Opponent Review

Computerized Adaptive Testing in Kinanthropology: Monte Carlo Simulations Using the Physical Self-Description Questionnaire

Ph.D. Candidate: Martin Komarc

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Physical Education and Sport

Charles University

Prague, Czech Republic

Prepared by

Lawrence M. Scheier, Ph.D.

Doctoral Defense Commission

Doctoral Thesis Reviewer

June 14, 2017

LARS Research Institute, Inc.
Scottsdale, AZ USA

Executive Summary. Much of the cornerstone features of human thought including deliberation, reasoning, and even our emotions are hidden from others in the deep recesses of our private “mind.” The cascade of fleeting mental experiences that we use to create impressions of reality are not “observed” by others although we may share our most private thoughts through action and language. Put quite simply, we don’t see a person’s self-esteem but we infer they have high or low self-esteem from their actions and what they share through social interaction and communication. Likewise, we don’t see a person’s personality, but we infer they are “extroverted,” “conscientious,” or “agreeable” by their actions and communication. Nowhere is the distinction between observed and unobserved more important than in the world of testing and assessment, which frequently encounters the need to assess different aspects of mental and psychological functioning using self-report questionnaire items. The study of kinanthropology, which encompasses learning about physical self-esteem, routinely applies these same assessment techniques. Using these simple tools, test experts frequently ask people whether they “flexible and coordinated,” or “good in sports” to name a few examples. Questionnaire items provide test experts with the tools to make “inferences” about a person’ likes, dislikes, and their affinities. It is the collective nature of these questionnaire items, whether or not they feel coordinated, flexible or athletic that enables test developers to paint a more vivid picture of physical self-concept, even if they are unable to directly observe a person’s interior mental functioning or their very private thoughts.

Studies in kinanthropology have also shed light on the composition of physical self-concept, with test developers recognizing that it is not just made up of “one thing.” Height and weight can easily be recorded using standard and objective metrics. For strength, we can assess the sheer kilos of weight a person can lift in a clean and jerk motion or the number of empty barrels of beer they can heave onto a platform above their head. We can also strap a rope to their body and ask them to pull a two-ton tractor 25 meters, or lift giant 50 lb. kettle bells onto a platform; all examples of strength challenges that require muscle and brawn. Unfortunately, these examples of sheer strength do little to tell us how “strong a person feels” or how “flexible and coordinated” they perceive themselves. The same problem arises if we test a person’s endurance or their athletic ability through various athletic challenges, for example competing in a triathlon. Again, even the toughest endurance challenge does little to inform us how ‘athletic’ or competitive a person may feel. To gauge a person’s athletic ability or whether they feel “coordinated” we often resort to asking them ‘questions’ on a survey; questions that require the individual truthfully, and upon self-reflection, attest to their athletic skills. Using their answers, we can then make “inferences” as to their physical self-concept. These inferences form the basis of empirically reifying a latent “attribute” of physical self-concept.

Questionnaires that assess all the myriad different “facets” of a person’s physical self-concept requires a large number of items. There are drawbacks to using lengthy surveys, especially with youth, where the imposition of using lots of questions can often be burdensome. The demands of sitting for prolonged periods of time at a desk and answering many questions can induce fatigue or lead to errant answers. In some cases, the examinee can lose focus, resulting in skipped questions. The overall effect of an examinee shifting their attention or experiencing fatigue diminishes the value or accuracy of the questionnaire. The loss of crucial information undermines our ability to gain insight into a person’s true ability or sense of physical prowess. This presents a conundrum to test developers because they need many items to reliably assess physical self-concept. An ideal situation would present a reduced set of items that collectively and “efficiently” assess physical self-concept without loss of information.

Computer adaptive testing or CAT, offers one solution to finding a reduced set of items that can efficiently assess physical self-concept. Historically, CAT came about because of developments in mathematical modeling, prompted in part by the advent of faster and more efficient computers. CAT requires enhanced algorithmic processing for conducting complex matrix algebra and this requires computational speed and increased memory capacity. Bringing together the worlds of CAT, physical self-concept, and recent developments in psychometric theory is the focus of this candidate’s doctoral dissertation. Having read this work, it is clear that the candidate does a wonderful, if not masterful, job of combining these advances to artfully demonstrate the utility and benefits of applying CAT procedures. Set within the framework of simulated data, the candidate builds on recent advances in psychometric theory including developments in item response theory to empirically confirm that a 70-item questionnaire assessing physical self-concept can be shortened considerably without loss of crucial information. Various simulated tests are proposed to assess the adequacy of CAT methods under different distributional assumptions, using different scale reliabilities, latent trait estimation procedures, and invoking different starting and stopping item selection rules; all representing major extensions of CAT in the field of kinanthropology.

I am honored to have been able to review this dissertation. The writing is impressive, artful, cogent, and illustrative. The range of ideas tested in the dissertation, the clear objectives and well-stated rationale all summarily address major psychometric and statistical concerns. The dissertation presents an important interface of theory and application, which can have a profound influence on the world of testing and assessment. Indeed this work paves the way for applications that extend far beyond physical self-concept to include a host of “attributes” that we summarily use to describe people’s personality, abilities, competencies, performance, and behavior. I am confident that the results of this study will surely contribute to and advance the field of computer adaptive testing as well as establish important new developments in kinanthropology, assessment, psychometrics, and statistical modeling.

With warmest regards,

Lawrence M. Scheier, Ph.D.

Literature Review

This doctoral thesis presents a very detailed, comprehensive, and thoughtful review of the foundations of psychometric theory (classical test theory or CTT), weaving in more recent advances in modern test theory, and including further refinements that capitalize on developments in item response theory. The dissertation begins with a description of several challenges facing test developers including problems assessing latent traits with fallible empirical indicators, and further discusses the concepts of sufficiency, reliability and validity. The discussion of historical paths to modern test theory (owing to the work of Charles Spearman) is accurate, polished, and very insightful. The candidate elaborates the strengths and drawbacks of CTT including that it is population dependent (i.e., specific features of the population including underlying homogeneity of behavior) and relies on certain assumptions that are contextually dependent or are wholly unrealistic. The review continues with the introduction of modern test theory sparked by Lord's seminal publication on latent traits and Rasch's work with Item Response Theory (IRT). The discussion continues by introducing several benefits of IRT including the introduction of the logit scale, treatment and conditional dependence of measurement error, as well as factors that affect measurement precision in test development. Overall, the candidate correctly points out that IRT offers several opportunities that CTT does not and these advances have paved the way for the development of computerized adaptive testing; offering new ways to structure tests in simplified briefer formats.

Item Response Theory. The candidate then elaborates the functional and mathematical basis of IRT, a major component to the current focus on computerized adaptive testing (CAT). Both unidimensional and multidimensional models are briefly discussed, with the ensuing focus addressing primarily unidimensional models. The discussion continues with the differentiation of dichotomous (yes/no) and polytomous (more than two response options) IRT models. There is a very nice buildup to this material including both mathematical expressions (formula) and graphic figures showing the different types of Rasch models (1-PL vs. 2-PL). The discussion of the strengths and weaknesses of the different Rasch models is both accurate and erudite. Extensions of the 2-PL model to handle polytomous scoring (more than 2 ordered response categories) is introduced and neatly segues into the graded response model (GRM). The candidate goes on to delineate the different model parameters including item difficulty and item discrimination (and a guessing parameter in the 3-PL model) and showcases the various strategies for plotting item parameters using the item characteristic curve, operating characteristic curve, and with the 2-PL polytomous model, the item category response function.

The discussion then ventures into the partial credit model and the rating scale model (divided-by-total models), both alternatives to the graded response model (difference polytomous IRT model) and that modify the item category threshold parameter. The candidate then discusses the assumptions required for the unidimensional IRT models including unidimensionality of the latent trait being assessed and the requirement of 'local independence.' Certain conditions may not be met and the candidate mentions several remedies and technical solutions when a scale is determined to be multidimensional or when local independence may not be achieved (i.e., when items are intentionally clustered a condition frequently encountered with tests of reading comprehension).

Estimation Methods. The discussion then moves to the two most frequently used parameter estimation methods used in IRT including maximum likelihood and Bayesian estimation methods. Both strengths and drawbacks of these approaches are discussed as are various alternative estimation methods like weighted likelihood estimation (WLE) for cases of perfect or zero scores. This discussion is then broadened with the inclusion of conditions requiring simultaneous estimation of the item parameters and latent trait (when both are unknown). This includes joint maximum likelihood, conditional maximum likelihood, and marginal maximum likelihood estimation methods. The differences, strengths, and limitations of these different estimation methods are thoroughly discussed (i.e., distributional assumptions that produce biased estimates). The candidate continues by introducing several concerns revolving around estimation of standard errors of the theta (θ) estimates. This is an important area as, even with the best of measures, there will be some error in the estimation of a latent trait and this can influence the utility of CAT methods. Large sampling variance is equated with imprecision in the estimate of the latent trait, which translates into less information about the ability or trait in question.

Computerized adaptive testing. We then are introduced to the historical and conceptual origins of computerized adaptive testing. To understand the utility of CAT, we also need to recognize that many tests utilize linear fixed-length formats that present the same number of items in the same order to test takers. The candidate illustrates the Binet-Simon intelligence test, using this opportunity to illustrate perhaps the earliest forms of CAT (albeit used primarily in clinical settings, requiring immediate scoring, and involving manipulation of physical objects), given that subtests may or may not be administered depending on the examinee’s performance. Regardless of what test is being discussed, the use of fixed-length format tests has its drawbacks including lack of flexibility, the demands on test takers, and the large number of items required to provide full coverage of the trait being assessed.

Next is a discussion of the basic beginnings of mass-administered “adaptive testing” with the use of two-stage testing, flexilevel testing, or pyramidal adaptive testing, all of which involve some type of “routing” or channeling to sections of a test or more difficult items based on performance and involve also specific testing algorithms (start, continuation, and stopping rules). There are limitations with fixed-branching adaptive testing that are remedied by IRT-based variable-branching adaptive testing. The strengths of the variable-branching procedure are discussed including consideration of item characteristics, control over the precision of latent trait estimates, greater and more precise control over item selection, and level of measurement precision (termination can be based on a specified degree of reliability).

Testing algorithms. The candidate next introduces the role of “testing algorithms” in IRT-based CAT, an area that receives considerable attention in the simulation analyses. This involves discussion of the start, continuation and stopping rules applied in CAT. Because at the beginning of a testing situation there is so little information to rely on for estimating theta (i.e., the latent trait), start rules can rely on the population mean in the absence of additional information. They can also use the Fisher information function, using a random initial theta value to avoid overexposure of the first item and then select an item that efficiently informs theta. Drawbacks to the ML estimation are also mentioned (ML does not provide accurate estimates in the beginning stages of testing with perfect or zero scores) as are drawbacks to the Bayesian approach that requires specification of priors (relying on priors with too limited information can introduce bias).

Item selection. The Fisher information function is also the basis for formulating “continuation rules” but also suffers from problems related to the “attenuation paradox” (flattened likelihood function) and other test information deficits that accompany early stages of testing. The weighted Fisher information function addresses some of these deficits. The candidate also discusses the Kullback-Leibler information selection method, which uses the difference between the true theta value for the test taker and the current theta estimate, a divergence measure that can guide the item selection process. Several alternative Bayesian approaches to item selection are also discussed, most of which rely on information drawn from a posterior theta distribution (combining likelihood functions for theta and prior theta distributions).

Stopping rules. Stopping rules are also discussed including rules for fixed-length CAT (reaching a pre-specified number of items) and variable-length CAT (reaching a pre-specified value for the standard error of theta). The next section elaborates additional factors that can affect item selection including item pool, content balancing, and exposure control. The size of the item pool is driven by the “stakes’ involved when using the test and whether decisions are going to be made that may be irreversible, as is often encountered with “selection” procedures in the military or acceptance to a highly specialized training program or University. Moreover, the quality of items matters to ensure equiprecise measurement. Content balancing (ensuring adequate representation of subdomains or subtests) is also discussed as is the requirement for unidimensionality of the latent trait, and exposure control (e.g., maintaining test security so that items do not become overused and available to the public). The candidate then discusses item selection and trait estimation methods for CAT showcasing the need for simulation studies to examine performance (i.e., bias) of various parameters under different “real world” conditions.

Methods

The detailed and very systematic introduction paves the way for the candidate to lay out an organized plan to test various features of CAT with regard to the Physical Self-Description Questionnaire, a well-known paper-and-pencil assessment of physical self-concept. The plan includes using a simulation study to test various CAT features with a calibrated item bank under various conditions (i.e., testing algorithms that include item selection, latent trait estimation methods, start and stopping rules). The candidate defends why a Monte Carlo simulation is being used in contrast to a post-hoc simulation, the former using stochastically generated responses, whereas the latter relies on real responses.

Simulation exercise. The 70-items in the PSDQ provide the “item pool” for the simulation and the candidate provides a very thorough description of the scale development, its contents, and psychometric information. The first test involves both EFA and CFA of the 70 items to ensure the instrument meets the unidimensional factor structure requirements for a 2-PL GRM. The candidate then provides a detailed overview of the simulation procedure and its requirements including specification of the CAT algorithmic components: Standard normal distribution and uniform distribution, item parameters (discrimination and threshold parameters), decision rules for starting, initial theta estimation methods, continuing rules, and stopping rules (based on a measurement precision cut-off value corresponding to SE values of .23, .32, .39, and .45 with corresponding reliabilities of .95, .90, .85, and .80, respectively), latent trait estimation (MLE, EAP with uniform prior, and EAP with standard normal prior), item selection methods (unweighted Fisher information and fixed-point Kullback-Leibler divergence-based method). The combinations outlined above produces the 2 x 3 x 2 x 4 matrix with 48 simulation conditions. The candidate further describes the means of evaluating the simulation results using ANOVA with a defined measure of bias for the CAT latent trait estimates and in concert with the actual test length (number of items administered to achieve the various benchmarks), and correlations of the CAT theta estimates with the true values of theta.

Results

The first test results pertain to the dimensionality of the PDSQ. This analysis checks the dimensionality using stochastically generated item parameters and latent trait values generated from the Monte Carlos simulations. A total of 1000 latent trait values were generated following a standard normal distribution and 1000 following a uniform distribution. Models specifying 1 and 2 factors were tested using the Mplus software with Exploratory Factor Analysis (EFA) using the WLSMV adjusted estimation method and Geomin rotation. Various fit indices indicated a suitable fit for the 1-factor model with negligible improvement specifying a 2-factor model.

Graphic presentation of the mean number of administered items for the CAT procedures indicate that the average number of items administered ranged from 22 to 35 regardless of the theta distribution and for each of the CAT factors that were varied (estimation method and type of distribution). This number decreased to between 14 and 18 with high test accuracy, and further decreased to between 4 and 10 with a stopping rule where the SE = .45. ANOVA results for the test length (number of items administered) indicate that among the different simulation conditions only latent trait estimation method, theta distribution and stopping rule were significantly different ($p < .001$). Effect sizes (η^2) show that the stopping rule accounted for the most variance (30.2%). The interaction of latent trait estimation method and theta distribution was significant (p < .001) as was the interaction of theta distribution and stopping rule, albeit both of these model terms provided trivially small effects.

Further graphical analysis shows that more items had to be administered with high levels of the latent trait and with increased measurement precision (SE = .23). Still, the item pool required was half of the original 70 items. The candidate further points out that performance of the CAT simulation produced fairly economical findings regardless of the latent trait estimation method or theta distribution (MLE vs. EAP and standard normal vs. uniform prior). Additional evidence suggests the need for administering additional items when latent trait values (theta) were in the high end of the distribution (i.e., high trait values require additional items).

Additional analyses indicated that bias in the latent trait estimates increased almost monotonically with the magnitude of the SEs (lower reliability). Bias differed very little with choice of estimator, albeit the EAP estimator with the uniform true prior distribution (vs. the standard normal prior) produced high values of bias with higher SEs. Choosing different item selection methods produced very little difference in the number of items that had to be administered. ANOVAs conducted on the bias parameter indicated that stopping rule (SEs) again accounted for the most variance among the main effects (6.7%) and although accounting for trivially small amounts of variance, the interaction of latent trait estimation method and stopping rule SE, the interaction of latent trait estimation method and theta distribution, and the interaction of theta distribution and stopping rule SE were all significant. Further plots of bias as a function of estimated distributions (uniform and standard normal true theta) indicate that CAT loses its efficacy at extreme levels of physical self-concept on the latent trait continuum.

Correlations between the estimated latent trait values (theta) and generated true latent trait values were impressively high under the conditions of high measurement precision (SE = .23) irrespective of the other factors varied in the simulation. Importantly, the magnitude of these correlations were still appreciably high under the other conditions varied in the simulation underscoring the usefulness of the PDSQ CAT administration. Additional information reveals that the association between latent trait estimates and estimates based on the full PDSQ are relatively high, suggesting that even with curtailed set of items using CAT the method produces veridical estimates of physical self-concept.

Discussion and Conclusion

Based on the information presented in the results section and in concert with the methods tested in this thesis, the candidate presents a very nicely written and well developed discussion raising many good points regarding the usefulness of CAT methods. Put quite simply, the goals of CAT are to improve testing efficiency by finding out, under varying conditions, whether a shorter test can produce valid and unbiased estimates of a latent trait, in this case, physical self-concept. In their entirety, the findings from this study reinforce the utility of CAT methods under several latent trait estimation methods, item selection algorithms, distributional properties, and stopping rules. The candidate does a nice job of systematically summarizing a vast pool of statistical findings, always reinforcing that CAT has utility even under conditions demanding high measurement precision (SE = .23). The simulation analyses continually reinforce there was minimal bias produced with varying conditions, resulting in time and cost savings when implementing CAT procedures with instruments used in kinanthropology. The discussion of potential limitations to the study is relevant and sufficient, owing mainly to the use of simulation to test various permutation of CAT. Overall, the discussion remains a true rendition of the study findings and presents this material in a cogent, scientific manner.

Commentary

This is a very well organized and detailed dissertation that shows a lot of incredible hard work and lessons learned. The candidate reinforces that he maintains an excellent grasp of the material including outstanding psychometric knowledge, a strong foundation in classical and modern test theory, a sound working knowledge of simulation techniques, and shows a willingness to integrate diverse statistical methods. Furthermore, the candidate shows a complete awareness of the implications of his work for the field of kinanthropology as well as CAT. I particularly enjoyed reading the Introduction, which provided a very well rounded picture of the field including a historically accurate review of CTT and IRT. The rich and precise explanations of the different latent trait estimation methods, distributional properties, item selection algorithms, and stopping rule permutations applied in the simulations was incredibly concise and well written. The candidate has a clear intellectual grasp of the material, has shown an incredible propensity to conduct independent research, and has acquired very important computational, psychometric, and statistical skills; skills that have tremendous application even beyond the world of CAT. It is worth noting that I am confident this work represents only beginning of a very illustrious career in academia and that the candidate has moved through his graduate studies with an eye cast toward a scholarly career.

Editorial Comments:

Page 21: “When a set of items correspond to the model **than** “specific objectivity” means …”

Should be “**then** “specific objectivity means …”

Page 33: the citations beginning with Baker (165) through Brock and Leiberman (1970) should not have “;” rather each citation should be separated by a “,”.

Page 39: The 2nd paragraph last sentence “can be found in (Baker & Kim, 2004) is incorrect for APA Style and should be, “can be found in Baker and Kim (2004).”

Page 42: “This it is also possible in 2-PL model …” should be “in **a** 2-PL model …”

Page 45: Moreover comparison of examinees taking the same test ...” Moreover, comparison of …” (Needs a comma “,” after moreover)

Page 50: “use pre-specified fixed patterns of item selection procedure …” grammatically incorrect and should be “use pre-specified fixed patterns of item selection procedure**s** …”

Page 71: “process of implementation of CAT …” should be “process of *implementing* CAT …” a sentence structure that avoids the use of double “infinitives (of … of).

Page 80: The Table 3 is missing a header for the Factors (above loadings), which should read “1 factor and 2 factors” (see the continuation page on Page 81 where you do have the correct heading).

Page 83: “indicating no effect of these model terms …” is an incorrect conclusion, since both terms were significant (p < .001) but trivially small.

Page 95: “Moreover this rather substantial reduction …” requires a “,” after Moreover, …

Page 96: “when measuring high levels of ~~the~~ physical self-concept.” Delete “the” in this sentence.