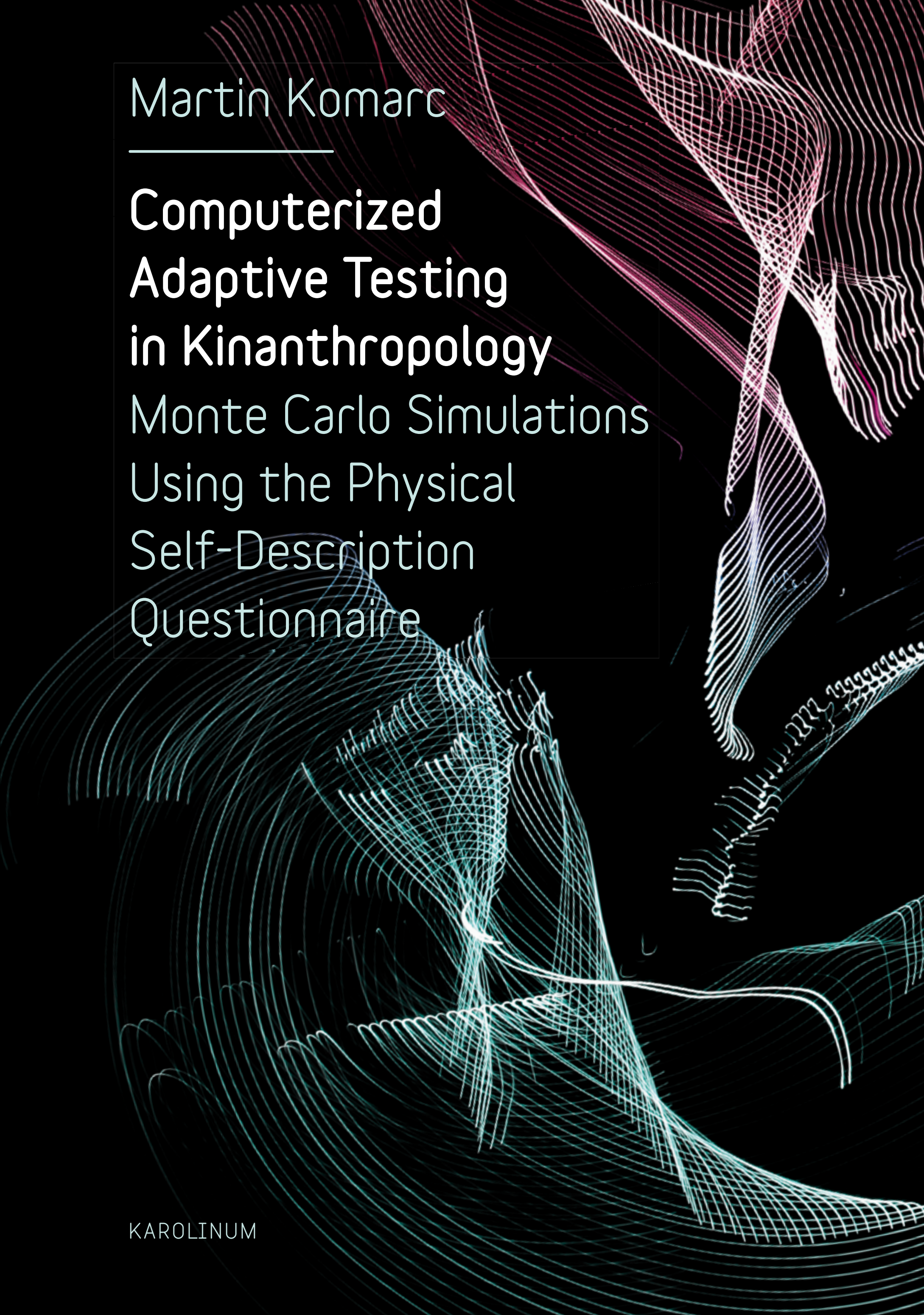
Martin Komarc

Computerized Adaptive Testing in Kinanthropology

Monte Carlo Simulations Using the Physical Self-Description Questionnaire



KAROLINUM

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

Martin Komarc

Reviewers:
RNDr. Patrícia Martinková, Ph.D.
prof. RNDr. Jan Hendl, CSc.

Published by Charles University
Karolinum Press
Cover design Jan Šerých
Set by Stará škola (staraskola.net)
First Edition

ISBN 978-80-246-3918-5
ISBN 978-80-246-3984-0 (pdf)

Charles University
Karolinum Press 2019

www.karolinum.cz
ebooks@karolinum.cz

Contents

Acknowledgements

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

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

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

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

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

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

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

This study was supported by the following projects and grants: UNCE 32 PRVOUK 39, PROGRES Q19, SVV 2016-260346, SVV 2017-260446, GAUK 962214 and GAUK 110217.

1. Brief introduction to measurement (in Kinanthropology)

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

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

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

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

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

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

2. Historical Paths to Modern Test Theory

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

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

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

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

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

3. Groundwork for Item Response Theory

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

Development of IRT was perhaps slowed by its computational complexity, which has been greatly facilitated by the increased computational capacity and speed of modern computers. The advent of powerful and relatively inexpensive computers introduced in the 1980s paved the way for IRT to be "the most dominant theory for test construction in all major testing organizations or agencies such as the Educational Testing Service (ETS) and American College Testing (ACT)" (Zhu, 2006, p. 53). The first systematic treatment of IRT in Kinanthropology is generally credited to Spray (1987), who introduced its advantages and described its practical applications in the measurement of psychomotor behavior. Since this introduction, many successful applications in Kinanthropology have followed (see Wood & Zhu, 2006 for review). In the past few years, researchers in the Czech Republic have used IRT successfully to address several kinanthropological research questions (e.g., Čepička, 2004; Štochl, 2008, 2012); questions that would be difficult – if not impossible – to answer within the CTT framework.

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

4. Item Response Theory (IRT)

4.1 Introduction

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

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

4.2 Unidimensional dichotomous IRT models

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

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

Here $P\left(\theta_j\right)$ indicates the probability of examinee *j* with latent trait θ responding to a keyed (correct, or trait indicating) category of item *i* with difficulty *b*. This model was first introduced by Danish mathematician George Rasch (1960) and is commonly referred to as dichotomous Rasch model or one-parameter logistic (1-PL) model in the literature (Embretson & Reise, 2000). The θ parameter of examinee *j* is theoretically unrestricted, and is quite similar to the well-known z-score, scaled to mean of 0 and standard deviation of 1 (it usually ranges from −3 to 3 in typical population)[1]. Larger values of θ indicate higher latent trait levels. Individuals with higher values of the latent trait are more likely to get the item correct, or generally to give a positive (keyed, trait indicating) response to a test item. The

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