CHARLES UNIVERSITY
FACULTY OF PHYSICAL EDUCATION AND SPORT

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

EXTENDED SUMMARY OF DOCTORAL THESIS

Author: Martin Komarc
Supervisor: Doc. PhDr. Jan Štochl, MPhil., PhD.

March 2017

INTRODUCTION

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.

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.

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).

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.

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. Item response theory (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.

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).

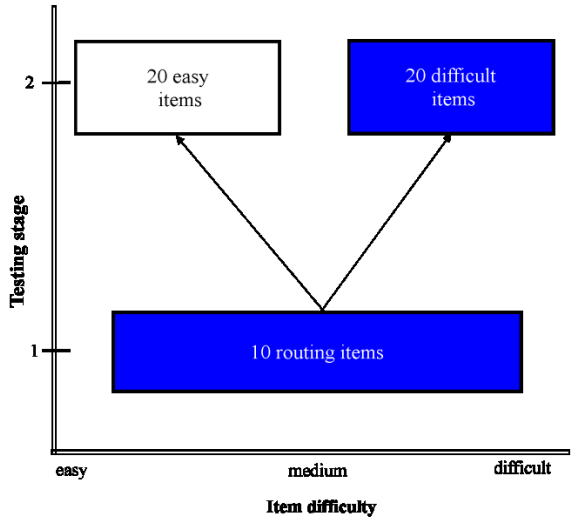

Figure 12 – Example of two-stage adaptive testing format

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.

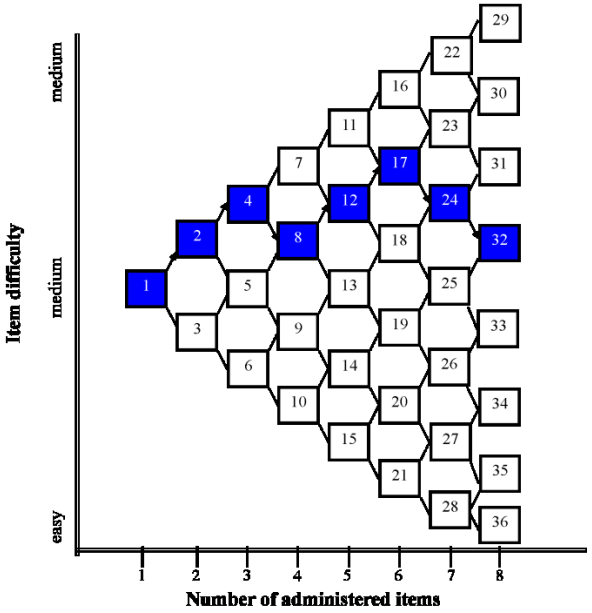

Figure 13 – Example of pyramidal adaptive testing format

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).

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 Physical self description questionnaire (PSDQ). The simulation study 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.

METHODS

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.

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

General description of the item pool

The 70-item PSDQ provided the item pool for the current simulation study. The PSDQ 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, Richards, Johnson, Roche, & Tremayne, 1994).

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 Grade response model (GRM) was used to estimate each item's discrimination and threshold parameters.

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.

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):

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.

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. 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.

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:

Latent trait (θ) estimation methods

The latent trait was estimated using one of the following methods: a) maximum likelihood estimation (MLE), b) expected a priori (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.

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. 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).

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.

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.

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 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 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 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, and another item is selected for the updated latent trait estimate. This process is repeated until a specified stopping rule was.

Analysis of simulation results

All simulations were performed in the *R* (R Core Team, 2013) statistical software using the *catIrt* package (Nydick, 2014). 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

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

- Mean absolute bias

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

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; 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 $\eta^2 > 0.14$.

RESULTS

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.

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.

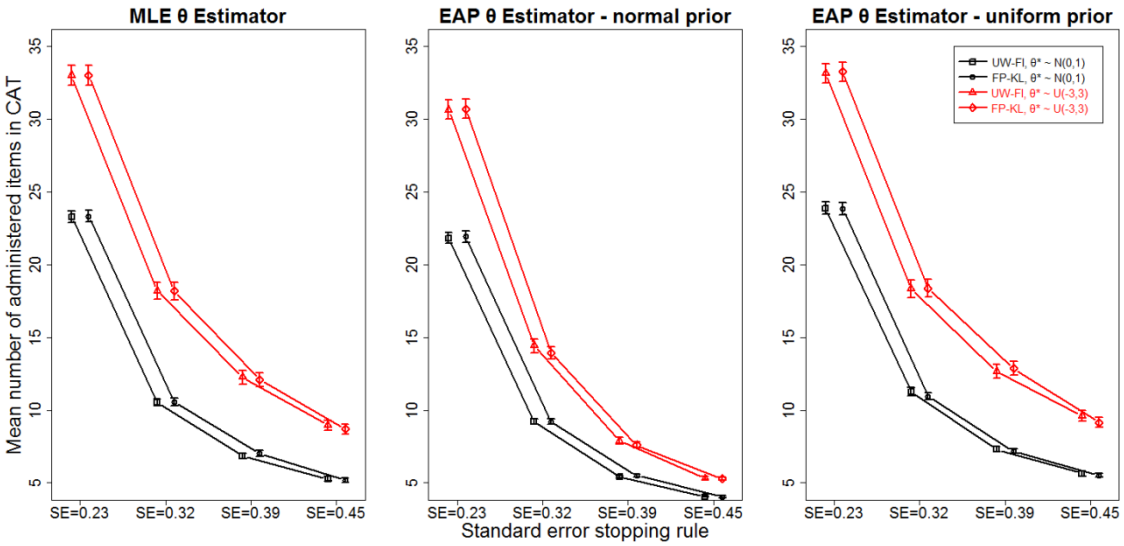

Figure 17 – Mean number of administered items from PSDQ in CAT simulations by level of measurement precision. Note: error bars represent standard error of the mean; shifts on x-axis within a particular SE are artificial to make all means visible.

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 ($\eta^2 = 0.051$, $p < 0.001$).

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.

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

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).

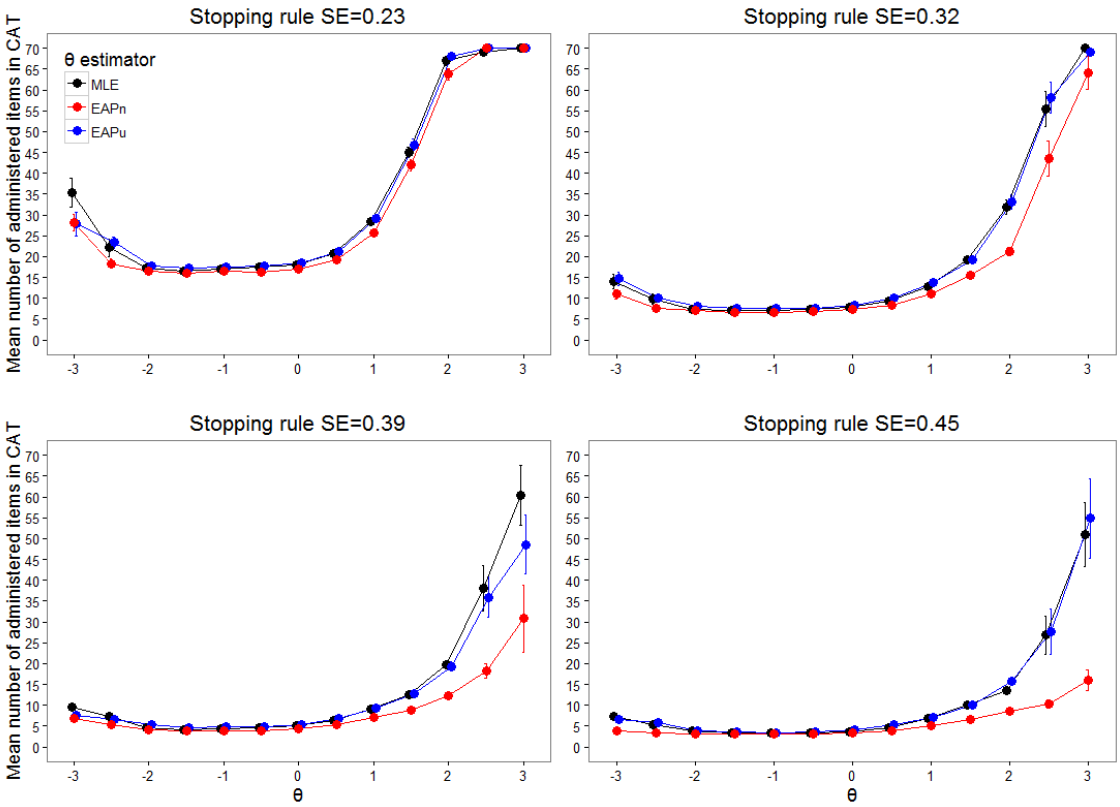

Figure 18 – Mean number of administered items from PSDQ (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for standard normal true latent trait (θ^*~ N(0,1)) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform prior; error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

The observation is a result of the distribution of the PSDQ items threshold and discrimination parameters and is therefore related to the item pool information function (see Appendix). 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).

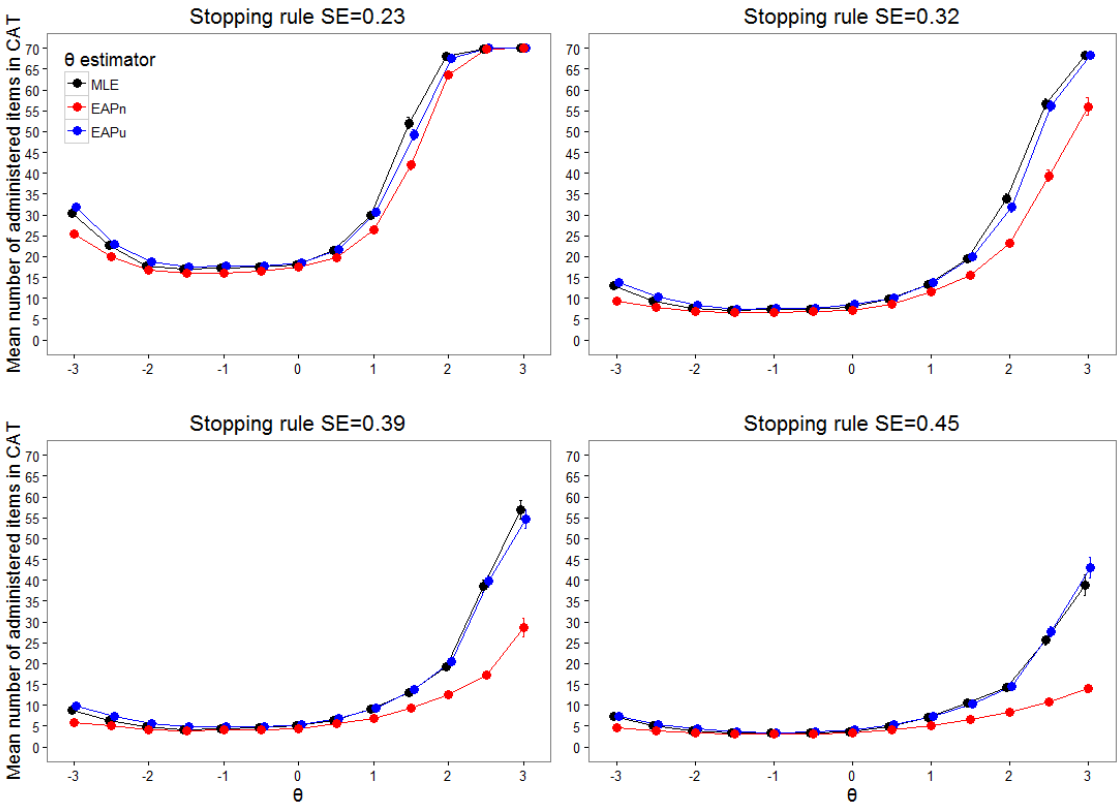

Figure 19 – Mean number of administered items from PSDQ (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for uniform true latent trait (θ^*~ U(-3,3)) distribution. Note: EAPn = EAP estimation with standard normal prior distribution; EAPu = EAP estimation with uniform prior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

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.

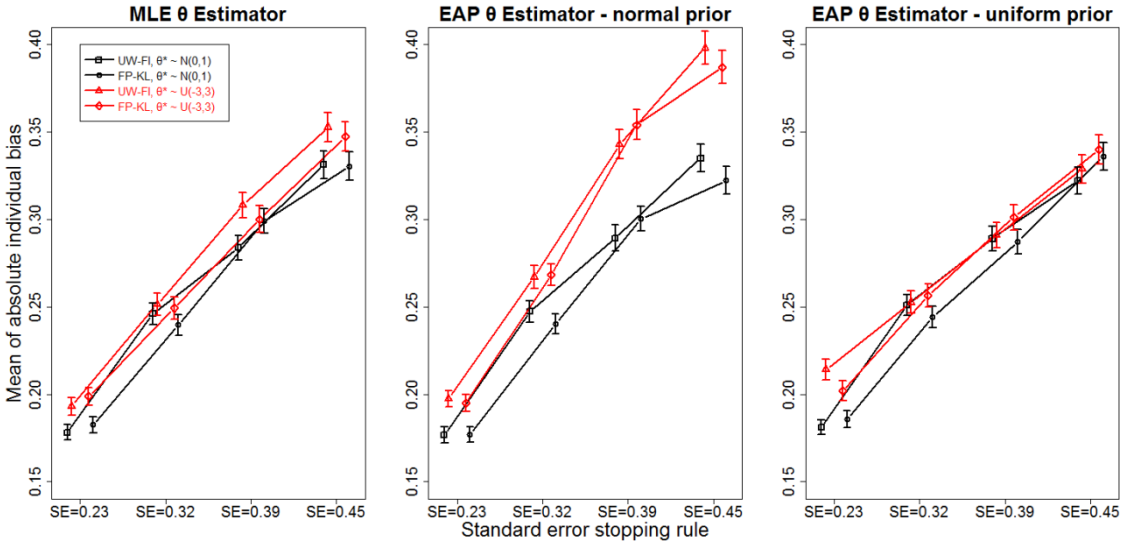

Figure 20 – Mean of absolute individual bias of CAT latent trait estimates by level of measurement precision. Note: error bars represent standard error of the mean; shifts on x-axis within a particular SE are artificial to make all means visible.

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 $\alpha = 0.05$ 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.

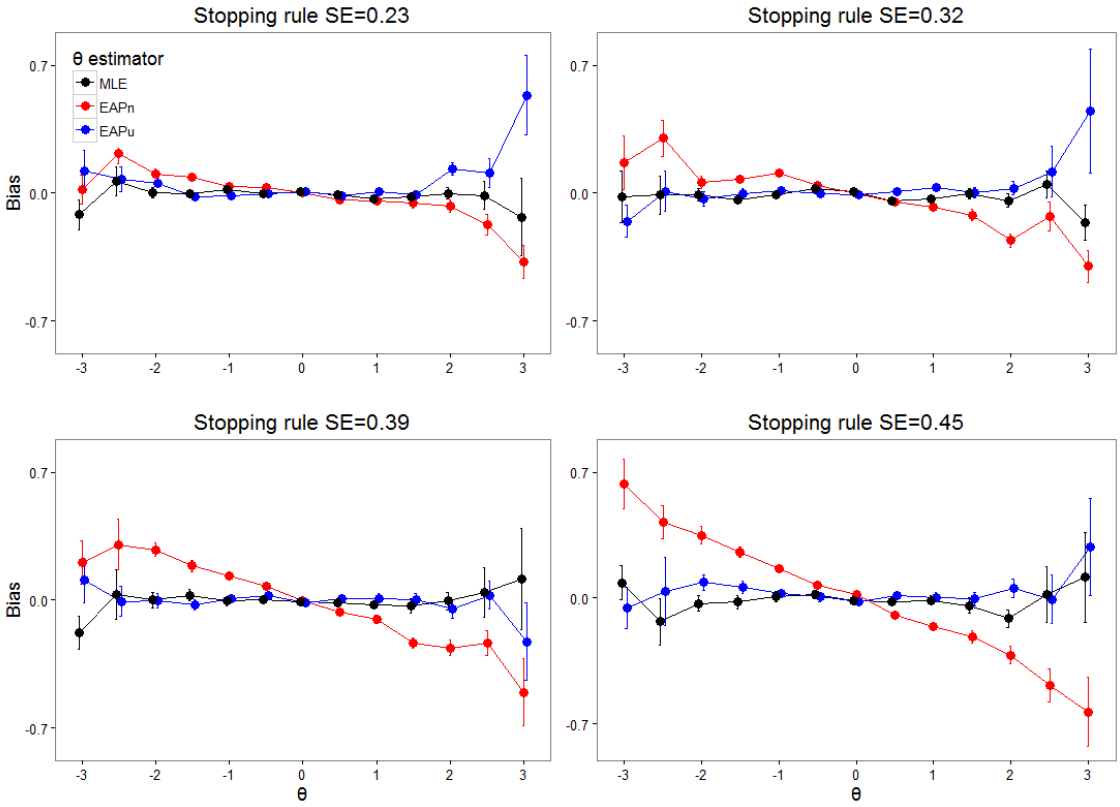

Figure 21 – Individual bias of CAT latent trait estimates (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for standard normal true latent trait ($\theta^* \sim N(0,1)$) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform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

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.

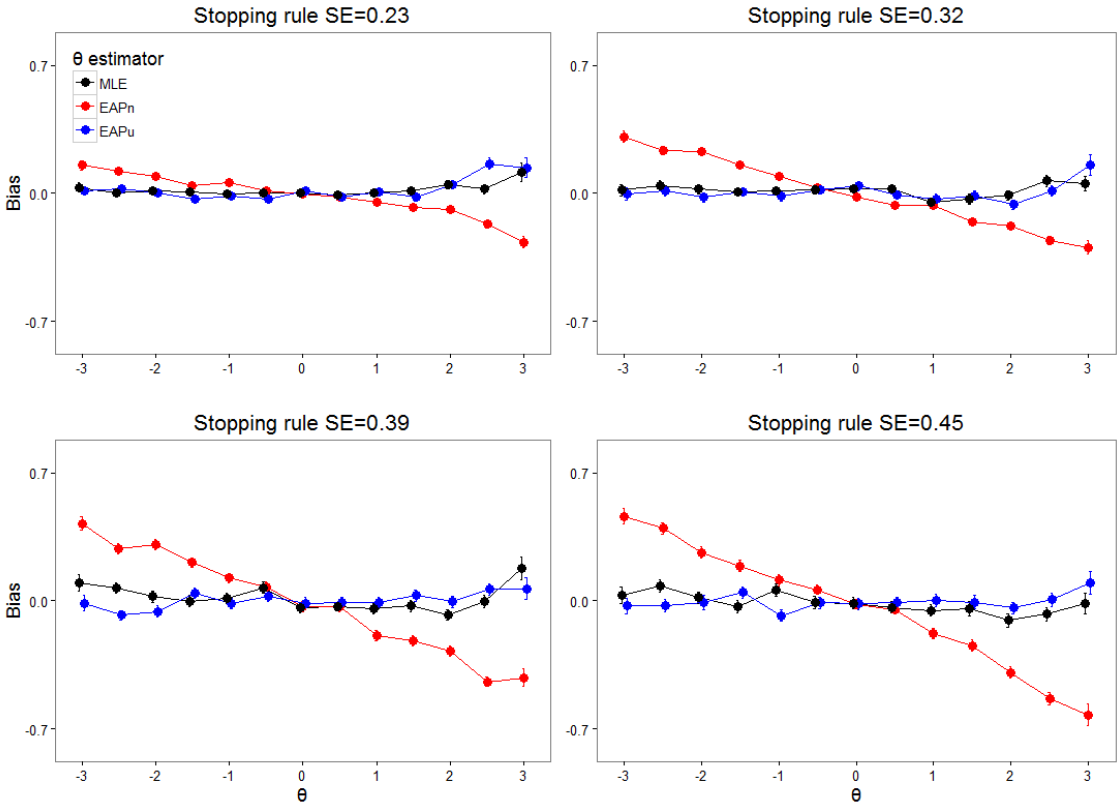

Figure 22 – Individual bias of CAT latent trait estimates (Y axis) as a function of CAT latent trait estimates ($\hat{\theta}$; X axis) for uniform true latent trait (θ^*~ U(-3,3)) distribution. Note: EAPn = EAP estimation with standard normal prior; EAPu = EAP estimation with uniform prior; error bars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

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.

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 $\theta^* \sim N(0,1)$				SE stopping rule for $\theta^* \sim U(-3,3)$			
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^*.

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.

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

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). The practical applicability of CAT was 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. 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). 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, Berger, & Tan, 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; Wang & Wang, 2001, 2002; Weiss, 1982). Notably, and in contrast to findings reported by Chen, Hou, Fitzpatrick, and Dodd (1997) or Chen, Hou, and Dodd (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, Dodd, & Choi, 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[1] (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 administration in real testing conditions. Nevertheless the present findings provide very encouraging support for the use of CAT in Kinanthropology.

[1] 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.

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.

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

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

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.

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.

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

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.

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

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.

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.

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/.

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

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.

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.

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:

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

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.

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

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/.

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.

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.

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/

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

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.

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.

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., & 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.

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

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

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

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

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.

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., & 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.

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.

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

APPENDIX

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