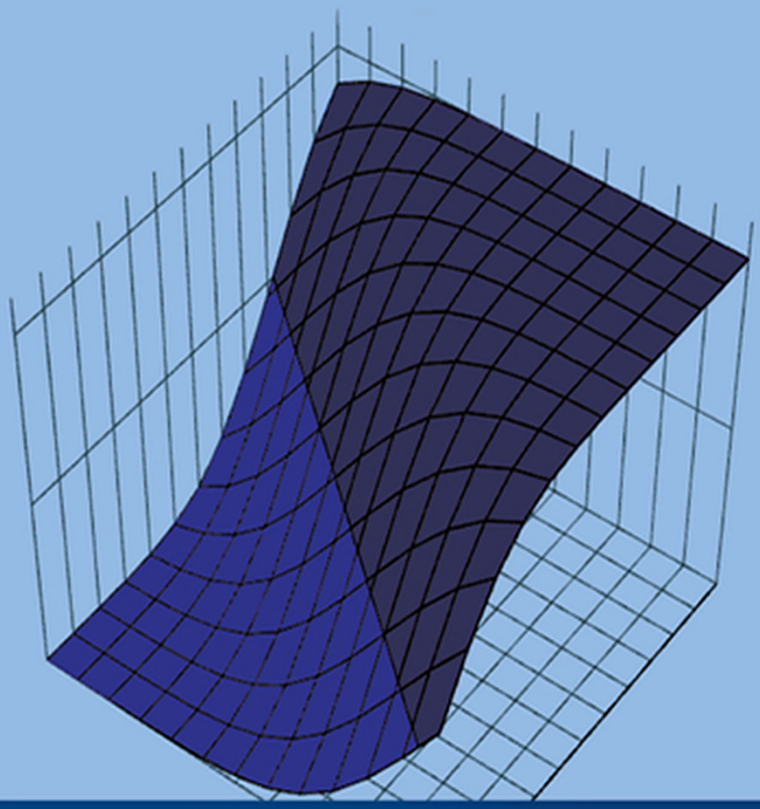
VOLUME I



THE WILEY HANDBOOK OF PSYCHOMETRIC TESTING:

A MULTIDISCIPLINARY REFERENCE *ON* SURVEY, SCALE *AND* TEST DEVELOPMENT

Edited by
Paul Irwing | Tom Booth | David J. Hughes

WILEY Blackwell

The Wiley Handbook of Psychometric Testing

The Wiley Handbook of Psychometric Testing

A Multidisciplinary Reference on Survey, Scale and Test Development

Volume One

Edited by

Paul Irwing
Tom Booth
David J. Hughes

WILEY Blackwell

This edition first published 2018

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Irwing, Paul, editor.
Title: The Wiley handbook of psychometric testing : a multidisciplinary reference on survey, scale and test development / edited by Paul Irwing, Manchester University, UK, Tom Booth, The University of Edinburgh, Edinburgh, UK, David J. Hughes, Manchester Business School, Manchester, UK.
Description: First Edition. | Hoboken : Wiley, 2018. | Includes bibliographical references and index. |
Identifiers: LCCN 2017041032 (print) | LCCN 2017061203 (ebook) | ISBN 9781118489826 (pdf) | ISBN 9781118489703 (epub) | ISBN 9781118489833 (cloth : alk. paper) | ISBN 9781119121176 (pbk. : alk. paper)
Subjects: LCSH: Psychometrics. | Psychological tests.
Classification: LCC BF39 (ebook) | LCC BF39 .W545 2018 (print) | DDC 150.28/7–dc23
LC record available at https://lccn.loc.gov/2017041032

Cover image: Printed with permission from Anna Brown
Cover design by Wiley

Set in 10/12pt Galliard by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Notes on Contributors to Volume 1

Amanda N. Baraldi is Assistant Professor of Psychology at Oklahoma State University. Dr. Baraldi received her doctorate in Quantitative Psychology from the Arizona State University in 2015. Dr. Baraldi's current research interests include missing data analyses, methods for assessing mediation, longitudinal growth modelling, and health and prevention research.

Tom Booth is Lecturer in Quantitative Research Methods in the Department of Psychology, University of Edinburgh. His primary methodological interests are in generalized latent variable modelling. His applied work covers individual differences, organizational, and health psychology.

Li Cai is Professor in the Advanced Quantitative Methodology program in the UCLA Graduate School of Education and Information Studies. He also serves as Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). In addition, he is affiliated with the UCLA Department of Psychology. His methodological research agenda involves the development of latent variable models that have wide-ranging applications in educational, psychological, and health-related domains of study.

Eva Ceulemans is Professor of Quantitative Data Analysis at the Faculty of Psychology and Educational Sciences, University of Leuven, Belgium. Her research focuses on the development of new techniques for modelling multivariate time series data and multigroup data, and exploring individual or group differences therein. To this end, she often combines general principles of cluster analysis with dimension reduction (principal component analysis, factor analysis) and/or regression.

Fang Fang Chen received her M.S. in psychology from Peking University, her doctoral training in Social and Quantitative Psychology from Arizona State University, and completed her post-doctoral training at the University of North Carolina at Chapel Hill. Her methodological work focuses on measurement invariance and the bifactor model. Dr. Chen was an assistant professor of psychology at the University of Delaware, and now is a senior research biostatistician at the Nemours Center for the HealthCare Delivery Science.

David J. Ciuk is an assistant professor in the Department of Government at the Franklin & Marshall College. His academic interests center on public opinion and political psychology. His research aims to build a better understanding of the attitude formation process in the mass public. More specifically, he looks at how morals and values, policy information, and political identity affect political attitudes. He is also interested in survey experimental designs, measurement, and various public health issues.

Christine E. DeMars serves at James Madison University as a professor in the department of graduate psychology and a senior assessment specialist in the Center for Assessment and Research Studies. She teaches courses in Item Response Theory, Classical Test Theory, and Generalizability Theory, and supervises Ph.D. students. Her research interests include applied and theoretical topics in item response theory, differential item functioning, test-taking motivation, and other issues in operational testing.

Conor V. Dolan is Professor at the VU University, Amsterdam. His research interests include: covariance structure modelling, mixture analyses, modelling of multivariate intelligence test scores, and modelling genotype-environment interplay.

Susan Embretson is Professor of Psychology at the Georgia Institute of Technology. She has been recognized nationally and internationally for her programmatic research on integrating cognitive theory into psychometric item response theory models and into the design of measurement tasks. She has received awards from the National Council on Measurement and Education; American Educational Research Association; and the American Psychological Association Division for research and theory on item generation from cognitive theory.

Craig K. Enders is a Professor in the Department of Psychology at UCLA where he is a member of the Quantitative program area. Professor Enders teaches graduate-level courses in missing data analyses, multilevel modelling, and longitudinal modelling. The majority of his research focuses on analytic issues related to missing data analyses and multilevel modelling. His book, *Applied Missing Data Analysis*, was published with Guilford Press in 2010.

David J. Hughes is an Organisational Psychologist at Manchester Business School. His research interests centre on individual differences and can be broken down into three main areas: the theory and measurement of individual differences, individual differences at work, and individual differences in financial behavior. He is interested in psychometric test evaluation and his statistical interests revolve around generalized latent variable models, in particular structural equation modelling, factor models, and other models appropriate for multivariate data with complex structures.

Paul Irwing is Professor of Psychometrics at the Manchester Business School. He chaired the Psychometrics at Work Research Group and is a director of the psychometric publishing company E-metrixx. He has authored two research and two commercial psychometric measures. He is known for research on sex differences and pioneering work on the general factor of personality. His current research concerns the 11+ Factor Model of Personality, and the newly proposed individual difference of Personality Adaptability.

William G. Jacoby is Professor of Political Science at Michigan State University and Editor of the *American Journal of Political Science*. He is the former Director of the Inter-university Consortium for Political and Social Research (ICPSR) Summer Program in Quantitative Methods of Social Research and former Editor of the Journal

of Politics. Professor Jacoby's areas of professional interest include mass political behavior and quantitative methodology (especially scaling methods, measurement theory, and statistical graphics).

Robert I. Jennrich began his work with the development of the first internationally used statistical software package BMD. His main field is statistical computing. He contributed to the development of nearly half of the 25 programs in this package. Because of this, he was a member of the 1972 Soviet-American Scientific Exchange Delegation on Computing. Since these early days, he has published papers on stepwise linear and non-linear regression, stepwise discriminant analysis, goodness-of-fit testing for covariance structure analysis, and a number of papers in factor analysis, primarily on rotation. He has recently been nominated for a lifetime achievement award, which he unfortunately didn't get.

Ken Kelley is Professor of Information Technology, Analytics, and Operations (ITAO) and the Associate Dean for Faculty and Research in the Mendoza College of Business at the University of Notre Dame. Professor Kelley's work is on quantitative methodology, where he focuses on the development, improvement, and evaluation of statistical methods and measurement issues. Professor Kelley's specialties are in the areas of research design, effect size estimation and confidence interval formation, longitudinal data analysis, and statistical computing. In addition to his methodological work, Professor Kelley collaborates with colleagues on a variety of important topics applying methods. Professor Kelley is an Accredited Professional Statistician™ (PStat®) by the American Statistical Association, associate editor of *Psychological Methods*, and recipient of the Anne Anastasi early career award by the American Psychological Association's Division of Evaluation, Measurement, & Statistics, and a fellow of the American Psychological Association.

Keke Lai is Assistant Professor of Quantitative Psychology at University of California, Merced. His research interests include structural equation modelling and multilevel modelling.

Stanley Lemeshow earned his Ph.D. at UCLA, and his MSPH at UNC. He has coauthored three textbooks: Applied Logistic Regression; Applied Survival Analysis; and Sampling of Populations – Methods and Applications. His honors include: the Wiley Lifetime Award (2003); UCLA School of Public Health Alumni Hall of Fame (2006); Fellow of the AAAS (2003); Distinguished Graduate Alumnus (Biostatistics) – UNC Graduate School Centennial (2003); Fellow of the ASA (1995); and the Statistics Section Award of the APHA (1995).

Urbano Lorenzo-Seva is a professor in the Department of Psychology at Universitat Rovira i Virgili, Spain. He is the coauthor of FACTOR, a free-shared software to compute exploratory factor analysis, and has published numerous articles related to this subject. His research interests include the development of new methods for exploratory data analysis, and applied psychometric research. He has taught data analysis at university level and in short courses for many years.

Bo Lu earned his Ph.D. in Statistics from the University of Pennsylvania. He is an associate professor of Biostatistics at the Ohio State University. His research expertise includes causal inference with propensity score based adjustment for observational data, survey sampling design and analysis, statistical models for missing data. He has been PIs for multiple NIH and AHRQ grants and served as the lead statistician for the Ohio Medicaid Assessment Survey series since 2008.

Rob R. Meijer is Professor in Psychometrics and Statistics at the University of Groningen, the Netherlands. His work is in item response theory and applications of testing. He is also interested in educational assessment and educational and personnel selection.

Dylan Molenaar is Assistant Professor at the Department of Psychology, University of Amsterdam. His research interests include: item response theory, factor analysis, and response time modelling.

Kristin M. Morrison is an advanced Ph.D. student at the Georgia Institute of Technology. She has published and presented work on item generation at various conferences and meetings. Other research interests include computer adaptive testing, multistage testing, educational research, and cognitive complexity.

Irini Moustaki is Professor of Social Statistics at the London School of Economics and Political Science. Her research interests are in the areas of latent variable models and structural equation models. Her methodological work includes treatment of missing data, longitudinal data, detection of outliers, goodness-of-fit tests, and advanced estimation methods. Furthermore, she has made methodological and applied contributions in the areas of comparative cross-national studies and epidemiological studies on rare diseases. She has coauthored two books on latent variable models. She was elected Editor in Chief of the journal *Psychometrika* in November 2014.

Stanley A. Mulaik is Emeritus Professor in the School of Psychology, Georgia Institute of Technology. His work has broadly focussed on latent variable models and the underlying philosophy of causality. He is the author of the influential texts, *Foundations of Factor Analysis* and *Linear Causal Modeling with Structural Equations.*

Aja L. Murray is a Research Associate in the Violence Research Centre, Institute of Criminology, University of Cambridge. Her research interests include psychometrics and childhood and adolescent mental health developmental.

Jorge N. Tendeiro is Assistant Professor at the University of Groningen, the Netherlands. He has research interests within item response theory (IRT), with a large focus on person-fit analyses. Besides (co)authoring several papers on this topic, he is also the author of the PerFit R package. He is currently extending person-fit approaches to a broader type of IRT, which include the unfolding model.

Marieke E. Timmerman is a professor in multivariate data analysis at the Heymans Institute for Psychological Research at the University of Groningen, The Netherlands. Her research focuses on the development of models for multivariate data with complex structures, to achieve an understanding of the processes underlying these, mainly psychological, data. Her research interests are latent variable modelling, data reduction methods, including multiset models, and classification. She serves as an associate editor of *Psychometrika*.

Zugui Zhang obtained his Ph.D. in biostatistics from the University of Iowa. He is the lead Biostatistician of the Value Institute of Christiana Care Health System, Assistant Professor at Thomas Jefferson University, and Joint Professor at the University of Delaware. He has a broad background in biostatistics, health outcomes research, quality of life, public health, epidemiology, and health economics. He has worked extensively on large international and national, multicenter, randomized clinical trials, and observational studies.

Preface

The existence of this volume owes itself to both opportunity, and many hours of coffee fuelled conversation, the general form of which would run; "have you seen X's critique of Y…? Did you see the special issue on X…? Wouldn't it be great if we could do something to help improve methods knowledge within psychology?" Given our collective interest in individual difference psychology, our musings would often be triggered by a favourite conversation – the inappropriate application of principal components analysis. We all taught, and continue to teach, research methods and statistics at our institutions; we all develop and evaluate psychometrics in our research, but the idea of something larger – a conference, a journal special edition, a book – continued to surface. There are many excellent papers, books, and courses that cover research methods relevant for psychometrics but, we thought, not a single resource that brings cutting-edge knowledge from the journals in our fields together in an accessible manner.

So, imagine our delight and trepidation when Paul was invited to edit a handbook on psychometric testing. We could finally put our coffee-shop "wisdom" to the test! Although, "Who reads books anymore? Surely, it will be a ton of work editing such a volume? Come to think of it, we're not knowledgeable enough in half of the areas we need to cover… How will we ensure we get expert authors?" We had many questions and doubts and committing to producing this book was not a lightly taken decision. Nevertheless, we returned consistently to one question: how else can we do our small part in improving the availability of cutting-edge methodological knowledge within the field of psychometrics? We had decided (or at least Tom and David had, Paul took some convincing!) that together we would try to produce a book which covered the core topics in psychometrics, a book that harnessed the work of excellent authors, a book that we would like to have, a book that, if used, would see methodological awareness grow, and statistical practice improve.

At the outset of planning and preparing this book, all three of us were at the University of Manchester's Business School; Tom and David were Ph.D. students, and Paul, our supervisor, a Reader of Organisational Psychology. At the time of writing this preface, almost five years later, Ph.Ds are a distant memory, Tom and David have moved on several times (or moved on and then back to Manchester in David's case), and Paul is now a professor. This book, *The Handbook*, has been a labour of love and frustration throughout these five years, five years that have not only seen workplace changes but

also a quite remarkable series of injuries and ill-health, several relegations for our relative football (soccer) teams, oh, and a wedding!

Now, here we are, a complete volume that looks remarkably close to our initial proposal and we are delighted with it. Our intention, succinctly described in the letters we sent to authors was to write chapters on key topics in psychometrics and that:

> Each chapter should cover the fundamentals. However, we favour a three-tiered approach, which covers: (1) historical and standard approaches, including all the core material, and then moves onto (2) a discussion of cutting-edge issues and techniques, together with a section on (3) how to do it, which should contain a worked example. These chapters should address real issues faced by both practitioners and researchers.

We hope with the help of our contributors that we have achieved our goal. We hope that a journey started with coffee-shop musings and reflections has led to the production of a useful resource for students, academics, and practitioners alike.

The biggest strength of *The Handbook* undoubtedly lies in the calibre of the authors who have contributed. Every chapter (modestly, with the exception of our own!) has been written by a field expert with specialist knowledge whose work we have admired and used to inform our own practice. Not only are our authors experts, they are also diverse with regard to nationality (e.g., Netherlands, U.K., U.S.A.), profession (e.g., academics, commercial test developers), and field (e.g., psychology, statistics, education, politics). This diversity was no accident. Approaching the topic of psychometrics from the perspective of one discipline would never showcase the range of theoretical and statistical advances that we hoped to convey to our readers. We wish to say a very large thank you to each and every author for sharing their expertise and for their patience throughout this process.

Beyond our excellent contributors, we would also like to acknowledge the help of our families, friends, and students, for their willingness to put up with constant references and sometimes play second fiddle to *The Handbook* over the last few years. Finally, we would also like to extend our deep gratitude to all the people at John Wiley & Sons who have helped us in this process (and there have been many).

Thank you all, and thank you to you as readers for picking this book up off the shelf. We hope it is useful.

David, Tom, and Paul.

Introduction

Aims and Scope

The principal aim of this Handbook was to provide researchers and practitioners from different academic and applied fields with a single practical resource covering the core aspects of psychometric testing. Psychometrics can be translated as mental measurement, however, the implication that psychometrics is confined to psychology is highly misleading. Virtually every conceivable discipline now uses questionnaires, scales, and tests developed from psychometric principles, and this book is therefore intended for a multidisciplinary audience. The field of psychometrics is vibrant with new and useful methods and approaches published frequently. Many of these new developments use increasingly sophisticated models and software packages that are easy to misunderstand. We have strived to make the chapters in this Handbook both intellectually stimulating, and practically useful, through the discussion of historical perspectives, cutting-edge developments, and providing practical illustrations with example code. Thus, each chapter provides an accessible account and example of the current state-of-the-art within the core elements of psychometric testing. We hope that this book is useful for those who develop, evaluate, and use psychometric tests.

Section and Chapter Structure

In structuring the chapters and sections of this Handbook, we attempted to approximate the process of test development. In Part I, the chapters cover core topics surrounding the foundations of test development. Here, we provide a macro view of the process of test development (Chapter 1); outline the broad differences in the classical and modern test theory approaches (Chapter 2); broach topics of the development and nature of item sets (Chapters 3 and 7); and address fundamental topics in study design (Chapters 4, 5, and 6). Chapter 1 is probably the most logical place to start, since this provides a context for most of the other chapters as to their role in test development.

In Part II, we consider the primary psychometric tools for analyzing item pools and identifying plausible scales. Here, we consider the fundamentals of both the common

factor (Chapters 8, 10, 11, and 12) and item response (15 and 16) approaches. Chapter 9 sits somewhat at the intersection of the classic and modern test approaches in discussing estimation of categorical item factor models. Part II also provides introductory coverage of multidimensional scaling (MDS: Chapter 14), which has been a highly influential psychometric tool in fields such as political science (see Chapter 28), but is less commonly used in psychometric evaluations in fields such as psychology. The remaining chapters in Part II deal with a number of slightly more advanced, but highly important topics. These chapters address nonnormality (Chapter 13), Bayesian approaches to scaling (Chapter 17) and the modelling of forced choice item formats (Chapter 18). Each of these chapters covers something of an "up and coming" area of psychometrics based upon advancements in computation that now allow us to model more complex data appropriately (as opposed to forcing data into models which presuppose unmet assumptions). Chapter 18, which covers forced choice items, is also particularly valuable for those who test in high-stakes scenarios (i.e., employee selection).

Part III addresses the topic of test scores and also deals with the process of linking and equating test scores. The purpose of psychometric tools is more often than not to understand where someone stands on a given latent trait versus other individuals. Often, we desire scores to represent this standing. Here then, we deal with fundamental topics in score estimation and evaluation, from simple sum scores to advanced item response estimates (Chapters 19 and 20). But what happens when we develop a new version of a test? Or we attempt to develop parallel forms? Or we want to try and relate individuals who have taken different tests? These needs are very common in both applied and academic analyses in education, business, and psychology, and all are concerned with the topic of score linking and equating (Chapters 19 and 21).

Part IV is concerned with the evaluation of scales from a statistical and theoretical perspective. Chapters 23 and 24 provide state of the art treatments of the classic topics of reliability and validity, respectively. Chapter 22 concerns the evaluation of the strength of general and specific factors using bi-factor models. Chapter 25 uses multitrait-multimethod analyses to explore the proportion of measured variance attributable to the construct and the measurement tool.

So, we have developed some items, collected some data, established our best set of measured constructs, and evaluated the quality of the scales. But does our test operate in the same way for all groups of people? This question is critically important for reasons of accuracy and fairness and can be approached through the application of the analytic methods discussed in Part V. Here, we deal with tools for modelling and understanding the measurement properties of psychometric tools across groups from a common factor (Chapter 26) and item response (Chapter 27) perspective.

Finally, in Part VI, we step away from topics related to the immediate development of tests, and we consider the role psychometrics has played, and may play in the future, in theoretical and practical arenas. In Chapters 28 and 29, we consider the substantial role psychometric tools and analyses have played in shaping the fields of political science and personality psychology. Lastly, we introduce recent work concerning the relationship between network and latent variable approaches to understanding behavior (Chapter 30). Both Chapters 29 and 30 provide critical appraisals of analytic tools common to the psychometrics world (i.e., factor analysis) and point the way to potential avenues of progress. Reviewing the contributions of psychometric testing and

considering how future Handbooks of Psychometric Testing might look felt like a fitting way to close this volume.

Mathematical and Statistical Foundations

Our aim for this Handbook was to make the content as accessible as possible for as many individuals as possible, both practitioners and academics. You do not need to be a mathematician or statistician to read this book. Equations are kept to the minimum required to provide satisfactory explanations of the methods under discussion. Where equations are presented by authors, they are broken down and described verbally to add clarity.

However, it would be remiss of us as editors to try and claim that this handbook is going to be an easy read for all who pick it up. The topic under discussion is statistical and thus there is some technical content. The degree of technical content varies across chapters inline with the mathematical complexity of the topics discussed.

So, what statistical and mathematical knowledge is required? With regard to statistics, we have assumed certain background knowledge. Modern psychometrics depends, amongst other things, on knowledge of structural equation modelling (SEM), and in particular confirmatory factor analysis (CFA). However, both of these are already covered by many excellent texts. For example, the early chapters in Little (2013) provide an excellent introduction to both topics, while Brown (2015) provides arguably one of the most useful comprehensive treatments of CFA, and Bollen (1989) still represents the most definitive advanced treatment of both. SEM depends on an understanding of multiple regression. Probably one of the best and most neglected books on regression is Pedhazur (1997), which in fact provides a comprehensive coverage of everything you need to know about the basics of multivariate statistics.

With regards to mathematics knowledge, to understand the methods conceptually and use the practical examples as guides to analyze your own data, not very much is required. But to appreciate fully the topics under discussion here, a basic knowledge of calculus, algebra, and perhaps most importantly, matrix algebra (e.g., Fieller, 2015; Khuri & Searle, 2017) is required. It would also be valuable, as is true for any statistical topic, for readers to have some grounding in probability and significance testing. A number of chapters also extend into Bayesian statistics where a slightly deeper appreciation of probability theory may be necessary (DeGroot & Schervish, 2012). Chapter 17 contains a short introduction to concepts from Bayesian analysis, but this is intended more as a refresher than as a comprehensive treatment of the fundamental of Bayes. In terms of a comprehensive and accessible introduction to mathematical statistics Larsen and Marx (2011) is hard to better.

We have resisted the temptation to provide introductory chapters or appendices on these core statistical and mathematical topics for two reasons. First, there are a multitude of high quality introductory texts on these topics (see previously). It is also important to point out that there are now a huge number of excellent (and free) web resources on these topics, and the reader who feels they need a refresher is encouraged to explore this route (no need to spend money unnecessarily) whilst bearing in mind that finding the right mathematics or statistics text is often a personal thing. Second, the Handbook is intended to have a diverse target audience whose needs will vary greatly. To cover each of these topics, for a diverse audience, would have required us to turn what is already a large book, into a behemoth perfectly suited to act as doorstop for the *Black Gate of Mordor*.

Software

The contributors to the Handbook have made use of a variety of different statistical software packages, some freely available, others proprietary. The two most popular tools across chapters are the R statistical programming language and MPlus. R has a large number of advantages as a statistical tool, the details of which we will not get into here. However, perhaps the two most commonly cited and important are that it is free, and it is highly flexible. However, with this flexibility comes a requirement for some knowledge of programming and coding languages – with great power comes great responsibility.

In recent years, MPlus has established itself as one of the primary proprietary tools for conducting general latent variable analyses, quickly incorporating new methodologies, providing excellent help facilities and abundant online resources. MPlus is again a flexible tool that is very user-friendly. However, the program does come at a cost, with a full single user license for University affiliates costing approximately £720 ($895) at the time of writing.

Popular general-purpose statistical packages such as SPSS, SAS, or STATA, are used, but less frequently. This is not to say that these packages have no capability with respect to the types of analysis discussed in this Handbook but often they do lack some of the nuances needed to conduct state of the art analyses. This is perhaps the reason that authors have also made use of a variety of additional programs including, MIRT, *flex*MIRT, IRTPro, Factor, LISREL, EQS, MATLAB, WinBUGS, and more. What this shows is the huge variety and widespread availability of tools for empirical analysis.

So, which software should you use? In making this decision, one of the key things to consider is which models are best suited to which software. We hope the Handbook helps in this endeavor in two ways. First, our empirical examples show directly some of what can be done in certain packages. Second, we hope the level of presentation of the technical details of analyses will allow the readers to tackle the supporting documentation of software to gain a deeper understanding of what is going on "under the hood."

The available tools vary in a number of other ways, too. For example, the means by which analyses are conducted varies from coding languages (both program-specific and general) to graphical user interfaces with and without diagrammatical capabilities. Whilst drop-down menus are great, we would recommend writing code. Not only does this clarify understanding but it also allows you to specify all modelling options rather than resting on software defaults. Perhaps the most pragmatically relevant variation lies in monetary cost, and many readers will likely be limited to proprietary software available from their institutions or freely available software. Thankfully, for scientific progress, more tools than ever are now free! Indeed, more than a third of our chapters used free software and every analysis presented in this handbook could be conducted in this manner, which we will now discuss.

Chapter Code

For a majority of the chapters in this handbook, software code has been made available for the analyses, and some of the figures, presented in the chapters. In most cases, this is in the form of a Code Appendix at the end of the chapter. For a smaller subset of chapters, the code has been integrated into the body of the chapter. Chapter dependent, the

detail of code provided varies, as does the level of annotation but in each case, it directly links to the practical examples presented.

Some chapters have made use of proprietary software, which if the reader does not have access to, obviously limits its usability. Here we wish to emphasize once again the abundance of freely available software for psychometric analyses. With very few exceptions, the analyses presented in this book can be conducted using a relatively limited number of R-packages. For those unfamiliar with R, we would strongly recommend investing some time to learn the basics of the program.

The packages *psych* (Revelle, 2016), *mirt* (Chalmers, 2012), *OpenMx* (Neale, et al. 2016; Pritikin, Hunter, & Boker, 2015; Boker et al., 2017), and *lavaan* (Rosseel, 2012), can be used to complete a vast majority of the common factor and IRT analyses presented in this Handbook. Other useful packages include *qgrpah* (Epskamp, Cramer, Waldorp, Schittmann, & Borsboom, 2012) for network models and plots, *smacof* (de Leeuw, & Mair, 2009) for multidimensional scaling, *rjags* (Plummer, 2016) for Bayesian analysis, and *mice* (van Buuren & Groothuis-Oudshoorn, 2011) for missing data analysis. Collectively, we hope that the code provided and access to the free R packages noted makes the Handbook a genuinely valuable practical tool.

How to Use the Book

As has been outlined here, this Handbook has been compiled with multiple audiences in mind and as such we anticipate the contents to be used in a variety of ways. Coarsely, our section structure can be seen as representing the process of developing and evaluating a psychometric test. Read from start to finish the chapters represent a comprehensive introduction to the process of test development, analysis, and revision. Readers interested in evaluating an extant scale for which they have collected some data will, dependent on their focus, find most value in the contents of sections two through four. Here we suggest that the readers treat the chapters and the associated reference lists as the start point for in depth consideration of a topic. Whilst we cover historical perspectives, state-of-the-art methods, and provide practical examples and code, the chapters of this handbook do not contain everything one might need to know.

In either case, whether a reader is interested in the start to finish process of scale development, or in methods for a specific purpose, the chapter content allows the reader to focus on either classical (common factor) or modern (IRT) approaches to most questions. For the reader unsure of which approach they wish to take, we would encourage them to read Chapter 2, and to consider their research focus in light of the type of information each approach has to offer. In some cases, this is overlapping, in others complementary, and so the choice is not always clear cut. Equally, whilst positioned towards the end of the book, Chapter 24, might be an interesting place to start because the treatment of "validity" aims to provide a coherent model to organize test development and evaluation procedures and references out to other chapters wherever relevant.

Practical hands on experience is a valuable part of the learning process. We hope that the code provided and the information on different software packages will provide a framework that will allow readers to apply analyses to their own data. The code is not a tutorial, and some knowledge of the different statistical packages will be needed.

We hope the Handbook is enjoyable and useful. Enjoy.

References

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., …& Manjunath, B. G. (2017) *OpenMx 2.7.12 user guide*. Available at http://openmx.psyc.virginia.edu/docs/OpenMx/latest/OpenMxUserGuide.pdf (accessed October 2017).

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons, Inc.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: The Guilford Press.

Chalmers, P. R. (2012). *mirt*: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. URL http://www.jstatsoft.org/v48/i06/ (accessed October 2017).

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics*. Boston, MA: Pearson Education.

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, *31*, 1–30. URL http://www.jstatsoft.org/v31/i03/ (accessed October 2017).

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*, 1–18. URL http://www.jstatsoft.org/v48/i04/ (accessed October 2017).

Fieller, N. (2015). *Basics of matrix algebra for statistics with R*. Boca Raton, FL: Chapman Hall/CRC.

Khuri, A. I., & Searle, S. R. (2017). *Matrix algebra useful for statistics* (2nd ed.). Chichester, UK: Wiley-Blackwell.

Larsen, R. J., & Marx, M. L. (2011). *An introduction to mathematical statistics and its applications* (5th ed.). Boston, MA: Pearson Education.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NJ: The Guilford Press.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M… and Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *80*, 535–549.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Boston, MA: Wadsworth.

Plummer, M. (2016). *rjags: bayesian graphical models using MCMC. R package version 4-6*. https://CRAN.R-project.org/package=rjags (accessed October 2017).

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for Item Factor Analysis. *Educational and Psychological Measurement*, *75*, 458–474.

Revelle, W. (2016) *psych: Procedures for personality and psychological research*, Northwestern University, Evanston, IL, USA, https://CRAN.R-project.org/package=psych Version = 1.6.12.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. URL http://www.jstatsoft.org/v48/i02/ (accessed October 2017).

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. URL http://www.jstatsoft.org/v45/i03/ (accessed October 2017).

The Wiley Handbook of Psychometric Testing

Part I
Practical Foundations

1

Test Development

Paul Irwing and David J. Hughes

The purpose of this chapter is to explain how the psychometric principles outlined in the remaining chapters of this Handbook can be applied in order to develop a test. We take a broad definition both of what constitutes a test and what is understood by test development. This is because the principles of psychometric testing are very broad in their potential application. Among others, they can apply to attitude, personality, cognitive ability, interest, and diagnostic measures. For the purposes of this chapter all such measures will be referred to as **tests**. Psychometrics is broad in another sense: It applies to many more fields than psychology; indeed, biomedical science, education, economics, communications theory, marketing, sociology, politics, business, and epidemiology, among other disciplines, not only employ psychometric testing, but also have made important contributions to the subject. Our definition of a test is broad in another sense: It encompasses everything from a simple attitude scale, say to measure job satisfaction, to comprehensive test batteries such as the Woodcock–Johnson IV battery of cognitive tests (Schrank, Mather, & McGrew, 2014). Of course, not every aspect of test development applies to both, but the overlap is considerable.

It may be useful to distinguish the different levels of complexity involved in test development. In the simplest case, the test comprises just one scale, but more usually a test is comprised of multiple scales (**single scale versus test battery**). A second distinction is between tests comprised of similar as opposed mixed types of scales (**scale similarity**). For example, the European Social Survey measures multiple constructs but all are attitude scales. However, some instruments may combine assessments of mixed scale types; for example, cognitive ability, personality, and attitudes. A third dimension concerns whether the test is intended to sample the entire spectrum of a domain, or whether it is focused on specific aspects (**broad** versus **narrow spectrum**). For example, it would not be feasible for a selection test to reliably measure all facets of either personality or cognitive ability. The point being that some form of systematic choice procedure is required such as job analysis or meta-analysis (Hughes & Batey, 2017). Fourth, there is the issue of **team size**. There is a very big difference from the situation in which a single investigator takes responsibility for the major

The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, First Edition. Edited by Paul Irwing, Tom Booth, and David J. Hughes.

portion of test development, and the situation in which there is a large team with diverse skill sets, which would be common when developing commercial tests. The MAT^{80} (Irwing, Phillips, & Walters, 2016), which we use later to demonstrate test development procedures, is a test battery with a mixed scale that combines personality and ability scales, involved a small test development team, and requires systematic selection of specific facets.

There are already many publications of relevance to the topic of test development. Probably the most useful single source is "The Standards for Educational and Psychological Testing" (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014). However, as its name implies, this tells you what needs to be done, but not how to do it. There is now a very useful *Handbook of Test Development* (Downing & Haladyna, 2006), which largely specializes in the design of educational and ability tests. Of almost equal use are textbooks on questionnaire and survey design (Brace, 2005; De Vaus, 2014; Foddy, 1996; Oppenheim, 1992). Perhaps what none of these books quite do is link modern psychometrics to test development, which is the aim of this chapter and the whole Handbook.

We begin with a comprehensive model of the stages of test development, and then discuss the major considerations that apply at each stage. We will leave the reader to decide which of these stages apply to their own situation, depending on the type and purpose of the test. Table 1.1 outlines a 10-stage model of test development. There are a number of stage models of test development in existence (e.g., Althouse, n.d.; Downing, 2006) and, to a degree, such models are arbitrary in the sense that which tasks are grouped into a stage and the order of stages is probably more for explanatory convenience rather than a description of reality. In practice, tasks may actually be grouped and undertaken in many different combinations and orders, with many tasks undertaken iteratively. Nevertheless, a stage model provides a systematic framework in which to discuss the tasks that must be undertaken, although not all tasks are relevant to all types of test development.

Table 1.1 Stages of test development.

Stages and substages
1 Construct definition, specification of test need, test structure.
2 Overall planning.
3 Item development.
a. Construct definition.
b. Item generation: theory versus sampling.
c. Item review.
d. Piloting of items.
4 Scale construction – factor analysis and Item Response Theory (IRT).
5 Reliability.
6 Validation.
7 Test scoring and norming.
8 Test specification.
9 Implementation and testing.
10 Technical Manual.

Construct Definition, Specification of Test Need, and Structure

The motivation for test development often stems from a practical concern: can we help children learn, can we identify effective managers, can we identify those at risk of mental distress? However, while motivation may provide impetus, it is not the formal starting point of test development. The formal starting point for all test development is to generate a construct definition, which broadly is a definition of what is to be measured. An initial construct definition should be as clear as possible but will often be somewhat broad. For example, one might decide that a measure of cognitive ability, or leader potential, or anxiety is required (perhaps in order to address our previous practical concerns). From this point, one can define these constructs (using extant work or a newly generated definition as appropriate) and conduct a systematic literature review to identify existing tests and find out more about the nature of the target construct. This review should help the developer to refine their construct definition. For example, if you were interested in measuring cognitive ability an initial definition might be incomplete or very high level (e.g., ability to acquire and use knowledge or the speed and accuracy of information processing). However, based on a literature review, one could choose to devise sufficient tests to provide coverage of all second-order factors of cognitive ability contained within the Cattell–Horn–Carroll model (McGrew, 2009). This was broadly the strategy used in the development of the Woodcock–Johnson IV (McGrew, 2009).

However, relying solely on extant models might not always be the most useful strategy for at least two reasons, which we will explore using the Five Factor Model (FFM) of personality (Costa & McCrae, 1995). First, because the FFM is so widely accepted, there already exist a large number of tests based on this model and the question then arises as to what is the need for another identical test. Second, although the FFM is widely accepted, it seems unlikely that it is the final word on personality. Some argue that there are facets of personality out with the sphere of the FFM (Paunonen & Jackson, 2000), including some aspects of abnormal personality (Mathieu, Hare, Jones, Babiak, & Neumann, 2013). Of course, the NEO-PIR was not designed to measure abnormal personality, but there are strong arguments that broad-spectrum measures of personality should cover both the normal and abnormal (Markon, Krueger, & Watson, 2005). This may seem like a disadvantage but of course, from the point of view of a test developer, it is an opportunity. There is much more value in a new test that does something that an old test does not.

There may of course be many reasons for developing a test. There may be a need for research on a topic, but no extant measure suitable to carry out the research. For example, knowledge is a very important aspect of human behavior, but until about the year 2000 there were no standardized tests of knowledge (Irwing, Cammock, & Lynn, 2001; Rolfhus & Ackermann, 1999). Outside of research: diagnosis, assessment, and development, employee selection, market research, licensing and credentialing (e.g., the examinations that qualify one to practice as an accountant or lawyer) represent other broad categories of test needs. Broadly, there is a need for a test if your systematic literature review reveals that a test does not currently exist, current tests are judged to be inadequate, or there are tests, but not ones suitable for the particular population or use to which the test is to be put. Certainly, many instances of copycat tests exist, but I am not aware that this strategy has generally proven to be a recipe for a successful test.

Generally, successful tests are developed due to some combination of three circumstances:

1 Theoretical advances (NEO PI-R: Costa & McCrae, 1995; 16 PF: Conn & Rieke, 1994; VPI: Holland, Blakeney, Matteson, & Schnitzen, 1974; WAIS: Wechsler, 1981);
2 Empirical advances (MMPI: Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989);
3 A practical (market) need (SAT: Coyle & Pillow, 2008; GMAT: Oh, Schmidt, Shaffer, & Le, 2008).

If the developer does not make a test based on theoretical advance, empirical advance, or a gap in the market and instead duplicates a test, or more realistically produces a test that shares a name with another but has subtle differences in content, then the result is construct proliferation and the well-documented problems commonly referred to as the Jingle-Jangle fallacy (Hughes, Chapter 24; Shaffer, DeGeest, & Li, 2016).

Theoretical and empirical advances

Theoretical advancements (often driven by prior empirical discoveries) undoubtedly provide the reason for the development of many tests. Briefly, the test developer must develop a theoretical framework, which is in some respect new and sounder than previous frameworks, or utilize existing theoretical frameworks that current tests have not exploited. A full discussion of the nature of theoretical advances is well beyond the practical bounds of this chapter because it will be unique for every construct. That said, the history of the development of the FFM is highly instructive as to the process whereby theory evolves from an interaction between theoretical and empirical developments (Block, 1995; John, & Srivastava, 1999, see later). Also, pivotal to test development is the evolution of tight construct definitions, which also emerges from the interaction between theory and empirical work.

Systematic domain mapping

Perhaps the most obvious example of an interaction between theoretical and empirical advance comes in the form of systematic domain mapping. Very simply, a systematic domain map consists of all construct-relevant content (e.g., every aspect of the domain of personality) mapped onto a theoretically supported structure. This serves as a precursor to developing a systematic taxonomy of the domain that ideally identifies all primary level and higher-level constructs and provides the basic material from which test items can be constructed.

The history of testing suggests ways in which this can be achieved. Although all attempts to map a domain suffer from practical and statistical limitations. For example, the total number of possible personality items is sizable and collecting data on so many items is difficult as is subsequent analysis. For instance, factor analysis cannot handle the size of data matrix that would be required, meaning that in practice the total domain needs to be divided into manageable chunks based on a subjective grouping (see Booth & Murray, Chapter 29). The process of grouping items inevitably means that

some constructs which span the subjective groupings or sit at the interface between two are not sufficiently captured. Nevertheless, the development of the FFM, for example, is instructive both as to how domain mapping can be achieved and also the potential flaws in this process. Actually, the history of the development of the FFM is complex (Block, 1995; John, Angleiter, & Ostendorf, 1988), but a simplified account of the principles of its development will suffice for our purposes. Arguably, the development of the FFM stems from the lexical hypothesis, which is comprised of two major postulates. The first states that those personality characteristics that are most important in peoples' lives will eventually become a part of their language. The second follows from the first, stating that more important personality characteristics are more likely to be encoded into language as a single word (John et al., 1988). If true, then in principle, if all words that describe personality were incorporated into a questionnaire, and a large population were rated as to the extent these words apply to them, then a factor analysis of this data would provide the facet and higher-order structure of personality. In practice, despite claims to the contrary, for various practical reasons this has never been done, but something like it has (e.g., Ashton, Lee, & Goldberg, 2004). Personality research is now at a stage at which there are many respected measures of personality and the next step might be to administer all known measures of personality to a large population and, guided by theoretical developments, factor analyze the resultant data set in order to provide a new and more comprehensive taxonomy of personality (Booth, 2011; Woods & Anderson, 2016).

What this example illustrates is that successful test development often requires some form of systematic domain mapping, which is in at least some respects novel.

Practical (market) need

Of course, measures derived from a taxonomy or theory do not necessarily correspond to a practical need (beyond the need for measurement). Indeed, one difficulty with omnibus measures (such as the Woodcock–Johnson IV and the NEO PI-R) is that they rarely correspond to a direct market need. In the most part, this is because omnibus measures are often long and time-consuming to complete, resulting in equally long and detailed reports. Exhaustive reports concerned with all aspects of personality or cognitive ability can be difficult for laypersons to understand and use. Usually, the tests adopted by consumers are shorter and considered more user friendly. For example, despite being technically deficient (Hughes & Batey, 2017), the MBTI is among the most commonly used personality tests because it is relatively short, the results are easily communicated and understood, and therefore it can readily be used in a practical context. Probably therefore, marketable tests may be based on a systematic taxonomy but the actual constitution of the test will depend on additional considerations. In short, for a test to address a market need it should be both technically sound (in terms of theoretical grounding and psychometric properties) and practically useful.

The area of selection can help illustrate what some of these additional practical considerations might be. One starting point might be to identify the market for a selection test based on systematic market research. Let us imagine that the results of this research reveal there to be a large market for the recruitment of managers, not least because a large number of managers are employed, and secondly because their characteristics are often considered crucial to the success or failure of companies. How then could

we devise a test for managers? Traditionally, most test developers for a selection instrument would begin with a job analysis (Smith & Smith, 2005). This is still an essential step in the development of selection tests, however, since the late 1970s psychometric meta-analysis has become an important source of information to guide the development of selection instruments.

Meta-analysis

The main purpose of psychometric meta-analysis is to obtain parameter estimates, which are unbiased and corrected for measurement artifacts. Hunter and Schmidt (2004) is probably the most useful introduction to meta-analysis, although some more recent developments are contained in Borenstein, Hedges, Higgins, and Rothstein (2009). Meta-analysis has many potential applications to test development. For example, with regard to the construction of test batteries for employee selection, findings of meta-analyses identify which constructs predict future job performance and, therefore, which should be included (e.g., Judge, Rodell, Kliner, Simon, & Crawford, 2013; Schmidt, Shaffer, & Oh, 2008).

Psychometric meta-analysis averages the value of an effect size across studies in order to obtain a reliable summary estimate. The most important effect size in a selection context is the predictive validity, which is measured by the correlation between the score on the selection measure and some measure of job performance. The biggest problem with most estimates of predictive validity from single studies arises from sampling error, which is more considerable than is generally imagined. As sample size tends to infinity, so sampling error tends to zero and thus by amalgamating findings across studies, large meta-analyses effectively reduce sampling error to miniscule proportions. Standardly, psychometric meta-analysis also corrects for artifacts due to error of measurement, range restriction, imperfect construct validity (e.g., different measures of purportedly the same personality construct typically correlate at 0.4–0.6, see Pace & Brannick, 2010), use of categorical measurement, study quality, and publication bias. However, once these corrections are made, the confidence interval around the effect size estimate may still be large. This may indicate that the effect size is dependent on a third variable, usually referred to as a moderator. For example, cognitive ability predicts more strongly for complex jobs (Schmidt & Hunter, 1998) and in the case of personality, traits predict more strongly when they are relevant (e.g., Extraversion and sales; Hughes & Batey, 2017).

The findings of meta-analysis with regard to which cognitive abilities and FFM personality factors predict job performance are, within limits, fairly definitive (Schmidt & Hunter, 1998, 1998; Schmidt, Shaffer, & Oh, 2008). Virtually every meta-analysis that has investigated the issue has concluded that, for most jobs, general cognitive ability is the best predictor and the level of prediction increases in proportion to the cognitive demands of the job (Schmidt & Hunter, 1998). Moreover, it is generally contended that second-order factors of cognitive ability such as spatial, verbal, and memory add little incremental prediction (e.g., Carretta & Ree, 2000; Ree, Earles, & Teachout, 1994). Although it is a hotly contested issue, meta-analyses of the predictive validity of personality show virtually the opposite; that is, that personality largely does not offer blanket prediction of job performance across roles. Some have argued from this data that personality tests should not be used in selection (Morgeson et al., 2007), but many have also argued otherwise (e.g., Ones, Dilchert,