Subscores

This authoritative guide directs consumers and users of test scores on when and how to provide subscores and how to make informed decisions based on them. The book is designed to be accessible to practitioners and score users with varying levels of technical expertise, from executives of testing organizations and students who take tests to graduate students in educational measurement, psychometricians, and test developers. The theoretical background required to evaluate subscores and improve them is provided alongside examples of tests with subscores to illustrate their use and misuse.

The first chapter covers the history of tests, subtests, scores, and subscores. Later chapters go into subscore reporting, evaluating and improving the quality of subscores, and alternatives to subscores when they are not appropriate. This thorough introduction to the existing research and best practices will be useful to graduate students, researchers, and practitioners.

SHELBY HABERMAN is a statistician with substantial experience in educational testing. He currently works as an independent consultant. Previously he taught at the University of Chicago, Hebrew University, and Northwestern University and was later employed by the Educational Testing Service (ETS). He is a fellow of the Institute of Mathematical Statistics, the American Statistical Association, and the American Academy for the Advancement of Science and was the 2019 recipient of the National Council on Measurement in Education Award for Career Contributions to Educational Measurement. Haberman is the author of *The Analysis of Frequency Data*, *Analysis of Qualitative Data*, and *Advanced Statistics*.

SANDIP SINHARAY is a distinguished presidential appointee in the Research & Development division at the ETS in Princeton, New Jersey. He is the current editor of *Psychometrika* and a past editor of the *Journal of Educational Measurement* (2020–2022) and the *Journal of Educational and Behavioral Statistics* (2010–2014). He has received seven awards from the National Council on Measurement in Education. These awards include the Outstanding Service Award (2023), Bradley Hanson Award for Contributions to Educational Measurement (2018 and 2020), Annual Award for Outstanding Technical or Scientific Contribution to the Field of Educational Measurement (2015 and 2009), Jason Millman Promising Measurement Scholar Award (2006), and Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar (2005). He received the ETS Scientist award in 2008 and the ETS Presidential award twice. Sinharay has coedited two published volumes, including the *Handbook of Statistics*, Volume 26 on psychometrics (which he coedited with Prof. C. R. Rao), and authored or coauthored more than 100 articles in peer-reviewed journals and edited books.

RICHARD A. FEINBERG is a principal psychometrician at the National Board of Medical Examiners and also teaches a course on research methods at the Philadelphia College of Osteopathic Medicine. Feinberg has received several awards and recognitions, including the National Council on Measurement in Education Jason Millman Promising Measurement Scholar Award (2018), the American Educational Research Association Division I Established Researcher Award (2017), and the Outstanding Research Publication Award (2022), and is a six-time winner of the Educational Measurement: Issues and Practice Cover Graphic/Data Visualization Competition. He has coedited one published volume and authored numerous articles across a broad range of statistical applications in the educational and psychological literature.

HOWARD WAINER is an independent statistician and author with experience in educational testing and data visualization. He has taught at the University of Chicago, Princeton University, and the Wharton School of the University of Pennsylvania. He was employed by the ETS from 1980 until 2001 and was the Distinguished Research Scientist at the National Board of Medical Examiners from 2001 until 2016. Wainer is a fellow of the American Statistical Association and American Educational Research Association, and the author or editor of 25 previous books.

Subscores

A Practical Guide to Their Production and Consumption

SHELBY HABERMAN
Independent consultant

SANDIP SINHARAY
Educational Testing Service

RICHARD A. FEINBERG
National Board of Medical Examiners

HOWARD WAINER
Independent consultant

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009413688

DOI: 10.1017/9781009413701

First published 2024

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-41368-8 Hardback
ISBN 978-1-009-41366-4 Paperback

Additional resources for this publication at www.cambridge.org/9781009413688

To Paul W. Holland, who has, for decades, provided wise counsel and guidance to us and others who toil in the same fields

To my late wife Penny and to my children Shoshanah, Chasiah, Sarah, Milcah, Boaz, and Devorah (SH)

To my parents, Puspamita and Radhanath Sinharay, who taught me that anything is possible except reporting subscores for unidimensional tests; to my siblings Joydeep and Manidipa for their continuous support and encouragement; and to my wife, Lopamudra, whose multidimensional skills have helped me become who I am (SS)

To my family, Betsy, Nathan, and Daniella Feinberg, for being genuinely excited whenever I have something to share (RF)

To Linda; to Laurent, Lyn, Koa, and Sophie; and to Sam – all of whom continue to bring joy and enlightenment (HW)

Contents

Preface *page* ix
Acknowledgments xv

1 **Introduction** 1
Background and historical context on what subscores are, using the US census and testing programs in the military as clarifying examples.

2 **How Are Subscores Reported?** 18
Different modalities for communicating subscore information, across varied testing purposes and audiences, are illustrated through excerpts from real score reports. Connections to general score reporting best practices and development are discussed.

3 **When and How Should Subscores Be Reported?** 45
A comprehensive review of the statistical requirements, methods to quantify value, and how to determine if subscores are worth reporting.

4 **A Survey to Explore the Conditions under Which Subscores Have Added Value** 88
Actual data from 36 operational testing programs are used to assess the extent to which reported subscores have added value.

5 **What to Do When Subscores Do Not Have Added Value: Augmentation, No Subscore Reporting, and Some Other Options** 110
Several approaches are recommended for what can be done when subscore information is needed yet the original subscores lack sufficient quality to be reported.

6 Coda: Lessons Learned, Conclusions, and Recommendations 136
Guidance for how subscores should be reported is discussed as well as what might be done when subscores ought not be reported. Advice is also given to help practitioners respond to pressure from various stakeholders when subscores would be misleading to report.

Appendix 150

Glossary 154
References 158
Name Index 169
Subject Index 172

Preface

Subscores have long been popular for intelligence tests. Psychologists have used results from different types of comparisons of the subscores on the Weschler Adult Intelligence Scale–Revised, Stanford–Binet Intelligence Scale, and so on to interpret performance on those tests. These analyses are often referred to as *profile* or *scatter* analyses. Subscores on educational assessments were prevalent but received scant attention during the twentieth century.

However, the first decade of the twenty-first century saw a change. There was a huge surge of interest in diagnostic scores for educational assessments, possibly due to the No Child Left Behind Act of 2001, which demanded, among other things, that students receive diagnostic reports to allow teachers to address their specific academic needs. Testing organizations responded by accelerating their recently begun or already ongoing efforts of reporting subscores and other types of diagnostic scores. Measurement researchers responded to the demand by suggesting various ways of computing diagnostic scores.

Despite good intentions, these efforts on subscore reporting had their limitations. Many of the reported subscores were based on only a handful of items, thus possessing dubious psychometric properties. Unfortunately, there was not much guidance for practitioners on when and how to report subscores. We began to feel the need to address one or more of these issues during the period from the late 1990s to 2010.

How We Became Involved with Subscores

The seed for the book was sown in the late 1990s when Kathy Sheehan reached out to her colleague Howard Wainer at the Educational Testing Service (ETS) with a knotty problem associated with a teachers' licensing examination. The score users wanted subscores; however, Kathy was hesitant to comply

because many of the potentially reportable subscores were based on only a few items. She felt that it would be unethical to supply scores that had unacceptably low reliability. Howard subsequently ran into his colleague Nick Longford at lunch, who suggested using the other items on the test to help stabilize the subscores and thought that this was a made-to-order application for an empirical Bayes approach. Nick sketched the idea on a napkin. Howard did the necessary algebra and handed it over to the remarkable Xiaohui Wang, who had just joined ETS as a data analyst. Xiaohui used the afternoon to write the code and run the program on Kathy's data to compute what are now called augmented subscores. Their reliabilities were stunningly high. It is now known that augmented subscores computed from any largely one-dimensional test will have high reliability regardless of the number of items formally constituting those subscores. But in the 1990s, the result looked too good to be true. This led Howard and Xiaohui to check the algebra and code several times. They later published three landmark papers on augmented subscores between 1998 and 2001. David Thissen insisted on referring to the augmented subscores as Wainer subscores in 2001, following Stigler's law of eponymy that nothing is ever named after the right person (which Stigler attributed to Robert Merton).

Despite this breakthrough, there remained the need for a widely accepted method for the parsimonious characterization of the psychometric quality of subscores. Enter Shelby Haberman, who joined the staff at ETS in 2002 shortly after Howard left ETS for the National Board of Medical Examiners (NBME).

Shelby, who joined the faculty of the University of Chicago in 1970 along with Howard Wainer (though they hardly knew each other in Chicago), became involved with subscores shortly after he arrived at ETS. He learned of ETS's interest in supplying diagnostic scores despite the possibility of their feeble psychometric properties. There were several ongoing projects, both from internal researchers and from external vendors, focused on producing subscores for ETS tests despite the very real questions about the prospective utility of these subscores. In response, Shelby tried to provide criteria for the utility of subscores and for methods to improve them. He published a paper in 2008 that suggested a simple yet elegant approach that summarizes a comprehensive evaluation of the psychometric quality of a subscore in a single number. That paper is widely regarded as a breakthrough in research on subscores and, as of January 2023, had been cited 265 times. Sandip Sinharay joined ETS in 2001, partially because his doctoral supervisor (Hal Stern – currently at the University of California–Irvine) knew of a job vacancy at ETS from his friend Howard Wainer. Sandip joined Shelby's research team for the study of subscores in the mid-2000s. Sandip's initial goal was disseminating to ETS

colleagues and outsiders the ideas behind Shelby's brilliant subscore-related papers that were highly technical and thus ill-suited for the faint of heart. Shelby and Sandip went on a vigorous spree of examining various aspects of subscore reporting in a series of papers, often involving ETS colleagues like Gautam Puhan and Kevin Larkin. Their research is documented in more than 20 technical reports, journal articles, and chapters in edited volumes. This earned them the euphonious appellation of the subscore police from their esteemed colleague Paul Holland and also the 2009 annual award for technical or scientific contribution to the field of educational measurement from the National Council on Measurement in Education.

In the meantime, Howard Wainer – who left ETS to join NBME in 2001 and was busy working on testlets, measurement problems in medical licensing tests, and several books on various topics – read Shelby's 2008 paper and figured that the next logical step would be the equating of subscores, which would make an excellent dissertation topic. Howard mentioned this topic to his NBME colleague Mike Jodoin, who contacted Richard A. Feinberg, then both a staff member at NBME and a graduate student at the University of Delaware. Rich was in the process of considering different plausible dissertation ideas. In 2010, Richard met with Howard to further discuss the idea, which instigated a productive research agenda between the two of them. Rich began communicating with Sandip and Shelby, who told him that ETS colleagues were already investigating equating of subscores (two papers on the topic were published in 2011). This led Howard and Rich to think carefully about other important practical questions that could add value to the rapidly growing subscore literature. Richard's dissertation evolved and led to several papers on previously unexplored and practical issues related to subscores.

The Origin of This Book

In July 2021, after completing work on his 25 book and passing beyond the biblically prescribed life span of 3 score and 10 years, Howard realized that he probably did not have many books left in him. He was looking for the right topic for a book that would have a lasting impact on operational testing and would allow him to "go out with a bang." He alit on subscores as the topic that he was looking for. He contacted Sandip, Richard, and Shelby, all of whom thought that a book on subscores was a great idea. The book project would also provide an opportunity for the four authors, who had narrowly missed collaborating in the past (e.g., when Howard left ETS immediately after Sandip joined ETS and a year before Shelby joined ETS),

to team up on a topic of common interest. The four of us then contacted Lauren Cowles of Cambridge University Press, who shared our enthusiasm, and the ball started rolling. The book was to be a coherent compendium of all of our past research as well as other new, relevant research. Thus, the version of the book that you now hold in your hand includes new material that we developed to bridge gaps in the literature that were exposed once we had collected prior work into a coherent whole. The section on canonical scores in Chapter 3 is one prime example of this, as are many of the figures in Chapters 4 and 5. Finally, the summary recommendations for practice appearing throughout the book are new.

The Need for This Book and the Purpose That We Intend It to Serve

> If there's a book that you really want to read, but it hasn't been written yet, then you must write it.
>
> Toni Morrison (1981)

Given the popularity of subscores and the high variability of their utility that is encountered in practice, the authors felt the need for a single, coherent, and authoritative source that provides a succinct and easily accessible summary of best practices supported by existing research. Because such a source did not exist, we felt compelled to follow Professor Morrison's advice and write it. We believe that this book will provide direction to both producers and consumers of psychological and educational test scores about when and how subscores should be provided, how to interpret reported subscores appropriately, and how to make informed decisions based on the subscores. The book will provide graduate students and researchers with a thorough introduction to existing research on subscores and guide them to future research topics related to subscores. We hope that it will contribute to improved understanding and more sensible and ethical use of subscores among psychometricians, test developers, institutional and individual users of assessment results, and management of testing organizations that currently report subscores or are considering their use. Thirteen years after Rich and Howard started their quest to find unsolved problems on subscores, many of the same problems still exist, which suggests a larger conversation is needed between psychometricians, test developers, governance stakeholders, and score recipients regarding how to estimate a subscore's value, its limitations, appropriate modalities of reporting, and so on. We hope that this book can contribute to that conversation.

A Brief Outline of This Book

We begin with an introductory chapter that provides a brief account of the long history of tests, subtests, scores, and subscores. Chapter 2 focuses on the reporting of subscores and provides examples of the various ways in which subscores are reported for several large-scale operational tests. Chapter 3 focuses on the importance of evaluating the quality of subscores and suggests various ways of assessing the psychometric quality of subscores. Chapter 4 includes the results from a survey regarding the quality of subscores using data from 34 operational tests. Chapter 5 provides some alternatives when subscores are demanded but do not satisfy contemporary standards of psychometric quality and consequently ought not, in good conscience, be provided. Chapter 6 includes concluding remarks and our recommendations for practice.

Acknowledgments

We would like to take this opportunity to acknowledge the help of, and express gratitude to, the individuals and organizations who contributed in various ways to the publication of the book.

First, organizational thanks are due to the ETS that employed three of the authors when they first became involved with subscores. The authors are especially grateful to Amit Sevak, Walt McDonald, and Kurt Landgraf, present and former presidents of ETS – their wise leadership was instrumental in providing continued support for basic research. Henry Braun, Drew Gitomer, Ida Lawrence, Kadriye Ercikan, and John Mazzeo, the former and current vice presidents of research at ETS, ensured the smooth running of ETS allocation projects that provided support for the research. Their enthusiasm and wise counsel were appreciated then and now. Paul Holland, Daniel McCaffrey, and Matthew Johnson have all provided valuable support, encouragement, and advice.

Second, the authors are grateful to the NBME and its then president, Donald Melnick, and current president, Peter Katsufrakis, who supported the work of Howard and Richard over many years. The enthusiasm of the NBME senior vice presidents Michael Jodoin and Ye Tong for the work in this book is much appreciated.

Third, we are indebted to Katherine Castellano for creating the cover graphic and to the following colleagues for their assistance, encouragement, review, and valuable suggestions: Neil Dorans, Kevin Larkin, Charles Lewis, Yi-Hsuan Lee, Nick Longford, John Mazzeo, Gautam Puhan, Yasuyo Sawaki, Kathy Sheehan, Xiaohui Wang, Jonathan Weeks, and Lili Yao (all currently or formerly at ETS); Mike Jodoin (formerly at ETS and now NBME); Amanda Clauser, Brian Clauser, Jerusha Henderek, Daniel Jurich, Francis O'Donnell, and Kimberley Swygert (current NBME employees); April Zenisky (at the University of Massachusetts, Amherst); Ronald Hambleton (formerly at the

University of Massachusetts, Amherst); and David Thissen (at the University of North Carolina).

Fourth, the authors would like to express gratitude to Cambridge University Press (the outlet Isaac Newton chose to publish his *Principia*) and the mathematics editor Lauren Cowles and her assistant Arman Chowdhury for ready acceptance of the book proposal, encouragement, and help in making this book the best it could be.

Fifth, any long-term project accumulates debts to many others whose help was important. Most important are Clare Dennison and Amala Gobiraman, whose work ethic and keen sense of organization kept everything in order.

We would like to end by thanking Shenghai Dai (Washington State University), Xiaolin Wang (Pearson VUE), and Dubravka Svetina (Indiana University–Bloomington) for preparing the *R* package *subscore* and Alexander Robitzsch for preparing the *R* package *sirt* that can be used to implement the aforementioned approach of Haberman that is the cornerstone of this book. The packages have facilitated the application of the authors' research to a variety of problems and immensely helped us while writing the book.