent trait estimate using the Newton-Raphson method for the 2-PL model could be obtained as follows:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{n} a_i \left[y_i - P_i(\hat{\theta}_s)\right]}{\sum_{i=1}^{n} a_i{}^2 \left[P_i(\hat{\theta}_s)(1 - P_i(\hat{\theta}_s)\right]} \quad (13)$$

where $\hat{\theta}_s$ is the latent trait estimate for a given response pattern within iteration *s*; $P_i(\hat{\theta}_s)$ is the probability of responding to a keyed (correct, or trait indicating) category of item i (i = 1, 2, …, n) at the given $\hat{\theta}_s$; y_i is the response to the item *i* (1 for keyed/correct/trait indicating response, 0 otherwise) and a_i is the discrimination parameter of a particular item i.

A disadvantage of ML estimation (MLE) is that it does not have a maximum for so-called perfect (maximum possible simple sum) or zero scores (minimum possible simple sum). In that case, MLE returns θ's equal to + infinity for perfect scores and – infinity for zero scores. Warm (1989) proposed a weighted version of the ML estimation in which the mean instead of the mode of the likelihood function is used as the θ estimate. Weighted likelihood estimation (WLE) is asymptotically equivalent to MLE, however it overcomes the unbounded nature of the MLE in case of zero or perfect score (S. Wang & Wang, 2001; T. Wang & Vispoel, 1998). Also Bayesian estimation, another popular approach for latent trait estimation, is able to correct for this MLE shortcoming.

In a Bayesian estimation approach, prior information about the θ distribution $f(\theta)$ is incorporated into the likelihood function $L(\boldsymbol{y}|\theta)$, yielding a posterior distribution:

$$\mathrm{p}(\boldsymbol{y}|\theta) \propto f(\theta)L(\boldsymbol{y}|\theta) \quad (14)$$

which subsequently serves as a basis to provide estimates of a latent trait (Segall, 2010). Using the mean or the mode of the posterior distribution $\mathrm{p}(\boldsymbol{y}|\theta)$ as a θ point estimate, defines two different Bayesian approaches, that is: expected a posteriori (EAP) and Bayes modal estimation (BME – also known as the maximum a posterior estimation (MAP)) respectively (Hambelton et al., 2010).

There are two classes of prior distribution commonly used in Bayesian estimation – a noninformative prior and an informative prior distribution $f(\theta)$ respectively. Since the $f(\theta)$ is simply treated as an "item" whose "ICC" is multiplied with another item ICC's, a noninformative prior yields a posterior distribution, which is proportional to the likelihood function and this is the reason why the BME with uniform prior is equivalent to the MLE (Wainer & Mislevy, 2000). The process of incorporating the prior information about the latent trait distribution by multiplying it along with another item ICC's is illustrated in Figure 8 – here two types of the prior distributions (noninformative – uniform distribution vs. informative – standard normal distribution) labeled as $f(\theta)$ were used to obtain the posterior distribution $\mathrm{p}(\boldsymbol{y}|\theta)$, while an examinee has responded to keyed/correct/trait-indicating category in both test items.

Figure 8 shows that the selection of a particular prior distribution influences the posterior distribution and consequently also affects the θ estimate. Another important point to make, not easily observed from the picture, is that the upper asymptote of the likelihood function for the noninformative prior goes to infinity when an examinee has answered all the test items in the same direction (in case of polytomous items when all the responses are either in the first or in the last category of the test items). Therefore, similarly to MLE, the Bayesian estimation with noninformative (uniform) prior is not appropriate in situations of perfect or zero scores.

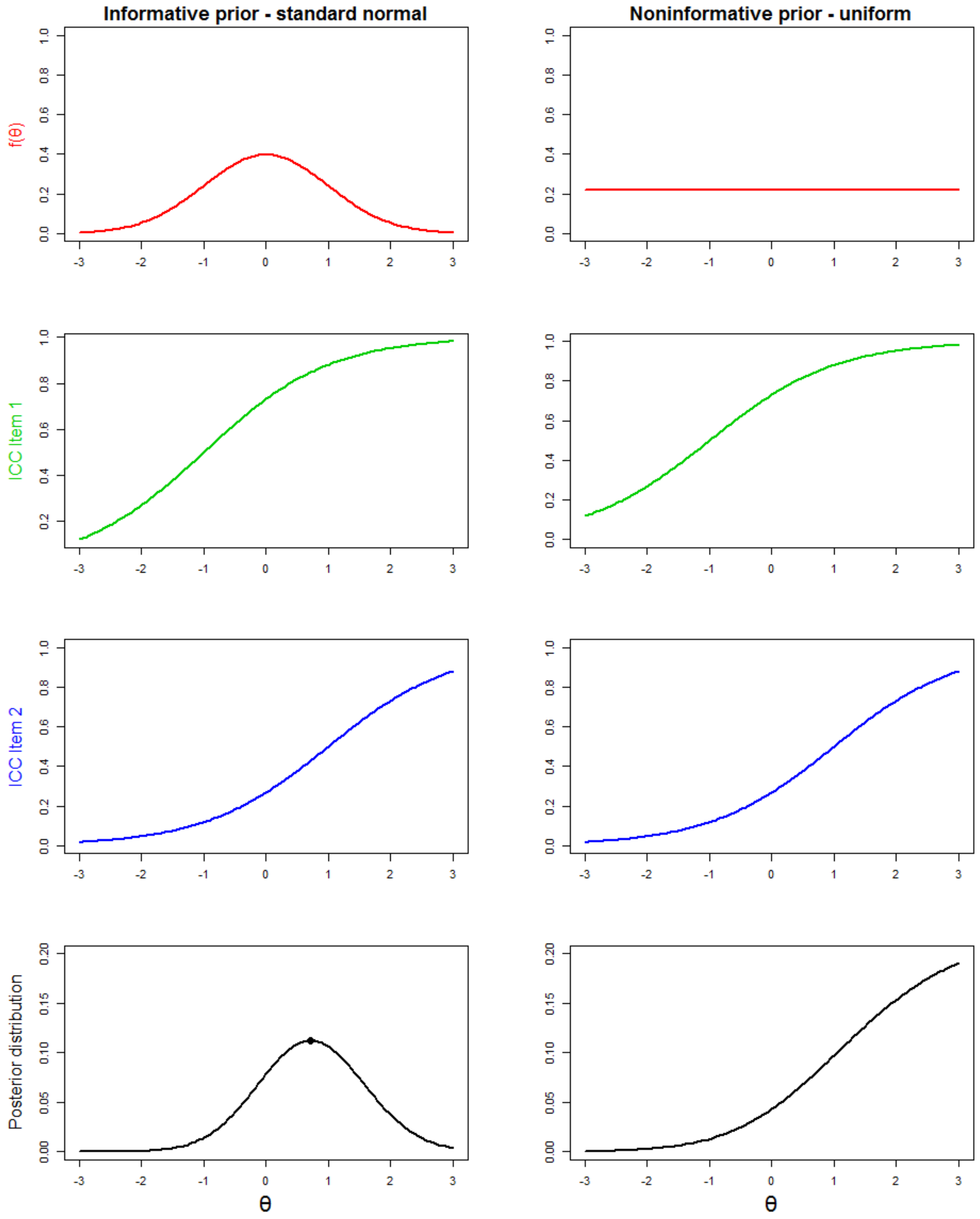


Figure 8 – Bayesian estimation of latent trait (θ) using two different prior distributions

4.5.2 Item parameters estimation

When latent trait parameters are known, it is possible to apply the MLE principles outlined above also for item parameter estimation – the task now becomes one of finding the maximum of the likelihood function of unknown item parameters given the observed responses and latent trait values (Hambelton et al., 2010).

Unfortunately, it is never the case in practical testing situations (at least during the initial phases of test administration), that either item parameters or latent trait parameters are known and thus the challenge becomes the simultaneous estimation of both sets of parameters. There are three commonly used methods for item parameter estimation that address this challenge: joint maximum likelihood (JML), conditional maximum likelihood (CML), and marginal maximum likelihood (MML) estimation (Baker & Kim, 2004; Wainer & Mislevy, 2000). The first type, JML estimation, uses an iterative two-step process while maximizing the joint likelihood function for both the examinees (latent trait) and the items. The joint function can be obtained simply by multiplying the likelihood function in Equation 12 across N examinees (de Ayala, 2009):

$$L = \prod_{j=1}^{N} \prod_{i=1}^{n} P(y_i|\theta) \quad (15)$$

In the first step of JML, the provisional θ estimates (various functions of the raw score may be used) are used in order to estimate the item parameters. In the second step the provisional θ estimates are updated treating the item parameters from the first step as known. These two steps are repeated until a convergence criterion is satisfied, that is until the successive change in updated parameter values for both the examinees and items are trivially small.

Simultaneous estimation of examinees and items parameters in JML is known to have several problems (e.g., de Ayala, 2009; Wainer & Mislevy, 2000), mainly with the inconsistency of the parameter estimates (Andersen, 1973). The inconsistency is caused by the fact that the main interest of JML rests with the estimation of a limited number of item parameters in the presence of many θ parameters, which are considered as nuisance parameters in this context (Eggen, 2000; Mair & Hatzinger, 2007; Mair, Hatzinger, & Maier, 2010). Addressing this concern led to the development of estimation methods that eliminate the nuisance unknown θ parameters, taking them out of the likelihood function by using either the simple raw score in CML, or relying on distributional assumptions about θ (e.g., standard normal distribution) in MML (Hambelton et al., 2010). Whereas the CML fulfills the Rasch's requirement of specific objectivity (McDonald, 1999; Rasch, 1960), this approach is suitable for the Rasch model and its extensions only, where the raw score provides a sufficient statistic for θ estimation. On the other hand, the precision of item parameter estimates in MML is influenced by the chosen θ distribution and it is obvious that an incorrect distributional assumption will produce biased estimates. Both the CML an MML yield consistent item parameters and if the θ distribution is correctly specified in MML, both methods provide asymptotically equivalent estimates (Pfanzagl, 1994). Once the item parameter estimates have been obtained using CML or MML respectively, they can be used to obtain latent trait parameters by treating item parameters as known, just as in the second step of JML outlined above (Hambelton et al., 2010). When considering the estimation of model parameters (examinee's latent trait parameters and item parameters) based on CML or MML, it should be noted that with CML, unlike MML, examinees whose raw scores are either zero or perfect are eliminated from the entire estimation process (Molenaar, 1995). Therefore, the MML estimation procedure is recommended if one wants to obtain finite latent trait estimates for such examinees (Mair & Hatzinger, 2007; Mair et al., 2010).

Besides the various MLE methods discussed above, Bayesian procedures can also be used for the purpose of item parameter estimation. Similarly as in the case of the latent trait estimation, Bayesian estimation of item parameters rests on using the prior distribution for both the item and θ parameters and incorporates priors into the likelihood function in order to obtain the posterior distribution for item parameters. A detailed description and corresponding mathematical formulas for Bayesian item parameter estimation can be found in (Baker & Kim, 2004).

4.6 Information and standard error of the θ estimates

Another important concept in IRT, which is also crucial for IRT-based computerized adaptive testing (see section 6, 'Testing algorithms in unidimensional IRT-based computerized adaptive testing'), is the item and test information function respectively. Generally, the amount of information in the field of statistics was defined by Sir R. A. Fisher as the reciprocal of the sampling variance of the estimated parameter (Baker, 1965). In the context of IRT we are mainly interested in the sampling variance of θ estimates and "the larger this variance the less precise the estimate of θ and the less information one has as to an examinee's unknown ability level" (Baker & Kim, 2004, p. 70)

For example the sampling variance $\sigma_e^2(\hat{\theta}_j)$ of θ estimates in case of dichotomous items could be obtained as:

$$\sigma_e^2(\hat{\theta}_j) = \left[\sum_{i=1}^{n} \frac{[P_i'(\theta_j)]^2}{[P_i(\theta_j)][1 - P_i(\theta_j)]}\right]^{-1} \quad (16)$$

where $P_i(\theta_j)$ is the probability of a latent trait indicating response to an item i (i = 1,..., n) given the particular model (1PL or 2PL model, respectively) and $P_i'(\theta_j)$ is the first derivative of $P_i(\theta_j)$ with respect to θ. Since the first derivative for dichotomous 1PL and 2PL model is $P_i'(\theta_j) = a_i[P_i(\theta_j)][1 - P_i(\theta_j)]$, the Equation 16 can be simplified to:

$$\sigma_e^2(\hat{\theta}_j) = \left[\sum_{i=1}^{n} a_i^2[P_i(\theta_j)][1 - P_i(\theta_j)]\right]^{-1} \quad (17)$$

where a denotes the discrimination parameter of an item i. By taking the square root of the Equation 16 or 17 we can get an IRT-based precision index called the standard error of the latent trait estimate – $SE(\hat{\theta}_j)$. Similar to the standard error of measurement (SEM) in CTT, $\pm SE(\hat{\theta}_j)$ around the point θ_j estimate constitutes the 68% confidence interval (de Ayala, 2009). However unlike SEM in CTT, which is considered a global measure of error for an estimate, $SE(\hat{\theta}_j)$ in IRT varies as a function of θ.

According to Fisher's definition mentioned above, the amount of information, which is provided by a set of items for a given θ_j estimate, can then be obtained by taking the reciprocal of the Equation 17:

$$I(\theta) = \sum_{i=1}^{n} a_i^2[P_i(\theta_j)][1 - P_i(\theta_j)] \quad (18)$$

where $I(\theta)$ is the test information function. The sum in the Equation 18 reflects that, individual items contribute independently to the test information function, and therefore for any given θ_j estimate it is possible to calculate the contribution of information for any particular item in an item pool:

$$I_i(\theta) = a_i^2[P_i(\theta_j)][1 - P_i(\theta_j)] \quad (19)$$

where $I_i(\theta)$ is the item information function and it follows that $I(\theta) = \sum I_i(\theta)$. Since the discrimination parameter a_i for 1-PL (Rasch) model is assumed to be 1 for all items within a test battery, the maximum of the provided information will always be equal among all the items for Rasch models. Figure 9 illustrates an item and a test information function for a fictional test battery consisting of three items with difficulties: b_1 = -2; b_2 = -1.5; b_3 = 2, which were estimated using a Rasch model (e.g.: a_i = 1).

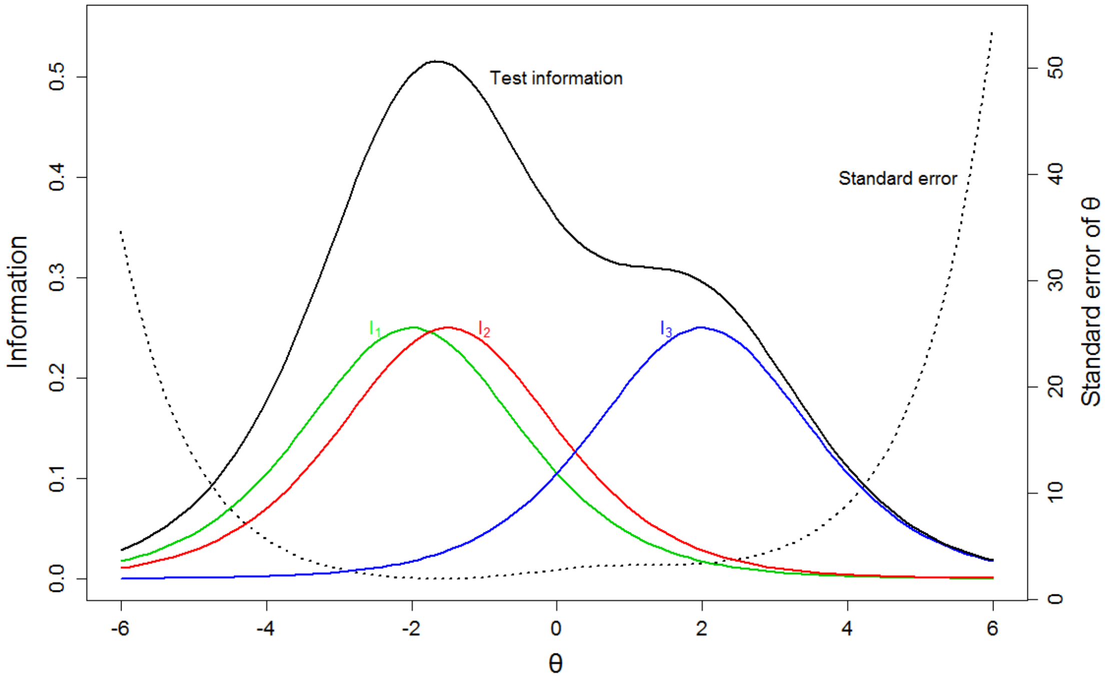


Figure 9 – Test and item information functions for three dichotomous items with a = 1, b_1 = -2, b_2 = -1.5, b_3 = 2

As Figure 9 depicts, the maximum of information for all three items is indeed the same, peaked at the particular item difficulty and in Rasch models it will always be of value 0.25 when $P_i(\theta_j) = 1 - P_i(\theta_j)$. Moreover the test information function in Figure 9 indicates that the fictional three-item test would provide considerably more information for estimating low levels of the latent trait (the maximum is at $\theta \approx -1.7$) – caused by the higher number of items with low difficulty in the fictional example test. The situation may be slightly different in a 2-PL model where item discrimination parameters a_i are also estimated, thus producing results not quite identical to the 1-PL model. To illustrate the difference between these respective models, Figure 10 portrays item and total information functions for items with the same difficulty parameters as in the previous example (b_1 = -2; b_2 = -1.5; b_3 = 2), but with different discrimination parameters (a_1 = 0.6; a_2 = 1.1; a_3 = 1.5). Even though the maximum of the provided information by the particular item is centered at the item's difficulty just as in the case of the 1-PL model, the maximum now is different and its value is dictated by the item's discrimination parameter.

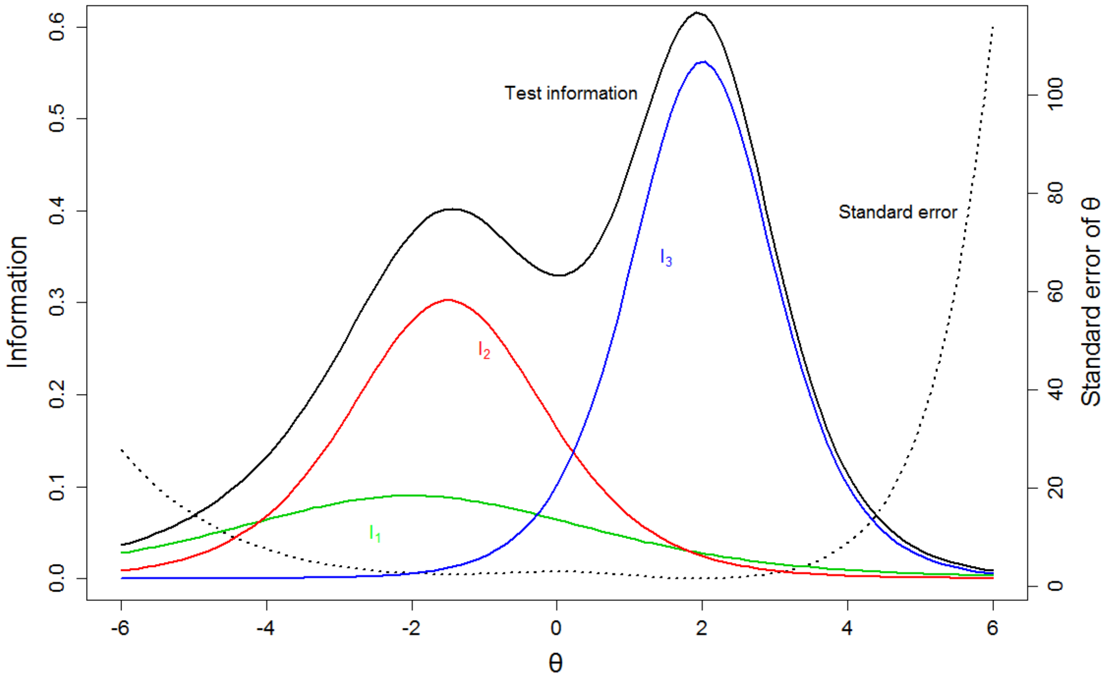


Figure 10 – Test and item information functions for three dichotomous items with a_1 = 0.6, a_2 = 1.1, a_3 = 1.5, and b_1 = -2, b_2 = -1.5, b_3 = 2

Thus it is also possible in 2-PL model that such a fictional three-item test would provide the maximum information for estimating high levels of the latent trait (the maximum is at $\theta \approx +1.9$), regardless of the test items difficulty parameters distribution.

It is also worth noting that with polytomous items and IRT models it is also possible to evaluate the amount of information provided by each item's response category. For instance, when the items follow a graded response format, Samejima (1969) defines the item's category information function as:

$$I_{ik}(\theta) = \frac{[P_{ik}{}'(\theta)]^2}{P_{ik}} \quad (20),$$

and it follows that

$$I_i(\theta) = \sum_{k=0}^{k} I_{ik}(\theta) \quad (21).$$

In case of Samejima's GRM (which will be used later in the empirical part) and given the Equation 5, it is possible to rewrite the Equation 20 as:

$$I_{ik}(\theta) = \frac{[P_{ik}{}'(\theta)]^2}{P_{ik}} = \frac{\left[P_{ik}^*{}'(\theta) - P_{i(k+1)}^*{}'(\theta)\right]^2}{P_{ik}^*(\theta) - P_{i(k+1)}^*(\theta)} \quad (22)$$

where $P_{ik}^*{}'(\theta)$ is the first derivative of $P_{ik}^*(\theta)$. Since the GRM is an extension of the 2-PL model, the formula for the first derivative presented above is applicable also here, that is $P_{ik}^*{}'(\theta) = a_i[P_{ik}^*(\theta)][(1 - P_{ik}^*(\theta)]$.

Even though it is not a common practice, the logic behind the item's category information function can be presented on a dichotomous item. A dichotomous item could be considered as a special case of a polytomous item with number of possible response categories k = 2 and so both the categories theoretically contribute to the item information function. This is illustrated in Figure 11 for a fictional dichotomous item assessing cognitive functioning (response categories: wrong answer vs. right answer) with a_1 = 1 and b_1 = -1.

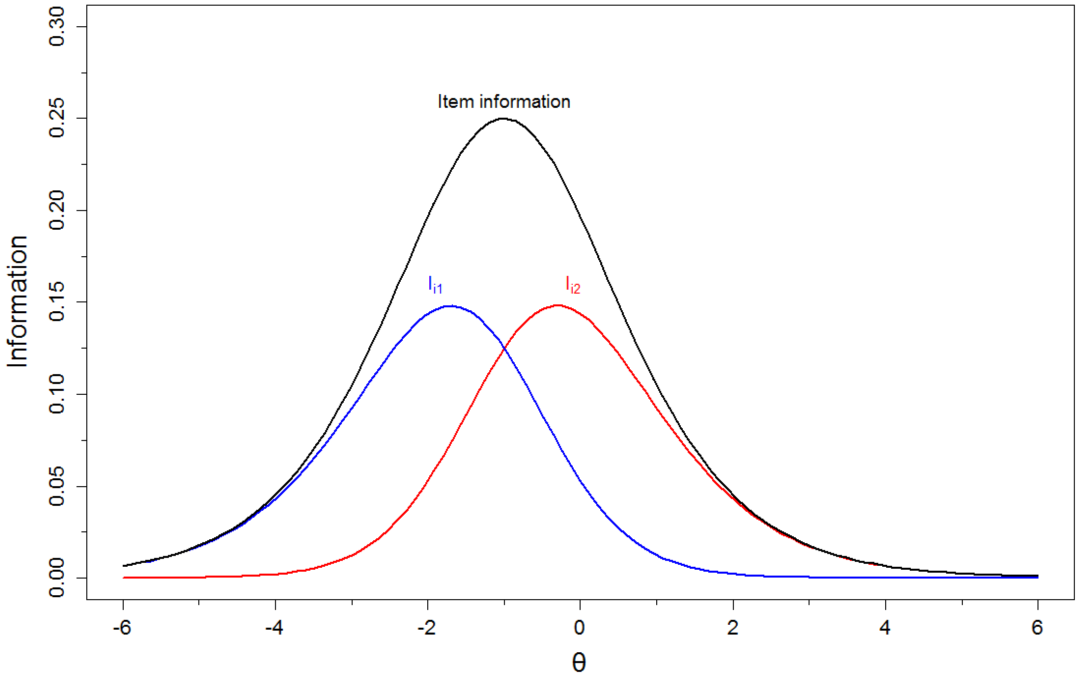


Figure 11 – Item's category information functions (I_{i2} – right answer, I_{i1} – wrong answer) for a dichotomous item with $a = 1$, $b = -1$

Information functions are especially useful in designing measurement instruments (test batteries, inventories, and questionnaires) which by necessity contain specific characteristics. For example, it is possible to create an instrument that will be suitable to measure a latent trait along a wide range of the continuum (e.g., norm-referenced or non-mastery testing, see Baumgartner, 2007; Zhu, 2007), or possibly an instrument that will be very precise in estimating the latent trait around some predetermined level (e.g. criterion-referenced or mastery testing, see Baumgartner, 2007; de Ayala, 2009; Zhu, 2007).

5 COMPUTERIZED ADAPTIVE TESTING (CAT): HISTORICAL AND CONCEPTUAL ORIGINS

Measurement instruments including questionnaires, inventories, test batteries, achievement tests, and surveys commonly used in the social and behavioral sciences, have traditionally been designed for administration in a linear (fixed-length) format (Becker & Bergstorm, 2013). This conventional measurement approach presents the same set and sequence of test items to each test taker, usually in a defined time frame, for instance during final exams after completion of a semester of sport physiology. This methodology has obvious advantages and disadvantages. One of the advantages is the possibility of administering the test to a large group of examinees at the same time (mass-administered testing – see DuBois, 1970), which also maximizes uniformity of the testing situation (all test takers experience the same context and events surrounding the test administration) and also reduces cost when compared to individual testing (Wainer, 2000). Moreover comparison of examinees taking the same test is simple and straightforward (Štochl, Böhnke, Pickett, & Croudace, 2016a; Wainer & Mislevy, 2000) and is for the most part what makes fixed-length linear assessments so attractive and popular for practical research activities.

Historically speaking, the advent of both World War I and II was instrumental in the transition from individual oral testing to mass-administered paper-and-pencil testing. Test instruments used in the area of intelligence research before the wars were administered on a case-by-case basis and to only one person at a time. Many of the items in these test instruments required oral responses from examinees, individual timing or manipulation of materials (i.e., building blocks). One of the most popular individual tests was the Binet-Simon Scale (Binet & Simon, 1905) developed to measure a person's mental level (or mental age – see Anastasi, 1976). The original scale consisted of 30 sub-tests or problems ordered according to their difficulty. In contrast to a linear fixed-length test, an administration of a particular sub-test in the Binet-Simon Scale was based on the examinee's actual ability. That is if an examinee passed a sub-test with a particular known difficulty level, then a sub-test with a higher difficulty could be administered subsequently. Conversely, in the event that an examinee failed a particular sub-test, also with a known difficulty level, the testing procedure could be terminated. Each individual would therefore be tested only over a specific range of ability suited to his or her intellectual level. Fairly complicated administration and scoring of sub-tests in the Binet-Simon Scale, however, requires a highly trained and experienced examiner. Moreover the scoring procedure for an individual intelligence test must be done immediately following administration of a particular sub-test, since the process of how the testing procedure unfolds is entirely driven by the examinee's responses to previously administered sub-tests.

The United States entry into World War I in 1917 necessitated testing a large number of recruits for military service and meeting the demands of both cost- and time-efficient test administration (DuBois, 1970; Wainer, 2000). The military's role in the transition from individual to group testing is heralded by development of the Army Alpha and Army Beta tests – two instruments that inaugurated large-scale mental testing (Wainer, 2000). The Alpha and Beta tests were developed by a group of psychologists led by then president of the American Psychological Association, R. M. Yerkes and used by the U.S. military for personnel selection and classification (Anastasi, 1976). The Army Alpha was designed for general routine testing of intelligence, while the Army Beta was a non-language alternative employed with illiterates and with non-English-speaking recruits. Subsequent to the wartime uses of mental testing, mass-administered testing using a linear fixed-length format also became popular by American Universities who could use screening tests to supplement the collegiate admission process (van der Linden, 2008). This was the context for developing the Scholastic Aptitude Test (SAT), which played a screening and selection role similar to the Army Alpha and Army Beta tests, but for entrance to college rather than the military. Since both the Alpha and Beta (as well as other group tests like SAT) were designed as mass testing instruments using multiple-choice formats based on the work of Otis (see Wainer, 2000), they not only enabled the simultaneous measurement of large groups of recruits but also simplified the instruction and administration procedures. The role of examiner was therefore greatly facilitated in the group testing when compared to the more laborious and time consuming individual testing situations.

Although easy and efficient to administer, a linear testing format is often time-consuming (from an examinee perspective) and thus may place considerable burden on the test taker (Štochl et al., 2016a). In order to effectively measure the full breadth of a particular latent trait, a measurement instrument has to contain items (i.e., empirical indicators) whose level of difficulty covers the entire spectrum of the specified latent trait continuum. For example an instrument assessing scholastic achievement must contain some relatively easy items earmarked for less proficient examinees, items of moderate difficulty targeting average examinees, and items of extreme difficulties for examinees that possess high proficiency (Wainer, 2000). The biggest limitation of the traditional group testing using a linear fixed-length format is its lack of flexibility, since every examinee is routinely tested on all of the items included in a test. Canvassing all of the latent trait levels with such a wide range and large number of items, linear testing can weaken a test's reliability by introducing undesirable incidental variables (e.g., boredom, lack of concentration or frustration), and increase the possibility of 'guessing' by individuals with lower levels of the latent trait (Wainer, 2000). These and related factors undermine the effectiveness of the testing process itself (de Ayala, 2009).

As the field of testing and assessment continued to unfold, researchers tried to combine several of the advantages associated with both individual and group testing. This fostered several innovative approaches and techniques that were proposed in the 1960's and 1970's. Major interest has focused on possibilities of mass-administered test that would be tailored to individuals based on their actual performance. In other words, psychometricians and test developers tried to provide a basis for mass-administered adaptive testing, in which the role of the test administrator would be greatly simplified despite the fact that the testing process is individualized according to the examinee's actual performance in the test in question.

The development of IRT in the middle to later portion of the 20th century has provided a sound theoretical background for mass-administered adaptive testing. Relatively slow computers at that time, unable to handle matrix algebra and complex computations involved in IRT models within a reasonable time, however, hindered researchers from taking advantage of the full potential of modern test theory. Early practical applications of group-administered adaptive testing were therefore mainly implemented in a traditional paper-and-pencil environment without using a specific mathematical model (e.g. IRT model) for the purpose of item selection and latent trait estimation. Examples of such an approach include two-stage testing (Cronbach & Gleser, 1965), the flexilevel test (Lord, 1971) or the pyramidal adaptive testing (Larkin & Weiss, 1975) among others. Figure 12 illustrates a simple hypothetical example of the two-stage test format.

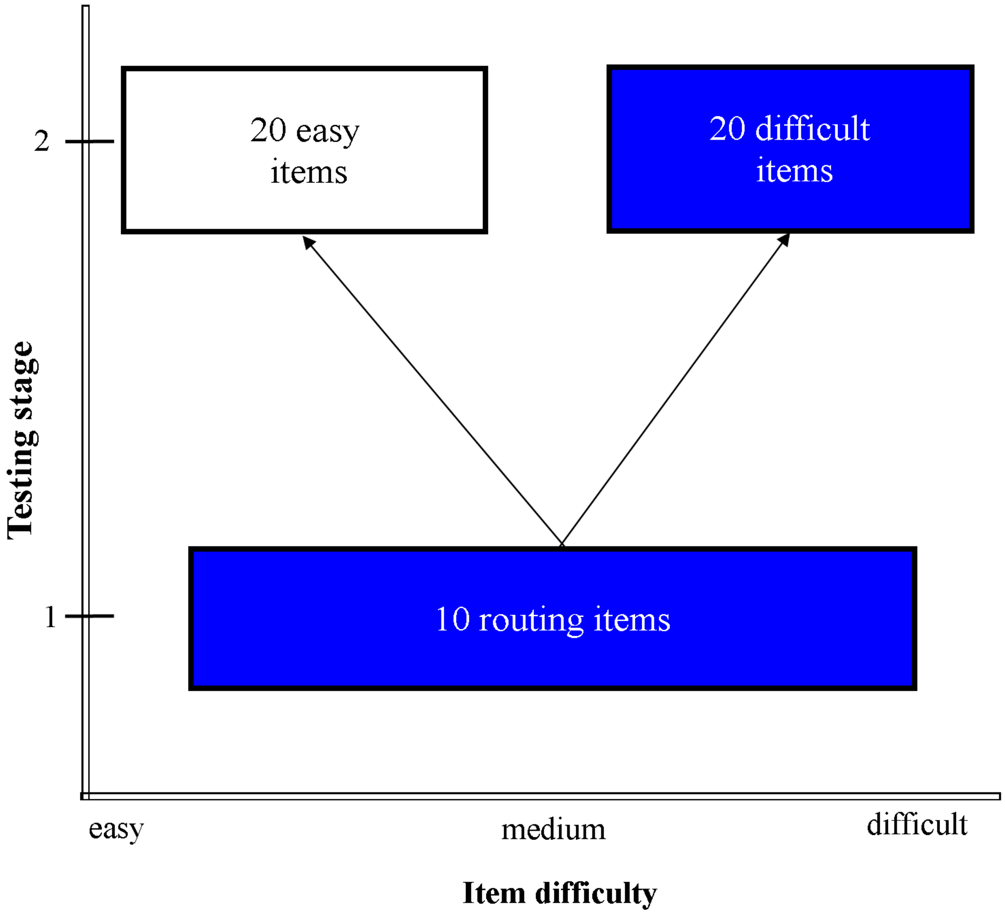


Figure 12 – Example of two-stage adaptive testing format

It should be noted that every test administration is driven by a specific testing algorithm, which defines the testing process in terms of how to begin, how to continue, and how to terminate the testing (Thissen & Mislevy, 2000). For instance, in standard linear testing formats, all examinees begin by responding to a particular test item and then continue until they have responded to all of the items in the test. In the example given in Figure 12, a two-stage testing format, all test takers start by responding to 10 designated ‘routing’ items, whose difficulties span a wide range of the latent trait being assessed. Based on the test taker’s responses to the routing items (whether they perform poorly or do well), each examinee is then channeled respectively to receive one of two 20-item sets, each of which contains items with different proficiency or difficulty levels (easy vs. difficult). By adapting the item difficulties in the second stage according to an examinee’s performance in the first stage, the two-stage format shortens the testing procedure from the test taker perspective. Using the format presented in Figure 12, each examinee has to respond to only 30 items, although the entire test contains 50 items.

Figure 13 shows a slightly different adaptive testing approach, called a 'pyramidal' test. In this case, test items are adapted to comport with each examinee's actual performance, albeit again without using any particular mathematical model in the decision tree, nor in the latent trait estimation.

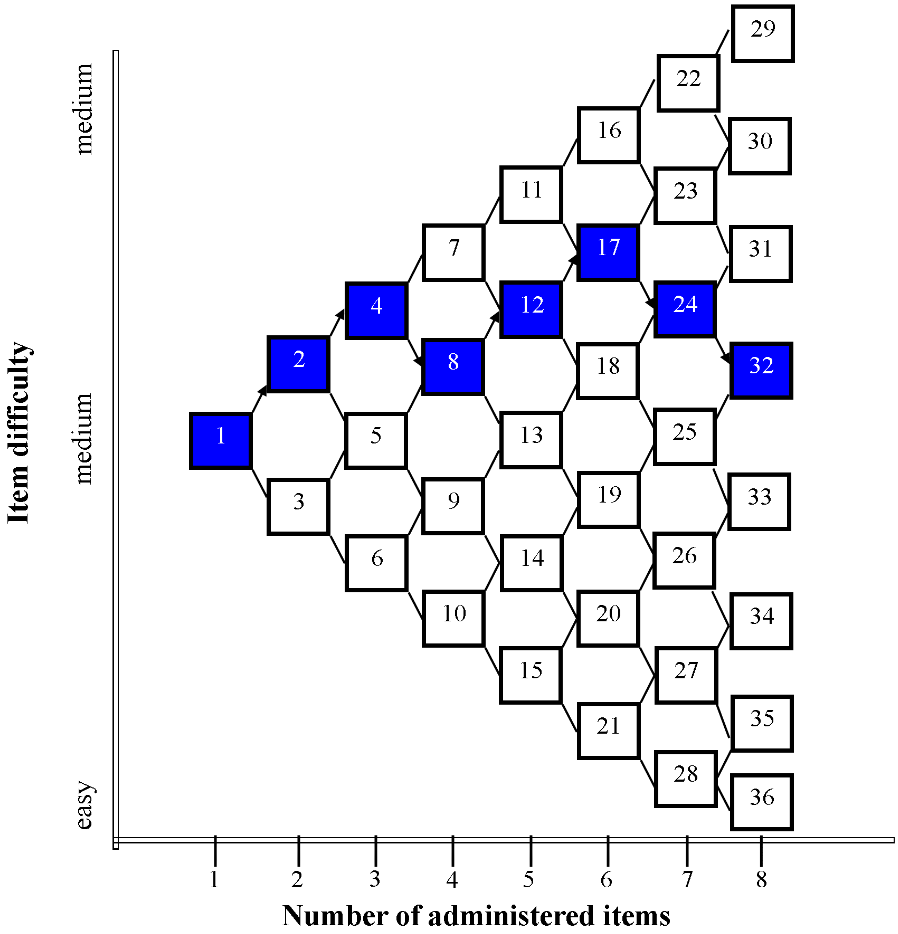


Figure 13 – Example of pyramidal adaptive testing format

As Figure 13 depicts an item with intermediate difficulty is administered to each test taker first. In the case of providing a "correct" response, the examinee is channeled to a more difficult item in sequence item by item. In the case where the examinee provides an incorrect answer, they are channeled to an easier item. This process is repeated until the examinee has responded to 8 items. Lord (1971) developed the flexilevel test, which is basically a variation to both of the abovementioned formats (two-stage, pyramidal). A detailed description of the proposed flexilevel testing algorithm is not essential for the present discussion. The important thing is that in the flexilevel format, like the other two formats, each examinee responds to only a specific subset of items from the complete test, and as they progress through the testing format the actual responses to the selected items are taken into account.

Generally, all adaptive testing formats discussed above, as well as other formats that do not rely on an explicit mathematical model, also referred to as fixed-branching adaptive testing formats (de Ayala, 2009; Patience, 1977), use pre-specified fixed patterns of item selection procedure to match the test to the examinee's level of the latent trait (Reckase, 1989). Fixed-branching testing formats, however, are suboptimal with regard to both the item selection and trait estimation. Variable-branching adaptive testing formats, on the other hand, typically use an IRT model as a theoretical and mathematical base to address the issues of item selection and trait estimation in a more methodologically rigorous way. Unique features of IRT-based variable-branching adaptive testing eliminate some of the problems inherent in fixed-branching adaptive techniques. For example, difficulties of the test items are expressed in the same metric as the latent trait estimates in IRT-based adaptive testing, allowing for a more precise and flexible definition of item selection than fixed-branching algorithms. Moreover, in addition to the difficulties, the item selection process in IRT-based variable-branching testing can take into account other very useful item characteristics (discrimination, guessing parameter). Unlike the fixed-branching adaptive procedures, the IRT-based variable-branching techniques provide a means for the researcher/examiner to control the precision of the trait estimates. Thus, instead of specifying a number of items to be administered just as in fixed-branching procedures, one can specify a required level of measurement precision as a test termination criterion within IRT-based variable-branching testing. In other words, an IRT-based testing process using variable-branching approach can be terminated as soon as a particular degree of reliability is obtained (de Ayala, 2009; Urry, 1977). This approach provides a means to achieve genuine equiprecise measurement where error of measurement is distributed uniformly along the latent continuum.

Because of the extensive computations involved in the process of item selection and trait estimation, variable-branching adaptive testing has been (almost) exclusively implemented on computers. The first practical applications of variable-branching adaptive formats based on the modern test theory were therefore delayed until inexpensive but powerful computers became available to the research community. The fast processing speed (and ability to handle complex matrix algebra algorithms) provided a means for immediate, real-time item selection and trait estimation leading the way to full implementation of IRT-based computerized adaptive testing with real-world applications (Gershon & Bergstorm, 2006). One of the first computerized adaptive tests to be developed by the Naval Personnel Research and Development Center in the mid 1980's, was the Armed Services Vocational Aptitude Battery (Wainer, 2000). This pioneering effort was shortly afterwards followed by the implementation of a CAT version of 1) the National Council of State Boards of Nursing licensing exam and 2) the Graduate Record Examination (van der Linden & Glas, 2010). Use of the CAT has increased substantially since that time, not only in education (Weiss & Kingsbury, 1984) and psychology (Waller & Reise, 1989), but more recently in the field of health-related outcomes (Fayers, 2007). In contrast to other behavioral and social sciences, application of CAT in Kinanthropology has been minimal with only a few published exceptions (Zhu, 1992; Zhu, Safrit, & Cohen, 1999).

Use of the CAT approach has been popular methodology particularly to establish equiprecise measurement, especially given that it reduces examinees burden by shortening a testing procedure. Moreover, Wainer (2000), Gershon and Bergstorm (2006) and van der Linden and Glas (2010) summarize some other benefits of a CAT that were well received especially during its early applications:

1. The possibility for examinees to schedule tests at their convenience since the tests could be offered continuously in real time.
2. Tests are taken in a more comfortable setting and with fewer people around than in large-scale paper-and-pencil administrations. Individuals can work at their own pace, and the speed of response (from point of item presentation on a screen) can be used as additional information in assessing proficiency.
3. Electronic processing of test data and reporting of scores are considerably faster. Computer-based tests allow for easy updating of test content and fast centralization of data.
4. Test security is usually improved. There are no paper documents to be copied, and computerization allows for greater randomization in test items presented to any single examinee.
5. Wider range of questions and test content can be put to use. Graphical images and multimedia presentations enable the measurement of concepts not possible with text-only formats.

6 TESTING ALGORITHMS IN UNIDIMENSIONAL IRT-BASED CAT

As already mentioned, every administration of a test follows a specific testing algorithm, which determines the precise order of test items administered to the examinee and when to terminate the testing process. Thissen and Mislevy (2000) identify three general steps in development of a testing algorithm: 1) how to START, 2) how to CONTINUE and 3) how to STOP. Table 2 summarizes the specific issues that need to be addressed within each of the three steps in the context of the IRT-based CAT.

Table 2 – CAT algorithm

Algorithm step	Procedure
1. STARTING	Initial latent trait estimation
	Selecting/administering the first (or first few) item/s
2. CONTINUING	Selecting/administering "the best" item for provisional latent trait estimate
	Processing the response to obtain interim latent trait estimate
3. STOPPING	Checking whether a specified termination criterion is satisfied
	Reporting final latent trait estimate

The development of a CAT thus requires a series of integrated decisions that address “how initial and interim ability estimates will be calculated, how items will be selected based on those estimates, and how the final ability estimate will be derived” (van der Linden & Pashley, 2010, p. 5). Figure 14 portrays the schematic structure for a CAT algorithm, modified according to several sources (de Ayala, 2009; Štochl, Böhnke, Pickett, & Croudace, 2016b; Thissen & Mislevy, 2000). A more detailed description of some methods and procedures (item selection, trait estimation, possible stopping rules) that are used in each of the three major steps (starting, continuing, and stopping) will be discussed next. For the purpose of illustration, this material only discusses methods and procedures suitable for unidimensional unconstrained CAT (that is CAT without using exposure control and content balancing methods – see sections

6.4.2 ‘Content balancing’ and 6.4.3 ‘Exposure control’ for more information). Readers interested in multidimensional, constrained, testlet or mastery CAT are referred to several published works: Eggen (2010); Glas & Vos (2010); Mulder & van der Linden (2010); Segall (2010); van der Linden (2010); Vos & Glas (2010).

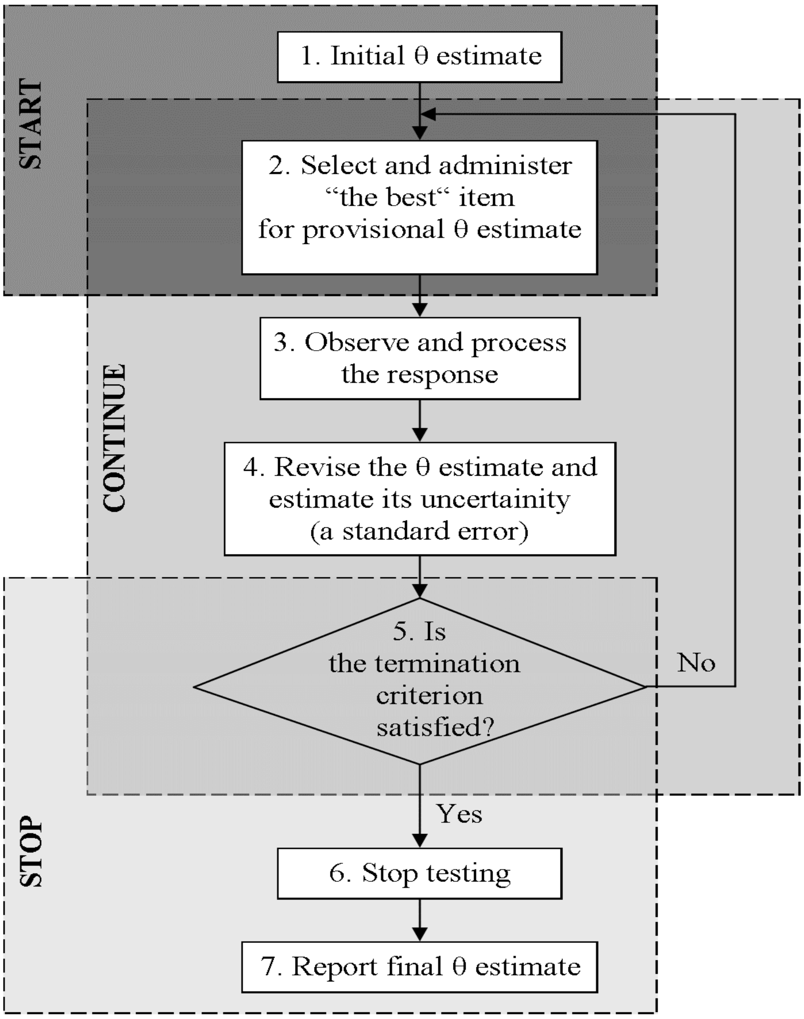

START
CONTINUE
STOP
1. Initial θ estimate
2. Select and administer "the best" item for provisional θ estimate
3. Observe and process the response
4. Revise the θ estimate and estimate its uncertainity (a standard error)
5. Is the termination criterion satisfied?
No
Yes
6. Stop testing
7. Report final θ estimate

Figure 14 – Schematic structure for CAT algorithm

6.1 Starting

In IRT-based adaptive testing formats, items are generally selected and administered to an examinee based on that individual’s previous responses, which serves to maximize the information for the most current θ estimate. However such an approach is not applicable at the beginning of a testing process, since the examinee has provided no responses, thus making it impossible to estimate the latent trait θ. In practice, the initial θ level for a particular examinee is therefore arbitrarily chosen by the CAT administrator. Frequently, the mean in the population of the tested persons is used as the initial θ value for each examinee. Indeed, if there is no other information about the particular examinee available, the population mean is therefore the most reasonable choice (Thissen & Mislevy, 2000). However additional background information about an examinee including demographic factors (e.g., age, race, gender) as well as prior test results also may be used to generate the initial θ guesstimate (van der Linden & Pashley, 2010). For example, the mean of some more narrowly specified population may be used as the initial θ value for a particular test taker. Needless to say, the more accurate the initial θ guesstimate for a given examinee (and the greater precision used to generate this estimate), the faster the CAT converges to his or her 'true' θ estimate (de Ayala, 2009).

Once the decision on the initial θ values has been made, it is possible to select and administer the first psychometrically optimal item. The most widely used item selection approach is based on the Fisher information function (defined by the example provided in Equation 19 or 21) and selects the item with the highest Fisher information at a provisional level of θ. However, in situations where the initial provisional θ value is the same for each examinee, then the same optimal item would be administered to all examinees first. Several approaches are used commonly to avoid overexposure of the first item and the related problems of CAT security and validity. For example, a random initial θ value within a reasonable range (e.g., from -1 to 1) is assigned to each examinee and then the most informative item is selected first. Alternatively, an adaptive algorithm may randomly select an item from, for example, 10 most informative items or possibly from an entire item pool. Fortunately, the choice of initial θ value and the selection of the first item have only negligible, if any, impact on the final CAT estimates even in tests containing as few as 20 - 30 items (Lord, 1977; van der Linden & Pashley, 2010).

Once an examinee's response to the first item is obtained this information is then used to generate the revised θ estimate. Generally both the likelihood and the Bayesian methods described in previous sections (see section 4.5.1 Latent trait estimation) can be used for the θ estimation. Recall, that there are two alternatives of both the likelihood and the Bayesian estimation methods currently in use: Maximum Likelihood Estimation (MLE – uses the maximum of the likelihood function as the θ estimate) vs. Weighted Likelihood Estimation (WLE – uses a mean of the weighted likelihood function as the θ estimate) and Expected a Posteriori (EAP – uses a mean of the posterior distribution as the θ estimate) vs. Bayesian Modal Estimation (BME – uses the mode of the posterior distribution as the θ estimate).

As discussed earlier, however, the MLE does not provide finite θ estimates for so-called perfect (maximum possible simple sum) or zero scores (minimum possible simple sum). Since such a score is likely to occur after the first administered item, MLE often cannot be used at the beginning stages of the testing process. Several approaches have been proposed in cases of the perfect or zero scores when using MLE within a CAT paradigm. One of them is a step-size procedure, in which the initial θ value is decreased (in case of incorrect responses), or increased (in case of correct responses) by a given fixed value (e.g.: 1 logit) until the finite estimates are available (until the mixture of correct and incorrect responses is observed).

In a Bayesian approach (EAP and BME), the θ point estimate is based on the posterior distribution defined in (14), which is simply the product of the likelihood function and the prior distribution. One advantage of Bayesian estimation methods is that they can be used even in situations when the response to the first administered item falls into the first or the last item category. However both the EAP and BME require specification of the prior distribution and using an inappropriate prior can bias the provisional θ estimates and thus lead to selection of suboptimal items, especially with short tests (van der Linden & Pashley, 2010). Moreover, recall that for a uniform prior, the BME is equivalent to the MLE and therefore cannot be used at the beginning of the testing process.

6.2 Continuing

After the initial phase of the adaptive testing, the algorithm requires instruction how to continue, in other words, which item in the test sequence will be administered next. As noted above, the most popular approach is to select the item with the highest Fisher information function (see section 4.6 'Information and standard error of the θ estimates') at the provisional θ point estimate. It should not be surprising, however, that the θ point estimates at the beginning of the testing procedure have relatively large errors and therefore item selection based on maximizing the Fisher information may favor "the best" item at the wrong θ location. Lord and Novick (1968; see also van der

Linden, Pashley, 2010) refer to this situation as ‘attenuation paradox’, which is illustrated in Figure 15.

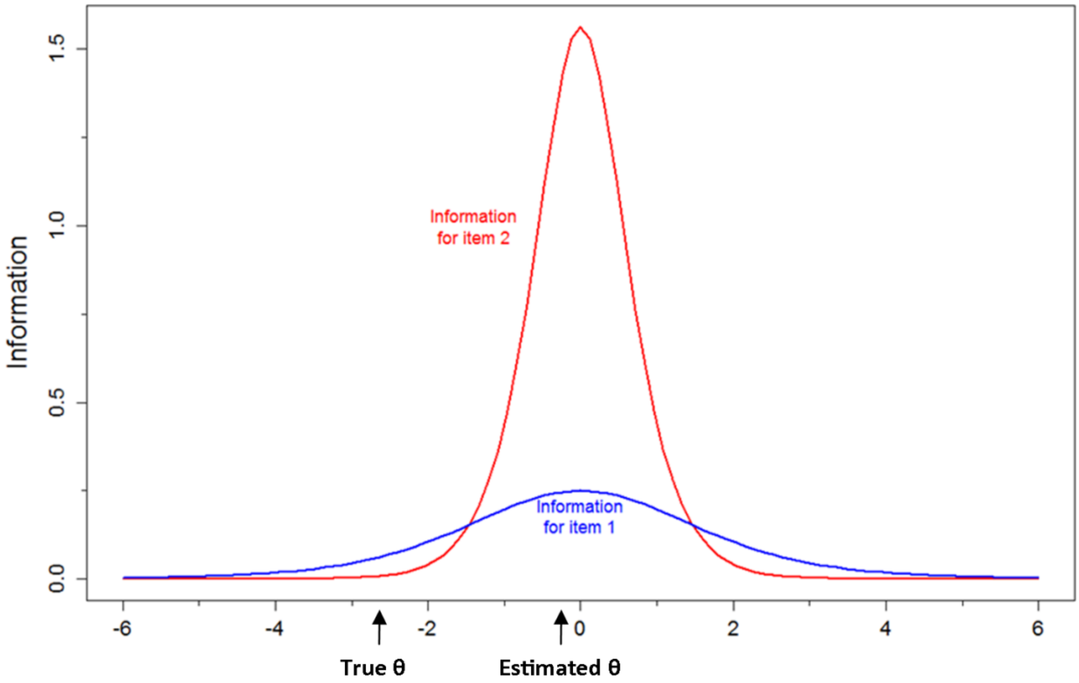


Figure 15 – Attenuation paradox in CAT item selection (adopted from van der Linden & Pashley, 2010)

It should be obvious that if the error of the θ point estimate is large, the true θ value may be considerably different from the current estimate. This also means that the best item for the current θ estimate (Item 2 in Figure 15) may be less informative at the true θ value. To deal with uncertainty in obtaining θ estimates (especially in the beginning stages of testing), Veerkamp and Berger (1997) introduced another item selection method based on weighted Fisher information. These authors proposed to weight each item’s Fisher information $I_i(\theta)$ defined in (19) or (21):

$$I_i^w(\theta) = w\, I_i(\theta) \quad (23)$$

The weight w in (23) is simply the likelihood function for observed responses defined in (12), which “expresses the plausibility of various values of θ given the data” (van der

Linden & Pashley, 2010, p. 15). The item with the highest integral (Int_i) of the weighted Fisher information:

$$Int_i = \int_{-\infty}^{\infty} I_i^w(\theta)d(\theta) \quad (24),$$

is then administered to the particular examinee. Figure 16 illustrates the difference between the traditional item selection based on maximizing item's information at the ML point estimate and the item selection method using the weighted Fisher information.

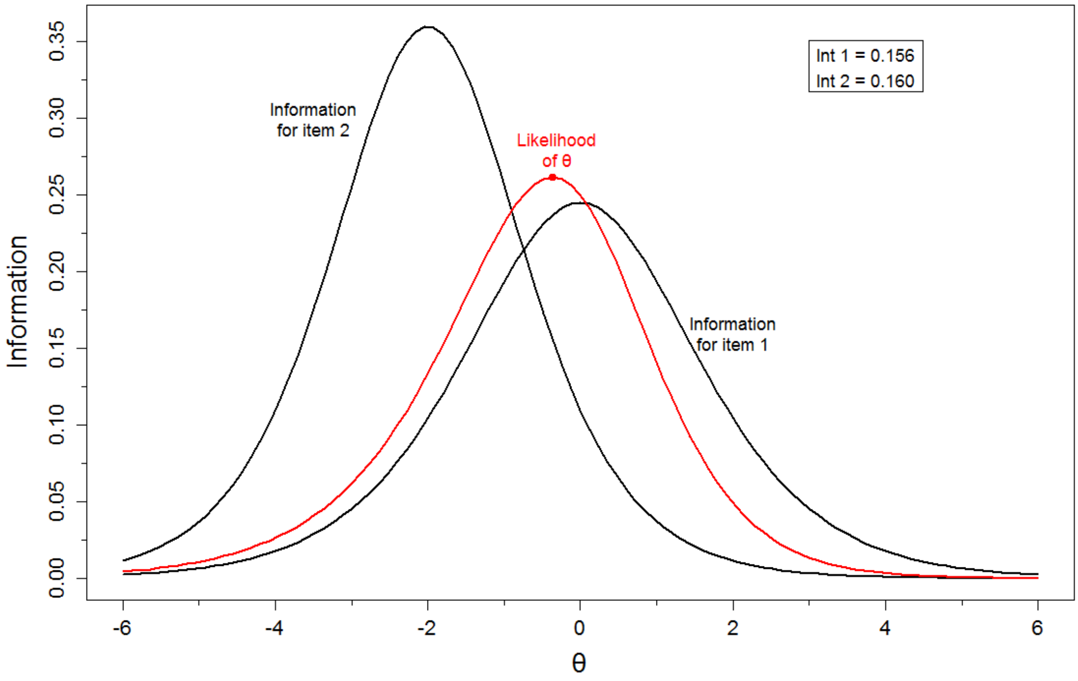


Figure 16 – Incorporating likelihood of θ within the weighted Fischer information item selection approach

For the provisional ML θ point estimate in Figure 16 (which is the highest point of the likelihood function for θ and is situated at θ = -0.37), item 1 provides more information (I_1 = 0.237 as compared to item 2 with I_2 = 0.159) and therefore would be selected for administration within the traditional item selection approach. However if the likelihood of all possible θ values is used to weight the item's information in Figure 16, the integral of the weighted Fisher information would be larger for item 2 (Int_{i2} = 0.160 vs.

Int$_{i1}$ = 0.156 for item 1) and consequently, item 2 would be administered if using the weighted information approach. Especially in the beginning of the testing process, when the likelihood function tends to be flat, a substantial weight is placed on item's information values away from the current θ point estimate (van der Linden & Pashley, 2010).

Chang and Ying (1996) proposed another class of item selection methods based on Kullback-Leibler (KL) information. Generally, KL information measures a difference between two distributions and the larger the KL information, the more different the two distributions are making it easier to discriminate between them (van der Linden & Glas, 2010). The rationale of the Kullback-Leibler selection methods within CAT is to compare distributions associated with the true θ value of the test taker (θ^*) and the current θ estimate ($\hat{\theta}$). For a dichotomous item *i* (*i* = 1,..., n) with parameters based on the IRT model defined in (3), the KL divergence measure (*KL$_i$*) would be computed as follows:

$$KL_i(\hat{\theta},\theta^*) = P_i(\theta^*)\log\left[\frac{P_i(\theta^*)}{P_i(\hat{\theta})}\right] + [1-P_i(\theta^*)]\log\left[\frac{1-P_i(\theta^*)}{1-P_i(\hat{\theta})}\right] \quad (25),$$

where P_i is the probability of a trait indicating response for a given θ (see section 4.2 Unidimensional dichotomous IRT models). In case of the polytomous items and the GRM (see section 4.3 'Unidimensional polytomous IRT models'), which will be used in the empirical part of the thesis, KL divergence information simplifies to:

$$KL_i(\hat{\theta},\theta^*) = \sum_{k=0}^{k} P_{ik}(\theta^*)\log\left[\frac{P_{ik}(\theta^*)}{P_{ik}(\hat{\theta})}\right] \quad (26),$$

where P_{ik} is the probability of responding to a particular category *k* (*k* = 0, 1,…, K) of an item i. Because the true values θ^* are unknown in practical testing situations, computation of the KL divergence information uses a reasonable interval around the most current trait estimate $[\hat{\theta}-\delta, \hat{\theta}+\delta]$, while δ may vary or may be fixed during CAT. The term $\hat{\theta}+\delta$ is used subsequently in lieu of the true θ^* and the current estimate is set to $\hat{\theta} = \hat{\theta}-\delta$ in Equations 25 and 26. Several KL divergence-based criteria then can be specified, which can be used to guide the item selection process within CAT (Nydick, 2014):

- The Pointwise KL divergence criteria based on the comparison of KL information at two points:

$$crit_i = KL_i\left(\hat{\theta} - \delta, \hat{\theta} + \delta\right) \quad (27)$$

  If δ is kept constant over the CAT procedure, the item selection method is referred to as Fixed Pointwise KL (FP-KL) selection. Often δ is set to vary as a decreasing function of the number of items administered to an examinee so far and such an approach defines Variable Pointwise KL (VP-KL) item selection method.

- The integral KL divergence criteria based on the integration of the KL information across a small area:

$$crit_i = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} KL_i(\hat{\theta} - \delta, \hat{\theta} + \delta) d(\theta) \quad (28)$$

  Again, keeping δ constant defines Fixed Integral KL (FI-KL) selection method while setting δ as a decreasing function of the number of items administered to an examinee is typical in Variable Integral KL (VI-KL) item selection.

In all of the KL item selection methods (FP-KL, VP-KL, FI-KL, VI-KL) the item with the highest *crit*$_i$ is administered to an examinee at any particular step in the CAT process.

Alternatively, Bayesian approaches may be used in the item selection process during CAT. The Bayesian item selection methods involve various characteristics of a posterior θ distribution, which is a combination of a θ likelihood function and a prior θ distribution. The Bayesian methods include: Minimum Expected Posterior Variance (MEPV, see also approximate Bayes procedure introduced by Owen, 1969); Maximum Posterior Weighted Information (MPWI, van der Linden, 1998); Maximum Expected Information (MEI, van der Linden & Pashley, 2000) and Maximum Expected Posterior Weighted Information (MEPWI, van der Linden & Pashley, 2010). The detailed description of these methods is beyond the scope of the current review, however the interested reader is referred to van der Linden and Pashley (2010); Mulder and van der Linden (2010); van der Linden (1998) for more information.

6.3 Stopping

The process of selecting an optimal item and updating the interim latent trait estimate based on the response to a selected item is repeated until a prespecified criterion is met. There are two general classes of termination criteria in CAT, which in turn define two different types of CAT: fixed-length CAT and variable-length CAT. In fixed-length CAT, the testing is terminated when a pre-specified number of items has been administered to the examinee. As a result the same number of items is presented to all examinees, which facilitates the interpretation and explanation of the results to the test takers (Boyd, Dodd, & Choi, 2010; Thissen & Mislevy, 2000). The disadvantage of the fixed-length CAT stopping rule is that examinees are measured with varying degree of precision – a strategy that is a typical characteristic of standard linear testing formats.

The standard (unconstrained, non-mastery) variable-length CAT termination criteria involve a standard error stopping rule or a minimum information stopping rule (Boyd et al., 2010). Using the standard error stopping rule, the CAT process for a particular examinee is terminated once standard error of his or her trait estimate's drops below some pre-specified value. The standard error stopping rule leads to different number of administered items to each examinee, however it provides the opportunity for equiprecise measurement (Weiss, 1982) – the main assumption of CTT. The minimum information stopping rule terminates testing when there are no more items in the pool with sufficient (predetermined) information for the most current trait estimate. Although this method avoids the administration of low discriminatory (uninformative) items (Boyd et al., 2010), it has been shown to result in many nonconvergent cases due to the low number of administered items (de Ayala, 1989; Dodd, Koch, & de Ayala, 1989).

In practice, termination of a CAT is almost always based on the combination of fixed-length and variable-length stopping rules, because the testing algorithm may select and administer all items from the item pool before the variable-length criterion has been satisfied (Thissen & Mislevy, 2000).

6.4 Practical issues related to item selection in CAT

Although, in theory, adaptive testing is a relatively simple concept, there are several critical issues associated with constructing, implementing, and practically maintaining CAT procedures. Three critical areas that can affect the item selection procedure and consequently influence the entire CAT process are discussed next: item pool, content balancing, and exposure control.

6.4.1 Item pool

In adaptive testing, each examinee is (at least theoretically) presented with a unique test composed of different sub-set of items drawn from a larger item pool (Flaugher, 2000; Thissen, Reeve, Bjorner, & Chang, 2007; Veldkamp & van der Linden, 2010; Zhou & Rackase, 2014). The practical utility of a CAT largely depends on the relative size of an item pool and the psychometric quality of the items contained in the pool.

The size of an item pool is usually driven by the stakes involved in a test (Wise & Kingsbury, 2000). Stakes refers to how important the test is and whether the results are used to make critical if not irreversible decisions, for instance, admissions to a highly selective university, the military, or some professional organization, where a decision supporting rejection is irreversible. It is not uncommon for high-stakes CAT, based on dichotomous IRT models, that the item pool contains more than 1000 items (Nogami & Hayashi, 2010). On the other hand, it has been shown, that in case of polytomously scored items, effective CAT can be performed with as few as 30 items in a pool (Dodd & de Ayala, 1994; Dodd et al., 1989; Dodd, Koch, & de Ayala, 1993). Such small item pools may however be appropriate only for specific purposes – for example in the measurement of attitudes or health outcomes, where item content and exposure specifications are not necessarily needed (Boyd et al., 2010). If the testing process requires content balancing and/or item exposure control to maintain content validity and/or test security, even an item pool containing 200 polytomous items may not be sufficient (Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Pastor, Dodd, & Chang, 2002). Recently, some authors have reported acceptable properties of CAT while using item pools containing 34 dichotomous items (Lu, Lien, & Hsieh, 2015), 16 polytomous items (Petersen et al., 2016), and 50 polytomous items (Devine et al., 2016) when measuring balance, pain, and anxiety, respectively. It is obvious that the item pool size may vary greatly in different fields (e.g., education vs. psychology) and under specific circumstances (high-stakes vs. screening tests).

Psychometric characteristics of the items contained within the item pool also influence the quality of a CAT. One of the advantages of CAT – an equiprecise measurement (precision of the measurement is kept constant along the whole latent scale) – can only be realized with a pool containing high-quality items for all possible levels of the latent trait. In other words, the ideal item pool information distribution for a CAT is the uniform distribution (Nogami & Hayashi, 2010). Moreover there needs to be a good match between the item pool information distribution and the distribution of the trait in the population being tested. Such a requirement would be certainly fulfilled if the item pool information function was uniformly distributed over a sufficiently broad range of the latent continuum (theoretically from minus infinity to plus infinity).

Psychometric characteristics of the items used in current CAT are usually determined by particular IRT model parameters and thus an appropriate IRT model should be selected for calibration of the individual items contained in the item pool. Even when a proper IRT model is used to describe the examinees responses, obtained item parameters are subject to some estimation error. Fortunately, errors in the item parameter estimates do not have serious consequences for CAT as long as the calibration sample is sufficiently large enough (van der Linden & Pashley, 2010).

6.4.2 Content balancing

The IRT model frequently employed in designing CAT assumes unidimensionality, in other words a collection of items should measure only one theoretical latent construct. In practice however, the theoretical construct of interest may be “composed” of subdomains that the test designer believes should represent a single general latent construct. For example, even when assessing a general construct of mathematical ability, a test designer may consider that mathematical ability is composed of two sub-components consisting of algebra and geometry knowledge. Consequently a test (or item pool) supposed to capture mathematical ability should contain items assessing both algebra and geometry. Content balancing methods within a unidimensional CAT ensure that the item selection algorithm adheres to the desired content of a test (Gershon & Bergstorm, 2006).

Introduction of content balancing in CAT was mainly motivated by the fact that item selection procedures based purely on statistical criteria (discussed above) may lead to undesirable item usage. Statistically motivated item selection procedures favor the items with high discrimination parameters and thus the assembled test may not cover all desired content areas equally well (Leung, Chang, & Hau, 2003; Zheng, Chang, & Chang, 2013). Many strategies of content balancing for application with CAT have been proposed to solve this problem. For example the simplest one – a rotation method – is based on systematic rotation of test content areas from which an item is then selected according to some item selection criteria (e.g.: item with maximum information function is selected within a given content area). Another popular method proposed by (Kingsbury & Zara, 1989) compares desired proportions of test content areas with actual proportions of content areas during the CAT (e.g.: 70% of items should represent algebra knowledge and 30% of items should represent geometry knowledge in order to guarantee the content validity of a test measuring general mathematical ability). The next “best” item is then selected from a content area with the highest discrepancy between the current and the desired content areas proportions. Leung, Chang, and Hau (2000) proposed a modification of Kingsbury and Zara’s approach, in which all the content areas not satisfying the desired administration percentage are considered in the item selection mechanism.

Other, perhaps more sophisticated and/or complex content balancing methods (weighted deviations model, modified multinomial model, CAT with shadow tests) were developed to meet specific practical testing goals. The description of these methods is beyond the scope of this review – an interested reader is therefore referred to Leung et al. (2003); Stocking and Swanson (1993); van der Linden (2010); van der Linden and Chang (2003) or van der Linden and Reese (1998), among others.

6.4.3 Exposure control

One of the limitations of unconstrained item selection methods is that it may create an undesirable use of the item pool, including overuse of items with high discriminatory power (Thissen & Mislevy, 2000). Exposure control techniques prevent certain items from being administered to a large number of examinees in order to maintain the test security and reliability (Leung et al., 2003). Exposure control also ensures better item pool utilization, however, at least theoretically, it decreases the efficiency of the adaptive testing process (van der Linden & Glas, 2010). Exposure control is especially important in high-stakes CATs that are being utilized for admission, licensure, selection, and so on. On the other hand, the exposure control is generally not an issue in low-stake CATs such as personality and attitudes assessment, health measures or quality of life surveys (Boyd et al., 2010; Gershon & Bergstorm, 2006).

Exposure control methods used in early applications of CAT were based mainly on randomization. For example the randomesque exposure control method (Kingsbury & Zara, 1989) selects a specified number of optimal items and then one of the items is randomly administered to a particular examinee. Another randomization method, "modified within 0.1 logits" (Davis & Dodd, 2003, 2008) selects a specified number of optimal items (e.g., most informative) at three points on the latent scale: a) one-third of items at the interim trait estimate, b) one-third of items at interim trait estimate plus 0.1 logits and c) one-third of items at interim trait estimate minus 0.1 logits. Again, one of the items selected using this method is then administered to the examinee. Conditional exposure methods, including the Sympson-Hetter (Sympson & Hetter, 1985) procedure, control the exposure rate of the items by using an exposure parameter for each item in the pool. The value of the exposure control parameter is determined empirically (through CAT simulations) to limit the usage of an item to a priori desired/defined maximum allowed exposure rate (Thissen & Mislevy, 2000). Chang and Ying (1999) proposed an exposure control method based on dividing the item pool into several strata according to item discrimination parameters. A major rationale of this *a*-stratified method is that items with low discriminatory power are likely to be more efficient over a broader range of the latent scale (van der Linden & Pashley, 2010). This is especially useful at the beginning of the testing, when the error of the latent trait estimate tends to be large. Therefore within the *a*-stratified method, items are selected from strata with low discriminatory items in early stages of the testing, while highly discriminating items are saved for later stages where the trait estimate becomes more accurate (Boyd et al., 2010; Leung et al., 2003).

For more comprehensive and detailed description of exposure control methods used in contemporary CAT see for example Boyd et al. (2010) or van der Linden and Glas (2010).

6.5 Evaluation of item selection and trait estimation methods used in computerized adaptive testing algorithms

Practical guidelines related to item selection and trait estimation methods used in CAT algorithm are not yet well established. Van der Linden and Pashley (2010) noted that both posterior-based (Bayesian) and likelihood-based item selection and latent trait estimation methods lead to identical asymptotic results, converging to the true value of θ (see also Chang & Ying, 2009). However the identical asymptotic properties of the methods may not hold in practical implementation of CAT, when much shorter tests are usually administered. This is perhaps the motivation behind using simulation studies that compare latent trait estimators and item selection criteria.

For example, Veerkamp and Berger (1997) showed that for dichotomous items the weighted Fisher information item selection was superior to the traditional item selection approach based on maximizing the information at a given point estimate. Comparison of the weighted and the traditional Fisher information item selection approach showed there was only a trivial difference in measurement efficiency while using polytomous generalized PCM (van Rijn, Eggen, Hemker, & Sanders, 2002). The difference between the two approaches was observed particularly at the beginning of the CAT process, when the number of administered items was relatively small (e.g., < 10) (Boyd et al., 2010).

The Kullback-Leibler item selection methods were compared to the traditional and the weighted Fisher information approach while using generalized PCM (Veldkamp, 2003) and nominal response model (Passos, Berger, & Tan, 2007) respectively. The simulation studies indicated that even when administering a relatively short test (e.g. 15 to 20 items) the three item selection methods produce very similar results. For instance, there was 85% to 100% overlap in selected items and the mean square error difference of the θ estimate was negligible across all three item selection methods (Boyd et al., 2010).

Van der Linden (1998) reported bias and mean square error functions when comparing traditional Fisher information approach and four Bayesian item selection procedures (MEPV, MPWI, MEI and MEPWI) for dichotomous test items. Bias functions for all of the methods were almost identical as soon as at least 20 items were administered. Similarly, differences in mean square errors when administering 20 items did not exceed 0.15 across the whole latent continuum for all item selection criteria.

Thus even though some authors (van der Linden & Glas, 2010) recommend using the Bayesian item selection methods in any case, their practical benefits over the traditional Fisher information methods may not be realized unless the test is very short (e.g. 5 items).

Choi and Swartz (2009) examined the properties of both Fisher information based (maximum Fisher information, likelihood weighted Fisher information) and Bayesian (MEPV, MPWI, MEI and MEPWI) item selection procedures under polytomous GRM. They concluded that for polytomous items, the advantages of Bayesian selection methods demonstrated with dichotomous items may be masked in many practical settings given the usual polytomous item pool size. This finding led the authors to suggest that for item banks with a relatively small number of polytomous items (e.g. 30 to 70 items), any of the different estimation methods considered in their study are appropriate.

Several other studies also focused on the comparison of latent trait estimation methods used in CAT. For example, using RSM, Chen, Hou, Fitzpatrick, and Dodd (1997) reported highly congruent latent trait estimates when using MLE, EAP with normal prior and EAP with uniform prior distributions, respectively. Moreover the CAT latent trait estimates based on these methods highly correlated with MLE estimates obtained from using all of the items in the pool. Chen, Hou, and Dodd (1998) reached very similar conclusions.

S. Wang and Wang (2001, 2002) conducted an extensive comparison of MLE, WLE, BME and EAP estimation methods for unconstrained CAT using GRM and generalized PCM, respectively. The authors specified a number of various conditions with respect to test length, item pool size and stopping rules. It has been shown that WLE performed systematically better than MLE across all conditions and for both polytomous ITR models. Using fixed-length stopping rules the WLE method outperformed the BME and the EAP estimation procedures. On the other hand, in variable-length CAT (e.g., test is terminated when a prespecified precision is reached) EAP and BME led to lower number of administered items when compared to WLE and MLE.

Štochl et al. (2016b) evaluated how several different combinations of CAT settings may affect the efficacy of adaptive testing when measuring general psychological distress using a bi-factor model. The authors concluded that using MLE, EAP and BME latent trait estimators as well as using Fischer and Kullback-Leibler item selection methods led to very similar results. The conclusion was replicated in another study, where Štochl et al. (2016b) used CAT for assessing longitudinal change in psychological distress.

Obviously, there are many different approaches that can be used in CAT item selection and latent trait estimation. The results of simulation studies to date certainly provide important insight into the optimal functioning of these methods; however, since in practice CAT applications can differ greatly in many ways (required length of a test, size and quality of the item pool, the size and composition of a calibration sample, content balancing and/or exposure control requirements, to name a few), simulation results may not be applicable to all possible practical situations. Thus, when a test developer wants to implement a new CAT, it is recommended "to study a few feasible arrangements in order to identify a suitable, though not necessarily optimal, solution for a planned adaptive testing program" (van der Linden & Pashley, 2010, p. 28).

7 EMPIRICAL PART – PROBLEM STATEMENT

Computerized adaptive testing (CAT) represents a novel, efficient, and cost-effective approach to test administration. Previous chapters have outlined the potential advantages and applications of this methodology, highlighting in particular the psychometric advantages and improvements in testing efficiency. Although CAT was initially developed with an emphasis on ability and achievement testing (Wainer, 2000), its application and popularity has garnered considerable attention in medicine as well as the social and behavioral sciences. Medical applications include measurement of patient-reported outcomes (Štochl et al., 2016a, 2016b; Zheng et al., 2013), and both education and psychology are replete with examples of CAT used to advance assessment of aptitude (Verschoor & Straetmans, 2010), attitudes (Hol, Vorst, & Mellenbergh, 2005; Koch, Dodd, & Fitzpatrick, 1990), personality (Reise & Henson, 2000; Simms et al., 2012) and even in settings involving clinical diagnostic assessments (Fliege et al., 2005; Fliege et al., 2009). Despite its growing popularity in various disciplines, the same widespread use of CAT has not witnessed similar growth in Kinanthropology (Gershon & Bergstorm, 2006). This noted gap in the literature is a driving force in the current exploration of CAT with the Physical Self-Description Questionnaire (PSDQ – Marsh, Richards, Johnson, Roche, & Tremayne, 1994), a well-validated assessment frequently used to measure physical self-concept in the field of Kinanthropology.

In general, the application of CAT offers the unique possibility of improving both the efficiency and psychometric qualities of a test. For instance, CAT can lead to reduced test length, reductions in administration cost, improved reliability, and reduced response burden, all of which contribute to the commercial uptake and utilization of a test. Notwithstanding these potential gains, several critical issues need to be considered prior to the practical implementation and continued use of CAT. For instance, still unresolved is whether a CAT approach is advantageous over traditional fixed-length paper-and-pencil testing when assessing physical self-concept. Related to this, little is known whether transforming a test to a CAT format reduces test length without loss to measurement precision (Embretson & Reise, 2000). Moreover, it is not known whether item reduction would be significant for practical testing purposes. And if so, it remains to be seen whether the measurement process is precise enough allowing to make valid conclusions about examinees' performance. Switching to CAT administration without addressing these and relevant ‘psychometric’ questions “runs the danger of being inefficient at least and legally indefensible at the worst” (Thompson & Weiss, 2011, p. 2). Fortunately, many such problems can promptly be addressed applying well known psychometric techniques with simulation studies (Thompson & Weiss, 2011; van der Linden & Glas, 2010). The present thesis addresses several of the issues raised above using an empirical example with Monte Carlo simulation; thus advancing the use of CAT with applications in testing and psychosocial assessments.

8 AIMS AND HYPOTHESES

The current thesis introduces the use of CAT applied to the field of Kinanthropology. The overall utility of CAT is demonstrated empirically via a controlled simulation study demonstrating how CAT shortens administration of a self-report fixed-length questionnaire routinely used to assess physical self-concept. Related to this first aim, the present study also evaluates the efficiency of different parameter estimation and item selection methods commonly encountered with CAT. This latter refinement offers the potential to assess the influence of varying distributional properties and test administration features on measurement efficiency and precision using CAT methodology.

Specifically, in the empirical part of the thesis, I present findings from CAT simulation of the PSDQ. The simulation study, which is described in the subsequent chapters, aimed to compare a) the number of administered items from PSDQ (test length) and b) accuracy of estimated latent levels of physical self-concept, while using a variety of latent trait estimation methods, items selection algorithms, stopping rules, and distributional properties. The specific study hypotheses include:

a) Kullback-Leibler divergence-based and Fisher information-based item selection methods will both produce similar number of administered items from the PSDQ,
b) the expected a posteriori trait estimation method will lead to a smaller number of administered items than the maximum likelihood latent trait estimation method,

c) using the uniform true latent trait distribution will lead to higher number of administered items from the PSDQ than using the standard normal true latent trait distribution, and

d) bias of the estimated latent levels of physical self-concept will be similar across the latent trait estimation methods (expected a posteriori vs. maximum likelihood estimation method) as well as across the item selection methods (Kullback-Leibler vs. Fisher information selection method) used in the simulation study.

9 METHODS

Many important issues related to the practice of CAT such as estimation of the test length or score precision can be addressed with simulation studies. In this chapter I describe a simulation study that was designed to evaluate the usefulness of CAT when applied to the PSDQ.

Recall from previous chapters that for any application of CAT there are two broad components: (1) calibrated item bank and, (2) testing algorithms in the CAT system (including starting point of the CAT, item selection and latent trait estimation procedures, and termination of a test – see Table 2 and Figure 14). Therefore, the process of implementation of CAT on the one hand includes creating an item bank (or possibly utilizing the existing item bank – e. g., a validated paper-and-pencil instrument) and on the other hand rendering a decision on the corresponding model (usually an IRT model) according to which the item parameters (difficulty or threshold parameters and discrimination parameters) are obtained. An important step toward achieving the first CAT component (calibrated item bank) requires conducting analyses to determine the underlying dimensionality of the assessment, since IRT-based CAT assumes unidimensionality of the items in a test or generally in the item pool (unless multidimensional IRT models are employed). Once the item bank has been developed and calibrated, testing algorithms that will guide the adaptive testing process itself (how to select the first and next item/s, how to estimate the interim and the final trait level, when to stop testing) need to be specified in order to make a CAT administration practically usable. Especially at this point, simulation studies are essential to determine the most appropriate settings (starting point, termination criterion, …) and methods (item selection method, latent trait estimation method). As Thompson and Weiss (2011) noted, a CAT utilizing arbitrary chosen specifications and methods "without adequate research in form of these simulation studies is substantially less defensible" (pp. 5) and may, for example, "result in examinee scores that are simply not as accurate as claimed, providing some subtraction from the validity" (p. 2).

Typically, CAT simulation studies operate on the basis of simulating the process of adaptive administration using a calibrated item bank under varying conditions (e.g. different item selection method, termination criterion). Since the adaptive administration is guided by the answers/responses to previous items contained in an item bank (see Figure 14), the means of obtaining the answers/responses defines two general types of simulation studies that are predominantly used in CAT research: (1) Monte Carlo simulation studies, and (2) post-hoc simulation studies (Nydick & Weiss, 2009; Thompson & Weiss, 2011). These approaches differ because Monte Carlo simulations utilize stochastically generated responses to the items based on the item parameters, latent trait estimates (θ) and assumed IRT model, whereas in post-hoc simulations examinee's real responses are used. In post-hoc CAT simulations, an adaptive algorithm simulates the item selection procedure, but the responses to the items in the item pool are already known. Monte Carlo simulations are based on the fact that IRT models provide an estimate of the exact probability of responding to a specific item category for a given value of θ. Thus having a calibrated item bank and a vector of θ's allows researchers to easily generate an entire response data matrix, which can be used for the purpose of the simulated adaptive test administration. Moreover, values of the θ parameters in Monte Carlo simulations can be randomly generated to satisfy a specific distribution (e.g. standard normal distribution), which facilitates the evaluation of how well CAT performs under certain distributional assumptions that may correspond to the test behavior of a specific population.

The current thesis uses a Monte Carlo simulation to evaluate the efficiency and accuracy of a CAT administration using the PSDQ. A real item bank calibrated with an IRT model was used and responses to test items during the adaptive administration were generated based on known item parameters and latent trait values (θ). The latent trait values (θ) were in this case simulated from a desired distribution and served as true values of physical-self description latent construct for 'hypothetical' examinees (simulees). Then the process of adaptive testing – that is in simplified form: selecting "the best" item for the most current θ estimate, revising the θ estimate based on the response to the selected item, and checking whether a criterion for the test termination is satisfied – was simulated using several different CAT algorithm specifications. The next section outlines the integral CAT components (calibrated item bank and testing algorithms) as well as the CAT simulation procedures.

9.1 Item pool, IRT model used for item calibration, dimensionality analysis

9.1.1 General description of the item pool

The 70-item PSDQ provided the item pool for the current simulation study. The PSDQ (see Appendix A) was designed to measure adolescents' (12 years and older) physical self-concept (see Shavelson, Hubner, & Stanton, 1976, for theoretical background, scale construction, and preliminary psychometric evidence). Each PSDQ item employs a six-point Likert-type scale (i.e., false, mostly false, more false than true, more true than false, mostly true, and true); with items scaled in the direction of higher physical self-concept. The PSDQ is comprised of 11 subscales (i.e., health, coordination, physical activity, body fat, sport competence, physical self, appearance, strength, flexibility, endurance/fitness, and self-esteem), all of which have been shown to have acceptable reliabilities (Cronbach's α ranged from 0.81 to 0.94, see Flatcher & Hattie, 2004; Marsh et al., 1994). Construct validation studies using the PSDQ provide evidence of a higher-order factor structure, with 11 first-order dimensions and one second-order dimension reflecting physical self-concept (Marsh, 1996a, 1996b; Marsh & Redmayne, 1994; Marsh et al., 1994).

9.1.2 Item calibration

Flatcher and Hattie (2004) provided empirical estimates for item parameters needed for an IRT-based CAT simulation. Their study involved an Australian sample of high school students (N = 868, ages 13 to 17 years) engaged in sports activities. A GRM defined in Equation 4 and 5 was used to estimate each item's discrimination and threshold parameters, which are provided in Appendix A.

9.1.3 Dimensionality analysis

A reasonable prerequisite of estimating the IRT parameters by a GRM requires that only one general latent factor (dimension) accounts for the association between all 70 test items. In order to test this unidimensional assumption, Flatcher and Hattie (2004) factor analyzed composite subscale scores for each of the 11 PSDQ sub-domains using exploratory factor analysis (EFA). The results of the EFA supported the existence of one general latent factor of physical self-concept that accounted for 47% of the total item variance. A confirmatory factor analysis (CFA) applied to the same 11 PSDQ subscale scores also showed that a single factor solution produced an adequate model fit (RMSEA = 0.032, see Flatcher & Hattie, 2004); lending further support to a unidimensional factor structure for the PSDQ.

9.2 CAT simulation design and specifications

A Monte Carlo simulation was conducted to evaluate the performance of a CAT administration of the PDSQ described above. This type of CAT simulation requires both the latent trait values in addition to the item parameter estimates from the calibration study at hand. Moreover, specific details of the CAT algorithmic component need to be defined. The whole process can be outlined as follows (see also Štochl et al., 2016b):

9.2.1 Step 1. Simulate latent trait values (true θ)

Two samples of 1000 latent trait values (θ) randomly drawn from a) the standard normal distribution N(0,1) and b) the uniform distribution U(-3,3) were obtained. The simulated latent trait values represent the true values of the latent physical self-concept (θ^*) in a sample of 'hypothetical' examinees.

9.2.2 Step 2. Supply item parameters for the intended item pool

Discrimination and threshold parameter estimates from the calibration study need to be provided for the 70 items in the PDSQ – these estimates are listed in Appendix A. The item parameters together with θ^*'s simulated in previous step are used to obtain stochastic responses to the selected items during the simulated CAT administration of the PSDQ (see 9.2.4 Step 4).

9.2.3 Step 3. Set CAT algorithm options

In this step, the algorithmic component of CAT needs to be specified – that is the decision rule indicating how to start (selection of the first item, initial θ estimation method, number of items for a starting phase of the testing), continue (item selection method, θ estimation method), and how/when to stop (termination criterion) the testing process need to be specified. Even though Monte Carlo studies offer a great opportunity to compare different CAT methods and specifications, the manipulated options should be carefully selected to prevent a rapid increase of the simulated conditions (Štochl et al., 2016a). In the current simulation, the following settings and methods were used:

9.2.3.1 Latent trait (θ) estimation methods

The latent trait was estimated using one of the following methods: a) MLE, b) EAP with uniform prior distribution, and c) EAP with standard normal prior distribution. The MLE and EAP were chosen because the aim was to compare the traditional likelihood-based latent trait estimation method with a Bayesian method, the latter which combines the likelihood with prior distribution. To evaluate the effect of the prior distribution on the efficacy of CAT (i.e. number of administered items and accuracy of the latent trait estimates) an informative (standard normal) and an non-informative (uniform) prior within the EAP estimation were selected.

9.2.3.2 Item selection methods

Two item selection methods were adopted in the current simulation: a) unweighted Fischer information (UW-FI) method, and b) fixed-point Kullback-Leibler (FP-KL) divergence-based method (see section 6.2 for details). The δ value within the FP-KL selection procedure was set to 0.1. Both methods select items at a particular (most current) point estimate of the latent trait. At each step of the CAT only the single best item according to a given criterion was considered for the administration. With regard to item selection, UW-FI and FP-KL were selected in order to compare traditional item selection approach (based on Fisher information) with the more recently proposed procedure (based on Kullback-Leibler divergence).

9.2.3.3 Stopping rules

The termination criterion based on the measurement precision cutoff was used in the current CAT simulation since this approach offers the opportunity of creating equiprecise measurement (Weiss, 1982). Equiprecise measurement refers to a situation where the test information is uniformly distributed and thus the reliability of the latent trait estimates is the same for all test takers. In such a case a global measure of reliability which is used within CTT (reliability is a constant within CTT) becomes justified. Number of administered items can vary for each examinee to reach equiprecise measurement within a CAT approach.

In CTT (in the case of standardized values with mean of 0 and SD = 1), the relation between standard error (SE) and reliability can be formalized as $SE = \sqrt{1 - reliability}$. The selected cutoff values of SEs which represent latent trait estimate reliabilities of a) ≈ 0.95, b) ≈ 0.90, c) ≈ 0.85 and d) ≈ 0.80, are therefore equal to a) 0.23, b) 0.32, c) 0.39 and d) 0.45 respectively. Thus the simulated CAT administration continued until the standard error of the θ estimate dropped below the selected cutoff value or until all 70 items from the PSDQ were administered.

9.2.3.4 Overall conditions in CAT simulations

The specifications described above produced a 2 (simulated θ^* distribution: standard normal distribution, uniform distribution) × 3 (latent trait estimation methods: MLE, EAP with standard normal prior, EAP with uniform prior) × 2 (item selection methods: UW-FI, FP-KL) × 4 (stopping rules: SE = 0.23, SE = 0.32, SE = 0.39, SE = 0.45) matrix with 48 overall simulation conditions. Within all of the conditions the initial θ value was kept constant for all hypothetical examinees, the step-size estimation procedure was used for the first two items, and at least 3 items had to be administered before the test was terminated.

9.2.4 Step 4. Simulate CAT administration

Within all of the 48 CAT simulation design conditions, an adaptive administration of the PDSQ was simulated for every single randomly generated true latent trait (θ^*) value (from 9.2.1 'Step 1'). Within the starting phase of each CAT simulated administration, the initial θ level was set to 0 logits (the mean of the distributions) and thus the same item was always administered first. Using the parameters of the selected item and the particular true θ^* value, the stochastic response is obtained and the initial θ value is updated based on the response. To obtain a stochastic response, a uniform random number u_{ij} from U(0,1) is generated for each item/simulated θ^* combination and compared to the model-generated probabilities of responding to a given item category to create a scored response. For instance, in a GRM with a three-category response format for a single item, if $P_{i1}(\theta_j) = 0.7$ and $P_{i2}(\theta_j) = 0.2$ then $P_{i3}(\theta_j) = 0.1$. If the generated random number $u_{ij} < P_{i1}(\theta_j)$ then the scored response for the particular simulated true θ^*_j is the first response category; if $P_{i1}(\theta_j) < u_{ij} < [1 - P_{i3}(\theta_j)]$ then the scored response fits the second response category and if $u_{ij} > [1 - P_{i3}(\theta_j)]$ then the response fits the third response category for a particular item.

A step-size procedure (see section 6.1) was used to "estimate" the latent trait for the first two administered items. Specifically, if a simulated response was in the selected item's first or in the selected item's last response category, the θ value was decreased by 1 logit or increased by 1 logit respectively, otherwise it was held constant. For the updated θ estimate after two administered items, the next item is selected from the item pool (see 9.2.3.2 'Item selection methods') and a stochastic response is obtained again. Given the response, the new θ estimate is calculated, now using one of the latent trait estimation methods listed in step 3 (see 9.2.3.1 'Latent trait estimation methods'), and another item is selected for the updated latent trait estimate. This process is repeated until a specified stopping rule was satisfied (see 9.2.3.3 'Stopping rules').

9.3 Analysis of simulation results

All simulations were performed in the R (R Core Team, 2013) statistical software using the *catIrt* package (Nydick, 2014). The corresponding syntax used in the current study can be found in Appendix B.

The performance of the CATs was evaluated with respect to: a) the number of administered items and b) proximity of CAT-estimated latent trait values ($\hat{\theta}$) to the true simulated latent trait values (θ^*) as well as to latent trait estimates based on the full PSDQ ($\hat{\theta}^{PSDQ}$). To assess such measurement accuracy, the following indices were used:

- Individual latent trait bias

$$Bias(\hat{\theta}_j) = \hat{\theta}_j - \theta_j^* \quad (29),$$

- Mean absolute bias

$$Bias(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\left|\hat{\theta}_j - \theta_j^*\right| \quad (30).$$

In addition, Pearson's correlation coefficient was computed to evaluate the relationship between $\hat{\theta}$ and θ^* and between $\hat{\theta}$ and $\hat{\theta}^{PSDQ}$ for each of the CAT simulation conditions.

A 2 (simulated θ^* distribution) × 3 (latent trait estimation methods) × 2 (item selection methods) × 4 (stopping rules) way ANOVA was used to assess the effect of various simulation conditions on both the test length and absolute bias of the CAT latent trait estimates. Consistent with other related IRT-based CAT studies (Guyer &

Weiss, 2009; Nydick, 2013; Nydick & Weiss, 2009; S. Wang & Wang, 2001, 2002), and given the design of the current study (resulting in N = 48000 observations and thus providing extremely high statistical power), ANOVA was used descriptively to indicate the amount of variance accounted for by each factor in the Monte Carlo simulation. Each ANOVA model specified both main and two-way interaction effects with the eta-squared η^2 statistic used to express effect sizes. The effect size η^2 was interpreted according to Cohen's (1988) recommendations: no effect if $\eta^2 < 0.01$, small effect if $0.01 < \eta^2 < 0.06$, medium effect if $0.06 < \eta^2 < 0.14$, and large effect if $\eta^2 > 0.14$.

10 RESULTS

This section presents results of the CAT Monte Carlo simulations. These simulations aim to evaluate the usefulness of an adaptive administration of the Physical Self-Description Questionnaire. I begin with presenting an assessment of dimensionality as this is prerequisite to unidimensional CAT simulation studies. Subsequently, I present findings related to the overall test length (number of administered items) for the different simulation conditions mentioned previously. The results of these simulations are presented in tabular form, as plots, and statistical information (variance decomposition) from the ANOVA procedures. I then describe the accuracy of CAT for different simulation conditions while presenting bias of the $\hat{\theta}$ estimates and also the correlations of the CAT $\hat{\theta}$ estimates with the true θ^* values and $\hat{\theta}^{PSDQ}$ estimates (based on all items in the PSDQ) respectively.

10.1 Dimensionality

As mentioned earlier, in CAT Monte Carlo simulations item responses are stochastically generated, using item parameters and latent trait values (θ). To support the use of the generated response data in the current CAT simulation, the dimensionality analysis was conducted using the PSDQ item parameters from Appendix A, and two sets of θ's following different distribution. A set of 1000 latent trait values (θ) was simulated following the standard normal distribution N(0,1) and a set of 1000 latent trait values (θ) from a uniform distribution U(-3,3) was drawn (note that the values were different from those defined in section 9.2.1). For both θ distributions, 1000 datasets were generated using the same procedure as described in section 9.2.4.

The generated datasets were subject to Exploratory Factor Analysis (EFA) using the *Mplus* statistical software (Muthén & Muthén, 1998-2016). Models with 1 and 2 factors respectively, were fitted to the simulated responses in each EFA. Model parameters were estimated using weighted least squares means and variance adjusted (WLSMV) estimation method and geomin (oblique) rotation was employed in the EFA with two factors (Muthén & Muthén, 1998-2016).

The results of 1000 EFA replications (means and SEMs) are summarized in Table 3. By all accounts, the fit of the unidimensional model was adequate for both θ distributions with mean RMSEA values close to 0 and CFI and TLI approaching 1. The two-factor EFA only slightly improved the model fit (the one- and two-factor models are parameter-nested models). More importantly, all factor loadings on the first factor within the two-factor solution were higher than the loadings on the second factor (mean values of the factor loadings on the second factor did not exceed the value of 0.1). Interestingly, the distribution of the estimated factor loadings for the first factor was almost identical within both the one-factor and the two-factor solution. These results clearly indicate that the second factor captures residual pieces of variance net of the first component (i.e., couplets that lack substantive meaning), lending further support to a unidimensional model. This finding clearly satisfies the assumption of the simulated response data required for the intended simulation of unidimensional IRT-based CAT.

Table 3 – Results of 1000 exploratory factor analysis replications of simulated responses (mean of point estimates with corresponding mean standard errors in the brackets)

	Normal θ distribution			Uniform θ distribution		
	1 factor	2 factors		1 factor	2 factors	
ChiSquare	2428.5 (39.3)	2326.5 (37.4)		2441.4 (37.9)	2340.3 (35.9)	
df	2345 (0)	2276 (0)		2345 (0)	2276 (0)	
p	0.145 (0.119)	0.256 (0.158)		0.110 (0.097)	0.199 (0.136)	
RMSEA	0.006 (0.002)	0.004 (0.002)		0.006 (0.001)	0.005 (0.002)	
CFI	0.998 (0.001)	0.999 (0.001)		0.999 (0.000)	0.999 (0.000)	
TLI	0.998 (0.001)	0.999 (0.001)		0.999 (0.000)	0.999 (0.000)	
SRMR	0.027 (0.001)	0.025 (0.001)		0.017 (0.001)	0.016 (0.000)	
Loadings	f1	f1	f2	f1	f1	f2
Item 1	0.11 (0.03)	0.11 (0.03)	0.09 (0.07)	0.20 (0.03)	0.20 (0.03)	0.09 (0.07)
Item 2	0.10 (0.03)	0.10 (0.03)	0.08 (0.07)	0.18 (0.04)	0.18 (0.04)	0.09 (0.07)
Item 3	0.21 (0.04)	0.21 (0.04)	0.10 (0.08)	0.36 (0.03)	0.36 (0.04)	0.10 (0.08)
Item 4	0.23 (0.03)	0.23 (0.03)	0.09 (0.07)	0.40 (0.03)	0.40 (0.03)	0.09 (0.07)
Item 5	0.27 (0.04)	0.27 (0.04)	0.09 (0.07)	0.46 (0.03)	0.46 (0.03)	0.09 (0.07)
Item 6	0.22 (0.04)	0.22 (0.04)	0.09 (0.07)	0.38 (0.03)	0.38 (0.03)	0.09 (0.07)
Item 7	0.23 (0.04)	0.23 (0.04)	0.10 (0.07)	0.39 (0.03)	0.39 (0.03)	0.10 (0.07)
Item 8	0.36 (0.03)	0.36 (0.03)	0.08 (0.06)	0.58 (0.02)	0.58 (0.02)	0.07 (0.06)
Item 9	0.47 (0.03)	0.47 (0.03)	0.07 (0.06)	0.70 (0.02)	0.70 (0.02)	0.06 (0.05)
Item 10	0.52 (0.03)	0.52 (0.03)	0.07 (0.06)	0.75 (0.02)	0.75 (0.02)	0.06 (0.04)
Item 11	0.64 (0.02)	0.64 (0.02)	0.05 (0.04)	0.85 (0.01)	0.85 (0.01)	0.04 (0.03)
Item 12	0.69 (0.02)	0.69 (0.02)	0.05 (0.04)	0.88 (0.01)	0.88 (0.01)	0.04 (0.03)
Item 13	0.61 (0.02)	0.61 (0.02)	0.06 (0.05)	0.82 (0.01)	0.82 (0.01)	0.04 (0.03)
Item 14	0.72 (0.02)	0.72 (0.02)	0.05 (0.04)	0.89 (0.01)	0.89 (0.01)	0.03 (0.03)
Item 15	0.49 (0.03)	0.49 (0.03)	0.07 (0.06)	0.72 (0.02)	0.72 (0.02)	0.06 (0.05)
Item 16	0.50 (0.03)	0.50 (0.03)	0.07 (0.05)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 17	0.42 (0.03)	0.42 (0.03)	0.08 (0.06)	0.66 (0.02)	0.66 (0.02)	0.06 (0.05)
Item 18	0.57 (0.03)	0.57 (0.03)	0.07 (0.05)	0.79 (0.01)	0.79 (0.02)	0.05 (0.04)
Item 19	0.69 (0.02)	0.69 (0.02)	0.06 (0.05)	0.88 (0.01)	0.88 (0.01)	0.04 (0.03)
Item 20	0.65 (0.02)	0.65 (0.02)	0.06 (0.05)	0.87 (0.01)	0.87 (0.01)	0.04 (0.03)
Item 21	0.52 (0.03)	0.52 (0.03)	0.08 (0.06)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 22	0.49 (0.03)	0.49 (0.03)	0.08 (0.06)	0.72 (0.02)	0.72 (0.02)	0.06 (0.05)
Item 23	0.52 (0.03)	0.52 (0.03)	0.07 (0.06)	0.76 (0.02)	0.76 (0.02)	0.06 (0.04)
Item 24	0.55 (0.03)	0.55 (0.03)	0.07 (0.06)	0.77 (0.02)	0.77 (0.02)	0.06 (0.05)
Item 25	0.54 (0.03)	0.54 (0.03)	0.07 (0.05)	0.77 (0.02)	0.77 (0.02)	0.06 (0.04)
Item 26	0.47 (0.03)	0.47 (0.03)	0.08 (0.06)	0.69 (0.02)	0.69 (0.02)	0.07 (0.05)
Item 27	0.68 (0.02)	0.68 (0.02)	0.05 (0.04)	0.88 (0.01)	0.88 (0.01)	0.04 (0.03)
Item 28	0.75 (0.02)	0.75 (0.02)	0.05 (0.04)	0.91 (0.01)	0.91 (0.01)	0.03 (0.03)
Item 29	0.68 (0.02)	0.68 (0.02)	0.05 (0.04)	0.87 (0.01)	0.87 (0.01)	0.04 (0.03)
Item 30	0.81 (0.01)	0.81 (0.01)	0.04 (0.03)	0.94 (0.01)	0.94 (0.01)	0.03 (0.02)
Item 31	0.70 (0.02)	0.70 (0.02)	0.05 (0.04)	0.90 (0.01)	0.90 (0.01)	0.03 (0.03)
Item 32	0.83 (0.01)	0.83 (0.01)	0.04 (0.03)	0.95 (0.00)	0.94 (0.01)	0.03 (0.02)
Item 33	0.59 (0.02)	0.59 (0.02)	0.06 (0.05)	0.81 (0.01)	0.81 (0.01)	0.05 (0.04)
Item 34	0.65 (0.02)	0.65 (0.02)	0.06 (0.05)	0.85 (0.01)	0.85 (0.01)	0.04 (0.03)
Item 35	0.76 (0.01)	0.76 (0.02)	0.05 (0.04)	0.92 (0.01)	0.92 (0.01)	0.03 (0.02)
Item 36	0.75 (0.02)	0.75 (0.02)	0.05 (0.04)	0.91 (0.01)	0.91 (0.01)	0.03 (0.03)
Item 37	0.74 (0.02)	0.74 (0.02)	0.05 (0.04)	0.90 (0.01)	0.90 (0.01)	0.04 (0.03)
Item 38	0.57 (0.02)	0.57 (0.02)	0.07 (0.05)	0.80 (0.01)	0.80 (0.01)	0.05 (0.04)
Item 39	0.60 (0.02)	0.60 (0.02)	0.06 (0.05)	0.83 (0.01)	0.83 (0.01)	0.05 (0.04)
Item 40	0.58 (0.02)	0.58 (0.02)	0.06 (0.05)	0.81 (0.01)	0.81 (0.01)	0.05 (0.04)

Table 3 (*continued*)

	Normal θ distribution			Uniform θ distribution		
	1 factor	2 factors		1 factor	2 factors	
Loadings	f1	f1	f2	f1	f1	f2
Item 41	0.56 (0.02)	0.56 (0.02)	0.07 (0.05)	0.80 (0.01)	0.80 (0.01)	0.05 (0.04)
Item 42	0.52 (0.03)	0.52 (0.03)	0.07 (0.06)	0.75 (0.02)	0.75 (0.02)	0.06 (0.04)
Item 43	0.61 (0.02)	0.61 (0.02)	0.06 (0.04)	0.84 (0.01)	0.84 (0.01)	0.04 (0.03)
Item 44	0.48 (0.03)	0.48 (0.03)	0.07 (0.05)	0.72 (0.02)	0.72 (0.02)	0.06 (0.05)
Item 45	0.55 (0.02)	0.55 (0.02)	0.06 (0.05)	0.78 (0.01)	0.78 (0.01)	0.05 (0.04)
Item 46	0.63 (0.02)	0.63 (0.02)	0.06 (0.04)	0.84 (0.01)	0.84 (0.01)	0.04 (0.03)
Item 47	0.50 (0.03)	0.50 (0.03)	0.07 (0.05)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 48	0.47 (0.03)	0.47 (0.03)	0.08 (0.06)	0.69 (0.02)	0.69 (0.02)	0.07 (0.05)
Item 49	0.64 (0.02)	0.64 (0.02)	0.06 (0.04)	0.85 (0.01)	0.85 (0.01)	0.04 (0.03)
Item 50	0.44 (0.03)	0.44 (0.03)	0.07 (0.06)	0.68 (0.02)	0.68 (0.02)	0.06 (0.05)
Item 51	0.50 (0.02)	0.50 (0.03)	0.07 (0.05)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 52	0.56 (0.02)	0.56 (0.02)	0.06 (0.05)	0.79 (0.01)	0.79 (0.01)	0.05 (0.04)
Item 53	0.31 (0.03)	0.31 (0.03)	0.08 (0.06)	0.51 (0.03)	0.51 (0.03)	0.08 (0.06)
Item 54	0.51 (0.03)	0.51 (0.03)	0.07 (0.06)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 55	0.72 (0.02)	0.72 (0.02)	0.05 (0.04)	0.89 (0.01)	0.89 (0.01)	0.03 (0.03)
Item 56	0.60 (0.02)	0.60 (0.02)	0.06 (0.05)	0.82 (0.01)	0.82 (0.01)	0.04 (0.03)
Item 57	0.65 (0.02)	0.65 (0.02)	0.05 (0.04)	0.87 (0.01)	0.87 (0.01)	0.04 (0.03)
Item 58	0.68 (0.02)	0.68 (0.02)	0.05 (0.04)	0.88 (0.01)	0.88 (0.01)	0.04 (0.03)
Item 59	0.58 (0.02)	0.58 (0.02)	0.07 (0.05)	0.83 (0.01)	0.83 (0.01)	0.05 (0.04)
Item 60	0.63 (0.02)	0.63 (0.02)	0.06 (0.04)	0.86 (0.01)	0.86 (0.01)	0.04 (0.03)
Item 61	0.71 (0.02)	0.71 (0.02)	0.05 (0.04)	0.90 (0.01)	0.90 (0.01)	0.03 (0.02)
Item 62	0.62 (0.02)	0.62 (0.02)	0.06 (0.04)	0.84 (0.01)	0.84 (0.01)	0.04 (0.03)
Item 63	0.49 (0.03)	0.48 (0.03)	0.07 (0.05)	0.72 (0.02)	0.72 (0.02)	0.06 (0.05)
Item 64	0.51 (0.03)	0.51 (0.03)	0.07 (0.05)	0.74 (0.02)	0.74 (0.02)	0.06 (0.04)
Item 65	0.49 (0.03)	0.49 (0.03)	0.08 (0.06)	0.71 (0.02)	0.71 (0.02)	0.07 (0.05)
Item 66	0.65 (0.03)	0.65 (0.03)	0.07 (0.06)	0.83 (0.01)	0.83 (0.01)	0.05 (0.04)
Item 67	0.69 (0.02)	0.69 (0.02)	0.05 (0.04)	0.87 (0.01)	0.87 (0.01)	0.04 (0.03)
Item 68	0.74 (0.02)	0.74 (0.02)	0.05 (0.04)	0.90 (0.01)	0.90 (0.01)	0.04 (0.03)
Item 69	0.60 (0.03)	0.60 (0.03)	0.08 (0.06)	0.79 (0.01)	0.79 (0.01)	0.06 (0.05)
Item 70	0.54 (0.02)	0.54 (0.02)	0.07 (0.05)	0.77 (0.01)	0.77 (0.01)	0.06 (0.04)

Note: RMSEA = Root Mean Square Error of Approximation (Steiger & Lind, 1980); CFI = Comparative Fit Index (Bentler, 1990); TLI = Tucker Lewis Index (Tucker & Lewis, 1973); SRMR = Square Root Mean Residual (Hu & Bentler, 1998, 1999).

10.2 Number of administered items in CAT simulation

Figure 17 shows the average number of administered PSDQ items for different CAT estimation methods, items selection procedures, termination criteria, and generated true latent trait (θ^*) distributions. On average between 22 and 34 items were administered regardless of θ^* distribution, item selection and latent trait estimation methods, when high measurement precision was required (termination criterion SE = 0.23, which corresponds to reliability of 0.95). The average number of administered items decreased rapidly (between 14 and 18 items) when the CAT stopping rule was set to SE = 0.32 (reliability of 0.90). A further reduction in desired level of measurement precision conforming to a SE of 0.39 and 0.45 (reliability of 0.85 and 0.80, respectively), showed that the number of items administered to meet this benchmark was far less; however, the change was not as steep as with a smaller SE and higher precision level (see Figure 17). Interestingly, when a relatively low, but widely accepted level of measurement precision was specified (stopping rule of SE = 0.45), only 4 to 10 items from the 70-item PSDQ were administered on average.

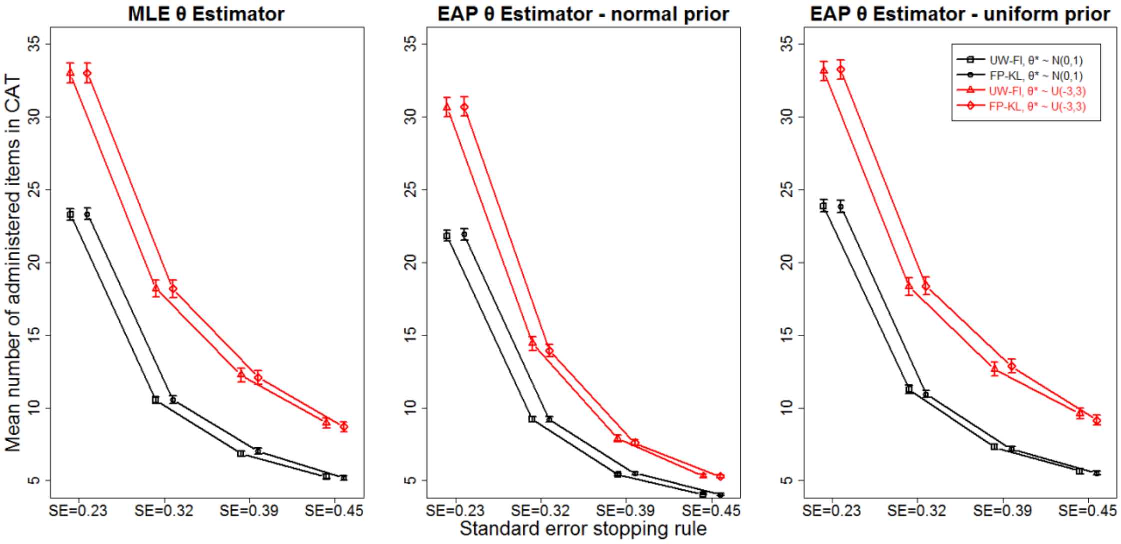


Figure 17 – Mean number of administered items from PSDQ in CAT simulations by level of measurement precision. Note: error bars represent standard error of the mean; shifts on x-axis within a particular SE are artificial to make all means visible.

Results displayed on Figure 17 indicate that the latent trait estimation methods were similarly effective while the two item selection methods were virtually identical across simulation conditions. For each combination of the latent trait estimator and the stopping rule, standard normal distribution of the generated θ^* led to lower number of administrated items.

Table 4 shows the analysis of variance (ANOVA) results to examine the effect of different simulation conditions on test length. As depicted, most of the variability in the number of administered items across the simulation conditions was accounted for by desired level of measurement precision and the θ^* distribution. Specifically, 30.2% of the test length total variability in the current simulation is due to stopping rule ($\eta^2 = 0.302$, $p < 0.001$). Therefore, specifying different values of the standard error (SE) stopping rule will have a large effect on the efficacy of the PSDQ CAT administration. In case of the θ^* distribution, which accounted for most of the remaining variance (5.1%), the effect size was relatively small ($\eta^2 = 0.051$, $p < 0.001$).

Table 4 – ANOVA results for number of administered items in CAT simulation (n = 48000)

Source	df	F	p	η^2
Main Effects				
Latent trait estimation method	2	246.0	0.000	0.010
θ^* distribution	1	2552.5	0.000	0.051
Stopping rule SE	3	6923.8	0.000	0.302
Item selection method	1	0.4	0.554	0.000
2-way Interaction Effects				
Latent trait estimation method * Item selection method	2	0.0	0.981	0.000
Latent trait estimation method * Stopping rule SE	6	1.9	0.078	0.000
Latent trait estimation method * θ^* distribution	2	40.7	0.000	0.002
Stopping rule SE * Item selection method	3	0.2	0.915	0.000
θ^* distribution * Item selection method	1	0.1	0.710	0.000
θ^* distribution * Stopping rule SE	3	148.8	0.000	0.009
Error	47975			

Note: df – degrees of freedom, F – F-statistics, p – p-value, η^2 – effect size

Turning to the remaining ANOVA main effects, the different estimation methods accounted for a significant portion of model variance ($p < 0.001$); however the overall effect this had on the number of administered items was almost negligible ($\eta^2 = 0.010$). The only nonsignificant main effect was associated with item selection methods ($p = 0.554$). The effect size of the item selection methods on the test length ($\eta^2 < 0.001$) is trivially small based on Cohen's (1988) guidelines. Although two out of six ANOVA interaction effects were statistically significant at the conventional $\alpha = 0.05$ level, both produced relatively small effect sizes ($\eta^2 < 0.01$), indicating no effect of these model terms on the test length.

It is worth noting that the efficacy of the PSDQ CAT administration, in terms of test length, varied greatly as a function of the CAT estimated latent trait ($\hat{\theta}$) values. This is further demonstrated in Figure 18 and Figure 19 for the standard normal true latent trait ($\theta^* \sim N(0,1)$) and the uniform true latent trait ($\theta^* \sim U(-3,3)$) distributions, respectively. Given the nonsignificant finding and likewise the negligible effect size observed in the ANOVA model for the item selection methods on test length, only different latent trait estimators and standard error stopping rules are compared in Figures 18 and 19.

As both Figures 18 and 19 reveal, generally more items were administered when estimating higher latent levels of physical self-concept (e.g., $\hat{\theta} > 1.5$ logits) for each stopping rule criterion. For instance, when high measurement precision was desired (SE stopping rule was set to SE = 0.23) approximately 15 to 35 items (saving at least half of the item pool) on average were administered where the range for $\hat{\theta}$ was between -3 to 1 logits. In contrast, 63 to 70 items were needed when latent trait levels were much higher ($\hat{\theta} \geq 2$ logits), regardless of the θ^* distribution and latent trait estimator (see the upper left portion of the Figures).

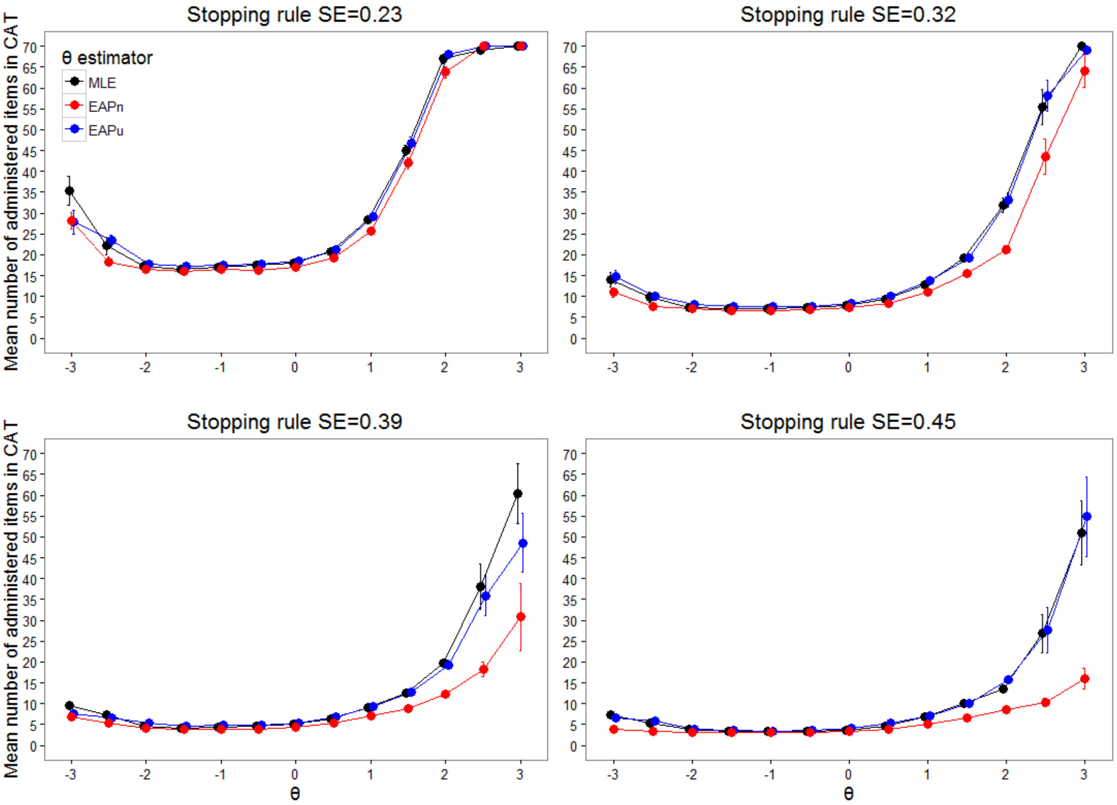


Figure 18 – Mean number of administered items from PSDQ (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for standard normal true latent trait ($\theta^* \sim$ N(0,1)) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform prior; error bars represent standard deviation

The observation is a result of the distribution of the PSDQ items threshold and discrimination parameters (see Appendix A) and is therefore related to the item pool information function (see Appendix C). The PSDQ items threshold parameters are mostly located on the negative side of the physical self-concept latent continuum, providing less information for high latent trait values, which produces the demand for more items in the test administration.

Even for situations requiring much lower measurement precision (stopping rule SE = 0.45), a relatively high number of items was administered on average for the latent trait estimates about $\hat{\theta}$ = 3 logits. This was especially true for MLE and EAP with uniform prior estimators, where 40 to 55 items were needed regardless the θ^* distribution (see the lower right parts of the Figure 18 and 19).

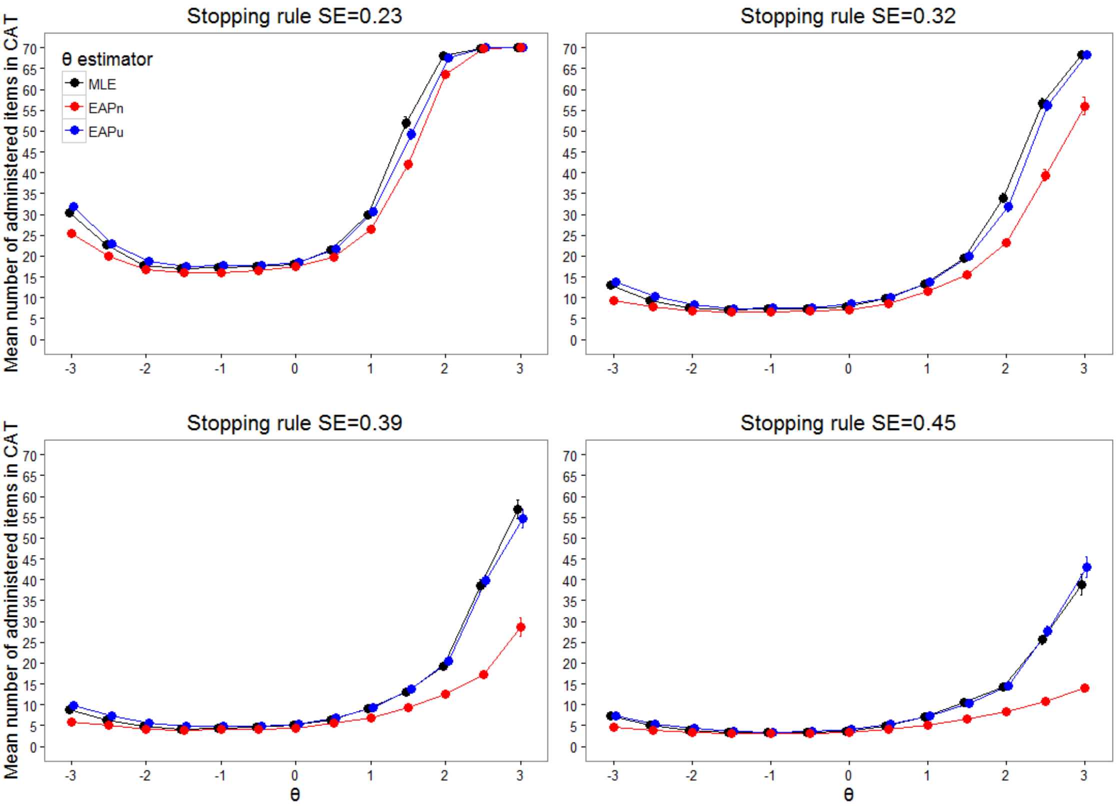


Figure 19 – Mean number of administered items from PSDQ (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for uniform true latent trait ($\theta^* \sim$ U(-3,3)) distribution. Note: EAPn = EAP estimation with standard normal prior distribution; EAPu = EAP estimation with uniform prior distribution; error bars represent standard deviation

Interestingly, at the same precision level (SE = 0.45), the EAP latent trait estimator with standard normal prior distribution required only about 15 items even for $\hat{\theta}$ = 3 logits. Generally, the performance of the MLE and EAP with uniform prior was very similar at each latent trait value across all termination criteria as well as across both θ^* distributions. The different efficacy of the EAP with standard normal prior at the higher extremes of the physical self-concept latent continuum starts to be apparent as soon as the stopping rule SE equals to 0.32 (equivalent to reliability of 0.90) and increases with decreasing level of the required measurement precision. These results indicate that the PSDQ CAT administration may not necessarily bring the expected benefits (reducing testing time and respondent burden) when measuring students with high trait values of physical self-concept. The efficacy of the PSDQ CAT administration for the higher latent trait values (e.g., $\hat{\theta} \geq 1.5$ logits) in terms of test length may be improved however, by employing EAP estimation with informative prior, especially if the standard error of the latent trait estimate SE ≥ 0.39 is acceptable.

10.3 Bias of the CAT latent trait estimates

This section explores fundamental issues of concern that revolve around the performance of the PSDQ CAT administration with respect to test accuracy. Accuracy is evaluated using bias of the CAT latent trait estimates ($\hat{\theta}$) from generated true latent trait values (θ^*); where smaller absolute values of bias indicate better performance. Figure 20 graphically presents the average absolute values of individual bias for each simulation condition.

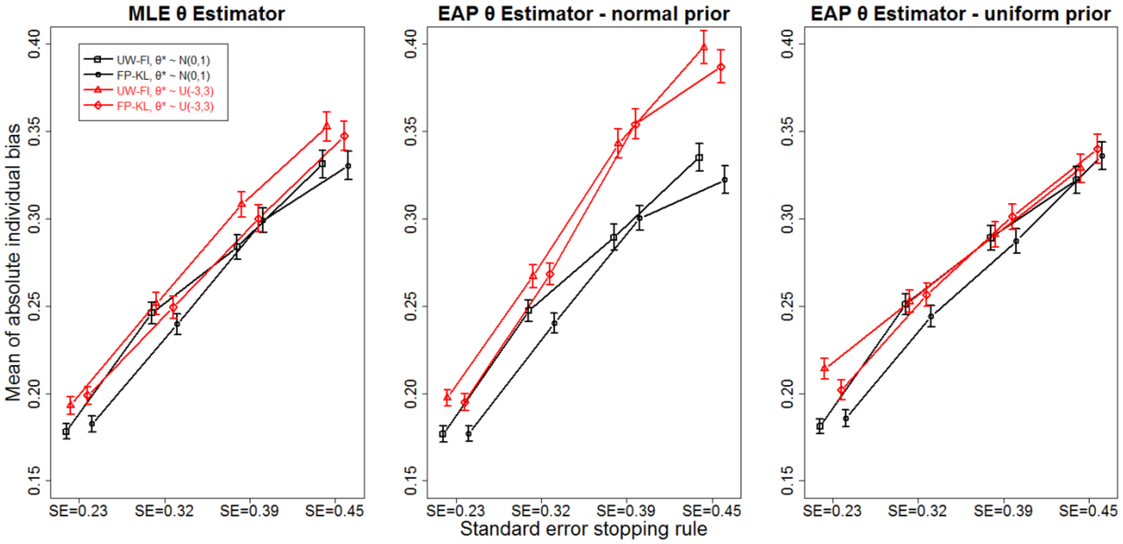


Figure 20 – Mean of absolute individual bias of CAT latent trait estimates by level of measurement precision. Note: error bars represent standard error of the mean; shifts on x-axis within a particular SE are artificial to make all means visible.

Not surprisingly, the absolute bias of the CAT latent trait estimates increased as the predefined measurement precision decreased, with mean values from 0.18 to 0.21

and from 0.32 to 0.40 logits for stopping rule SE = 0.23 and SE = 0.45 respectively. It should be noted however, that the bias dispersion was higher for the higher SE stopping rule values as well.

Likewise, when the same analysis was conducted with test length, the Fisher information-based and Kullback-Leibler divergence-based item selection methods led to almost identical results (see Figure 20). Interestingly, when the MLE or EAP estimator with uniform prior distribution was contrasted for the different measurement precision, the findings underscored very negligible differences in latent trait bias (refer to the left and right hand part of Figure 20). This was not true, however, when the EAP estimator with standard normal prior distribution was employed, these results underscoring that the uniformly generated true latent trait distribution led to higher values of absolute bias, especially when stopping rule was set to SE = 0.39 and 0.45. This finding indicates that specifying an incorrect informative prior with EAP estimation seems to be less plausible for obtaining CAT accuracy than specifying an uninformative prior or not specifying a prior at all (e.g., using MLE).

Table 5 summarizes the ANOVA results, evaluating the effect of various simulation conditions on absolute values of individual latent trait bias. The ANOVA was run with the main and the two-way interaction effects and eta-squared η^2 was used to determine the effect sizes.

Table 5 – ANOVA results for absolute individual bias of CAT latent trait estimates in CAT simulation (n = 48000)

Source	df	F	p	η^2
Main Effects				
Latent trait estimation method	2	19.91	0.000	0.001
θ^* distribution	1	121.11	0.000	0.003
Stopping rule SE	3	1145.43	0.000	0.067
Item selection method	1	0.11	0.742	0.000
2-way Interaction Effects				
Latent trait estimation method * Item selection method	2	0.37	0.691	0.000
Latent trait estimation method * Stopping rule SE	6	7.94	0.000	0.001
Latent trait estimation method * θ^* distribution	2	22.08	0.000	0.001
Stopping rule SE * Item selection method	3	1.01	0.385	0.000
θ^* distribution * Item selection method	1	0.06	0.813	0.000
θ^* distribution * Stopping rule SE	3	3.28	0.020	0.000
Error	47975			

Note: df – degrees of freedom, F – F-statistics, p – p-value, η^2 – effect size

Using α = 0.05 as the acceptable limit for statistical hypotheses testing, three main effect terms and three interactions significantly influenced the absolute individual bias of CAT theta estimates. All of the nonsignificant ANOVA terms were associated with item selection methods, with trivially small effect sizes (all $\eta^2 < 0.001$). Consistent with the findings from test length, the Fisher information-based and Kullback-Leibler divergence-based item selection methods are indistinguishable in their effectiveness with regard to systematic bias of the CAT latent trait estimates.

Among the statistically significant main effects, stopping rule explained most of the variance in absolute bias, however this effect was quite modest ($\eta^2 = 0.067$). Of the remaining significant main effects, the generated θ^* distribution, also produced a relatively small effect size ($\eta^2 = 0.003$) as did the estimation methods ($\eta^2 = 0.001$). The three significant interactions also explained a trivially small amount of model variance (each less than 0.1 %).

Figures 21 and 22 graphically display the magnitude of individual bias as a function of CAT estimated theta for the uniform and standard normal true theta distributions, respectively. Given the ANOVA results, the item selection methods are not factored into the comparison in Figures 20 and 21.

The values of individual latent trait bias varied between approximately -0.7 and 0.7 logits on average along the latent trait continuum, regardless of θ^* distribution, stopping rules, and latent trait estimation methods. However for latent trait estimates $-2 < \hat{\theta} < 2$, the bias estimate ranged only from about -0.35 to 0.35 logits. This again highlights the questionable effectiveness of PSDQ CAT administration for assessing the extreme levels of physical self-concept.

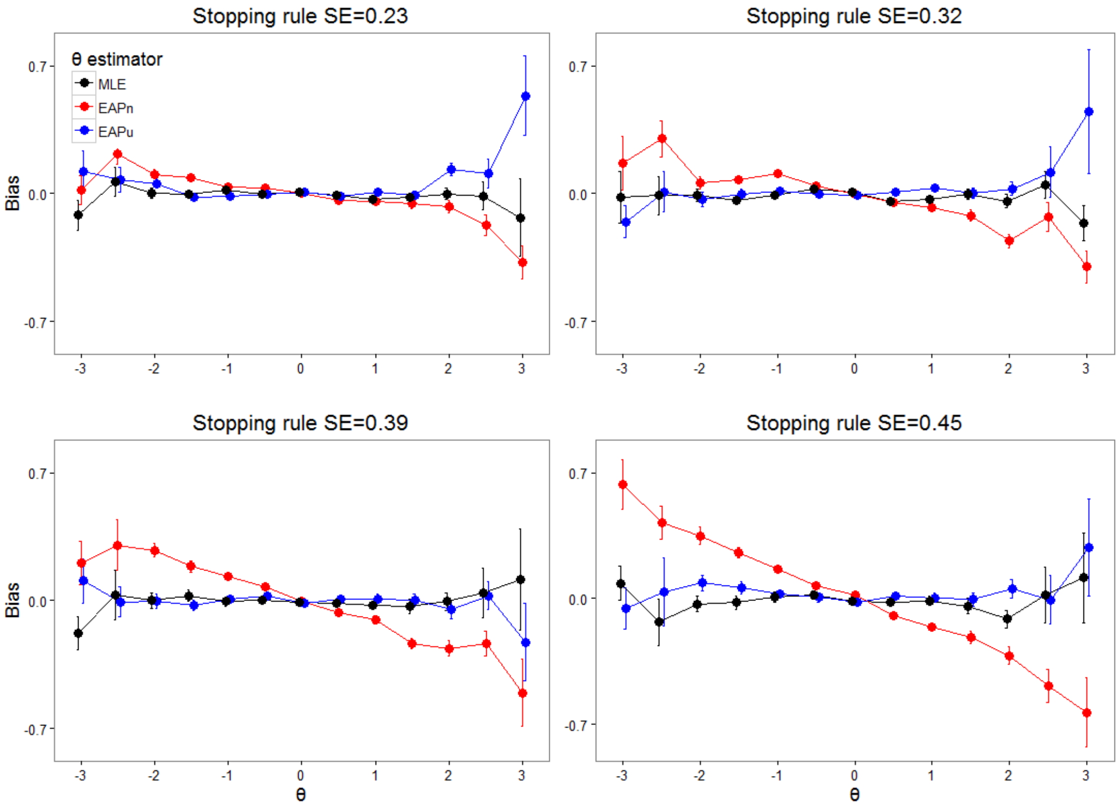


Figure 21 – Individual bias of CAT latent trait estimates (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for standard normal true latent trait ($\theta^{*}\sim$ N(0,1)) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform prior; error bars represent standard deviation

MLE and EAP estimation with uniform prior distribution produced very similar findings underscoring relatively small amounts of bias for the latent trait estimates along the latent trait continuum; and this was regardless of the specified test precision and θ^{*} distribution. Some small differences between the two estimation methods were observed at both positive and negative extremes of the $\hat{\theta}$ scale, especially in case of the standard normal true theta distribution. This could be caused, however, by the fact that in standard normal distribution there are far less observations at both tails than around the mean, and thus the computed mean values of bias at both extremes of the latent trait might not converge to the true (population) parameters. EAP estimation with standard normal prior led to a considerably different pattern of the bias estimates than the other two latent trait estimation methods. At each SE stopping rule, EAP estimation with standard normal prior produced obvious inward bias, indicating the tendency of $\hat{\theta}$ estimates to regress towards the prior mean.

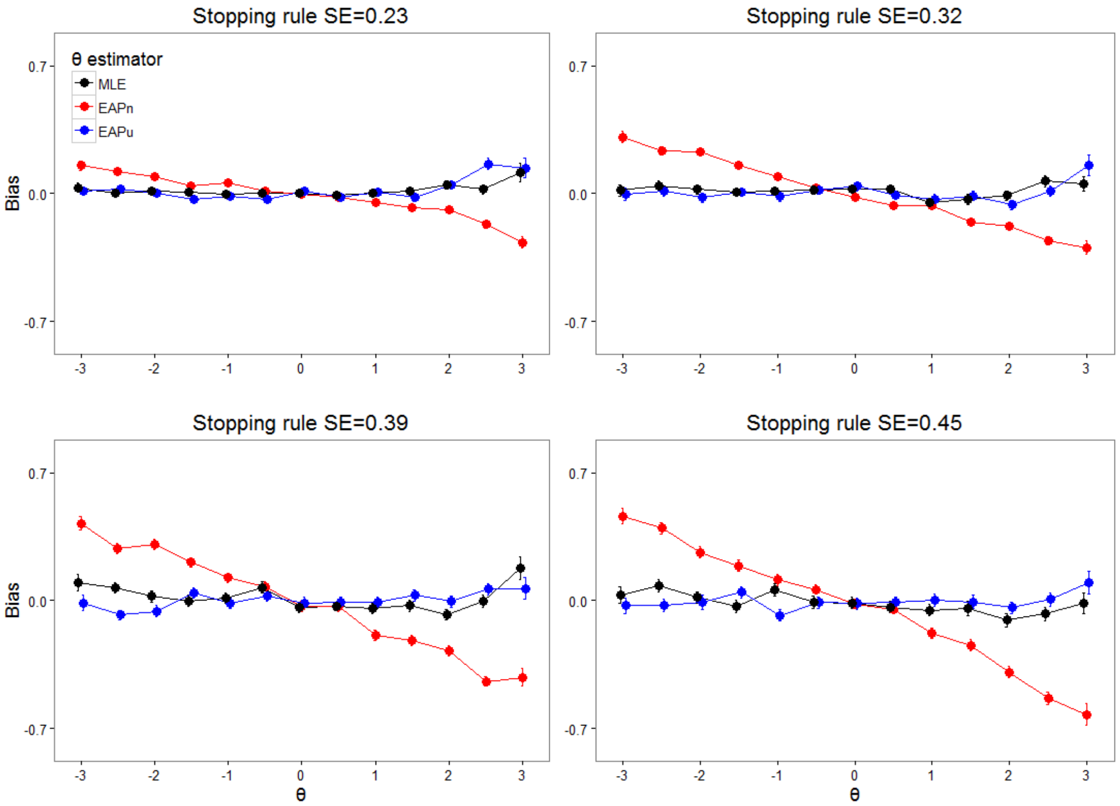


Figure 22 – Individual bias of CAT latent trait estimates (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for uniform true latent trait (θ^*~ U(-3,3)) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform prior; error bars represent standard deviation

10.4 Correlations

Table 6 shows the Pearson correlation coefficients between CAT estimated latent trait values ($\hat{\theta}$) and generated true latent trait values (θ^*) for various simulation conditions.

When high measurement precision was desired (SE = 0.23) the correlations were indeed high, ranging from 0.973 to 0.990, regardless the estimation procedure, item selection method as well as true latent trait distribution. As expected, the correlations decrease with decreasing level of measurement precision, however even for stopping rule of SE = 0.45 the correlations were still relatively high (from 0.907 to 0.972). This results point to the potential usefulness of the PSDQ CAT administration, because it produces latent trait estimates very close to the true (hypothetical) latent values of the physical self-concept, while saving a considerable portion of the item pool (from about 50% at SE = 0.32 to more than 90% at SE = 0.45 on average).

Table 6 – Correlations between CAT latent trait estimates ($\hat{\theta}$) and true latent trait values (θ^*)

θ	Item	SE stopping rule for θ* ~ N(0,1)				SE stopping rule for θ* ~ U(-3,3)			
estimator	selection	0.23	0.32	0.39	0.45	0.23	0.32	0.39	0.45
MLE	UW-FI	0.975	0.954	0.939	0.923	0.990	0.983	0.975	0.967
MLE	FP-KL	0.974	0.957	0.936	0.920	0.989	0.983	0.975	0.968
EAPn	UW-FI	0.974	0.950	0.927	0.907	0.990	0.982	0.972	0.963
EAPn	FP-KL	0.974	0.953	0.927	0.912	0.990	0.984	0.970	0.965
EAPu	UW-FI	0.976	0.956	0.939	0.926	0.988	0.983	0.978	0.972
EAPu	FP-KL	0.973	0.955	0.939	0.920	0.988	0.982	0.977	0.970

Note: EAPn = EAP estimation with standard normal prior distribution; EAPu = EAP estimation with uniform prior distribution

The correlations between CAT estimated latent trait values ($\hat{\theta}$) and generated true latent trait values (θ^*) were higher for uniformly distributed θ^* at each level of measurement precision. This is most likely the consequence of higher average number of administered items in CAT simulations for uniformly distributed θ^*. On the other hand, the two item selection methods employed in the simulations led to almost identical results also in terms of correspondence between $\hat{\theta}$ and θ^*. Likewise, using the different estimation procedures (MLE, EAP with normal prior distribution, and EAP with uniform prior distribution) did not produce any substantial differences in correlations between $\hat{\theta}$ and θ^*.

To provide a clearer picture of the relation between $\hat{\theta}$ and θ^*, scatterplots showing θ^* on x-axis and $\hat{\theta}$ on y-axis were produced for normally distributed (Figure 23) as well as for uniformly distributed (Figure 24) true latent trait values. Since the three latent trait estimation methods as well as the two item selection methods produced very similar correlations, the scatterplots are only presented for MLE in a combination with UW-FI item selection method.

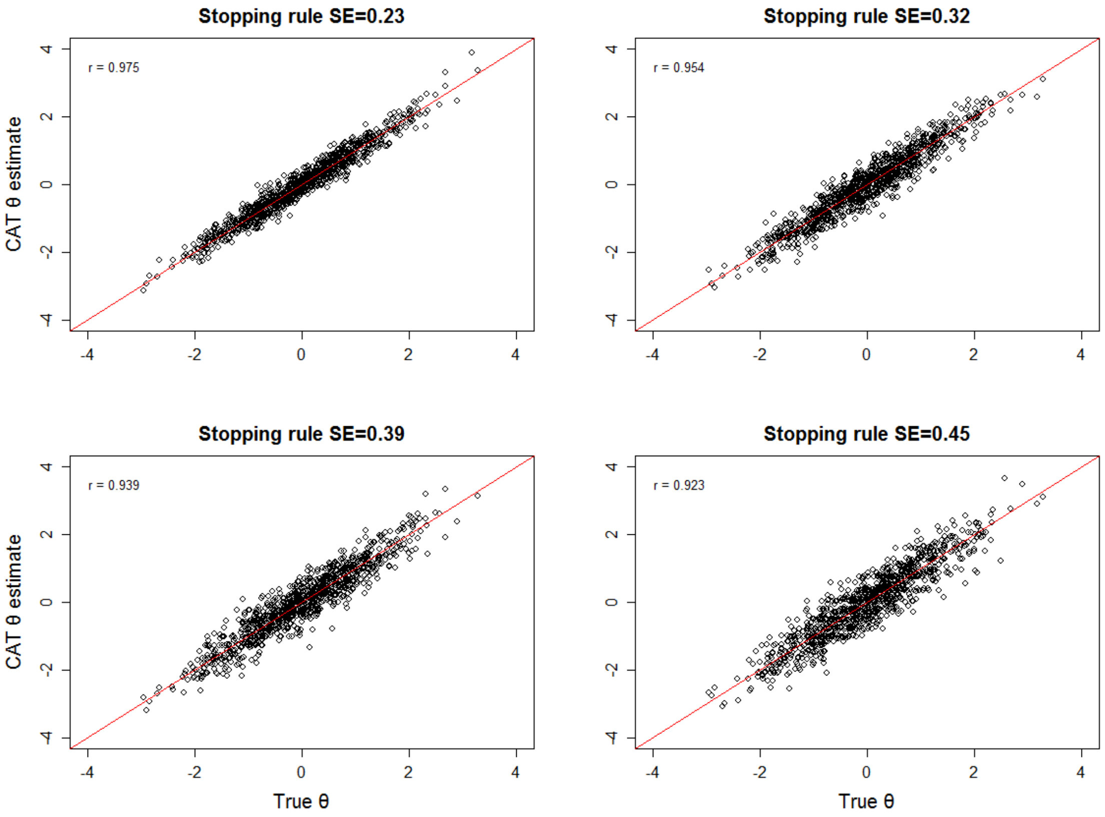


Figure 23 – Correlations between CAT latent trait estimates ($\hat{\theta}$; Y axis) and true latent trait values (θ^{*}; X axis) with standard normal ($\theta^{*}\sim$ N(0,1)) distribution

Both Figure 23 and 24 indicate no systematic bias of the CAT estimated latent trait values for all SE stopping rules (dots are symmetrically distributed along the red line, which indicates a perfect correlation).

Table 7 lists correlation between CAT latent trait estimates ($\hat{\theta}$) and estimates based on the full PSDQ ($\hat{\theta}^{PSDQ}$). These correlations assess the usefulness of PSDQ CAT administration as compared to the CTT approach of linear fixed-length testing.

Also in this case the correlations decreased with increasing value of the standard error stopping rule. Uniformly distributed θ^{*} produced higher correlations than the normally distributed θ^{*}, while only negligible differences were observed with regard to different estimation and item selection methods. Generally high values of the correlations in the Table 7 (0.922 to 0.997) indicate, that even when administration of a considerable number of PSDQ items is curtailed using CAT, it is possible to obtain almost the same estimates of physical self-concept as when the whole questionnaire is used.

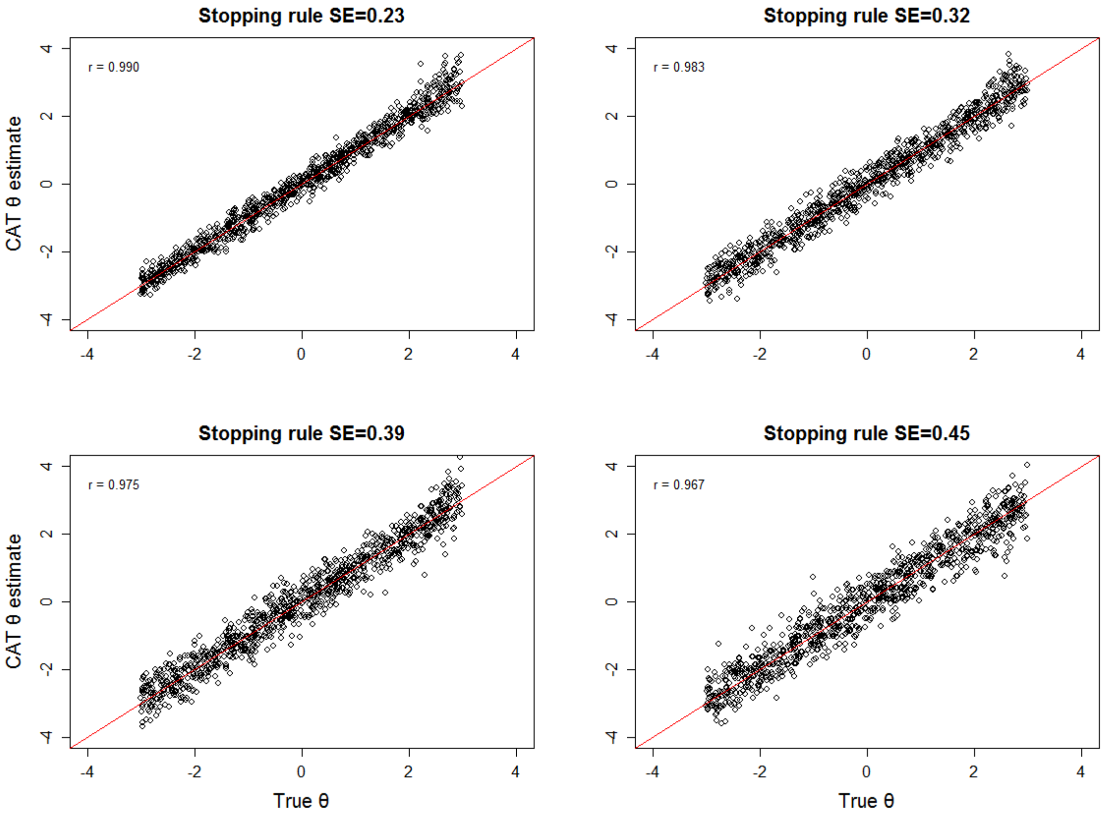


Figure 24 – Correlations between CAT latent trait estimates ($\hat{\theta}$; Y axis) and true latent trait values (θ^{*}; X axis) with uniform (θ^{*}~ U(-3,3)) distribution

Table 7 – Correlations between CAT latent trait estimates ($\hat{\theta}$) and full PSDQ latent trait estimates ($\hat{\theta}^{PSDQ}$).

θ	Item	SE stopping rule for θ^{*} ~ N(0,1)				SE stopping rule for θ^{*} ~ U(-3,3)			
estimator	selection	0.23	0.32	0.39	0.45	0.23	0.32	0.39	0.45
MLE	UW-FI	0.990	0.966	0.953	0.935	0.997	0.991	0.984	0.975
MLE	FP-KL	0.990	0.970	0.951	0.936	0.997	0.992	0.984	0.976
EAPn	UW-FI	0.987	0.964	0.941	0.922	0.997	0.989	0.979	0.971
EAPn	FP-KL	0.988	0.967	0.942	0.929	0.997	0.989	0.977	0.972
EAPu	UW-FI	0.991	0.973	0.953	0.939	0.997	0.991	0.986	0.980
EAPu	FP-KL	0.990	0.970	0.955	0.935	0.997	0.991	0.985	0.978

Note: EAPn = EAP estimation with standard normal prior distribution; EAPu = EAP estimation with uniform prior distribution

11 DISCUSSION

Computerized adaptive testing (CAT) represents a novel approach to test administration, and offers the unique possibility of vastly improving testing efficiency (Anastasi, 1976; van der Linden & Glas, 2010; Weiss, 1982). The use of CAT methodology is now a firm part of the landscape in both psychology and education, however, this approach is much less utilized in the field of Kinanthropology. Since many self-report assessments developed in psychology are now used in studies of physical education and athletic performance, it makes sense to determine the suitability of CAT methods in this area of inquiry (Gershon & Bergstorm, 2006). In this thesis, I first presented the theoretical and conceptual basis of CAT, and followed this with a brief overview of its historical origins. I then present the basic general principles of CAT, including a description of Item Response Theory (IRT), since IRT is almost exclusively used as a mathematical model in today's CAT applications (Wainer, 2000). The practical applicability of CAT was then evaluated using Monte-Carlo simulations of adaptive administration of the Physical Self-Description Questionnaire (PSDQ) – an instrument widely used to assess physical self-concept in the field of Kinanthropology.

For the Monte Carlo simulation of the PSDQ adaptive administration I used a real item pool (N = 70) calibrated with a Graded Response Model (GRM, see Samejima, 1969, 1997) and generated the responses to test items based on the item parameters and pre-specified true latent values (θ^*) of physical self-concept. The Monte Carlo simulation study was designed to compare the number of administered items from PSDQ (test length) and accuracy of estimated latent levels of physical self-concept, while using a variety of latent trait estimation methods (MLE, EAP with standard normal prior, and EAP with uniform prior distribution), items selection algorithms (UW-FI, and FP-KL), distributional properties (standard normal and uniform distribution of the true latent trait values) and stopping rules (standard error of latent trait estimate SE = 0.23, SE = 0.32, SE = 0.39, and SE = 0.45). Each of these frequently discussed CAT settings represents important elements that should be considered in the application of CAT, both in general (Thompson & Weiss, 2011) and specifically within the measurement of physical self-concept as it can be used in Kinanthropology.

The Monte Carlo simulation results showed that CAT can successfully be applied as a method of reducing test length when using the PSDQ to assess physical self-concept. For instance, CAT requiring widely acceptable measurement precision (SE = 0.45 which represents test reliability of 0.80) saved on average about 85% to 93% of administered items. Naturally, when increasing the required measurement precision, the average number of administered items increases. Notwithstanding, the CAT approach may be very useful in reducing response burden even for a relatively high benchmark of precision (SE = 0.23 which represents test reliability of 0.95), where on average implementation of this procedure can still result in a reduction of more than 50% of the items from the original questionnaire per respondent.

Moreover this rather substantial reduction in examinee response burden was achieved without any serious loss of information about the trait in question for simulated respondents. For example, with the PSDQ in hand, and using a CAT stopping rule SE = 0.45 (requiring test reliability of 0.80 along the latent continuum), where only 4 to 10 items were administered on average, the correlations between CAT estimated latent trait values ($\hat{\theta}$) and generated true latent trait values (θ^*) exceeded 0.90. This clearly shows that individually tailored selection of items from the PSDQ provides an unbiased estimate of the underlying latent trait using a much shorter test. The correlations between CAT latent trait estimates ($\hat{\theta}$) and the physical self-concept estimates based on all of the items in the PSDQ ($\hat{\theta}^{PSDQ}$) were even higher. This latter finding reflects more about the usefulness of a CAT application compared to the fixed-length linear testing. Others have noted that there are no clear cut-offs for expected correlation levels between CAT estimates and the full-length measure (Makransky, Dale, Havmose, & Bleses, 2016). However previous simulation studies using similar SE stopping rules as those employed in the current thesis reported correlations between 0.85 and 0.98 (e.g., Hula, Kellough, & Fergadiotis, 2015; Makransky, Mortensen, & Glas, 2013; Štochl et al., 2016b). The lowest correlations yielded by the current CAT simulation of the PSDQ were 0.922 and 0.987 for standard error stopping rules SE = 0.45 and SE = 0.23 respectively. This relatively high magnitude of association indicates considerable time and perhaps costs savings when CAT is used to administer the PSDQ. In essence, a test developer is able to obtain a very good “read” on the underlying latent trait of physical self-concept using a reduced set of items, rather than resorting to the full 70 items. Thus, in line with results of many other CAT studies (Devine et al., 2016; Makransky et al., 2016; Petersen et al., 2016; Štochl et al., 2016a, 2016b; Tseng, 2016), we can conclude that a CAT methodology leads to improved test efficiency, economy, and precision.

The same may not be true, however, when we discuss the expected benefits of CAT (i.e., reducing the respondent's burden) when measuring high levels of the physical self-concept. The lack of desired efficiency with high trait levels may be attributable to the original measurement properties of the PSDQ items, which provide more information for individuals with low physical self-concept (Flatcher & Hattie, 2004). Like the original fixed-length instrument, a CAT PSDQ administration would therefore be far less precise in detecting high levels of physical self-concept. Therefore, if the primary purpose is to detect and discriminate between examinees with low to average levels of physical self-concept, a CAT version of the PSDQ seems sufficient. Some authors (Nogami & Hayashi, 2010; Smits, Cuijpers, & van Straten, 2011) have argued, however, that for common CAT applications, the item pool information function should ideally follow a uniform distribution. Thus, to take the advantage of the CAT approach when assessing high levels of physical self-concept requires extending the PSDQ item pool with new items that have very high threshold parameters and provide greater coverage of the latent trait (see Appendix C). It should be noted, however, that this might not be an easy task in practice, since some authors reported problems in assessing high levels of physical self-concept and the problems appear to be inherent in the nature of the construct (Flatcher & Hattie, 2004).

Several authors have noted that simulation studies are essential in order to compare and evaluate different CAT algorithm specifications (e.g. latent trait estimation methods, item selection methods, stopping rules) and to identify a suitable combination of the settings for a given CAT (e.g., Thompson & Weiss, 2011; van der Linden & Pashley, 2010). Not surprisingly, the results of the current simulation revealed that the efficacy of the PSDQ CAT administration in terms of test length is greatly influenced by the desired value of the SE stopping rule. There are many situations where screening instruments are needed, whether they involve clinical settings or where time limitations come into play, and where parsimony in the number of items administered is a concern. In these situations, imposition of the SE = 0.45 stopping rule seems attractive. While CAT using this termination decision rule ensures the acceptable reliability (0.80) of the physical self-concept estimates along the whole latent continuum, on average only about 15% of items from the original PSDQ questionnaire is administered and imposition of this rule also yields very similar trait estimates as the traditional linear administration of the full PSDQ. However, when considering the question of which SE stopping rule would be optimal in a real PSDQ CAT administration, the appropriate value may vary as a result of the prioritization of parsimony versus accuracy in a given physical self-concept measurement (Makransky et al., 2016; Tseng, 2016).

With respect to item selection, both Kullback-Leibler divergence-based and Fisher information-based methods led to almost identical test length and produced similar levels of bias for latent trait estimates. Veldkamp (2003) reported very similar performance of these two item selection methods in polytomous IRT-based CAT using the generalized partial credit model (GPCM). In his study, Veldkamp (2003) found a relatively large amount of overlap in administered items (85% to 100%) between Fisher-based and Kullback-Leibler-based item selection methods, while the difference in measurement precision was negligible. Similarly, a simulation study by Passos et al. (2007) identified comparable performance of the two item selection methods using a nominal IRT model. More recently, Štochl et al. (2016a, 2016b) investigated the Kullback-Leibler divergence-based and Fisher information-based item selection methods in simulated CATs with real item pools designed to measure mental health in a community setting. These studies showed that the CAT item selection methods discussed here are practically indistinguishable in terms of CAT efficacy and accuracy. Thus in line with previous research it can be concluded that when assessing physical self-concept by the PSDQ adaptively, the more recently developed Kullback-Leibler divergence procedure may not deliver real benefits compared to the traditional item selection approach based on maximizing Fisher information [hypothesis a) was accepted].

Since selecting an appropriate estimation method is crucial to CAT procedure, the current simulation compared three latent trait estimation methods: the maximum likelihood estimation (MLE), expected a posteriori trait estimation with uniform prior (EAP-u), and expected a posteriori trait estimation with standard normal prior distribution (EAP-n). Generally, all of these estimation methods produced a similar number of PSDQ administered items. Moreover, regardless of latent trait estimation method, the CAT estimates of physical self-concept ($\hat{\theta}$) correlated similarly with true latent trait values (θ^{*}) as well as with estimated latent trait values based on the full PSDQ ($\hat{\theta}^{PSDQ}$). Some differences were nevertheless observed at the higher extremes of the physical self-concept latent continuum (e.g. $\hat{\theta} \geq 2$), where using EAP-n resulted in a reduced test length compared to the other latent trait estimation methods, especially when lower measurement precision was desired (e.g. stopping rule SE = 0.45). This reduction in a test length when estimating extreme levels of the latent trait however came at the cost of a slightly larger bias at both ends of the latent continuum as compared to the MLE and EAP-u. The 'inward' bias (reflecting regression to the prior mean) of the EAP-n method observed in the current simulation comports with many other studies evaluating the accuracy of latent trait estimation methods (Chang & Ying, 1999; van der Linden & Pashley, 2010; S. Wang & Wang, 2001, 2002; Weiss, 1982). Notably, and in contrast to findings reported by Chen et al. (1997) or Chen et al. (1998), the bias functions for EAP-u, which were comparable to those produced by MLE, did not indicate substantial inward bias. This indicates that employing an informative prior distribution with Bayesian latent trait estimation methods (e.g. EAP) in PSDQ CAT can lead to a shorter test, but also it may reduce test accuracy at both extremes of the latent trait. Although such an observation may be of a theoretical interest, it would seem to have only negligible effect in a practical CAT administration of the PSQD [hypothesis b) was rejected; hypothesis d) was accepted]. Moreover it should also to be emphasized, that choosing an inappropriate informative prior may seriously distort the precision of the latent trait estimates (Boyd et al., 2010; Mislevy & Stocking, 1989; Seong, 1990) and may adversely affect the test length (Štochl et al., 2016b; van der Linden & Pashley, 2010). This fact was highlighted also in the current study, where EAP-n in combination with uniformly generated true latent trait values resulted in a slightly higher bias of the physical self-concept than any other combination of estimation method and true latent trait distribution (EAP-n with normal true latent trait distribution; EAP-u with normal true latent trait distribution; EAP-u with uniform true latent trait distribution). In conclusion, the present simulation underscores that MLE remains the recommended estimation method for practical applications of CAT with the PSDQ.

When using CAT with Monte Carlo simulation a vector of true latent trait values needs to be specified by a researcher in order to obtain simulated responses to the test items. In the current study, two types of the hypothetical true latent physical self-concept distributions (standard normal vs. uniform) were compared with respect to the performance of the PSDQ CAT administration. Standard normal and uniform true latent trait distribution produced very similar bias of the physical self-concept CAT estimates. Employing generated true latent values of the physical self-concept with uniform distribution led to a higher number of administrated items [hypothesis c) was accepted], particularly for higher levels of measurement precision (e.g. stopping rules SE = 0.23 and SE = 0.32). Fortunately, a uniform distribution of physical self-concept is a very unlikely outcome when applied to an adolescent population, for which the PSDQ was developed (Marsh, 1996b; Marsh & Redmayne, 1994; Marsh et al., 1994). Therefore the average number of administered items in practical CAT applications for the PSDQ will likely be lower than indicated by the current results for uniformly distributed true latent trait values. In fact, the performance of CAT administration in a sample of youth drawn from the general population should resemble the results obtained using the standard normal true latent trait distribution – a more realistic distribution for physical self-concept in real-world conditions (Marsh, 1996a).

Even with the tremendous opportunity provided through CAT administration of the PSDQ, the present study also has several limitations. First, the findings relied exclusively on Monte-Carlo simulation resulting in the potential for real versus simulated CAT administration to produce different findings (Smits et al., 2011). This is mainly because the generated responses during CAT Monte-Carlo simulations follow precisely the IRT model used for item calibration (Štochl et al., 2016b). However examinee's real responses can vary considerably because of systematic or random error (Makransky et al., 2016). Fortunately, empirical examinations of these potential differences have shown little divergence in outcomes between real and simulated findings (Kocalevent et al., 2009).

Related to the previous limitation, the present study did not take into account the model misfit within the item calibration. The PSDQ item parameters used for the simulation were obtained from a published paper (Flatcher & Hattie, 2004) and the parameters were considered as true parameters. Flatcher and Hattie (2004) however reported relatively high standard errors of some item parameter estimates leading to the supposition that the departure of estimates from true item parameters could undermine validity of the CAT procedure (Wainer & Mislevy, 2000). According to van der Linden and Pashley (2010), ignoring errors of the item parameters estimates in CAT is a "strategy without serious consequences as long as the calibration sample is large" (p. 13). The sample used by Flatcher and Hattie (2004) for the PSDQ item calibration was relatively modest in size (N = 868) suggesting that re-calibration of the PSDQ items using larger samples may be required for future application of CAT when assessing physical self-concept.

In addition to the concerns raised above, conceptual differences may exist between the PSDQ CAT administration and the traditional fixed-length linear PSDQ assessment. The PSDQ was initially developed using principles comporting with a CTT framework and intended to measure 11 different specific sub-domains of general physical self-concept (Marsh et al., 1994). The present CAT simulation however used item parameters, which were calibrated using a unidimensional GRM[2] (Samejima, 1969). As a result, the model testing procedure assumed that adaptive administration of the PSDQ will adequately assess a single dimension of general physical self-concept. Although assessing a single dimension of general physical self-concept using the PSDQ may be legitimate for practical or research purposes (Flatcher & Hattie, 2004; Marsh, 1996a, 1996b; Marsh et al., 1994), some might argue that the general construct should tap all 11 proposed subdomains in order to represent the full nature of physical self-concept (Marsh & Redmayne, 1994; Shavelson et al., 1976). This might not be fulfilled when items within a CAT procedure are selected purely on the basis of statistical criteria – that is without applying content balancing methods. For example, it is very likely that using statistically motivated item selection procedures in PSDQ CAT administration (used in the current study), may lead to under-representation of the health subdomain items because these items provide relatively low amount of information along the latent continuum (Flatcher & Hattie, 2004). Future research should therefore explore whether application of content balancing methods using CAT with the PSDQ would be practically feasible and useful.

Despite these limitations, the current study has shown that CAT represents “a sophisticated method of delivering examinations” (Thompson & Weiss, 2011, p. 1) and improves the efficiency of a testing procedure. Using an assessment instrument commonly used in the field of Kinanthropology, the present study shows that CAT has a great potential for the assessment of physical self-concept and that the PSDQ is very well suited for this approach. Given the favorable results of the present simulation study, an interesting next step would be to evaluate the usefulness of the PSDQ CAT

[2] Unfortunately the item-level data from the original calibration (Flatcher & Hattie, 2004) were not available while conducting the present simulation study. It was therefore impossible to verify whether the unidimensional model is indeed the most suitable underlying description of the examinees’ responses to the PSDQ items.

administration in real testing conditions. Nevertheless the present findings provide very encouraging support for the use of CAT in Kinanthropology.

12 Conclusions

This thesis aimed to investigate the feasibility and usefulness of the adaptive administration of the Physical Self-Description Questionnaire while using a variety of item selection and latent trait estimation methods, distributional properties and test termination criteria. A Monte Carlo simulation study was designed to address the proposed aims. The main findings of the study can be briefly summarized as follows:

- CAT can successfully be applied as a method of reducing test length when measuring physical self-concept using the PSDQ items. Using a much shorter test, CAT provides latent trait estimates which are unbiased and correspond highly with the estimates based on administration of the whole questionnaire.
- More items with high positive threshold values should be incorporated into the PSDQ in order to improve the CAT efficiency when assessing the high levels of the physical self-concept.
- CAT using Kullback-Leibler divergence-based (FP-KL) and Fisher information-based (UW-FI) item selection methods respectively, led to almost identical average number of administered items from the PSDQ and produced very similar bias of the latent trait estimates. Either item selection method can therefore be recommended in further PSDQ CAT administrations.
- The maximum likelihood latent trait estimation (MLE), expected a posteriori estimation with uniform prior (EAP-u), and expected a posteriori estimation with standard normal prior distribution (EAP-n) were similarly effective with regard to the average number of administered items in PSDQ CAT. Some minor differences between these estimation methods were observed only at the higher end of the latent trait continuum, where EAP-n led to smaller average number of administered items but at the cost of higher bias of the latent trait estimates. Given the results of the present study the MLE may be recommended for future practical applications of CAT to assess physical self-concept.

REFERENCES

Anastasi, A. (1976). *Psychological testing.* (4 ed.). New York, NY: Macmillan Publishing.

Andersen, E. B. (1973). Conditional inferences for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*, 31-44.

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 123-154). New York, NY: Routledge.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus models. *Applied Psychological Measurement, 17*, 253-276.

Baker, F. B. (1965). Origins of item parameters X_{50} and β as a modern item analysis technique. *Journal of Educational Measurement, 2*, 167-180.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques.* (2. ed.). New York, NY: Marcel Dekker.

Baumgartner, T. A. (2006). Reliability and error of measurement. In T. M. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 27-52). Champaign, IL: Human Kinetics.

Becker, K. A., & Bergstorm, B. A. (2013). Test administration models. *Practical Assessment, Research, and Evaluation, 18*(14). Retrieved from http://pareonline.net/getvn.asp?v=18&n=14

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal research on drug use: Empirical findings and methodological issues* (pp. 267-302). New York, NY: Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 10*, 238-246.

Binet, S., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année psychologique, 2*, 411-463.

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. 392-479). Reading, MA: Addison-Wesley.

Blahuš, P. (1985). *Faktorová analýza a její zobecnění [Factor analysis and its generalization]*. Praha, CZ: SNTL.

Blahuš, P. (1996). Concept formation via latent variables modeling of motor abilities. *Kinesiology, 28*, 12-21.

Blahuš, P. (2010). *Methodology-based introduction to behavioral statistics, test theory and the latent factors model*. Unpublished manuscript.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory.* (pp. 33-50). New York, NY: Springer.

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., & Gibbons, R. (2010). Factor analisis of categorical item responses. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 155-184). New York, NY: Routledge.

Bock, R. D., & Lieberman, M. (1970). Fitting response model n dichotomously scored items. *Psychometrika, 35*, 179-197.

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 229-256). New York, NY: Routledge.

Bridgman, P. W. (1959). *The way things are.* Cambridge, MA: Harvard University Press.

Cohen, J. A. (1988). *Statistical power analysis for the behavioral sciences.* (2. ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* (2. ed.). Urbana, IL: University of Illinois Press.

Čepička, L. (2004). Assessing ball-handling skill in children using the Rasch analysis. *Journal of Human Movement Studies, 46*, 155-169.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*, 335-356.

Davis, L. L., & Dodd, B. G. (2008). Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. *Journal of Applied Measurement, 9*, 1-117.

Davis, L. L., Pastor, D. A., Dodd, B. G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement, 4*, 24-42.

de Ayala, R. J. (1989). A comparison of the nominal response model and the three.parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49*, 789-805.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* London, UK: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*, 109-117.

Devine, J., Fliege, H., Kocalevent, R., Mierke, A., Klapp, B. F., & Rose, M. (2016). Evaluation of computerized adaptive tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *Journal of Affective Disorders, 190*, 846-853.

Dodd, B. G., & de Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 301-317). Norwood, NJ: Ablex.

Dodd, B. G., Koch, W. R., & de Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd, B. G., Koch, W. R., & de Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*, 61-77.

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn and Bacon.

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika, 65*, 337-362.

Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373-388). New York, NY: Springer.

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

Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: the chalanges for health outcomes assessment. *Quality of Life Research, 16*, 187-194.

Flatcher, R. B., & Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise, 5*, 423-446.

Flaugher, R. (2000). Item Pools. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 37-60). Mahwah, NJ: Lawrence Erlbaum.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277-2291.

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computerized-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18*(1), 23-36.

Gershon, R. C., & Bergstorm, B. A. (2006). Computerized adaptive testing. In T. M. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 127-144). Champaign, IL: Human Kinetics.

Glas, C. A. W., & Vos, H. J. (2010). Adaptive mastery testing using a multidimensional IRT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 409-432). New York, NY: Springer.

Gulliksen, H. O. (1950). *A theory of mental tests.* New York, NY: John Wiley & Sons.

Guyer, R. D., & Weiss, D. J. (2009). Effects of early misfit in computerized adaptive testing on the recovery of theta. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/.

Hambelton, R. K., & Cook, L. L. (1983). The robustness of item response theory models and affects of length and samle size on precision of ability estimation. In D. Weiss (Ed.), *New horizons in testing* (pp. 31-49). New York, NY: Academic Press.

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

Hambelton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 21-42). New York, NY: Routledge.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2005). A randomized experiment to compare conventional, computerized, and computerized adaptive administration of ordinal polytomous attitude items. *Applied Psychological Measurement, 29*, 159-183.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hula, W., Kellough, S., & Fergadiotis, G. (2015). Development and simulation testing of a computerized adaptive version of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58*, 878-890.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Measurement in Education, 23*, 211-222.

Chang, H. H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response models with application to computerized adaptive tests. *The Annals of Statistics, 37*, 1466-1488.

Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of the maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*, 569-595.

Chen, S. K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT). *Educational and Psychological Measurement, 57*, 422-439.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*, 419-440.

Kingsbury, G. C., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Kocalevent, R., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., Kleiber, D., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology, 62*, 278-287.

Koch, W. R., Dodd, B. G., & Fitzpatrick, S. J. (1990). Computerized adaptive measurements of attitudes. *Measurement and Evaluation in Counseling and Development, 23*, 20-30.

Larkin, K. C., & Weiss, D. J. (1975). *An empirical investigation of two-stage and pyramidal adaptive ability testing.* (Research Report 75-1). Retrieved from Mineapolis: University of Mineapolis, Department of Psychology:

Leung, C. K., Chang, H. H., & Hau, K. T. (2000). *Content balancing in stratified computerized adaptive testing designs.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment, 2*, 3-15.

Loevinger, J. (1947) A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61, 4*.

Lord, F. M. (1952) A theory of test scores. *Psychometrika, Monograph: Vol. 7*.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lu, W. S., Lien, B. Y. H., & Hsieh, C. L. (2015). Psychometric properties of the Balance Computerized Adaptive Test in residents in long-term care facilities. *Archives of Gerontology and Geriatrics, 61*, 149-153.

Mair, P., & Hatzinger, R. (2007). CLM based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43.

Mair, P., Hatzinger, R., & Maier, M. J. (2010). Extended Rasch Modeling. The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory-based coomputerized adaptive testing version of the MacArthur-Bates Comunicative Development Inventory: Words and Sentences (CDI:WS). *Journal of Speech, Language, and Hearing Research, 59*, 281-289.

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facets scores with multidimensional adaptive testing: An illustration with the NEO PI-R. *Assessment, 20*, 3-13.

Marsh, H. W. (1996a). Construct validity of physical self-description questionnaire responses. *Journal of Sport and Exercise Psychology, 18*, 111-131.

Marsh, H. W. (1996b). Physical self-description questionnaire: stability and discriminant validity. *Research Quarterly for Exercise and Sport, 67*, 249-264.

Marsh, H. W., & Redmayne, R. S. (1994). A multidimensional physical self-concept and its relation to multiple components of physical fitness. *Journal of Sport and Exercise Psychology, 16*, 45-55.

Marsh, H. W., Richards, G. E., Johnson, S., Roche, L., & Tremayne, P. (1994). Physical self-description questionnaire: Psychometric properties and multitrait-multimethod analysis of relations with existing instruments. *Journal of Sport and Exercise Psychology, 16*, 45-55.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N. (2010). The partial credit model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 109-122). New York, NY: Routledge.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 529-544.

Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item response theory, and factor analysis: a unified framework. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 73-101). London, UK: Lawrence Erlbaum.

McDonald, R. P. (1967) Nonlinear factor analysis. *Psychometric Monographs, 15*.

McDonald, R. P. (1985). *Factor analysis and related methods.* New York, NY: Psychology Press.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations. recent developmerits, and applications* (pp. 3-14). New York, NY: Springer.

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. 77-102). New York, NY: Springer.

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351-363.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory.* (pp. 153-164). New York, NY: Springer.

Muthén, B., & Muthén, L. (1998-2016). *Mplus: Statistical analysis with latent variables (Version 7.4)*. Los Angeles, CA.

Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Nogami, Y., & Hayashi, N. (2010). A Japanese adaptive test of English as a foreign langueage: Development and operational aspects. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 191-213). New York, NY: Springer.

Nydick, S. W. (2013). *Multidimensional Mastery Testing with CAT.* (Doctoral dissertation), Faculty of the Graduate School, University of Minesotta.

Nydick, S. W. (2014). *catIrt: An R Package for Simulating IRT-Based Computerized Adaptive Tests. R package version 0.5-0.* Retrieved from https://CRAN.R-project.org/package=catIrt

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Thousands Oaks, CA: Sage.

Owen, R. J. (1969). *A Bayesian approach to tailored testing.* (Research Report 69-92). Retrieved from Princeton, NJ: Educational Testing Service:

Passos, V. L., Berger, M. P. F., & Tan, F. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement, 31*, 213-232.

Pastor, D. A., Dodd, B. G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using generalized partial credit model. *Applied Psychological Measurement, 26*, 147-163.

Patience, W. M. (1977). Description of components in tailored testing. *Behavioral Research Methods and Instrumentation, 9*, 153-157.

Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., Hjermstad, M. J., Kaasa, S., Loge, J. H., Velikova, G., Young, T., & GReoenvold, M. (2016). Development of an item bank for computerized adaptive test (CAT) measurement of pain. *Quality of Life Research, 25*, 1-11.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* New York, NY: Springer.

Popper, K. R. (2002). *The logic of scientific discovery*. New York, NY: Routledge.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ramsay, J. (1991). Kernal smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

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

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*, 11-15.

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

Roberts, J. S., Donoghue, J. R., & Laughlin, J. S. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 23*, 3-32.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph: Vol. 17*.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Samejima, F. (2010). The general gareded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77-108). New York, NY: Routledge.

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, NY: Springer.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct and interpretations. *Review of Educational Research, 46*, 407-441.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* (Vol. 5). London, UK: Sage.

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2012). Computerized adaptive assessment of personality disorder: Introducing the CAT-PD project. *Journal of Personality Assessment, 93*, 380-389.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188*, 147-155.

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

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72-101.

Spray, J. A. (1987). Recent developments in measurement and possible applications to the measurement of psychomotor behavior. *Research Quarterly for Exercise and Sport, 58*, 203-209.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of factors.* Paper presented at the Annual spring meeting of the Psychometric Society, Iowa city, IA.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 185-230). Mahwah, NJ: Lawrence Erlbaum.

Stocking, M. L., & Swanson, L. (1993). A method for severly constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the Annual Meeting of the Military Testing Association, San Diego.

Štochl, J. (2008). *Structure of motor symptoms of Parkinson's disease*. Prague, CZ: Karolinum.

Štochl, J. (2012). *Five essays on laterality*. Didcot: LAP.

Štochl, J., Böhnke, J., Pickett, K. E., & Croudace, T. J. (2016a). Computerized adaptive testing of population psychological distress: simulation-based evaluation of GHQ-30. *Social Psychiatry and Psychiatric Epidemiology, 51*, 895-906.

Štochl, J., Böhnke, J., Pickett, K. E., & Croudace, T. J. (2016b). An evaluation of computerized adaptive testing for general psychological distress: Combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Medical Research Methodology, 16.*

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 159-184). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43-76). New York, NY: Routledge.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen

(Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 101-134). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16*, 109-119.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*(1). Retrieved from http://pareonline.net/getvn.asp?v=16&n=1

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago, IL: University of Chicago Press.

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers and Education, 97*, 69-85.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1-10).

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Zeitschrift für Psychologie / Journal of Psychology, 216*, 3-11.

van der Linden, W. J. (2010). Sequencing an adaptive test battery. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 103-121). New York, NY: Springer.

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

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

van der Linden, W. J., & Chang, H. H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*, 107-120.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized* *adaptive testing: Theory and practice* (pp. 271-288). Dordrecht, DE: Kluwer Academic Publishers.

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, NY: Springer.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*, 393-411.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Maulman (Eds.), *New developments in psychometrics* (pp. 207-214). Tokyo, JP: Springer.

Veldkamp, B. P., & van der Linden, W. J. (2010). Designing item pools for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 231-246). New York, NY: Springer.

Verschoor, A. (2007). *Genetic algorithms for automated test assembly.* (Dissertation), University of Twente. Retrieved from https://pdfs.semanticscholar.org/4f54/050868db547617918a3122772b2056c0ce74.pdf

Verschoor, A., & Straetmans, G. J. J. M. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 137-150). New York, NY: Springer.

Vos, H. J., & Glas, C. A. W. (2010). Testlet-based adaptive mastery testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 389-408). New York, NY: Springer.

Wainer, H. (2000). Introduction and history. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 1-21). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2. ed.). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2. ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology, 57*, 1051-1058.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*, 317–331.

Wang, S., & Wang, T. (2002). Relative precision of ability estimation in polytomous CAT: A comparison under the generalized partial credit model and graded response model. *Advances in Psychology Research, 16*, 62-77.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika, 4*, 427-450.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

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

Wise, S. G., & Kingsbury, G. C. (2000). Practical issues in development and maintaining a computerized adaptive testing program. *Psicologia, 21*, 135-155.

Wood, T. M. (2006). Introduction. In T. M. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 3-8). Champaign, IL: Human Kinetics.

Wood, T. M., & Zhu, W. (2006). *Measurement theory and practice in kinesiology*. Champaign, IL: Human Kinetics.

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22*, 491-499.

Zhou, X., & Rackase, M. D. (2014). Optimal item pool design for computerized adaptive tests wtih polytomous items using GPCM. *Psychological Test and Assessment Modeling, 56*, 255-274.

Zhu, W. (1992). Development of a comuterized adaptive visual testing model. In G. Tenenbaum, T. Baz-Liebermann, & Z. Arati (Eds.), *Proceedings of the International Conference on Computer Application in Sport and Physical Education* (pp. 260-267). Natanya: Wingate Institute for Physical Education and Sport and the Zinman College of Physical Education.

Zhu, W. (2006). Constructing tests using item response theory. In T. M. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 53-76). Champaign, IL: Human Kinetics.

Zhu, W., Safrit, M. J., & Cohen, A. S. (1999). *FitSmart test user manal: High school edition.* Champaign, IL: Human Kinetics.

Appendices

Appendix A – IRT parameters (*a* – discrimination and *b*'s – thresholds) for the Physical Self-Description Questionnaire items (source: Flatcher & Hattie, 2004)

Subscale	Item	a	b_1	b_2	b_3	b_4	b_5
Health	1. When I get sick I feel so bad that I cannot even get out of bed.	0.19	-16.60	-8.92	-5.21	0.11	4.80
	12. I usually catch whatever illness (flu, cold, ect.) is going around.	0.17	-15.70	-10.80	-6.69	-1.20	4.36
	23. I am sick so often that I cannot do all the things I want to do.	0.38	-9.06	-7.41	-5.85	-4.48	-2.24
	34. I hardly ever get sick or ill.	0.40	-5.49	-3.20	-1.26	0.20	2.21
	45. I get sick a lot.	0.49	-7.38	-6.11	-4.70	-2.86	-0.91
	56. When I get sick it takes me a long time to get better.	0.39	-8.65	-6.72	-5.17	-3.06	-0.02
	67. I have to go to the doctor because of illness more than most people my age.	0.41	-8.70	-6.93	-5.23	-3.52	-1.80
	69. I usually stay healthy even when my friends get sick.	0.67	-4.30	-3.07	-1.89	-0.91	0.84
Coordination	2. I feel confident when doing coordinated movements.	0.93	-4.15	-2.95	-1.70	-0.59	1.14
	13. Controlling movements of my body comes easily to me.	1.07	-4.01	-3.04	-2.09	-0.86	0.60
	24. I am good at coordinated movements.	1.49	-2.80	-2.08	-1.33	-0.31	1.03
	35. I can perform movements smoothly in most physical activities.	1.72	-2.99	-2.06	-1.27	-0.27	0.85
	46. I find my body handles coordinated movements with ease.	1.35	-3.00	-2.17	-1.37	-0.28	1.03
	57. I am graceful and coordinated when I do sports and activities.	1.85	-2.63	-1.92	-1.10	-0.19	0.91
Physical activity	3. Several times a week I exercise or play hard enough to breathe hard (huff and puff).	0.97	-3.40	-2.62	-1.59	-0.70	0.38

	14. I often do exercise or activities that make me breathe hard.	1.01	-3.19	-2.06	-1.52	-0.57	0.49
	25. I get exercise or activity three or four times per week that makes me huff and puff and last at least 30 minutes.	0.79	-2.31	-1.55	-0.67	0.11	1.07
	36. I do physically active things (jogging, dancing, aerobics, swimming) at least three times a week.	1.18	-2.54	-1.81	-1.16	-0.57	0.09
	47. I do a lot of sports, dance, gym, or other physical activities.	1.65	-2.27	-1.82	-1.17	-0.62	0.10
	58. I do sports, exercise, dance or other physical activities almost every day.	1.50	-1.84	-1.34	-0.66	-0.16	0.61
Body fat	4. I am too fat.	1.05	-3.34	-2.58	-1.51	-0.80	-0.19
	15. My waist is too large.	0.97	-3.08	-2.34	-1.45	-0.57	-0.01
	26. I have too much fat on my body.	1.06	-2.92	-2.12	-1.28	-0.60	0.14
	37. I am overweight.	1.14	-2.82	-2.29	-1.67	-1.09	-0.52
	48. My stomach is too big.	1.12	-2.93	-2.20	-1.32	-0.66	-0.05
	59. Other people think that I am fat.	0.91	-3.42	-2.71	-1.96	-1.19	-0.57
Sport competence	5. Other people think I am good at sports.	1.63	-1.92	-1.43	-0.74	0.04	1.15
	16. I am good at most sports.	2.01	-2.47	-1.87	-1.16	-0.43	0.43
	27. Most sports are easy for me.	1.64	-2.78	-2.08	-1.22	-0.34	0.66
	38. I have good sports skills.	2.47	-2.15	-1.66	-1.18	-0.45	0.36
	49. I am better at sports than most of my friends.	1.74	-1.91	-1.33	-0.58	0.35	1.33
	60. I play sports well.	2.60	-2.27	-1.82	-1.33	-0.57	0.27
Global physical	6. I am satisfied with the kind of person I am physically.	1.28	-3.15	-2.41	-1.45	-0.60	0.59

	17. Physically, I am happy with myself.	1.48	-2.77	-2.31	-1.53	-0.67	0.30
	28. I feel good about the way I look and what I can do physically.	2.06	-2.38	-1.80	-0.96	-0.10	0.83
	39. Physically, I feel good about myself.	1.99	-2.43	-1.83	-1.09	-0.34	0.46
	50. I feel good about who I am and what I can do physically.	1.92	-2.69	-2.12	-1.39	-0.53	0.32
	61. I feel good about who I am physically.	1.20	-2.54	-2.05	-1.36	-0.48	0.39
Appearance	7. I am attractive for my age.	1.33	-1.76	-1.19	-0.12	1.09	2.11
	18. I have a nice looking face.	1.28	-2.20	-1.47	-0.35	0.95	2.34
	29. I’m better looking than most of my friends.	1.20	-1.69	-0.93	0.25	1.33	2.14
	40. I am ugly.	1.04	-2.92	-2.37	-1.48	-0.26	0.59
	51. I am good looking.	1.39	-1.97	-1.45	-0.45	0.75	1.88
	62. Nobody thinks that I’m good looking.	0.96	-3.16	-2.30	-1.22	0.06	1.01
Strength	8. I am a physically strong person.	1.17	-3.16	-2.47	-1.29	-0.03	1.43
	19. I have a lot of power in my body.	1.46	-3.16	-2.19	-1.15	-0.20	1.05
	30. I am stronger than most people my age.	1.02	-2.49	-1.59	-0.42	0.88	2.13
	41. I am weak and have no muscles.	0.92	-4.51	-3.65	-2.49	-1.31	-0.43
	52. I would do well in a test of strength.	1.48	-2.61	-1.76	-0.77	0.19	1.34
	63. I am good at lifting heavy objects.	0.86	-3.46	-2.47	-1.33	0.19	1.70
Flexibility	9. I am quite good at bending, twisting, and turning my body.	1.01	-3.00	-2.14	-1.08	-0.10	1.14

	20. My body is flexible.	1.18	-2.24	-1.48	-0.40	0.62	1.72
	31. My body is stiff and inflexible.	0.56	-6.03	-4.54	-2.91	-1.31	0.21
	42. My body parts bend and move in most directions well	1.03	-3.58	-2.71	-1.66	-0.68	0.72
	53. I think I am flexible enough for most sports.	1.84	-2.70	-1.96	-1.26	-0.36	0.57
	64. I think I would perform well on a test measuring flexibility.	1.32	-2.39	-1.55	-0.57	0.48	1.59
Endurance	10. I can run a long way without stopping.	1.51	-1.93	-1.41	-0.56	0.09	1.02
	21. I would do well in a test of physical endurance and stamina.	1.67	-2.17	-1.52	-0.61	0.29	1.28
	32. I could jog 5 kilometers without stopping.	1.24	-0.85	-0.35	0.27	0.75	1.38
	43. I think I could run a long way without getting tired.	1.44	-1.55	-1.03	-0.28	0.42	1.29
	54. I can be physically active for a long period of time without getting tired.	1.81	-2.10	-1.57	-0.82	-0.01	0.96
	65. I am good at endurance activities like distance running, aerobics, bicycling, swimming, or crosscountry skiing.	1.38	-2.07	-1.55	-0.74	0.10	0.97
Esteem	11. Overall, most things I do turn out well.	0.99	-4.36	-3.22	-2.11	-0.31	1.74
	22. I don’t have much to be proud of.	1.03	-3.25	-2.57	-1.72	-0.81	0.32
	33. I feel that my life is not very useful.	0.96	-3.59	-2.76	-1.99	-1.34	-0.50
	44. Overall, I am no good.	1.48	-2.91	-2.53	-2.09	-1.44	-0.72
	55. Most things I do, I do well.	1.71	-3.51	-2.39	-1.76	-0.55	0.71
	66. Overall, I have a lot to be proud of.	1.91	-2.68	-2.00	-1.37	-0.59	0.30
	68. Overall, I’m a failure.	1.32	-3.32	-2.83	-2.31	-1.48	-0.86
	70. Nothing I do ever seems to turn out right.	1.12	-3.31	-2.67	-1.90	-1.04	0.05

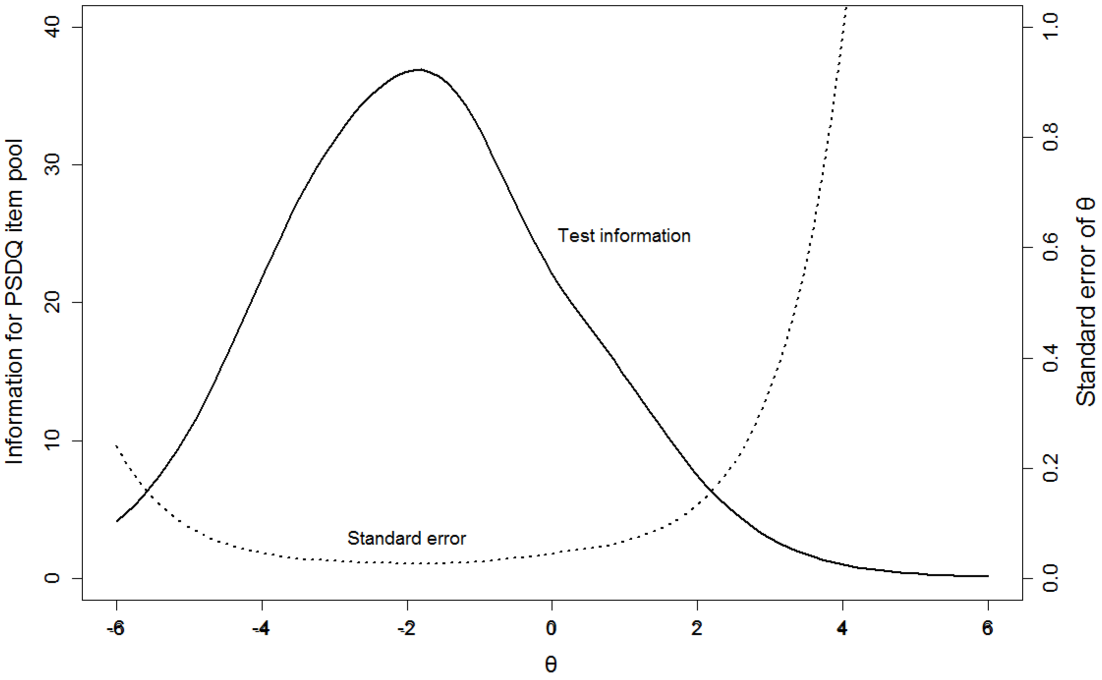
Appendix B – R code used for the simulation of the PSDQ CAT

```
# Loading IRT parameters for PSDQ items
load("J:/SelfDescriptionParameters.Rdata")

# Simulated true latent trait values
SimTrueTHETAn <- rnorm(1000)                          #standard normal
SimTrueTHETAu <- runif(1000, -3, 3)                   #uniform
SimTrueTHETA <- cbind(SimTrueTHETAn, SimTrueTHETAu)

# Values for standard error stopping rules
SEM <- c(0.23, 0.32, 0.39, 0.45)

# Choosing the item selection methods
Isel <- c("UW-FI", "FP-KL")

# Preparing table for simulation results
Results <- array(data = NA, dim=c(1000,5, 2,4,2,2,2),
dimnames = list(NULL,
c("CatTheta","CatLength", "TotalTheta", "TrueTheta", "catSEM"),
c("NormalTheta", "UniformTheta"),
c("SE_0.23","SE_0.32","SE_0.39","SE_0.45),
c("MLE", "EAP"),
c("UW-FI", "FP-KL"),
c("NormalPrior","UnifrormPrior")))



# CAT simulation
require(catIrt)

for (s in c(1:2)) {
  for (t in c(1:length(SEM))) {
    for (i in c(1:length(Isel))) {
```

```
CATresultsMLE <-
catIrt(SelfDescription, mod = "grm", resp=NULL, theta = SimTrueTHETA[,s],
catStart = list(init.theta = 0, n.start = 2, select = Isel[i], delta=0.1,
at = "theta", n.select = 1, score= "step", step.size = 1, leave.after.MLE = F),
catMiddle = list(select =  Isel[i], at = "theta", n.select = 1, delta=0.1,
score = "MLE", range = c(-6, 6), expos = "none"),
catTerm = list(term = c("fixed", "precision"), n.min = 3, n.max = 70,
score = "MLE",
p.term = list(method = "threshold",crit=SEM[t])),
ddist = dnorm, progress=TRUE)

CATresultsEAPn <-
catIrt(SelfDescription, mod = "grm", resp=NULL, theta = SimTrueTHETA[,s],
catStart = list(init.theta = 0, n.start = 2, select = Isel[i], delta=0.1,
at = "theta", n.select = 1,
score= "step", step.size = 1, leave.after.MLE = F),
catMiddle = list(select =  Isel[i], at = "theta", n.select = 1, delta=0.1,
score = "EAP", range = c(-6, 6), expos = "none"),
catTerm = list(term = c("fixed", "precision"), n.min = 3, n.max = 70,
score = "EAP",
p.term = list(method = "threshold",crit=SEM[t])),
ddist = dnorm, progress=TRUE)

CATresultsEAPu <-
catIrt(SelfDescription, mod = "grm", resp=NULL, theta = SimTrueTHETA[,s],
catStart = list(init.theta = 0, n.start = 2, select = Isel[i], delta=0.1,
at = "theta", n.select = 1,
score= "step", step.size = 1, leave.after.MLE = F),
catMiddle = list(select =  Isel[i], at = "theta", n.select = 1, delta=0.1,
score = "EAP", range = c(-6, 6), expos = "none"),
catTerm = list(term = c("fixed", "precision"), n.min = 3, n.max = 70,
score = "EAP",
p.term = list(method = "threshold",crit=SEM[t])),
ddist = dunif2, progress=TRUE)
```

```
# Saving results of the CAT simulation
Results[, 1, s,t,1,i,1] <- CATresultsMLE$cat_theta
Results[, 2, s,t,1,i,1] <- CATresultsMLE$cat_length
Results[, 3, s,t,1,i,1] <- CATresultsMLE$tot_theta
Results[, 4, s,t,1,i,1] <- CATresultsMLE$true_theta
Results[, 5, s,t,1,i,1] <- CATresultsMLE$cat_sem

Results[, 1, s,t,2,i,1] <- CATresultsEAPn$cat_theta
Results[, 2, s,t,2,i,1] <- CATresultsEAPn$cat_length
Results[, 3, s,t,2,i,1] <- CATresultsEAPn$tot_theta
Results[, 4, s,t,2,i,1] <- CATresultsEAPn$true_theta
Results[, 5, s,t,2,i,1] <- CATresultsEAPn$cat_sem

Results[, 1, s,t,2,i,2] <- CATresultsEAPu$cat_theta
Results[, 2, s,t,2,i,2] <- CATresultsEAPu$cat_length
Results[, 3, s,t,2,i,2] <- CATresultsEAPu$tot_theta
Results[, 4, s,t,2,i,2] <- CATresultsEAPu$true_theta
Results[, 5, s,t,2,i,2] <- CATresultsEAPu$cat_sem
```

Appendix C – Test information and corresponding standard error for the Physical Self-Description Questionnaire item pool



Appendix D – Example of R code used to create Figure 1

```
op <- par(mai= c(1.3,1.3,0.5,0.5))
a <- c(1,1,1)
b <- c(-1,0,1)
LineType <- c(1,2,6)
LineColour <- c(1,2,3)

for (i in c(1:3)) {
  curve(exp(x - b[i])/(1+exp(x - b[i])), -6, 6,
        xlab="θ", ylab="P(θ)",
        lwd = 2, lty=LineType[i],col=LineColour[i], cex.lab=1.5, cex.axis=1.5,
add=FALSE)
                  }
for (i in c(1:3)) {
  curve(exp(x - b[i])/(1+exp(x - b[i])), -6, 6,
        xlab="θ", ylab="P(θ)",
        lwd = 2, lty=LineType[i],col=LineColour[i], cex.lab=1.5, cex.axis=1.5,
add=TRUE)
                  }

legend(-6, 1, c("b1 = -1", "b2 =  0", "b3 =  1"),
       col = c(1, 2, 3),
       text.col = c(1, 2, 3),
       lwd = c(2, 2, 2),
       lty = c(1, 2, 6), pch = c(NA, NA, NA),
       merge = TRUE, bg = "white",
       cex=1.15, text.width=0.7, text.font=c(6,6,6),
       adj = c(0, 0.5))
```