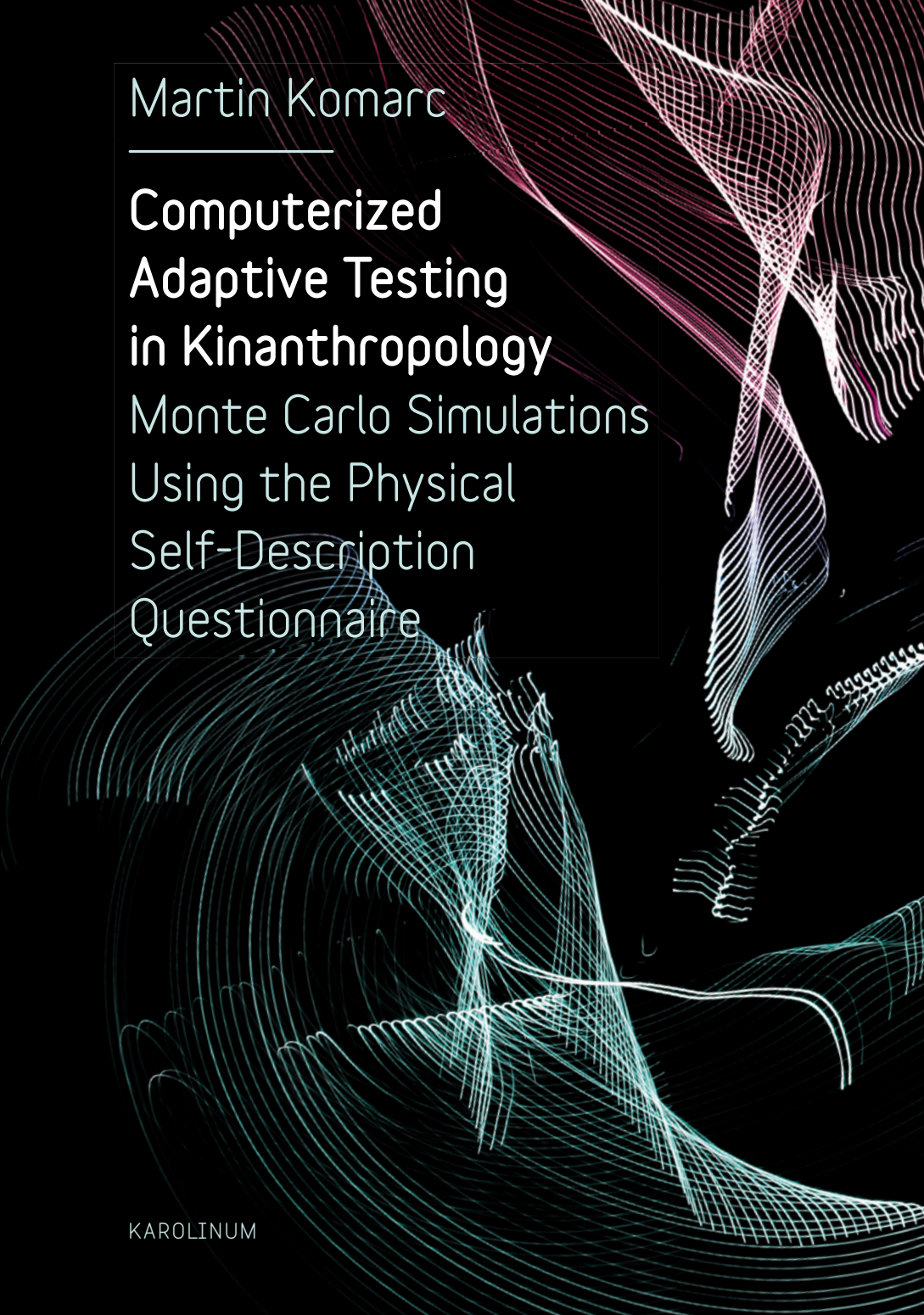

Martin Komarc

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.