Cross-Cultural Analysis

Intended to bridge the gap between the latest methodological developments and cross-cultural research, this interdisciplinary resource presents the latest strategies for analyzing cross-cultural data. Techniques are demonstrated through the use of applications that employ cross-national data sets such as the latest European Social Survey. With an emphasis on the generalized latent variable approach, internationally prominent researchers from a variety of fields explain how the methods work, how to apply them, and how they relate to other methods presented in the book. Syntax and graphical and verbal explanations of the techniques are included. A website features some of the data sets and syntax commands used in the book.

Applications from the behavioral and social sciences that use real data sets demonstrate:

- The use of samples from 17 countries to validate the resistance to change scale across these nations
- How to test the cross-national invariance properties of social trust
- The interplay between social structure, religiosity, values, and social attitudes
- A comparison of anti-immigrant attitudes and patterns of religious orientations across European countries

The second edition includes six new chapters and two revised ones presenting exciting developments in the literature of cross-cultural analysis, including topics such as approximate measurement invariance, alignment optimization, sensitivity analyses, a mixed-methods approach to test for measurement invariance, and a multilevel structural equation modeling approach to explain noninvariance.

This book is intended for researchers, practitioners, and advanced students interested in cross-cultural research. Because the applications span a variety of disciplines, the book will appeal to researchers and students in: psychology, political science, sociology, education, marketing and economics, geography, criminology, psychometrics, epidemiology, and public health, as well as those interested in methodology. It is also appropriate for an advanced methods course in cross-cultural analysis.

Eldad Davidov is Professor of Sociology at the University of Cologne, Germany and the University of Zurich, Switzerland.

Peter Schmidt is Professor Emeritus of Methodology of Social Research at the University of Giessen and Humboldt Research Fellow of the Polish Foundation for Basic Research.

Jaak Billiet is Professor Emeritus at the Katholieke Universiteit Leuven, Belgium.

Bart Meuleman is Associate Professor and head of department at the Centre for Sociological Research (CeSO) at the University of Leuven, Belgium.
The European Association of Methodology (EAM) serves to promote research and development of empirical research methods in the fields of the Behavioural, Social, Educational, Health and Economic Sciences as well as in the field of Evaluation Research. 
Homepage: http://www.eam-online.org

The purpose of the EAM book series is to advance the development and application of methodological and statistical research techniques in social and behavioral research. Each volume in the series presents cutting-edge methodological developments in a way that is accessible to a broad audience. Such books can be authored, monographs, or edited volumes.

Sponsored by the European Association of Methodology, the EAM book series is open to contributions from the Behavioral, Social, Educational, Health and Economic Sciences. Proposals for volumes in the EAM series should include the following: (1) Title; (2) authors/editors; (3) a brief description of the volume’s focus and intended audience; (4) a table of contents; (5) a timeline including planned completion date. Proposals are invited from all interested authors. Feel free to submit a proposal to one of the members of the EAM book series editorial board, by visiting the EAM website http://eam-online.org. Members of the EAM editorial board are Manuel Ato (University of Murcia), Pamela Campanelli (Survey Consultant, UK), Edith de Leeuw (Utrecht University) and Vasja Vehovar (University of Ljubljana).

Volumes in the series include

**Davidov/Schmidt/Billiet:** Cross-Cultural Analysis: Methods and Applications, 2011

**Das/Ester/Kaczmirek:** Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies, 2011

**Hox/Roberts:** Handbook of Advanced Multilevel Analysis, 2011

**De Leeuw/Hox/Dillman:** International Handbook of Survey Methodology, 2008

**Van Montfort/Oud/Satorra:** Longitudinal Models in the Behavioral and Related Sciences, 2007
Contents

About the Contributors ix
Preface xxi

Peter Schmidt, Eldad Davidov, Jaak Billiet, and Bart Meuleman

SECTION I  MGCFA and MGSEM Techniques

Chapter 1  Capturing Bias in Structural Equation Modeling ........ 3
  Fons J. R. van de Vijver

Chapter 2  Evaluating Change in Social and Political Trust in
  Europe ................................................................. 45
  Nick Allum, Sanna Read, and Patrick Sturgis

Chapter 3  Methodological Issues in Using Structural Equation
  Models for Testing Differential Item Functioning ........ 65
  Jaehoon Lee, Todd D. Little, and Kristopher J. Preacher

Chapter 4  Estimation and Comparison of Latent Means
  Across Cultures ...................................................... 95
  Holger Steinmetz

Chapter 5  Biased Latent Variable Mean Comparisons Due to
  Measurement Noninvariance: A Simulation Study ..... 127
  Alain de Beuckelaer and Gilbert Swinnen

Chapter 6  Testing the Invariance of Values in the Benelux
  Countries With the European Social Survey:
  Accounting for Ordinality ..................................... 157
  Eldad Davidov, Georg Datler, Peter Schmidt, and Shalom H. Schwartz
Chapter 7  Religious Involvement: Its Relation to Values and Social Attitudes ................................................................. 181
Bart Meuleman and Jaak Billiet

Chapter 8  Measurement Equivalence of the Dispositional Resistance to Change Scale................................................... 215
Shaul Oreg, Mahmut Bayazıt, Maria Vakola, Luis Arciniega, Achilles Armenakis, Rasa Barkauskiene, Nikos Bozionelos, Yuka Fujimoto, Luis González, Jian Han, Martina Hřebíčková, Nerina Jimmieson, Jana Kordačová, Hitoshi Mitsuhashi, Boris Mlačić, Ivana Ferić, Marina Kotrla Topić, Sandra Ohly, Per Øystein Saksvik, Hilde Hetland, Ingvild Berg Saksvik, and Karen van Dam

Chapter 9  Measurement Equivalence Testing 2.0 .......................... 245
William M. van der Veld and Willem E. Saris

SECTION II  Multilevel Analysis

Chapter 10 Perceived Economic Threat and Anti-Immigration Attitudes: Effects of Immigrant Group Size and Economic Conditions Revisited........................................ 283
Bart Meuleman

Chapter 11  A Multilevel Regression Analysis on Work Ethic........ 315
Hermann Dülmer

Chapter 12 Multilevel Structural Equation Modeling for Cross-cultural Research: Exploring Resampling Methods to Overcome Small Sample Size Problems ............... 347
Remco Feskens and Joop J. Hox

Chapter 13  Explaining Cross-National Measurement Inequivalence: A Bayesian Multilevel CFA With Random Loadings ................................................................. 363
Bart Meuleman and Elmar Schlüter
SECTION III  Latent Class Analysis (LCA)

Chapter 14 Testing for Measurement Invariance With Latent Class Analysis.............................................................. 393
Miloš Kankaraš, Guy Moors, and Jeroen K. Vermunt

Chapter 15 A Multiple Group Latent Class Analysis of Religious Orientations in Europe................................. 421
Pascal Siegers

Chapter 16 Testing for Invariance of Latent Classes: Group-as-Covariate Approach........................................... 449
Maksim Rudnev

SECTION IV  Item Response Theory (IRT)

Chapter 17 Using a Differential Item Functioning Approach to Investigate Measurement Invariance......................... 483
Rianne Janssen

Chapter 18 Using the Mixed Rasch Model in the Comparative Analysis of Attitudes.............................................. 501
Markus Quandt

Chapter 19 Random Item Effects Modeling for Cross-National Survey Data.......................................................... 529
Jean-Paul Fox and Josine Verhagen

SECTION V  New and Emerging Topics in Cross-Cultural Analysis

Chapter 20 Exact and Bayesian Approximate Measurement Invariance................................................................. 553
Daniel Seddig and Heinz Leitgöb
Chapter 21 Alignment Optimization: Estimation of the Most Trustworthy Means in Cross-Cultural Studies even in the Presence of Noninvariance ........................................ 571
   Jan Cieciuch, Eldad Davidov, and Peter Schmidt

Chapter 22 Sensitivity Analysis .......................................................... 593
   Daniel L. Oberski

Chapter 23 How Should Immigrants Adapt to Their Country of Residence? A Mixed Methods Approach to Evaluate the International Applicability of a Question From the German General Social Survey (ALLBUS) ............ 615
   Michael Braun and Timothy P. Johnson

Author Index .................................................................................... 633

Subject Index .................................................................................... 645
Contributors

Nick Allum is Professor of Research Methodology at the University of Essex. He obtained his PhD from the London School of Economics and Political Science and his graduate work was on the public perception of risk in relation to gene technology. Nick has published on risk perception and attitudes toward science and technology in numerous articles and books. He also researches and publishes on survey methodology, with a particular focus on measurement error.

Luis Arciniega is Professor of Organizational Behavior and Human Resources Management (HRM) at ITAM Business School in Mexico City. He received his PhD in Organizational Psychology from the University of Salamanca in Spain. His research interests focus on studying the influence of personal values on work attitudes and team processes. He has published in leading academic journals and presented several papers in international conferences. He worked as a practitioner in HRM for more than a decade at Televisa Group, one of the largest media corporations in the world and as a consultant for multinational organizations.

Achilles Armenakis is the James T. Pursell, Sr. Eminent Scholar in Management Ethics and Director of the Auburn University Center for Ethical Organizational Cultures. Achilles’ research has focused on diagnosis, implementation, and assessment of organizational change. He is currently involved in research on ethics-related topics in organizational change.

Rasa Barkauskiene is Associate Professor at the Department of Clinical and Organizational Psychology at Vilnius University, Lithuania. Her research interests include personality psychology and developmental psychopathology.

Mahmut Bayazit is Assistant Professor of Organizational Studies at Sabanci University, where he teaches courses on organizational behavior and leadership. He received his PhD from the School of Industrial and
Labor Relations, Cornell University. His research focuses on the social psychology of unions and the impact of leadership on employee motivation and attitudes. His research has been published in journals such as *Journal of Applied Psychology, Journal of Occupational Health Psychology*, and *Small Group Research*.

**Jaak Billiet** was Full Professor in Social Methodology at the Katholieke Universiteit Leuven, Belgium, and is now Professor Emeritus but with a research tasks. His main research interest in methodology deals with validity assessment and the modeling of measurement error in social surveys. His substantial work covers longitudinal and comparative research in the domains of ethnocentrism, political attitudes, and religious orientations. Recent publications have appeared in *Sociological Methods & Research, European Sociological Review, Social Science Research, Survey Research Methods*, and *Journal of Social Issues*.

**Nikos Bozionelos** is a member of the academic faculty of the Athens University of Economics and Business in the field of Organizational Behavior. His interests include cross-cultural psychology and management, careers, and emotion work. He is the Associate Editor for Career Development International and a member of the Steering Committee of the Careers Division of the Academy of Management. He belongs to the small proportion of academics with an individual impact factor (hI) above 10 (http://www.harzing.com/pop.htm).

**Michael Braun** is a project consultant at GESIS, the Leibniz Institute for the Social Sciences, and an adjunct professor at the University of Mannheim. His interests include the methodology of international comparative research as well as substantive research in the areas of family, work, and migration.

**Jan Cieciuch** is a project leader of the University Research Priority Program on Social Networks at the University of Zurich and an Associate Professor of Psychology at the Cardinal Wyszynski University in Warsaw, Poland. His interests are applications of structural equation modeling, especially in psychology, with focus on the investigation of human values and personality traits. Recent publications have appeared in the *Journal*.

**Georg Datler** is a Lecturer at the Institute of Sociology, University of Duisburg-Essen, Germany. His research interests are in political sociology, European integration, the methodology of the social sciences, and the application of structural equation modeling in a comparative perspective.

**Eldad Davidov** is Professor of Sociology at the University of Cologne, Germany, and the University of Zurich, Switzerland, and co-director of the University Research Priority Program “Social Networks.” His research interests are applications of structural equation modeling to survey data, especially in cross-cultural and longitudinal research. Applications include human values, national identity, and attitudes toward immigrants and other minorities.

**Alain de Beuckelaer** is Professor of Organizational Research Methods at Ghent University, Belgium, and a Tenured Faculty Member of Strategic Human Resource Management at Nijmegen School of Management, The Netherlands. His scientific work is interdisciplinary including management, psychology, and organizational behavior. Most of his research has dealt with, and deals with methodological issues in quantitative empirical (including cross-cultural comparative) research.

**Hermann Dülmer** is Assistant Professor of Sociology at the Institute of Sociology and Social Psychology of the University of Cologne. His substantive research interests include comparative values research, including value change, and electoral research, with a particular emphasis on right-wing extremism. His methodological interests focus on multilevel analysis including multilevel structural equation modeling and on factorial surveys.

**Ivana Ferić** earned her PhD in Social Psychology from the Department of Psychology at University of Zagreb. Since 1998 she has been employed at the Institute of Social Sciences Ivo Pilar in Zagreb, where she is now a research associate. She specialized in methodology of scientific research, and her current interests include human values, attitudes, and beliefs.
Remco Feskens is a researcher at the IQ Healthcare department at the University Medical Centre St. Radboud Nijmegen. His current research projects are multilevel research and survey methodology. Recent publications include “Difficult groups in survey research and the development of tailor-made approach strategies” (PhD thesis, Utrecht University) and publications in the *Journal of Official Statistics, Field Methods*, and *Survey Research Methods*.

Jean-Paul Fox is Associate Professor in the Department of Research Methodology, Measurement, and Data Analysis, at the University of Twente, Enschede, The Netherlands. His areas of specialization include developing and applying Bayesian techniques and statistical models in test theory and other scientific disciplines.

Yuka Fujimoto is Senior Lecturer at Deakin University within the School of Management and Marketing. Her research interest lies in diversity and justice-oriented decision-making processes and diversity inclusiveness in organizations. She is a recipient of best paper proceedings at Academy of Management and Asia Academy of Management conferences. She has also coauthored a textbook entitled *SHRM: Transforming theory into practice*.

Luis González is Professor of Human Resource Management and Work and Organizational Psychology in the Faculty of Economics and Management at the University of Salamanca, Spain. He received his PhD in Social Psychology from University of Salamanca. He has served as a Director of the Master on Human Resource Management of University of Salamanca. His research interests include the development of evaluation instruments in work and organizational psychology, methodological issues in structural equation modeling, work motivation and work redesign, organizational commitment, and human resources management.

Jian Han is Assistant Professor of Management at CEIBS. Her research interest mainly focuses on how to integrate human resources management with characteristics of different levels (individual, team, and corporate) in an effort to improve organizational performance. Dr. Han obtained her doctoral degree from the ILR School, Cornell University.
**Hilde Hetland** is Associate Professor at the Department of Psychology at the University of Bergen, Norway. Her research areas are organizational psychology and educational psychology. She has published articles in various international journals.

**Joop J. Hox** is Full Professor of Social Science Methodology at the Department of Methodology and Statistics of the Faculty of Social Sciences at Utrecht University. As Methodology Chair, he is responsible for the research, development, and teaching carried out at the faculty in the field of social science methods and techniques. His research interests focus on two lines of work: data quality in social surveys and multilevel modeling. In survey methodology, he has written articles on nonresponse problems and the effects of data collection mode and interviewers on various aspects of data quality. In multilevel modeling, he has written numerous articles, book chapters, and an introductory handbook, with a newly written monograph currently in press.

**Martina Hřebíčková** received her PhD from Charles University of Prague; she is currently Senior Scientist at the Institute of Psychology, Academy of Science of the Czech Republic in Brno. Her research interests include personality structure and measurement, with a special interest in the psycholexical and dispositional approach to trait psychology.

**Rianne Janssen** received her PhD in Psychology from the Katholieke Universiteit Leuven (Belgium). Currently, she is Assistant Professor at the Faculty of Psychology and Educational Sciences at the same university.

**Nerina Jimmieson** is Associate Professor and the Centre and Program Director for Organizational Psychology at The University of Queensland, Australia. She teaches undergraduate and postgraduate courses in organizational behavior, personnel selection, and organizational change. Her research interests are concerned with stress and coping in the workplace, employee adaptation to organizational change, and the role of employee attitudes and behaviors in the prediction of client satisfaction.

**Timothy Johnson** is Director of the Survey Research Laboratory and Professor of Public Administration at the University of Illinois at Chicago.
His research interests include cultural influences on survey-related response processes, and health behaviors in disadvantaged populations. During 2017–2018, he also served as President of the American Association for Public Opinion Research (AAPOR).

Miloš Kankaraš is a PhD student in the Department of Methodology and Statistics at Tilburg University, The Netherlands. He holds a MSc in Psychology from University of Belgrade, Serbia. His research interests are in the areas of cross-cultural comparative research, measurement equivalence, and attitudes and values.

Jana Kordačová holds the position of senior research fellow at the Institute of Experimental Psychology of the Slovak Academy of Sciences. Her long-term interest lies in functional and dysfunctional aspects of cognition and emotion relations (until late it was specifically irrational beliefs) and at present in the context of positive psychology as well. In 2006 to 2008 she was the leader of the VEGA research grant “Human strengths and their place in the experiencing of well-being and happiness.”

Jaehoon Lee is Assistant Professor of Educational Psychology at Texas Tech University. His research interests are primarily in the evaluation of structural equation modeling and item response theory with respect to practical measurement issues, including measurement equivalence, differential item functioning, and matching/anchoring. An additional interest is the use of Monte Carlo analysis for educational and organizational research.

Heinz Leitgöb is academic councilor at the University of Eichstätt-Ingolstadt (Germany), Institute of Sociology, and Lecturer at the University of Linz (Austria), Department of Empirical Social Research. He is co-chair of the European Working Group on Quantitative Methods in Criminology in the European Society of Criminology. His methodological research interests cover the comparability of survey-based measurements, the identification of survey mode effects, and nonlinear modeling. Topics of his research belong to the fields of criminology, educational inequalities, and labor market sociology. Recent publications have appeared in the Journal for Labour Market Research and International Review of Victimology.
Todd D. Little is Professor of Educational Psychology at Texas Tech University. He also serves as the Director of Texas Tech’s Institute for Measurement, Methodology, Analysis and Policy (IMMAP). His quantitative research interests include psychometrics, structural equation modeling, the analysis of repeated measures/longitudinal data, and cross-cultural data analysis.

Bart Meuleman is Associate Professor and head of department at the Centre for Sociological Research (CeSO) at the University of Leuven (Belgium). He currently (2017–2019) acts as President of the European Survey Research Association (ESRA). His main research interests are cross-cultural survey methodology and cross-national comparisons of value and attitude patterns.

Hitoshi Mitsuhashi is Associate Professor on the Faculty of Business and Commerce at Keio University. He received his PhD from the New York State School of Industrial and Labor Relations at Cornell University. His research interests include the evolutionary dynamics of interorganizational relations and the path dependency of organizational behavior.

Boris Mlačić earned his PhD in personality psychology from the Department of Psychology, Faculty of Philosophy at University of Zagreb. Since 1992 he has been employed at the Institute of Social Sciences Ivo Pilar in Zagreb, where he is now a Senior Research Associate. He was a recipient of the Croatian Annual National Award for Science for 1999. In his research he focuses on individual differences, the lexical approach in personality psychology, the big-five model, and personality development.

Guy Moors is Assistant Professor in the Department of Methodology and Statistics at Tilburg University, The Netherlands. He holds a PhD in social sciences from the Free University of Brussels. His publications are in the field of social demography, survey methodology, cross-cultural comparative research, attitudes and values research.

Daniel L. Oberski is Associate Professor at the Department of Methodology and Statistics, Utrecht University, The Netherlands. His research interests focus on latent variable modeling and measurement error in the social sciences. He is Associate Editor of the Journal of Survey Statistics and

Sandra Ohly is currently Assistant Professor in Industrial and Organizational Psychology at the University of Frankfurt. Her research focuses on creativity at work, proactive behavior, and suggestion making. She is also interested in the effects of time pressure on motivation and in emotions, organizational change, and resistance to change.

Shaul Oreg earned his PhD in organizational behavior from Cornell University. He is now Associate Professor at the School of Business Administration of the Hebrew University of Jerusalem. In his research he focuses on individual differences in social and organizational contexts. In particular, he studies traits and values and their proximal and distal effects, with a particular interest in the contexts of organizational change and leadership. He specializes in quantitative methods, including scale development, advanced factor analytic techniques, and cross-level analyses. He is co-author of Resistance to Innovation (University of Chicago Press, 2015) and co-editor of The Psychology of Organizational Change (Cambridge, 2013). His work on organizational change is featured in the upcoming Palgrave Handbook of Organizational Change Thinkers.

Kristopher J. Preacher is Professor of Quantitative Methods at Vanderbilt University. His research concerns the use of structural equation modeling and multilevel modeling to analyze longitudinal and correlational data. Other interests include developing techniques to test mediation and moderation hypotheses, and improving model selection in the application of multivariate methods to social science questions.

Markus Quandt is Senior Researcher and acting head of department at the GESIS Data Archive in Cologne, Germany. He has rich experience in dealing with comparative survey data from his involvement in the data integration work of the International Social Survey Programme. He has also worked in the analysis of political attitudes and electoral research.
Sanna Read has a PhD in Psychology and works as an Assistant Professorial Research Fellow at the London School of Economics and Political Science. Her research interests are motivation, social networks, and health in adulthood and old age.

Maksim Rudnev is a researcher at the Laboratory for Comparative Research of Mass Consciousness, National Research University Higher School of Economics in Moscow, Russia. The emphasis of his PhD thesis was on the topic of basic values, and he is interested in the methodology of large-scale comparative studies that focus on basic values and moral attitudes.

Ingvild Berg Saksvik is a PhD student at the Department of Psychology at the University of Bergen in Norway. Her research areas are personality, stress, sleep, and shift work.

Per Øystein Saksvik is Professor at the Department of Psychology, Norwegian University of Science and Technology, where he also received his PhD in 1991 in Occupational Health Psychology. He has seven years of experience as a researcher at the Institute of Social Research in Industry, Trondheim, Norway. He does research in occupational health and safety, organizational interventions, sickness absenteeism and presenteeism, and organizational change.

Willem E. Saris is a laureate of the Descartes Prize 2005, for the best collaborative research. In 2009 he received the Helen Dinerman Award by the World Association of Public Opinion Research in recognition of his extensive contributions to the field of public opinion research and development of survey research methods. In 2014 he received, from AAPOR, the Warren J. Mitofsky Innovators Award for the Survey Quality Predictor (SQP) 2.0 and its Contribution to the Improving Questionnaire Design. He is also the former president and founder of the European Survey Research Association and a former member of the central coordinating team (CCT) of the European Social Survey.

Elmar Schlüter obtained his PhD as a fellow of the German Science Foundation Research Training School “Group-focused enmity.” He is Professor of Sociology at the Justus-Liebig-University of Giessen, Germany,
with a specialization in methods of cross-national research methods. His methodological research interests are in the areas of structural equation modeling and multilevel modeling.

Peter Schmidt is Professor Emeritus of Methodology of Social Research at the University of Giessen and Humboldt Research Fellow of the Polish Foundation for Basic Research. His research interests are foundations and applications of structural equation modeling to survey data and experimental and quasi-experimental data, especially in cross-cultural and longitudinal research. Applications include human values, innovation, national identity and nationalism, and attitudes toward immigrants and other minorities.

Shalom H. Schwartz is the Sznajderman Professor Emeritus of Psychology at the Hebrew University of Jerusalem, Israel. His recent work concerns two topics: the nature and sources of basic human values and their role as bases of attitudes and behavior, and the nature and sources of cultural value orientations as expressions of and influences on the institutional structures, policies, and prevailing norms and practices in different societies.

Daniel Seddig is Academic Councilor at the University of Cologne (Germany), Institute of Sociology and Social Psychology and statistical consultant at the University of Zurich (Switzerland), Department of Psychology. He is co-chair of the European Working Group on Quantitative Methods in Criminology (EQMC) in the European Society of Criminology (ESC). His research interests are cross-contextual and longitudinal data analysis with structural equation models. Topics of his research are the value-attitude-behavior relationship, youth development and delinquency, and life-course criminology. Recent publications have appeared in *Crime & Delinquency* and *Sociological Methods & Research*.

Pascal Siegers graduated in social sciences from SciencesPo Bordeaux (France) and the University of Stuttgart (Germany). He is currently a postgraduate student at the research training group “Social Order and Life Chances in International Comparison” at the University of Cologne. His main research interests are the causes and consequences of alternative religiosity in contemporary Europe.
Holger Steinmetz is scientific assistant that the Faculty of Economics and Business Administration at the University of Giessen, Germany. His research interests are structural equation modeling, cross-cultural research, and work psychology.

Patrick Sturgis is Professor of Research Methodology in the Division of Social Statistics at the University of Southampton and Director of the UK National Centre for Research Methods. His research focuses on survey methodology, the dynamics of opinion formation and change, and the analysis of panel data.

Gilbert Swinnen is Professor Emeritus of Marketing at Hasselt University, Belgium. His scientific work within the marketing field has a strong emphasis on retailing and consumer behavior. He has also published numerous papers on multivariate statistical analysis techniques and data mining techniques.

Marina Kotrla Topić received her BA in psychology in 2002, and since 2003 she has been completing her PhD in Language Communication and Cognitive Neuroscience at the University of Zagreb. Since 2004 she has been employed at the Institute of Social Sciences Ivo Pilar in Zagreb.

Maria Vakola is an organizational psychologist and is currently working as an Assistant Professor at the Athens University of Economics and Business in Greece. Her main research interests focus on individual differences and reactions to organizational change. Maria has published in academic journals such as the *Journal of Applied Psychology*, *Journal of Organizational Change Management*, *Communications of the ACM*, and others.

Karen van Dam is Associate Professor of Personnel Psychology at Tilburg University, The Netherlands. She received her PhD at the University of Amsterdam. Karen’s research focuses on how employees adapt to changes in the work situation, including employability, employee learning, job changes, retirement, and resistance to change.

Fons J. R. van de Vijver holds a chair in Cross-Cultural Psychology at Tilburg University, The Netherlands; in addition, he is Extraordinary Professor at
North-West University, South Africa, and the University of Queensland, Australia. He is former Editor-in-Chief of the *Journal of Cross-Cultural Psychology* and former President of the International Association for Cross-Cultural Psychology. He has published over 500 publications, mainly in the domain of cross-cultural psychology (methods, intelligence, acculturation, and multiculturalism).

**William M. van der Veld** is a Senior Researcher at the Behavioral Science Institute. His research interests are scale development, factor analysis, model evaluation, survey methodology, and cross-cultural comparative research in the field of clinical psychology.

**Josine Verhagen** is a PhD student in the Department of Research Methodology, Measurement, and Data Analysis, at the University of Twente, Enschede, The Netherlands. Her research project is focused on the Bayesian modeling of heterogeneity for large-scale comparative research.

**Jeroen K. Vermunt** is Professor in the Department of Methodology and Statistics at Tilburg University, The Netherlands. He holds a PhD in social sciences from Tilburg University. He has published extensively on categorical data techniques, methods for the analysis of longitudinal and event history data, latent class and finite mixture models, and latent trait models.
In recent years, the increased awareness of researchers on the importance of choosing appropriate methods for the quantitative analysis of cross-cultural data can be clearly seen in the growing amount of literature on this subject (see Johnson, 2015, for an introduction of a special issue on the topic in the journal *Public Opinion Quarterly*). At the same time, the increasing availability of cross-national data sets, such as the European Social Survey (ESS), the International Social Survey Program (ISSP), the European Value Study (EVS), the World Value Survey (WVS), the European Household Panel Study (EHPS), and the Program for International Assessment of Students’ Achievements (PISA), just to name a few, allows researchers to engage in cross-cultural research now more than ever. Nevertheless, at present, most of the methods developed for such purposes are insufficiently applied, while several new ones are not well or not at all known, and their importance is often not recognized by substantive researchers in quantitative cross-national studies (for a discussion and a controversy on this issue, see Alemán & Woods, 2016). Thus, there is a growing need to bridge the gap between the methodological literature and applied cross-cultural research.¹ This edited volume is aimed toward this goal.

The purposes we try to achieve through this book are twofold. First, it should inform readers about the state of the art in the growing methodological literature on analysis of cross-cultural data. Because this body of literature is very large, our book pays a substantial amount of attention to strategies developed within the generalized latent variable approach (Skrodnal & Rabe-Hesketh, 2004). The methods presented in the book may be applied in diverse fields of comparative research such as sociology, political science, psychology and/or social psychology, business administration and marketing, economics, education, human geography, public health, and social epidemiology, just to name a few. Second, the book illustrates these methods by presenting interesting substantive applications using cross-national data sets and employing theory-driven empirical analyses. Our selection of contributors further reflects this structure. The authors represent established and internationally prominent, as well
as younger, researchers working in a variety of methodological and substantive fields in the social sciences.

The book is divided into five major sections that provide complementary strategies for analyzing cross-cultural data, all within the generalized latent variable approach: (1) multiple group confirmatory factor analysis (MGCFA), including the comparison of relationships and latent means, and the expansion of MGCFA into multiple group structural equation modeling (MGSEM), means and covariance structure analysis (MACS), and multiple indicator multiple causes (MIMIC) models; (2) multilevel analysis, including multilevel structural equation modeling; (3) latent class analysis (LCA); (4) item response theory (IRT); and (5) new developments.

The second edition of the book includes several innovations while trying to address the excellent suggestions made by Cho (2015) in his review of the first edition of this book. First, two chapters in the first section on MGCFA, one by Fons J. R. van de Vijver and another by Saris and van der Veld, were updated by the authors to include recent developments in the field. Second, a new chapter was added to the second section on multilevel analysis (Chapter 13 by Meuleman and Schlüter) and one to the third section on LCA (Chapter 16 by Rudnev). Third, we added a new, fifth section to the book with four exciting chapters discussing several contemporary developments in the analysis of cross-cultural data. Two chapters in this new section introduce and apply alternative methods to test measurement equivalence that can be applied if classical invariance tests fail to establish comparability; namely, approximate invariance testing using Bayesian estimation (Chapter 20 by Seddig and Leitgöb) and the alignment method (Chapter 21 by Cieciuch, Davidov, and Schmidt). In addition, this edition contains a new study by Oberski (Chapter 22) on the robustness of substantive findings in cases where researchers fail to establish measurement invariance. Braun and Johnson’s contribution (Chapter 23) completes the volume with an introduction to the use of mixed methods to assess measurement invariance by combining qualitative and quantitative techniques.

Whereas researchers in different disciplines tend to use different methodological approaches in a rather isolated way (e.g., IRT is commonly used by psychologists or education researchers; LCA by marketing researchers and sociologists; and MGCFA and multilevel analysis by sociologists, psychologists, and political scientists, among others), this book offers the
research community an integrated framework. In this framework, different cutting-edge methods are described, developed, applied, and linked, crossing ‘methodological borders’ between disciplines. The units include methodological as well as more applied chapters. Some chapters include a description of the basic strategy and how it relates to other strategies presented in the book. Other chapters include applications in which the different strategies are applied using real data sets to address interesting theoretically oriented research questions. A few chapters combine both aspects.

Some words about the structure of the book: Several orderings of the chapters within each unit were possible. We chose to organize the chapters from general to specific; that is, each unit begins with more general topics followed by later chapters focusing on more specific issues. However, the later chapters are not necessarily more technical or complex.

The first and largest unit focuses especially on MGCFA and MGSEM techniques and includes nine chapters (Chapters 1–9). The first chapter, “Capturing bias in structural equation modeling” by Fons J. R. van de Vijver, is a general discussion of how the models developed in cross-cultural psychology to identify and assess bias can be identified using structural equation modeling techniques. The second chapter, “Evaluating change in social and political trust in Europe” by Nick Allum, Sanna Read, and Patrick Sturgis, provides a nontechnical introduction to and application of MGCFA (including means and intercepts) to assess invariance. The method is demonstrated with an analysis of social and political trust in Europe in three rounds of the European Social Survey (ESS). The third chapter, “Methodological issues in using structural equation models for testing differential item functioning” by Jaehoon Lee, Todd D. Little, and Kristopher J. Preacher, discusses methodological issues that may arise when researchers conduct SEM-based differential item functioning (DIF) analysis across countries and shows techniques for conducting such analyses more accurately. In addition, they demonstrate general procedures to assess invariance and the mean differences of latent constructs across countries. The fourth chapter, “Estimation and comparison of latent means across cultures” by Holger Steinmetz, focuses on the use of MGCFA to estimate mean differences across cultures, a central topic in cross-cultural research. The author gives a clear and nontechnical introduction to latent mean difference testing, explains its presumptions, and illustrates its use
with data from the ESS on self-esteem. The fifth chapter, “Biased latent variable mean comparisons due to measurement noninvariance: A simulation study” by Alain de Beuckelaer and Gilbert Swinnen, presents a simulation study that assesses the reliability of latent variable mean comparisons across two groups when one latent variable indicator fails to satisfy the condition of measurement invariance across groups. The main conclusion is that noninvariant measurement parameters, and in particular a noninvariant indicator intercept, form a serious threat to the robustness of the latent variable mean difference test. The sixth chapter, “Testing the invariance of values in the Benelux countries with the European Social Survey: Accounting for ordinality” by Eldad Davidov, Georg Datler, Peter Schmidt, and Shalom H. Schwartz, tests the comparability of the measurement of human values in the second round (2004–2005) of the ESS across three countries, Belgium, The Netherlands, and Luxembourg, while accounting for the fact that the data are ordinal (ordered-categorical). They use a model for ordinal indicators that includes thresholds as additional parameters to test for measurement invariance. The general conclusions are that results are consistent with those found using MGCFA, which typically assumes the use of normally distributed, continuous data. The seventh chapter, “Religious involvement: Its relation to values and social attitudes” by Bart Meuleman and Jaak Billiet, focuses on the interplay between social structure, religiosity, values, and social attitudes. The authors use ESS (Round 2) data to compare these relations across 25 different European countries. Their study provides an example of how MGSEM can be used in comparative research. A particular characteristic of their analysis is the simultaneous test of both the measurement and structural parts in an integrated multigroup model. The eighth chapter, “Measurement equivalence of the dispositional resistance to change scale” by Shaul Oreg et al., uses confirmatory smallest space analysis (SSA) as a complementary technique to MGCFA. The authors use samples from 17 countries to validate the resistance to change scale across these nations. The last and ninth chapter in this unit, “Measurement equivalence testing 2.0” by William M. van der Veld and Willem E. Saris, illustrates how to test the cross-national invariance properties of social trust. The main difference to the second chapter is that here they propose a procedure that makes it possible to test for measurement invariance after correction for random and systematic measurement errors. In addition, they propose an alternative procedure to evaluate cross-national invariance that is implemented in a
software program called JRule. This software can detect local misspecifications in structural equation models taking into account the power of the test, which is not taken into account in most applications.

The second unit, which focuses on multilevel analysis, includes four chapters (10–13). The tenth chapter, “Perceived economic threat and anti-immigration attitudes: Effects of immigrant group size and economic conditions revisited,” by Bart Meuleman, demonstrates how two-level data may be used to assess context effects on anti-immigration attitudes. By doing this, the chapter proposes some refinements to existing theories on anti-immigrant sentiments and an alternative to the classical multilevel analysis. The eleventh chapter, “A multilevel regression analysis on work ethic” by Hermann Dülmer, uses multilevel analysis to reanalyze results on work ethic presented by Norris and Inglehart in 2004. This contribution illustrates the disadvantages of using conventional ordinary least squares (OLS) regression for international comparisons instead of the more appropriate multilevel analyses, by contrasting the results of both methods. The twelfth chapter, “Multilevel structural equation modeling for cross-cultural research: Exploring resampling methods to overcome small sample size problems” by Remco Feskens and Joop J. Hox, discusses the problem of small sample sizes on different levels in multilevel analyses. To overcome this small sample size problem they explore the possibilities of using resampled (bootstrap) standard errors. The thirteenth chapter, by Bart Meuleman and Elmar Schlüter, “Explaining cross-national measurement inequivalence: A Bayesian multilevel CFA with random loadings,” is a new contribution that demonstrates how a multilevel CFA with random loadings can be used to assess cross-cultural measurement equivalence. In addition, it illustrates how the method can be used to explain metric noninvariance using contextual variables with Bayesian estimation procedures.

The third unit, which focuses on LCA, includes three chapters (14–16). The fourteenth chapter, “Testing for measurement invariance with latent class analysis” by Miloš Kankaraš, Guy Moors, and Jeroen K. Vermunt, shows how measurement invariance may be tested using LCA. Because it can be used to model any type of discrete level data, LCA is an obvious choice when nominal indicators are used and/or it is a researcher’s aim to classify respondents in latent classes. The methodological discussion is illustrated by two examples. In the first example they use a multigroup LCA with nominal indicators; in the second a multigroup latent class
factor analysis with ordinal indicators. The fifteenth chapter, “A multiple group latent class analysis of religious orientations in Europe” by Pascal Siegers, tries to draw a comprehensive picture of religious orientations in 11 European countries by elaborating a multiple group latent class model that distinguishes between church religiosity, moderate religiosity, alternative spiritualities, religious indifferences, and atheism. Maksim Rudnev is the author of the sixteenth chapter, “Testing for invariance of latent classes: Group-as-covariate approach.” This chapter describes the group-as-covariate approach as one way of testing for the measurement invariance of latent classes. Using data on basic human values from two groups of countries (West and North European and Eastern European countries), he presents a practical application of this approach.

The fourth unit focuses on item response theory and includes three chapters (17–19). The seventeenth chapter, “Using a differential item functioning approach to investigate measurement invariance” by Rianne Janssen, describes how item response theory techniques may be used to test for measurement invariance. The author illustrates the procedure with an application using different modes of data collection: paper-and-pencil and computerized test administration. The eighteenth chapter, “Using the mixed Rasch model in the comparative analysis of attitudes” by Markus Quandt, explores the advantages and limitations of using Rasch models for identifying potentially heterogeneous populations using a practical application. This chapter could also be placed in the preceding LCA unit because of its use of latent class analysis as well. The nineteenth chapter, “Random item effects modeling for cross-national survey data” by Jean-Paul Fox and Josine Verhagen, shows how cross-national survey data can be properly analyzed using IRT with random item effects for handling measurement noninvariant items. Without the need of anchor items, differences of item characteristics across countries are explicitly modeled and a common measurement scale is obtained. The authors illustrate the method with educational data available from the PISA surveys.

Focusing on new developments, the fifth unit includes four new chapters (20–23). All of these contributions, like the new chapter of Meuleman and Schlüter in the second unit, provide valuable alternative options for researchers when cross-cultural measurement equivalence is not supported by the data. The twentieth chapter, “Exact and Bayesian approximate measurement invariance” by Daniel Seddig and Heinz Leitgöb,
presents the recently developed method of testing for approximate (rather than exact) measurement invariance. Using Bayesian analysis techniques, it demonstrates the application of this method in testing for cross-country approximate measurement invariance of attitudes toward granting citizenship rights to immigrants. It is an important alternative to more classical approaches testing for (exact) measurement invariance, when these methods fail to establish invariance, because the approximate measurement invariance method is more liberal and imposes less restrictive constraints on measurement parameters. The twenty-first chapter, “Alignment optimization: Estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance” by Jan Cieciuch, Eldad Davidov, and Peter Schmidt, presents the alignment method for testing for measurement invariance. This method is another important recent development and provides researchers with a valuable alternative test, particularly when a large number of groups is compared, when classical methods to test for measurement invariance fail to establish invariance. A practical application demonstrates how to use the method to test for measurement invariance of threat due to immigration. The twenty-second chapter, “Sensitivity analysis” by Daniel L. Oberski, evaluates the impact of freeing cross-group equality constraints on substantive conclusions. The chapter clarifies how to determine whether the parameters of interest are sensitive to violations of measurement invariance. The final and twenty-third chapter, “How should immigrants adapt to their country of residence? A mixed methods approach to evaluate the international applicability of a question from the German General Social Survey (ALLBUS)” by Michael Braun and Timothy P. Johnson, proposes a mixed methods approach to data analysis, where qualitative approaches are employed alongside quantitative procedures to test for measurement invariance. The chapter presents a practical application of mixed methods techniques to test for measurement invariance of attitudes toward immigrants using web surveys. Thus, it presents an alternative to common quantitative techniques to test for measurement invariance, allowing researchers to better understand the possible causes of noninvariance. Table 0.1 provides a summary and overview of all chapters, their goals, the methods they use, and the countries and data sets in their empirical applications.

In anthropology and cross-cultural psychology, it is common to distinguish between emic and etic approaches (see, e.g., Berry et al., 2011, or
<table>
<thead>
<tr>
<th>TABLE 0.1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overview</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Chapter Number, Author/s and Title</strong></td>
<td><strong>Topic, Statistical Method/s, and Goal of Analysis</strong></td>
</tr>
<tr>
<td>Section I</td>
<td>Countries and Data-Set</td>
</tr>
<tr>
<td>MGCFA and MGSEM Techniques</td>
<td></td>
</tr>
<tr>
<td>*1. Fons J.R. van de Vijver</td>
<td>Strengths and weaknesses of structural equation modeling (SEM) to test equivalence</td>
</tr>
<tr>
<td>* Capturing Bias in Structural Equation Modeling</td>
<td>in cross-national research</td>
</tr>
<tr>
<td>1. Theoretical framework of bias and equivalence</td>
<td>1.</td>
</tr>
<tr>
<td>2. Procedures and examples to identify bias and address equivalence</td>
<td>2.</td>
</tr>
<tr>
<td>3. Identification of bias using SEM</td>
<td>3.</td>
</tr>
<tr>
<td>4. Strengths, weaknesses, opportunities, and threat (SWOT) analysis</td>
<td>4.</td>
</tr>
<tr>
<td>2. Nick Allum, Sanna Read, and Patrick Sturgis</td>
<td>Analysis of social and political trust in European countries over time using SEM</td>
</tr>
<tr>
<td>* Evaluating Change in Social and Political Trust in Europe Using Multiple Group Confirmatory Factor Analysis with Structured Means</td>
<td>with structured means and multiple groups</td>
</tr>
<tr>
<td>1. Introduction to structured means analysis using SEM</td>
<td>1.</td>
</tr>
<tr>
<td>2. Application to the ESS data</td>
<td>2.</td>
</tr>
<tr>
<td>3. Jaehoon Lee, Todd D. Little, and Kristopher J. Preacher</td>
<td>Differential Item Functioning (DIF) and SEM based invariance testing</td>
</tr>
<tr>
<td>1. Introduction to the concept of factorial invariance</td>
<td>1.</td>
</tr>
<tr>
<td>2. Levels of invariance</td>
<td>2.</td>
</tr>
<tr>
<td>3. The concept of differential item functioning</td>
<td>3.</td>
</tr>
<tr>
<td>17 European countries; First three rounds of the ESS 2002, 2004, 2006</td>
<td>Two simulation studies</td>
</tr>
</tbody>
</table>
4. Holger Steinmetz  
*Estimation and Comparison of Latent Means Across Cultures*

Comparison of the use of composite scores and latent means in confirmatory factor analysis (CFA) with multiple groups (MGCFA), higher-order CFA, and MIMIC models
1. General discussion of observed means MGCFA, composite scores, and latent means
2. Application to ESS data measuring self-esteem in two countries using MGCFA

Two countries; First round of the ESS, 2002

5. Alain de Beuckelaer and Gilbert Swinnen  
*Biased Latent Variable Mean Comparisons Due to Measurement Non-Invariance: A Simulation Study*

Noninvariance of one indicator
MACS SEM with latent means and intercepts. Simulation study with a full factorial design varying:
1. the distribution of indicators
2. the number of observations per group
3. the non-invariance of loadings and intercepts
4. the size of difference between latent means across two groups

Two-country case; Simulated data

6. Eldad Davidov, Georg Datler, Peter Schmidt, and Shalom Schwartz  
*Testing the Invariance of Values in the Benelux Countries with the European Social Survey: Accounting for Ordinality*

Invariance testing of thresholds, intercepts, and factor loadings of values in the Benelux countries with MGCFA accounting for the ordinality of the data.
1. Description of the approach including MPLUS code
2. Comparison with MGCFA assuming interval data
3. Application to the ESS value scale

Three European countries; Second round of the ESS, 2004

7. Bart Meuleman and Jaak Billiet  
*Religious Involvement: Its Relation to Values and Social Attitudes. A Simultaneous Test of Measurement and Structural Models Across European Countries*

Effects of religious involvement on values and attitudes in Europe
MGCFAs and multigroup structural equation model (MGSEM)
1. Specification and test of measurement models
2. Specification of structural models

25 European countries; Second round of the ESS, 2004

8. Shaul Oreg et al.  
*Measurement Equivalence using Multi-Group Confirmatory Factor Analysis and Confirmatory Smallest Space Analysis: The Case of Dispositional Resistance to Change*

Resistance to change scale
MGCFAs and confirmatory SSA
Invariance of measurement, comparison over 17 countries using MGCFA, and confirmatory smallest space analysis (confirmatory SSA)

17 countries; Data collected in 2006–2007

(Continued)
### TABLE 0.1 (Continued)

<table>
<thead>
<tr>
<th>Chapter Number, Author(s), and Title</th>
<th>Topic, Statistical Method(s), and Goal of Analysis</th>
<th>Countries and Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. William van der Veld and Willem E. Saris, <em>Causes of Generalized Social Trust: An Innovative Cross-National Evaluation</em></td>
<td>Comparative analysis of the causes of generalized social trust with a correction of random and systematic measurement errors and an alternative procedure to evaluate the fit of the model using MGCFA/SEM and JRule software to detect model misspecifications taking into account the power of the test.</td>
<td>19 European countries; First round of the ESS, 2002</td>
</tr>
<tr>
<td>Section II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilevel Analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
12. Remco Feskens and Joop Hox

*Multilevel Structural Equation Modeling for Cross-cultural Research: Exploring Resampling Methods to Overcome Small Sample Size Problems*

Use of resampling methods to get accurate standard errors in multilevel analysis
1. MGCFA
2. SEM (with Mplus), a bootstrap procedure
3. MGSEM bootstrap procedure
Test of the use of bootstrap techniques for multilevel structural equation models and MGSEM

**13. Bart Meuleman and Elmar Schlüter

*Explaining Cross-National Measurement Inequivalence: A Bayesian Multilevel CFA with Random Loadings*

Developing a model that can be used to explain why specific indicators measuring ethnic citizenship conceptions function differently across cultural contexts.
1. Multilevel confirmatory factor analysis (MLCFA)
2. Bayesian estimation procedure (with Mplus)
Modeling cross-cultural variation in factor loadings and explaining their variation across groups. Presenting the results of a Monte Carlo simulation study evaluating the reliability and validity of the procedure.

Section III

Latent Class Analysis (LCA)

14. Miloš Kankaraš, Guy Moors, and Jeroen K. Vermunt

*Testing for Measurement Invariance with Latent Class Analysis*

Use of latent class analysis (LCA) for testing measurement invariance
a. Latent class cluster model
b. Latent class factor model
1. Identification of latent structures from discrete observed variables using LCA
2. Treating latent variables as nominal or ordinal
3. Estimations are performed assuming fewer distributional assumptions

15. Pascal Siegers

*A Multiple Group Latent Class Analysis of Religious Orientations in Europe*

Religious orientation in Europe
Multiple group latent class analysis (MGLCA)
Quantification of the importance of alternative spiritualities in Europe

26 European countries; First three rounds of the ESS, pooled data set, 2002–2006

31 countries; ISSP, 2003

(Continued)
<table>
<thead>
<tr>
<th>Chapter Number, Author(s), and Title</th>
<th>Topic, Statistical Method(s), and Goal of Analysis</th>
<th>Countries and Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>16. Maksim Rudnev</strong>&lt;br&gt;<strong>Testing for Invariance of Latent Classes: Group-As-Covariate Approach</strong></td>
<td>Testing for measurement invariance using LCA&lt;br&gt;1. LCA&lt;br&gt;2. Group-as-covariate approach&lt;br&gt;Describing and demonstrating one way to test for measurement invariance of latent classes. Demonstrating this approach using data on basic human values in two groups of countries – West &amp; North Europe and Eastern Europe.</td>
<td>11 West &amp; North European countries and 12 Eastern European countries; Fourth round of the ESS, 2008</td>
</tr>
<tr>
<td><strong>17. Rianne Janssen</strong>&lt;br&gt;<strong>Using a Differential Item Functioning Approach to Investigate Measurement Invariance</strong></td>
<td>Item response theory (IRT) and its application to testing for measurement invariance&lt;br&gt;IRT model used&lt;br&gt;a. strictly monotonical&lt;br&gt;b. parametric&lt;br&gt;c. dichotomous items&lt;br&gt;1. Introduction to IRT&lt;br&gt;2. Modeling of differential item functioning (DIF)&lt;br&gt;3. Application to a data-set</td>
<td>One country;</td>
</tr>
<tr>
<td><strong>18. Markus Quandt</strong>&lt;br&gt;<strong>Using the Mixed Rasch Model in the Comparative Analysis of Attitudes</strong></td>
<td>Use of a mixed polytomous Rasch model&lt;br&gt;1. Introduction to polytomous Rasch models&lt;br&gt;2. Their use for testing invariance of the national identity scale</td>
<td>Five countries; ISSP, 2003</td>
</tr>
<tr>
<td><strong>19. Jean-Paul Fox and A. Josine Verhagen</strong>&lt;br&gt;<strong>Random Item Effects Modeling for Cross-National Survey Data</strong></td>
<td>Random item effects modeling&lt;br&gt;Normal ogive item response theory (IRT) model with country specific item parameters; Multilevel item response theory (MLIRT) model&lt;br&gt;1. Properties and possibilities of random effects modeling&lt;br&gt;2. Simulation study&lt;br&gt;3. Application to the PISA data</td>
<td>40 countries; PISA-study 2003; Mathematics Data</td>
</tr>
</tbody>
</table>
Section V
New Developments

**20.** Daniel Sedig and Heinz Leitgöb

*Exact and Bayesian Approximate Measurement Invariance*

Presenting and testing for approximate measurement invariance
1. Approximate measurement invariance
2. Bayesian analysis techniques
Testing for cross-country approximate measurement invariance of attitudes toward granting citizenship rights to immigrants

six countries; ISSP, 2003

**21.** Jan Cieciuch, Eldad Davidov and Peter Schmidt

*Alignment Optimization: Estimation of the Most Trustworthy Means in Cross-Cultural Studies Even in the Presence of Noninvariance*

Presenting the alignment method for testing measurement invariance
1. Alignment optimization
2. MGCF

Outlining the main idea of alignment optimization in a nontechnical way, making some recommendations as to when and how it may be useful to apply it, providing an example of how to use it in Mplus, testing for measurement invariance of threat due to immigration

15 countries; seventh round of the ESS, 2014

**22.** Daniel L. Oberski

*Sensitivity Analysis*

Evaluating the impact of freeing cross-group equality restrictions on the parameters of substantive interest
1. MGCF using the package lavaan
2. Calculating the expected parameter change (EPC) using the package lavaan

Showing how to calculate the EPC with an empirical example on the measurement of internal political efficacy, and testing the robustness of estimations of the parameter of interest to violations of measurement invariance

15 countries; seventh round of the ESS, 2014

**23.** Michael Braun and Timothy P. Johnson

*How Should Immigrants Adapt to Their Country of Residence? A Mixed Methods Approach to Evaluate the International Applicability of a Question from the German General Social Survey (ALLBUS)*

Proposing a mixed-methods approach where qualitative approaches are employed alongside quantitative procedures to test for measurement invariance.
1. Web probing
2. Nonprobability online panels

Testing for measurement invariance of attitudes towards immigrants using a mixed-methods approach.

6 countries; Web surveys, 2011; German ALLBUS, 1980–2014

* A revised chapter
** A new chapter
van de Vijver, this volume). The etic approach is top-down and deductive, implying that general propositions or even laws exist that should be universally true. By way of contrast, the emic approach assumes that relations between constructs or even between indicators and constructs may vary across cultures. The first edition of the book suffered from some cultural universalism (etic) bias in the sense that it emphasized approaches that try to assess measurement invariance across cultures and looked for measurements that operate equally well across cultures or even universally. The assumption was that without such measurements, substantive comparisons across cultures may not be meaningful. The second edition of the book now pays more attention to approaches that can explain measurement noninvariance more systematically, thus allowing for an emic approach and for cross-cultural measurement differences, while toning down the cultural universalism bias. Several chapters in the second edition of the book demonstrate that even when full (exact) measurement invariance across cultures is not given, these cultures may be compared when certain conditions are met. This is a very important development in the methodological literature, and it allows researchers to conduct meaningful cross-cultural comparisons even in the absence of full (exact) measurement invariance.

Finally, for the second edition of the book, we contacted the authors of the existing and new chapters to ask them to update or include (if they had not done so) the data and syntax used in their chapters. This material may be found at www.routledge.com/9781138690271. By making this material available, we hope that readers will find the book more complete and consistent. In addition, we now include a subject index to help readers locate topics of interest more easily and thoroughly.

In sum, the second edition is valuable, because (as did the first edition) it presents, in a single framework, many approaches to analyze cross-cultural data in a rather nontechnical way, while illustrating these in empirical applications and providing both the data and the syntax for the interested reader. This can hopefully make the book useful both for statisticians or methodologists and applied researchers in various social science disciplines. Furthermore, and most relevant for the second edition, the book addresses the recent developments in cross-cultural research sketched above, and particularly tries to provide applied researchers with solutions on how to handle their data when invariance is not found. Thus,
the new edition of the book provides state of the art methods to analyze cross-cultural data.

The primary audience for the second edition of the book includes applied researchers within the social sciences, cross-cultural researchers, methodologists, and statisticians. More specifically, because the content of the substantive applications spans a variety of disciplines, and because the techniques may be applied to very different research questions, the book should be of interest to survey researchers, social science methodologists, cross-cultural researchers, as well as scholars and graduate and postgraduate students in the following disciplines: psychology, political science, sociology, communication, education, business administration and marketing, economics, human geography, criminology, epidemiology, and public health. Readers from more formal backgrounds such as statistics and methodology may be more interested in the purely methodological parts. Readers with less knowledge of mathematical statistics may be more interested in the applied parts. A secondary audience includes practitioners who wish to gain a better understanding of how to analyze cross-cultural data for their field of study. For example, many practitioners may want to use these techniques to analyze consumer data from different countries for marketing purposes. Clinical or health psychologists and epidemiologists may be interested in methods of how to analyze and compare cross-cultural data on, for example, addictions to alcohol or smoking or depression across various populations. The procedures presented in this volume may be useful for their work. Finally, the book is also appropriate for use as a textbook in an advanced methods course in cross-cultural analysis.

Last but not least we would like to thank all the reviewers for their work on the different chapters included in this volume and the contributors for their dedicated efforts evident in each contribution presented here. Their great cooperation enabled the production of this book. We would also like to thank Ms. Georgette Enriquez, the editor at Taylor & Francis and her editorial assistant, Mr. Brian Eschrich, as well as Denise File for her support through the process of the production of the second edition of the book. Finally, we would like to thank Lisa Trierweiler for the English proof of many of the chapters.

Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman
NOTES

1 A remarkable effort to reach this goal was the establishment of the Comparative Survey Design and Implementation (CSDI) Working Group by the late Janet Harkness. This group has been meeting annually to discuss important developments in the analysis and implementation of cross-cultural analysis (see also Harkness, Van de Vijver, & Mohler, 2003; Harkness et al., 2010).

2 We are indebted to Timothy P. Johnson, guest editor of a special issue of the journal Public Opinion Quarterly devoted to Cross Cultural Issues in Survey Methodology, for an excellent special issue on the topic and for including a very helpful review of the first edition of the book (Cho, 2015) in this special issue. We would like to thank Young Ik Cho for the excellent review.

3 We are also indebted to Bengt Muthén for fruitful discussions, suggestions, cooperation, and support in integrating both methods in the software package Mplus, as well as to Rens van der Schoot for enlightening discussions and support in applying the Bayesian approach.

4 The names of the other authors are: Mahmut Bayazıt, Maria Vakola, Luis Arciniega, Achilles Armenakis, Rasa Barkauskiene, Nikos Bozionelos, Yuka Fujimoto, Luis González, Jian Han, Martina Hřebíčková, Nerina Jimmieson, Jana Kordačová, Hitoshi Mitsuhashi, Boris Mlačić, Ivana Ferić, Marina Kotrla Topić, Sandra Ohly, Per Øystein Saksvik, Hilde Hetland, Ingvild Berg Saksvik, and Karen van Dam.

REFERENCES


Section I

MGCFM and MGSEM Techniques
Capturing Bias in Structural Equation Modeling

Fons J. R. van de Vijver

1.1 INTRODUCTION

Equivalence studies are coming of age. Thirty years ago there were few conceptual models and statistical techniques to address sources of systematic measurement error in cross-cultural studies (for early examples, see Cleary & Hilton, 1968; Lord, 1977, 1980; Poortinga, 1971). This picture has changed; in the last decades conceptual models and statistical techniques have been developed and refined. Many empirical examples have been published. There is a growing awareness of the importance of the field for the advancement of cross-cultural theorizing. An increasing number of journals require authors who submit manuscripts of cross-cultural studies to present evidence supporting the equivalence of the study measures. Yet the burgeoning of the field has not led to a convergence in conceptualizations, methods, and analyses. For example, educational testing focuses on the analysis of items as sources of problems of cross-cultural comparisons, often using item response theory (e.g., Emenogu & Childs, 2005). In personality psychology, exploratory factor analysis is commonly applied as a tool to examine similarity of factors underlying a questionnaire (e.g., McCrae, 2002). In survey research and marketing, structural equation modeling (SEM) is most frequently employed (e.g., Steenkamp & Baumgartner, 1998). From a theoretical perspective, these models are related; for example, the relationship of item response theory and confirmatory factor analysis (as derived from a general latent variable model) has been described by Brown (2015; see also Glockner-Rist &
Hojitink, 2003). However, from a practical perspective, the models can be seen as relatively independent paradigms; for a critical outsider the link between substantive field and analysis method is rather arbitrary and difficult to comprehend.

The present chapter relates the conceptual framework surrounding measurement problems that is developed in cross-cultural psychology (with input from various other sciences studying cultures and cultural differences) to statistical developments and current practices in SEM vis-à-vis multigroup testing. More specifically, I address the question of the strengths and weaknesses of SEM from a conceptual bias and equivalence framework. There are few publications in which conceptually based approaches to bias that are mainly derived from substantive studies are linked to statistically based approaches such as those developed in SEM. Thus, the current chapter adds to the literature by linking two research traditions that have worked largely independently in the past, despite the overlap in bias issues addressed in both traditions. The chapter deals with the question to what extent the study of equivalence, as implemented in SEM, can address all the relevant measurement issues of cross-cultural studies. The first part of the chapter describes a theoretical framework of bias and equivalence. The second part describes various procedures and examples to identify bias and address equivalence. The third part discusses the identification of all the bias types distinguished using SEM. The fourth part presents a SWOT analysis (strengths, weaknesses, opportunities, and threats) of SEM in dealing with bias sources in cross-cultural studies. Conclusions are drawn in the final part.

1.2 BIAS AND EQUIVALENCE

The bias framework is developed from the perspective of cross-cultural psychology and attempts to provide a comprehensive taxonomy of all systematic sources of error that can challenge the inferences drawn from cross-cultural studies (Poortinga, 1989; Van de Vijver & Leung, 1997). The equivalence framework addresses the statistical implications of the bias framework and defines conditions that have to be fulfilled before cross-cultural comparisons can be made in terms of constructs or scores.
Bias refers to the presence of nuisance factors (Poortinga, 1989). If scores are biased, the meaning of test scores varies across groups and constructs and/or scores are not directly comparable across cultures. Different types of bias can be distinguished (Van de Vijver & Leung, 1997).

1.2.1 Construct Bias

Construct bias exists if a construct differs across cultures, usually due to an incomplete overlap of construct-relevant behaviors. An empirical example can be found in Ho’s (1996) work on filial piety (defined as a psychological characteristic associated with being “a good son or daughter”). The Chinese concept, which includes the expectation that children should assume the role of caretaker of elderly parents, is broader than the Western concept.

1.2.1.2 Method Bias

Method bias is the generic term for all sources of bias due to factors often described in the methods section of empirical papers. Three types of method bias have been defined, depending on whether the bias comes from the sample, administration, or instrument. Sample bias refers to systematic differences in background characteristics of samples with a bearing on the constructs measured. Examples are differences in educational background, which can influence a host of psychological variables, such as cognitive test scores. Administration bias refers to the presence of cross-cultural conditions in testing conditions, such as ambient noise. The potential influence of interviewers and test administrators can also be mentioned here. In cognitive testing, the presence of the tester does not need to be obtrusive (Jensen, 1980). In survey research there is more evidence for interviewer effects (Lyberg et al., 1997). Deference to the interviewer has been reported; participants are more likely to display positive attitudes to an interviewer (e.g., Aquilino, 1994). Instrument bias is a final source of bias in cognitive tests that includes instrument properties with a pervasive and unintended influence on cross-cultural differences, such as the use of response alternatives in Likert scales that are not identical across groups (e.g., due to a bad translation of item anchors).
1.2.1.3 Item Bias

Item bias, or differential item functioning, refers to anomalies at the item level (Camilli & Shepard, 1994; Holland & Wainer, 1993). According to a definition that is widely used in education and psychology, an item is biased if respondents from different cultures with the same standing on the underlying construct (e.g., they are equally intelligent) do not have the same mean score on the item. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993; Sireci, 2011; Van de Vijver & Leung, 1997, 2011).

Item bias can arise in various ways, such as poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, and the influence of culture-specific nuisance factors or connotations associated with the item wording. Suppose that a geography test is administered to pupils in all EU countries that asks for the name of the capital of Belgium. Belgian pupils can be expected to show higher scores on the item than pupils from other EU countries. The item is biased because it favors one cultural group across all test score levels.

1.2.2 Equivalence

Bias has implications for the comparability of scores (e.g., Poortinga, 1989). Depending on the nature of the bias, four hierarchically nested types of equivalence (or rather inequivalence) can be defined: construct, structural or functional, metric (or measurement unit), and scalar (or full score) equivalence. These four are further described below.

1.2.2.1 Construct Inequivalence

Constructs that are inequivalent lack a shared meaning, which precludes any cross-cultural comparison. In the literature, claims of construct inequivalence can be grouped into three broad types, which differ in the degree of inequivalence (partial or total). The first and strongest claim of inequivalence is found in studies that adopt a strong emic, relativistic viewpoint, according to which psychological constructs are completely and inseparably linked to their natural context. Any cross-cultural comparison is then erroneous, as psychological constructs are cross-culturally inequivalent.
The second type is exemplified by psychological constructs that are associated with specific cultural groups. The best examples are culture-bound syndromes. A good example is Amok, which is specific to Asian countries such as Indonesia and Malaysia. Amok occurs among men and is characterized by a brief period of violent aggressive behavior. The period is often preceded by an insult and the patient shows persecutory ideas and automatic behaviors. After this period, the patient is usually exhausted and has no recollection of the event (Azhar & Varma, 2000). Violent aggressive behavior among men is universal, but the combination of triggering events, symptoms, and lack of recollection is culture specific. Such a combination of universal and culture-specific aspects is characteristic for culture-bound syndromes. Taijin Kyofusho is a Japanese example (Suzuki, Takei, Kawai, Minabe, & Mori, 2003; Tanaka-Matsumi & Draguns, 1997). This syndrome is characterized by an intense fear that one’s body is discomforting or insulting for others by its appearance, smell, or movements. The description of the symptoms suggests a strong form of a social phobia (a universal), which finds culturally unique expressions in a country in which conformity is a widely shared norm. Suzuki et al. (2003) argue that most symptoms of Taijin Kyofusho can be readily classified as social phobia, which (again) illustrates that culture-bound syndromes involve both universal and culture-specific aspects.

The third type of inequivalence is empirically based and found in comparative studies in which the data do not show any evidence for construct comparability; inequivalence here is a consequence of a lack of cross-cultural comparability. Van Leest (1997) administered a standard personality questionnaire to mainstream Dutch and Dutch immigrants. The instrument showed various problems, such as the frequent use of colloquialisms. The structure found in the Dutch mainstream group could not be replicated in the immigrant group.

1.2.2.2 Structural or Functional Equivalence

An instrument administered in different cultural groups shows structural equivalence if it measures the same construct(s) in all these groups (it should be noted that this definition is different from the common definition of structural equivalence in SEM; in a later section I return to this confusing difference in definitions). Structural equivalence has been examined for various cognitive tests (Jensen, 1980), Eysenck’s Personality Questionnaire
(Barrett, Petrides, Eysenck, & Eysenck, 1998; Bowden, Saklofske, Van de Vijver, Sudarshan, & Eysenck, 2016), and the five-factor model of personality (McCrae, 2002). Functional equivalence as a specific type of structural equivalence refers to identity of nomological networks (Cronbach & Meehl, 1955). A questionnaire that measures, say, openness to new cultures shows functional equivalence if it measures the same psychological constructs in each culture, as manifested in a similar pattern of convergent and divergent validity coefficients (i.e., nonzero correlations with presumably related measures and zero correlations with presumably unrelated measures). Tests of structural equivalence are applied more often than tests of functional equivalence. The reason is not statistical. With advances in statistical modeling (notably path analysis as part of SEM), tests of the cross-cultural similarity of nomological networks are straightforward. However, nomological networks are often based on a combination of psychological scales and background variables, such as socioeconomic status, education, and sex. The use of psychological scales to validate other psychological scales can lead to an infinite regression in which each scale in the network that is used to validate the target construct requires validation itself. If this issue has been dealt with, the statistical testing of nomological networks can be done in path analyses or in a MIMIC model (“Multiple Indicators, Multiple Causes”; Jöreskog & Goldberger, 1975; see also Kline, 2015), in which the background variables predict a latent factor that is measured by the target instrument as well as the other instruments studied to address the validity of the target instrument.

1.2.2.3 Metric or Measurement Unit Equivalence

Instruments show metric (or measurement unit) equivalence if their measurement scales have the same units of measurement, but a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement unit in each culture). Metric equivalence is found when a source of bias creates an offset in the scale in one or more groups, but does not affect the relative scores of individuals within each cultural group. For example, social desirability and stimulus familiarity influence questionnaire scores more in some cultures than in others, but they may influence individuals within a given cultural group in a fairly homogeneous way.
1.2.2.4 Scalar or Full Score Equivalence

Scalar equivalence assumes an identical interval or ratio scale in all cultural groups. If (and only if) this condition is met, direct cross-cultural comparisons be made. It is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different or equal.

1.3 BIAS AND EQUIVALENCE: ASSESSMENT AND APPLICATIONS

1.3.1 Identification Procedures

Most procedures to address bias and equivalence require only cross-cultural data with a target instrument as input; there are also procedures that rely on data obtained with additional instruments. The procedures using additional data are more open, inductive, and exploratory in nature, whereas procedures that are based only on data with the target instrument are more closed, deductive, and hypothesis testing. An answer to the question of whether additional data are needed, such as new tests or other data collection methods, such as cognitive pretesting, depends on many factors. Collecting additional data is the more laborious and time-consuming way of establishing equivalence; it is more likely to be used if fewer cross-cultural data with the target instrument are available, the cultural and linguistic distance between the cultures in the study are larger, fewer theories about the target construct are available, or the need is felt to develop a more culturally appropriate measure (possibly with culture-specific items).

1.3.1.1 Detection of Construct Bias and Construct Equivalence

The detection of construct bias and construct equivalence usually requires an exploratory approach in which local surveys, focus group discussions, or in-depth interviews with members of a community are used to establish which attitudes and behaviors are associated with a specific construct. The assessment of method bias also requires the collection of additional data, alongside the target instrument. Yet a more guided search is needed
than in the assessment of construct bias. For example, examining the presence of sample bias requires the collection of data about the composition and background of the sample, such as educational level, age, and sex. Similarly, identifying the potential influence of cross-cultural differences in response styles requires their assessment. If a bipolar instrument is used, acquiescence can be assessed by studying the levels of agreement with both the positive and negative items; however, if a unipolar instrument is used, information about acquiescence should be derived from other measures. Item bias analyses are based on closed procedures; for example, scores on items are summed and the total score is used to identify groups in different cultures with a similar performance. Item scores are then compared in groups with a similar performance from different cultures.

1.3.1.2 Detection of Structural Equivalence

The assessment of structural equivalence typically employs only quantitative procedures. Correlations, covariances, or distance measures between items or subtests are used to assess their dimensionality. Coordinates on these dimensions (e.g., factor loadings) are compared across cultures. Similarity of coordinates is used as evidence in favor of structural equivalence. The absence of structural equivalence is interpreted as evidence in favor of construct inequivalence. Structural equivalence techniques are helpful to determine the cross-cultural similarity of constructs, but they may need to be complemented by qualitative procedures, such as focus group discussions or cognitive interviews, to provide a comprehensive coverage of the definition of a construct in a cultural group. Functional equivalence, on the other hand, is based on a study of the convergent and divergent validity of an instrument measuring a target construct. The assessment of functional equivalence requires additional instruments.

1.3.1.3 Detection of Metric and Scalar Equivalence

Metric and scalar equivalence are also based exclusively on quantitative procedures; overviews of exact (conventional) procedures can be found in Kline (2015), while an overview of approximate invariance procedures can be found in Davidov, Meuleman, Cieciuch, Schmidt, and Billiet (2014), Muthén and Asparouhov (2012), and Van De Schoot, Schmidt, De
Beuckelaer, Lek, and Zondervan-Zwijnenburg (2015). SEM is often used to assess relations between items or subtests and their underlying constructs.

1.3.2 Examples and Applications

1.3.2.1 Examples of Construct Bias

An interesting study of construct bias has been reported by Patel, Abas, Broadhead, Todd, and Reeler (2001). These authors were interested in the question of how depression is expressed in Zimbabwe. In interviews with Shona speakers, they found that:

multiple somatic complaints such as headaches and fatigue are the most common presentations of depression. On inquiry, however, most patients freely admit to cognitive and emotional symptoms. Many somatic symptoms, especially those related to the heart and the head, are cultural metaphors for fear or grief. Most depressed individuals attribute their symptoms to “thinking too much” (kufungisisa), to a supernatural cause, and to social stressors. Our data confirm the view that although depression in developing countries often presents with somatic symptoms, most patients do not attribute their symptoms to a somatic illness and cannot be said to have “pure” somatisation.

(p. 482)

This conceptualization of depression is only partly overlapping with Western theories and models. As a consequence, Western instruments will have a limited suitability, particularly with regard to the etiology of the syndrome.

Few studies have been aimed at demonstrating construct inequivalence, but various studies have found that the underlying constructs were not (entirely) comparable and, hence, provide evidence for construct inequivalence. For example, De Jong, Komproe, Spinazzola, Van der Kolk, and Van Ommeren (2005) examined the cross-cultural construct equivalence of the Structured Interview for Disorders of Extreme Stress (SIDES), an instrument designed to assess symptoms of disorders of extreme stress not otherwise specified (DESNOS). The interview aims to measure the psychiatric sequelae of interpersonal victimization, notably the consequences of war, genocide, persecution, torture, and terrorism. The interview covers six clusters, each with a few items; examples are alterations in affect
regulation and impulses. Participants completed the SIDES as a part of an epidemiological survey conducted between 1997 and 1999 among large samples of survivors of war or mass violence in Algeria, Ethiopia, and Gaza. Exploratory factor analyses were conducted for each of the six clusters; the cross-cultural equivalence of the six clusters was tested in a multisample confirmatory factor analysis. The Ethiopian sample was sufficiently large to be split up into two subsamples. Equivalence across these subsamples was supported. However, comparisons of this model across countries showed a very poor fit. The authors attributed this lack of equivalence to the poor applicability of various items in these cultural contexts; they provide an interesting table in which they compare the prevalence of various symptoms in these populations with those in field trials to assess posttraumatic stress disorder that are included in the DSM-IV. The general pattern was that most symptoms were less prevalent in these three areas than reported in the manual and that there were large differences in prevalence across the three areas. Findings indicated that the factor structure of the SIDES was not stable across samples; thus construct equivalence was not shown. It is not surprising that items with such large cross-cultural differences in endorsement rates are not related to their latent constructs in a similar manner across cultures. The authors conclude that more sensitivity for the cultural context and the cultural appropriateness of the instrument would be needed to compile instruments that would be better able to stand cross-cultural validation. It is an interesting feature of the study that the authors illustrate how this could be done by proposing a multistep interdisciplinary method that accommodates universal chronic sequelae of extreme stress and accommodates culture-specific symptoms across a variety of cultures. The procedure illustrates how constructs with only a partial overlap across cultures require a more refined approach to cross-cultural comparisons, as shared and unique aspects have to be separated. It may be noted that this approach exemplifies universalism in cross-cultural psychology (Berry, Poortinga, Breugelmans, Chasiotis, & Sam, 2011), according to which the core of psychological constructs tends to be invariant across cultures but manifestations may take culture-specific forms.

As another example, it has been argued that organizational commitment contains both shared and culture-specific components. Most Western research is based on a three-componential model (e.g., Meyer & Allen, 1991; cf. Van de Vijver & Fischer, 2009) that differentiates between affective, continuance, and normative commitment. Affective commitment
is the emotional attachment to organizations, the desire to belong to the organization, and the identification with the organizational norms, values, and goals. Normative commitment refers to a feeling of obligation to remain with the organization, involving normative pressure and perceived obligations by important others. Continuance commitment refers to the costs associated with leaving the organization and the perceived need to stay. Wasti (2002) argued that continuance commitment in more collectivistic contexts, such as Turkey, loyalty and trust are important and strongly associated with paternalistic management practices. Employers are more likely to give jobs to family members and friends. Employees hired in this way will show more continuance commitment. However, Western measures do not address this aspect of continuance commitment. A meta-analysis by Fischer and Mansell (2009) found that the three components are largely independent in Western countries, but are less differentiated in lower-income contexts. These findings suggest that the three components become more independent with increasing economic affluence.

### 1.3.2.2 Examples of Method Bias

Method bias has been addressed in several studies. For instance, Fernández and Marcopulos (2008) describe how incomparability of norm samples made international comparisons of the Trail Making Test (an instrument to assess attention and cognitive flexibility) impossible: “In some cases, these differences are so dramatic that normal subjects could be classified as pathological and vice versa, depending upon the norms used” (p. 243). Sample bias (as a source of method bias) can be an important rival hypothesis to explain cross-cultural score differences in acculturation studies. Many studies compare host and immigrant samples on psychological characteristics. However, immigrant samples that are studied in Western countries often have lower levels of education and income than the host samples. As a consequence, comparisons of raw scores on psychological instruments may be confounded by sample differences. Arends-Tóth and Van de Vijver (2008) examined similarities and differences in family support in five cultural groups in the Netherlands (Dutch mainstreamers, Turkish-, Moroccan-, Surinamese-, and Antillean-Dutch). In each group, provided support was larger than received support, parents provided and received more support than siblings, and emotional support was stronger
than functional support. The cultural differences in mean scores were small for family exchange and quality of relationship, and moderate for frequency of contact. A correction for individual background characteristics (notably age and education) reduced the effect size of cross-cultural differences from .04 (proportion of variance accounted for by culture before correction) to .03 (after correction) for support and from .07 to .03 for contact. Thus, it was concluded that the cross-cultural differences in raw scores were partly unrelated to cultural background and had to be accounted for by differences in the background characteristics of the sample.

The interest in response styles is old. The first systematic study is due to Cronbach (1950). Yet the theoretical yield of the many decades of studies of response styles is rather meager. The view seems to be taken for granted that response styles are to be avoided, but there is no coherent framework that integrates response styles, nor are there well-developed cognitive models of how response styles affect response processes. Response styles have been associated with satisficing (Simon, 1956, 1979; see also Krosnick, 1991), which is a response strategy to reduce the cognitive load of responding to survey items by making shortcuts, such as choosing the midpoint of a scale. The study of response styles enjoys renewed interest in cross-cultural psychology. He and Van de Vijver (2013, 2015) have shown that extreme, acquiescent, midpoint, and socially desirable responding tend to be correlated and can be taken to be constituted by a single underlying factor, labeled the General Response Style Factor, which has shown cross-cultural stability. In a comparison of European countries, Van Herk, Poortinga, and Verhallen (2004) found that Mediterranean countries, particularly Greece, showed higher acquiescent and extreme responding than Northwestern countries in surveys on consumer research. They interpreted these differences in terms of the individualism versus collectivism dimension. In a meta-analysis across 41 countries, Fischer, Fontaine, Van de Vijver, and Van Hemert (2009) calculated acquiescence scores for various scales in the personality, social-psychological, and organizational domains. A small but significant percentage (3.1%) of the overall variance was shared among all scales, pointing to a systematic influence of response styles in cross-cultural comparisons. In a large study of response styles, Harzing (2006) found consistent cross-cultural differences in acquiescence and extremity responding across 26 countries. Cross-cultural differences in response styles are systematically related to various country characteristics. Acquiescence and
extreme responding are more prevalent in countries with higher scores on Hofstede’s collectivism and power distance, and on GLOBE’s uncertainty avoidance. Furthermore, extraversion (at country level) is a positive predictor of acquiescence and extremity scoring. Finally, she found that English-language questionnaires tend to evoke less extremity scoring and that answering items in one’s native language is associated with more extremity scoring. Cross-cultural findings on social desirability also point to the presence of systematic differences in that more affluent countries show, on average, lower scores on social desirability (Van Hemert, Van de Vijver, Poortinga, & Georgas, 2002).

Counterintuitive as it may sound, studies of the effects on corrections for response styles (including social desirability) have not unequivocally shown increments in the validity of cross-cultural differences. Some authors found changes after correction in the structure, mean levels, variance of personality measures, or the association with other variables (e.g., Danner, Aichholzer, & Rammstedt, 2015; Möttus et al., 2012; Rammstedt, Goldberg, & Borg, 2010), whereas others found negligible effects of response styles on personality measures both within and across cultures (e.g., He & Van de Vijver, 2015; Ones, Viswesvaran, & Reiss, 1996). More research is needed. Still, it is clear that a straightforward elimination of response styles may challenge the validity of cross-cultural findings and may eliminate true differences. McCrae and Costa (1983) have argued that response styles are part and parcel of someone’s personality. They demonstrated that response-style scores are positively related to agreeableness (one of the Big Five personality factors). In a recent study among South African adults, social desirability was found to be positively associated with conscientiousness, possibly because of the high desirability of traits like meticulousness in this context (Fetvadjiev, Meiring, Van de Vijver, Nel, & Hill, 2015).

Instrument bias is a common source of bias in cognitive tests. An example can be found in Piswanger’s (1975) application of the Viennese Matrices Test (Formann & Piswanger 1979). A Raven-like figural inductive reasoning test was administered to high-school students in Austria, Nigeria, and Togo (educated in Arabic). The most striking findings were the cross-cultural differences in item difficulties related to identifying and applying rules in a horizontal direction (i.e., left to right). This was interpreted as bias in terms of the different directions in writing Latin-based languages as opposed to Arabic.
Measurement mode (e.g., face-to-face interview, telephone interview, online questionnaire) can have an impact on response processes and, hence, on the validity of study results (De Leeuw, 2005; Groves & Kahn, 1979). Studies of the influence of computer administration have been conducted in psychology (Mead & Drasgow, 1993). Thus, Van de Vijver and Harsveld (1994) compared the performance of 163 applicants for the Dutch Royal Military Academy on the computerized version of the GATB (an intelligence test with many speeded subtests) to the performance of 163 applicants on the paper-and-pencil version. These two groups were matched according to age, sex, and general intelligence. A CFA invariance testing approach yielded evidence (only) for the invariance of a configural invariance model. Speeded subtests in which as many items had to be completed as possible in a fixed time were more affected by administration mode than knowledge items administered under untimed conditions. Mead and Drasgow (1993) attribute the differences in timed tests to the differential procedures to respond in paper-and-pencil and computer-assisted tests, to lack of computer experience (this factor may lose salience over time given the massive introduction of computers and various smart devices), and differences in evaluation apprehension (people may feel less inhibition to admit to undesirable behaviors when communicating with a computer than with an interviewer). In survey research, studies of mode effects have taken a slightly different direction. Gordoni, Schmidt, and Gordoni (2012) compared face-to-face and telephone modes on an attitude scale concerning social integration between Arabs and Jews in Israel. Conceptual models were used to derive hypotheses, such as more threat of disclosure and motivation in face-to-face interviews. Threat of disclosure was hypothesized to introduce an intercept difference between the modes, whereas motivation differences were expected to differentially affect measurement error. A multigroup MIMIC model was used to test the hypotheses. The disclosure effect was partially supported, while the motivation effect was fully supported. The study illustrates how mode effects can be addressed in a SEM framework. Finally, Jäckle, Roberts, and Lynn (2010) describe issues in designing mixed-modes international surveys (they specifically refer to the European Social Survey). They argue that most reported studies are inadequate because of incomplete comparability of samples to which the different modes were administered. In addition, they conducted a field study comparing face-to-face interviewing using showcards and
telephone interviewing in Hungary and Portugal. Their main conclusion was that mode effects are highly specific. These mode effects were not global (and hence could not be modeled using a single mode parameter), but involved specific anchors at specific questions, hence their use of partial proportional odds models. For example, telephone respondents were more likely to strongly agree that men should share responsibilities for their home and family and that the law should be obeyed whatever the circumstances. The authors conclude that, on the one hand, a theoretical framework is lacking to describe mode effects, and that, on the other hand, we do not yet have enough evidence to enable a conclusion when mode effects do or do not matter although the statistical models to analyze relevant data (proportional odds models) are available.

1.3.2.3 Examples of Item Bias

More studies of item bias have been published than of any other form of bias. All widely used statistical techniques have been used to identify item bias. Item bias is often viewed as an undesirable item characteristic that should be eliminated. As a consequence, items that are presumably biased are eliminated prior to the cross-cultural comparisons of scores. However, it is also possible to view item bias as a source of cross-cultural differences that is not to be eliminated but requires further examination (Poortinga & Van der Flier, 1988). The background of this view is that item bias, which by definition involves systematic cross-cultural differences, can be interpreted as referring to culture-specificities. Biased items provide information about cross-cultural differences on constructs other than the target construct. For example, in a study on intended self-presentation strategies by students in job interviews involving 10 countries, it was found that dress code yielded biased items (Sandal et al., 2014). Dress code was an important aspect of self-presentation in more traditional countries (such as Iran and Ghana) whereas informal dress was more common in more modern countries (such as Germany and Norway). These items provide important information about self-presentation in these countries, which cannot be dismissed as bias that should be eliminated. More generally, from the perspective of sciences that have culture as their focus of study, such as ethnography and cross-cultural psychology, it is difficult to understand the focus on finding invariance, where their disciplines target both differences and similarities (Berry et al., 2011).
The almost 50 years of item bias research after Cleary and Hilton’s (1968) first study have not led to accumulated insights as to which items tend to be biased. In fact, one of the complaints has been the lack of accumulation. Educational testing has been an important domain of application of item bias. Linn (1993), in a review of the findings, came to the sobering conclusion that no general findings have emerged about which item characteristics are associated with item bias; he argued that item difficulty was the only characteristic that was more or less associated with bias. More recently, Walzebug (2014) used a combination of quantitative and qualitative (interviews) procedures to identify bias in mathematics items administered to German fourth-graders of low and high socioeconomic strata. Her sociolinguistic analysis of biased items and interviews suggested that item bias was not related to item difficulty but mainly a consequence of the language register used in the items: “the language used at school contains specific speech variants that differ from the language experiences of low SES children” (p. 159). By providing a substantive explanation of bias (although it would be better called “method bias” than item bias), these findings are promising yet await replication. Roth, Oliveri, Sandilands, Lyons-Thomas, and Ercikan (2013) asked three experts to evaluate items of the French and English version of a science test, using think-aloud protocols. Previous statistical analyses had shown that half of the items were biased. There was some agreement among the (independently working) experts, and there was some agreement between the qualitative and quantitative findings, but the agreement was far from perfect. This study is an example of a rather general finding: The agreement of qualitative and quantitative procedures is often better than chance but far from impressive, highlighting the elusive nature of item bias. The item bias tradition has not led to widely accepted practices about item writing for multicultural assessment. One of the problems in accumulating knowledge from the item bias tradition about item writing may be the often specific nature of the bias. Van Schilt-Mol (2007) identified item bias in educational tests (Cito tests) in Dutch primary schools, using psychometric procedures. She then attempted to identify the source of the item bias, using a content analysis of the items and interviews with teachers and immigrant pupils. Based on this analysis, she changed the original items and administered the new version. The modified items showed little or no bias, indicating that she successfully identified and removed the bias source. Her study illustrates an effective,
though laborious, way to deal with bias. The source of the bias was often item specific (such as words or pictures that were not equally known in all cultural groups), and no general conclusions about how to avoid items could be drawn from her study.

Item bias has also been studied in personality and attitude measures. There are numerous examples in which many or even a majority of the items turned out to be biased. Church et al. (2011) administered the widely used Revised NEO Personality Inventory to college students in the United States, Philippines, and Mexico. Using confirmatory factor analysis, the authors found that about 40% to 50% of the items exhibited some form of item bias. If so many items are biased, serious validity issues have to be addressed, such as potential construct bias and adequate construct coverage in the remaining items.

A few studies have examined the nature of item bias in personality questionnaires. Sheppard, Han, Colarelli, Dai, and King (2006) examined bias in the Hogan Personality Inventory in Caucasians and African Americans who had applied for unskilled factory jobs. Although the group mean differences were trivial, more than a third of the items showed item bias. Items related to cautiousness tended to be potentially biased in favor of African Americans. Ryan, Horvath, Ployhart, Schmitt, and Slade (2000) were interested in determining sources of item bias in global employee opinion surveys. Analyzing data from a 36-country study involving more than 50,000 employees, they related item bias statistics (derived from item response theory) to country characteristics. Hypotheses about specific item contents and Hofstede’s (2001) dimensions were only partly confirmed; the authors found that more dissimilar countries showed more item bias. The positive relation between the size of global cultural differences and item bias may well generalize to other studies. Sandal et al. (2014) also found more bias between countries that are culturally further apart. If this conclusion would hold across other studies, it would imply that a larger cultural distance between countries can be expected to be associated with both more valid cross-cultural differences and more item bias. Bingenheimer, Raudenbush, Leventhal, and Brooks-Gunn (2005) studied bias in the Environmental Organization and Caregiver Warmth scales that were adapted from several versions of the HOME Inventory (Bradley, 1994; Bradley, Caldwell, Rock, Hamrick, & Harris, 1988). The scales are measures of parenting climate. Participants were around 4,000 Latino, African American, and European American parents living in
Chicago. Procedures based on item response theory were used to identify bias. Biased items were not thematically clustered.

Although I do not know of any systematic comparison, the picture that emerges from the literature on item bias is one of great variability in numbers of biased items across instruments and limited insight into what contributes to the bias.

### 1.3.2.4 Examples of Studies of Multiple Sources of Bias

Some studies have addressed multiple sources of bias. Hofer, Chasiotis, Friedlmeier, Busch, and Campos (2005), for instance, studied various forms of bias in a thematic apperception test, which is an implicit measure of power and affiliation motives. The instrument was administered in Cameroon, Costa Rica, and Germany. Construct bias in the coding of responses was addressed in discussions with local informants; the discussions pointed to the equivalence of coding rules. Method bias was addressed by examining the relation between test scores and background variables such as age and education. No strong evidence was found. Finally, using loglinear models, some items were found to be biased. As another example, Meiring, Van de Vijver, Rothmann, and Barrick (2005) studied construct, item, and method bias of cognitive and personality tests in a sample of 13,681 participants who had applied for entry-level police jobs in the South African Police Services. The sample consisted of Whites, Indians, Coloreds, and nine Black groups. The cognitive instruments produced very good construct equivalence, as often found in the literature (e.g., Berry et al., 2011; Van de Vijver, 1997); moreover, logistic regression procedures identified almost no item bias (given the huge sample size, effect size measures instead of statistical significance were used as criteria for deciding whether items were biased). The personality instrument (i.e., the 16 PFI Questionnaire, which is an imported and widely used instrument in job selection in South Africa) showed more structural equivalence problems. Several scales of the personality questionnaire revealed construct bias in various ethnic groups. Using analysis of variance procedures, very little item bias in the personality scales was observed. Method bias did not have any impact on the (small) size of the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably
in the Black groups. It was concluded that the cognitive tests were suitable as instruments for multicultural assessment, whereas bias and low internal consistencies limited the usefulness of the personality scales.

1.4 IDENTIFICATION OF BIAS IN STRUCTURAL EQUATION MODELING

There is a fair amount of convergence on how equivalence should be addressed in structural equation models. I mention here the often quoted classification by Vandenberg (2002; Vandenberg & Lance, 2000) that, if fully applied, has eight steps:

1. A global test of the equality of covariance matrices across groups;
2. A test of configural invariance (also labeled weak factorial invariance) in which the presence of the same pattern of fixed and free factor loadings is tested for each group;
3. A test of metric invariance (also labeled strong factorial invariance) in which factor loadings for identical items are tested to be invariant across groups;
4. A test of scalar invariance (also labeled strict invariance) in which intercepts for identical items are tested to be invariant across groups;
5. A test of invariance of unique variances across groups;
6. A test of invariance of factor variances across groups;
7. A test of invariance of factor covariances across groups;
8. A test of the null hypothesis of invariant factor means across groups.

The latter is a test of cross-cultural differences in unobserved means.

The first test (the local test of invariance of covariance matrices) is infrequently used, presumably because researchers are typically more interested in modeling covariances than merely testing their cross-cultural invariance, and the observation that covariance matrices are not identical may not be informative about the nature of the difference. The most frequently reported invariance tests involve configural, metric, and scalar invariance (steps 2 through 4). The latter three types of invariance address relations between observed and latent variables. As these involve
the measurement aspects of the model, they are also referred to as measurement invariance (or measurement equivalence). The last four types of invariance (steps 5 through 8) address characteristics of latent variables and their relations; therefore, they are referred to as structural invariance (or structural equivalence).

As indicated earlier, there is a confusing difference in the meaning of the term “structural equivalence,” as employed in the cross-cultural psychology tradition, and “structural equivalence“ (or structural invariance), as employed in the SEM tradition. Structural equivalence in the cross-cultural psychology tradition addresses the question of whether an instrument measures the same underlying construct(s) in different cultural groups and is usually examined in exploratory factor analyses. Identity of factors is taken as evidence in favor of structural equivalence which then means that the structure of the underlying construct(s) is identical across groups. Structural equivalence in the structural equation tradition refers to identical variances and covariances of structural variables (latent factors) of the model. So, whereas structural equivalence addresses links between observed and latent variables, structural invariance does not involve observed variables at all. Structural equivalence in the cross-cultural psychology tradition is much closer to what in the SEM tradition is between configural invariance and metric invariance (measurement equivalence) than to structural equivalence.

I now describe procedures that have been proposed in the structural equation modeling tradition to identify the three types of bias (construct, method, and item bias) as well as illustrations of the procedures; an overview of the procedures (and their problems) can be found in Table 1.1.

1.4.1 Construct Bias

1.4.1.1 Procedure

The structural equivalence tradition has started from the question of how invariance of any parameter of a structural equation model can be tested. The aim of the procedures is to establish such invariance in a statistically rigorous manner. The focus of the efforts has been on the comparability of previously tested data. The framework does not specify or prescribe how instruments have to be compiled to be suitable for cross-cultural comparisons; rather, the approach tests corollaries of the assumption that
TABLE 1.1
Overview of the Types of Bias and the Structural Equation Modeling (SEM) Procedures for Their Identification

<table>
<thead>
<tr>
<th>Type of Bias</th>
<th>Definition</th>
<th>SEM Procedure for Identification</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct</td>
<td>A construct differs across cultures, usually due to an incomplete overlap of construct-relevant behaviors.</td>
<td>Multigroup confirmatory factor analysis, testing configural invariance (identity of patterning of loadings and factors).</td>
<td>Cognitive interviews and ethnographic information may be needed to determine whether the construct is adequately captured.</td>
</tr>
<tr>
<td>Method</td>
<td>Generic term for all sources of bias due to factors often described in the methods section of empirical papers. Three types of method bias have been defined, depending on whether the bias comes from the sample, the administration, or the instrument.</td>
<td>Confirmatory factor analysis or path analysis of models that evaluate the influence of method factors (e.g., by testing method factors).</td>
<td>Many studies do not collect data about method factors, which makes the testing of method factors impossible.</td>
</tr>
<tr>
<td>Item</td>
<td>Anomalies at the item level; an item is biased if respondents from different cultures with the same standing on the underlying construct (e.g., they are equally intelligent) do not have the same mean score on the item.</td>
<td>Multigroup confirmatory factor analysis, testing scalar invariance (testing identity of intercepts when identical items are regressed on the latent variables; assumes support for configural and metric equivalence).</td>
<td>Model of scalar equivalence, prerequisite for a test of item bias, may not be supported. Reasons for item bias may be unclear.</td>
</tr>
</tbody>
</table>

the instrument is adequate for comparative purposes. The procedure for addressing this question usually follows the steps described before, with an emphasis on the establishment of configural, metric, and scalar invariance (weak, strong, and strict invariance).
1.4.1.2 Examples

Caprara, Barbaranelli, Bermúdez, Maslach, and Ruch (2000) tested the cross-cultural generalizability of the Big Five Questionnaire (BFQ), which is a measure of the five-factor model, in large samples from Italy, Germany, Spain, and the United States. The authors used exploratory factor analysis, simultaneous component analysis (Kiers, 1990), and confirmatory factor analysis. The Italian, American, German, and Spanish versions of the BFQ showed factor structures that were comparable: “Because the pattern of relationships among the BFQ facet-scales is basically the same in the four different countries, different data analysis strategies converge in pointing to a substantial equivalence among the constructs that these scales are measuring” (p. 457). These findings support the universality of the five-factor model. At a more detailed level the analysis methods did not yield completely identical results. The confirmatory factor analysis picked up more sources of cross-cultural differences. The authors attribute the discrepancies to the larger sensitivity of confirmatory models.

Another example comes from the values domain. Like the previous study, it addresses relations between the (lack of) structural equivalence and country indicators. Another interesting aspect of the study is the use of multidimensional scaling where most studies use factor analysis. Fontaine, Poortinga, Delbeke, and Schwartz (2008) assessed the structural equivalence of the values domain, based on the Schwartz value theory, in a dataset from 38 countries, each represented by a student and a teacher sample. The authors found that the theoretically expected structure provided an excellent representation of the average value structure across samples, although sampling fluctuation causes smaller and larger deviations from this average structure. Furthermore, sampling fluctuation could not account for all these deviations. The closer inspection of the deviations showed that higher levels of societal development of a country were associated with a larger contrast between protection and growth values. Studies of structural equivalence in large-scale datasets open a new window on cross-cultural differences. There are no models of the emergence of constructs that accompany changes in a country, such as increases in the level of affluence. The study of covariation between social developments and salience of psychological constructs is largely uncharted domain.

A third example from the values domain is provided by Cieciuch, Davidov, Vecchione, Beierlein, and Schwartz (2014). They tested the
invariance of a new instrument to measure human values, based on Schwartz’s theory of human values. A previous version of the instrument, the Portrait Values Questionnaire, has been applied in about 50 countries, but has never shown high levels of invariance in this heterogeneous set, although metric invariance was found for a version that was administered in a culturally more homogeneous set of countries involved in the European Social Survey (Davidov, Schmidt, & Schwartz, 2008). The new scale is based on a slightly modified conceptual structure and measures 19 values with three items each. Convenience samples of adults were drawn in Finland, Germany, Israel, Italy, New Zealand, Poland, Portugal, and Switzerland. Invariance was tested per value. Most values showed metric invariance in most countries, whereas partial and full scalar invariance was supported for half of the values. These results compare favorably to findings obtained with the original instrument.

Arends-Tóth and Van de Vijver (2008) studied associations between well-being and family relationships among five cultural groups in the Netherlands (Dutch mainstreamers, and Turkish, Moroccan, Surinamese, and Antillean immigrants). Two aspects of relationships were studied: family values, which refer to obligations and beliefs about family relationships, and family ties, which involve more behavior-related relational aspects. A SEM model was tested in which the two aspects of relationships predicted a latent factor, called well-being, that was measured by loneliness and general and mental health. Multisample models showed invariance of the regression weights of the two predictors and of the factor loadings of loneliness and health. Other model components showed some cross-cultural variation (correlations between the errors of the latent and outcome variables).

Van de Vijver (2002) examined the comparability of scores on tests of inductive reasoning in samples of 704 Zambian, 877 Turkish, and 632 Dutch pupils from the highest two grades of primary and the lowest two grades of secondary school. In addition to two tests of inductive reasoning (employing figure and nonsense words as stimuli, respectively), three tests were administered that assessed cognitive components assumed to be important in inductive thinking (i.e., classification, rule generation, and rule testing). SEM was used to test the fit of a MIMIC model in which the three component tests predicted a latent factor, labeled inductive reasoning, that was measured by the two tests mentioned. Configural invariance was supported, metric equivalence invariance was partially supported,
and tests of scalar equivalence showed a poor fit. It was concluded that comparability of test scores across these groups was problematic and that cross-cultural score differences were probably influenced by auxiliary constructs such as test exposure. Finally, Davidov (2008) examined invariance of a 21-item instrument measuring human values of the European Social Survey that was administered in 25 countries. Multigroup confirmatory factor analysis did not support configural and metric invariance across these countries. Metric equivalence was only established after a reduction of the number of countries to 14 and of the original 10 latent factors to 7.

1.4.2 Method Bias

1.4.2.1 Procedure

The study of method bias in SEM is straightforward. Indicators of the source of method bias, which are typically viewed as confounding variables, can be introduced in a path model, thus enabling the statistical evaluation of their impact. Examples of studies of response styles are given below, but other examples can be easily envisaged, such as including years of schooling, socioeconomic status indicators, or interviewer characteristics. The problem with the study of method bias is usually not the statistical evaluation but the availability of pertinent data. For example, social desirability is often mentioned as a source of cross-cultural score differences but infrequently measured; only when such data are available can an evaluation of its impact be carried out.

1.4.2.2 Examples

Various authors have addressed the evaluation of response sets, notably acquiescence and extremity scoring (e.g., Cheung & Rensvold, 2000; Mirowsky & Ross, 1991; Watson, 1992); yet there are relatively few systematic SEM studies of method bias compared to the numerous studies on other types of bias. Billiet and McClendon (2000) worked with a balanced set of Likert items that measured ethnic threat and distrust in politics in a sample of Flemish respondents. The authors found a good fit for a model with three latent factors: two content factors (ethnic threat and distrust in politics that are negatively correlated) with positive and negative slopes
according to the wording of the items, and one uncorrelated common style factor with all positive loadings. The style factor was identified as acquiescence, given that its correlation with the sum of agreements was very high. Welkenhuysen-Gybels, Billiet, and Cambré (2003) applied a similar approach in a cross-cultural study.

1.4.3 Item Bias

1.4.3.1 Procedure

Item bias in SEM is closely associated with the test of scalar invariance. It is tested by examining invariance of intercepts when an item is regressed on its latent factor (fourth step in Vandenbergs’s procedure). The procedure is different from those described in the differential item functioning tradition (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993). Although it is impossible to capture the literally hundreds of item bias detection procedures that have been proposed, some basic ideas prevail. The most important is the relevance of comparing item statistics per score level. The latter are usually defined by splitting up a sample in subsamples of respondents with similar scores (such as splitting up the sample in low, medium, and high scorers). Corollaries of the assumption that equal sum scores on the (unidimensional) instrument reflect an equal standing on the latent trait are then tested. For example, the Mantel–Haenszel procedure tests whether the mean scores of persons with the same sum scores are identical across cultures (as they should be for an unbiased item). The SEM procedure tests whether the (linear) relation between observed and latent variable is identical across cultures (equal slopes and intercepts). From a theoretical point of view, the Mantel–Haenszel and SEM procedures are very different; for example, the Mantel–Haenszel procedure is based on a nonlinear relation between item score and latent trait, whereas SEM employs a linear model. Also, both employ different ways to get access to the latent trait (through covariances in SEM and slicing up data in score levels in the Mantel–Haenszel procedure. Yet, from a practical point of view, the two procedures will often yield convergent results. It has been shown that using the Mantel–Haenszel is conceptually identical to assuming a Rasch model to apply to the scale and testing identity of item parameters across groups (Fischer, 1993). The nonlinear (though strictly monotonous) relation between item and latent construct score that
is assumed in the Rasch model will often not differ much from the linear relation assumed by SEM. Convergence of results is therefore not surprising, in particular when an item shows a strong bias.

It is an attractive feature of SEM that biased items do not need to be eliminated from the instrument prior to the cross-cultural comparison (as often done in analyses based on other statistical models). Biased items can be retained as culture-specific indicators. Partial measurement invariance allows for including both shared and nonshared items in cross-cultural comparisons. Scholderer, Grunert, and Brunsø (2005) describe a procedure for identifying intercept differences and correcting for these differences in the estimation of latent means; De Beuckelaer and Swinnen (2011) conducted a Monte Carlo study to investigate the impact of incomplete invariance on latent means estimation.

1.4.3.2 Examples

Two types of procedures can be found in the literature that address item bias. In the first and most common type, item bias is part of a larger exercise to study equivalence and is tested after configural and metric equivalence have been established. The second kind of application adds information from background characteristics to determine to what extent these characteristics can help to identify bias.

De Beuckelaer, Lievens, and Swinnen (2007) provide an example of the first type of application. They tested the measurement equivalence of a global organizational survey that measures six work climate factors in 24 countries from Western Europe, Eastern Europe, North America, the Americas, the Middle East, Africa, and the Asia-Pacific region; the sample comprised 31,315 employees and survey consultants. The survey instrument showed configural and metric equivalence of the six-factor structure, but scalar equivalence was not supported. Many intercept differences of items were found; the authors argued that this absence was possibly a consequence of response styles. They split up the countries into regions with similar countries or with the same language. Within these more narrowly defined regions (e.g., Australia, Canada, the United Kingdom, and the United States as the English-speaking region), scalar equivalence was found. A study by Prelow, Michaels, Reyes, Knight, and Barrera (2002) provides a second example. These authors tested the equivalence of the Children’s Coping Strategies Checklist in a sample of
European American, African American, and Mexican American adolescents from low-income inner-city families. The coping questionnaire consisted of two major styles, active coping and avoidant coping, each of which comprised different subscales. Equivalence was tested per subscale. Metric equivalence was strongly supported for all subscales of the coping questionnaire; yet intercept invariance was found in few cases. Most of the salient differences in intercept were found between the African American and Mexican American groups.

An example of the second type of item bias study has been described by Grayson, Mackinnon, Jorm, Creasey, and Broe (2000). These authors were interested in the question of whether physical disorders influence scores on the Center for Epidemiologic Studies Depression Scale (CES-D) among the elderly, thereby leading to false positives in assessment procedures. The authors recruited a sample of 506 participants aged 75 or older living in their community in Sydney, Australia. The fit of a MIMIC model was tested. The latent factor, labeled depression, was measured by the CES-D items; item bias was defined as the presence of significant direct effects of background characteristics on items (so, no cultural variation was involved). Various physical disorders (such as mobility disability and peripheral vascular disease) had a direct impact on particular item scores in addition to the indirect path through depression. The authors concluded that the CES-D score is “polluted with contributions unrelated to depression” (p. 279). The second example is provided by Jones (2003), who assessed cognitive functioning among African American and European American older adults (> 50 years) in Florida in a telephone interview. He also used a MIMIC model. Much item bias was found (operationalized here as differences in both measurement weights and intercepts of item parcels on a general underlying cognition factor). Moreover, the bias systematically favored the European American group. After correction for this bias, the size of the cross-cultural differences in scores was reduced by 60%. Moreover, various background characteristics had direct effects on item parcels, which was interpreted as evidence for item bias.

The two types of applications provide an important difference in perspective on item bias. The first approach only leads to straightforward findings if the null hypothesis of scalar equivalence is confirmed; if, as is often the case, no unambiguous support for scalar equivalence is found, it is often difficult to find reasons that are methodologically compelling for the lack of scalar equivalence. Therefore, the conclusion can then be
drawn that scalar equivalence is not supported and a close inspection of the deviant parameters will indicate which items are responsible for the poor fit. However, such an observation usually does not suggest a substantive reason for the poor fit. The second approach starts from a more focused search for a specific antecedent of item bias. As a consequence, the results of these studies are easier to interpret. This observation is in line with a common finding in item bias studies of educational and cognitive tests (e.g., Holland & Wainer, 1993): Without specific hypotheses about the sources of item bias, a content analysis of which items are biased and unbiased hardly ever leads to interpretable results as to the reasons for the bias. The literature on equivalence testing is still scattered and is not yet ready for a full-fledged meta-analysis of the links between characteristics of instruments, samples, and their cultures on the one hand, and levels of equivalence on the other hand; yet, it is already quite clear that studies of scalar equivalence often do not support the direct comparison of scores across countries. Findings based on SEM and findings based on other item bias techniques point in the same direction: Item bias is more pervasive than we may conveniently think and, when adequately tested, scalar equivalence is often not supported. The widespread usage of analyses of (co)variance, t tests, and other techniques that assume full score equivalence is not based on adequate invariance testing. The main reason for not bothering about scalar invariance prior to comparing means across cultures is opportunistic: Various studies have compared the size of cross-cultural differences before and after correction for item bias, and most of these found that item bias does not tend to favor a single group and, hence, that correction for item bias usually does not affect the size of cross-cultural differences (Van de Vijver, 2011, 2015). An alternative and better founded approach would be to rely on robustness studies such as those described by Oberski in this volume (see chapter 22).

1.5 STATISTICAL MODELING IN SEM AND BIAS: A SWOT ANALYSIS

After the description of a framework for bias and equivalence and a description of various examples in which the framework was employed, the stage is set for an evaluation of the contribution of SEM to the study of
bias and equivalence. The evaluation takes the form of a SWOT analysis (strengths, weaknesses, opportunities, and threats).

The main strength of SEM is the systematic manner in which invariance can be tested. There is no other statistical theory that allows for such a fine-grained, flexible, and integrated analysis of equivalence. No other older approach combines these characteristics; for example, a combination of exploratory factor analysis and item bias analysis using regression analysis could be used for examining the configural and scalar equivalence, respectively. However, the two kinds of procedures are conceptually unrelated. As a consequence, partial invariance is difficult to incorporate in such analyses. Furthermore, SEM has been instrumental in putting equivalence testing on the agenda of cross-cultural researchers and in stimulating the interest in cross-cultural studies.

The first weakness of equivalence testing using SEM is related to the large discrepancy between the advanced level of statistical theorizing behind the framework and the far from advanced level of available theories about cross-cultural similarities and differences. The level of sophistication of our conceptual models of cross-cultural differences is nowhere near the statistical sophistication available to test these differences. As a consequence, it is difficult to strike a balance between conceptual and statistical considerations in equivalence testing. The literature shows that it is tempting to use multigroup factor analysis in a mechanical manner by relying entirely on statistical, usually significance, criteria to draw conclusions about levels of equivalence. An equivalence test using SEM can easily become synonymous to a demonstration that scores can be compared in a bias-free manner. In my view, there are two kinds of problems with these mechanical applications of equivalence tests. First, there are statistical problems with the interpretation of fit tests. Particularly in large-scale cross-cultural studies, the lack of convergence of information provided by the common fit statistics, combined with the absence of adequate Monte Carlo studies and experience with fit statistics in similar cases, can create problems in choosing the most adequate model. In these studies it is difficult to tease apart fit problems due to conceptually trivial sample particulars that do not challenge the interpretation of the model as being equivalent and fit problems due to misspecifications of the model that are conceptually consequential. Second, equivalence testing in SEM can easily become a tool that, possibly inadvertently, uses statistical sophistication to compensate for problems with the adequacy of instruments or samples.
Thus, studies using convenience samples have problems of external validity, whatever the statistical sophistication used to deal with the data. Also, it is relatively common in cross-cultural survey research to employ short instruments. Such instruments may yield a poor rendering of the underlying construct and may capitalize on item specifics, particularly in a cross-cultural framework.

In addition to statistical problems, there is another and probably more salient problem of equivalence testing in a SEM framework: Sources of bias can be easily overlooked in standard equivalence tests based on confirmatory factor analysis, thereby reaching overly liberal conclusions about equivalence. Thus, construct inequivalence cannot be identified in deductive equivalence testing (i.e., testing in which only data from a target instrument are available, as is the case in confirmatory factor analysis). There is a tendency in the literature to apply closely translated questionnaires without adequately considering adaptation issues (Hambleton, Merenda, & Spielberger, 2005). Without extensive pretesting, the use of interviews to determine the accuracy of items, or the inclusion of additional instruments to check the validity of a target instrument, it is impossible to determine whether closely translated items are the best possible items in a specific culture. Culture-specific indicators of common constructs may have been missed. The focus on using identical instruments in many cultures may lead to finding superficial similarities between cultures, because the instrument compilation may have driven the study to an emphasis on similarities. The various sources of bias (construct, method, and items) cannot be investigated adequately if only data from the target instrument are available. Various sources of bias can be studied in SEM, but most applications start from a narrow definition of bias that capitalizes on confirmatory factor analysis without considering or having additional data to address bias. It should be noted that the problem of not considering all bias sources in cross-cultural studies is not an intrinsic characteristic of SEM (in line with my argument in earlier parts of the chapter), but a regrettable, self-imposed limitation in its use.

A first opportunity of equivalence testing using SEM is the pursuit of a closer link between design and analysis in large-scale assessment. More specifically, we can build on the frequently observed lack of scalar invariance in large cross-cultural studies to improve study design. A good example can be found in the attempts to deal with cross-cultural differences in response styles. It is a recurrent finding in PISA studies that within
each participating country there is a small, positive correlation between motivation (e.g., interest in math) and educational achievement in that domain. However, when data are aggregated at country level (so that each country constitutes one case), the correlation is much stronger and negative. This achievement–motivation paradox (He & Van de Vijver, 2016) is probably due to response styles. Whereas countries with a high performance (such as East Asian countries) tend to score low on motivation (modesty bias), some countries with a low performance tend to have high motivation scores (such as Central and South America, presumably due to extreme responding). Various procedures have been proposed and successfully tested. The simplest is called forced choice. Each item presents two or more alternatives; the total number of choices is identical for all respondents but individual differences are derived from preferences of certain types of choices. For example, in a forced-choice personality scale, each item describes two traits and respondents choose one of the traits that is more like them. Respondents would get a higher score on emotional stability if they indicate to prefer indicators of stability compared to other personality traits. In a PISA Field Trial, these forced-choice scales (which can be analyzed by IRT models; Brown, 2016) were able to resolve the paradox mentioned, although the forced-choice scales did not show scalar invariance (Kyllonen & Bertling, 2014).

A second opportunity is approximate invariance testing (see Seddig and Leitgöb’s chapter in this volume). This work goes back to the end of the 1980s when it was proposed that it may be useful to release some factor loadings or intercepts in invariance testing (Byrne, Shavelson, & Muthén, 1989); the procedure has become known as partial invariance testing. Asparouhov and Muthén (2009) proposed another relaxation in their exploratory structural equation modeling, in which items (or sub-tests) were allowed to have secondary loadings on nontarget factors. The procedure was found to yield cross-cultural invariance of factor loadings of the Eysenck Personality Questionnaire across 33 countries where confirmatory factor analysis of the same data did not show this type of invariance (Bowden et al., 2016). Several error components needed to be correlated, mainly to accommodate item polarity (i.e., positive or negative wording of the item vis-à-vis the target construct). More recently, two new procedures have been introduced that offer important scope for testing approximate invariance: Bayesian structural equation modeling (Muthén & Asparouhov, 2012; see also Seddig and Leitgöb’s chapter in
this volume) and the alignment method (Asparouhov & Muthén, 2014; see also Cieciuch et al.’s chapter in this volume). As an aside, it should be noted that these approximate invariance testing procedures make it easy to capitalize on specific item problems to improve fit (in much the same way as correlated errors are often used for improving fit without having a good, substantive reason for allowing the correlation). These procedures will be very useful when judiciously used, but may run the risk of producing non-replicable findings when used entirely to improve fit without any consideration of the substantive implications of the released invariance constraints.

A third opportunity is the further investigation of fit in large-scale studies. Confirmatory factor analysis has become the standard procedure for assessing invariance in large-scale studies such as the OECD’s Programme for International Student Assessment (PISA; www.oecd.org/pisa/), the Teaching and Learning International Survey (TALIS; www.oecd.org/edu/school/talis.htm), and the IEA’s Trends in International Mathematics and Science Study (TIMSS; www.iea.nl/timss_2015.html). However, ample experience with large datasets, in some cases involving more than 50 countries and 100,000 participants, has shown that fit statistics are difficult to evaluate in these studies. It is exceptional to find scalar invariance for any scale in these studies when evaluated by common fit criteria; the fit criteria may be too strict (Davidov et al., 2014). Both empirical and Monte Carlo studies are needed to gain more insight into the adequacy of fit criteria for large-scale assessment (Rutkowski & Svetina, 2014). The null hypothesis of invariance of all factor loadings and intercepts across all countries is not realistic and extremely unlikely to hold when many countries are involved. For example, the hypothesis implies that all items are equally relevant in all countries, that response scales are used in the same manner across countries, that there are no translation issues in any country, etc. Guidelines need to be developed that strike a balance between Type I and Type II errors: Is lack of fit (differences in loadings and intercepts) sufficiently large to make any practical difference? There is much literature in the clinical field that models the link between the (complement of the) two errors in receiver operating characteristic curves (ROC curve; Hsieh & Turnbull, 1996), which can be used here.

The main threat is that SEM procedures remain within the purview of SEM researchers. Usage of the procedures has not (yet?) become popular among substantive researchers. There is a danger that SEM researchers will keep on “preaching the gospel to the choir” by providing solutions to
increasingly complex technical issues without linking these to the questions of substantive researchers and determining how SEM can help to solve substantive problems and advance our theorizing.

1.6 CONCLUSIONS

Statistical procedures in the behavioral and social sciences are tools to improve research quality. This also holds for the role of SEM procedures in the study of equivalence and bias. In order to achieve a high quality, a combination of various types of expertise are needed in cross-cultural studies. SEM procedures can greatly contribute to the quality of cross-cultural studies, but more interaction between substantive and method researchers is needed to realize this potential. It is not a foregone conclusion that the potential of SEM procedures will materialize and that the threats of these procedures will not materialize. We need to appreciate that large-scale cross-cultural studies require many different types of expertise (Byrne & Van de Vijver, 2010, 2014); it is unrealistic to assume that there are many researchers who have all the expertise required to conduct such studies. Substantive experts are needed with knowledge of the target construct, next to cultural experts with knowledge about the construct in the target context, next to measurement experts who can convert substantive knowledge in adequate measurement procedures, next to statistical experts who can test bias and equivalence in a study. The strength of a chain is defined by the strength of the weakest link; this also holds for the quality of cross-cultural studies. SEM has great potential for cross-cultural studies, but it will be able to achieve this potential only in close interaction with the expertise from various other domains.

SUMMARY

This chapter discusses various forms of bias—construct, method, and item bias—that threaten the equivalence of cross-cultural measurements. It is shown how the various forms of bias can be identified using structural equation modeling (SEM).
REFERENCES


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. Organizational Research Methods, 5, 139–158. doi:10.1177/1094428102005002001


SUGGESTIONS FOR FURTHER READING


metric invariance is recommended because using ANOVA and regression methods on observed variables cannot tell us whether the measures are truly comparable across groups or time and we cannot, therefore, be confident that our comparative results reflect real differences or measurement error.

**SUMMARY**

The authors test the cross-national comparability of the social and political trust scales included in the ESS. By doing so, they provide a nontechnical introduction on invariance testing by means of multiple group CFA (including mean structures).

**REFERENCES**


SUGGESTIONS FOR FURTHER READING


APPENDIX 2.A: MPLUS INPUT FILE FOR MODEL 6, TABLE 2.3

```
TITLE: ESS trust cross-country cross-time model, excluding the Netherlands, Ireland, UK, Austria, and Portugal
DATA: FILE IS esstrust2long.dat;
VARIABLE: NAMES ARE ESSROUND IDNO trust fair help parliament politicians cntry1;
USEVARIABLES ESSROUND trust fair help parliament politicians;
GROUPING IS essround (1 = round1 2 = round2 3 = round3);
MISSING = ALL (999);
MODEL:
soctrust BY trust fair help;
poltrust BY parliament politicians;
Model round2:
[soctrust@0];
[poltrust] (1);
Model round3:
[poltrust] (1);
OUTPUT: SAMPSTAT RESIDUAL cinterval tech1;
```
methodological issues that arise when SEM is used to test DIF. Based on Monte Carlo simulations, recommendations on the preferred analytical strategy and scaling method are formulated.

NOTES

1 The invariance levels that are not discussed here include invariance of unique factor variances, invariance of latent construct variances/covariances, and invariance of latent construct means. For a detailed discussion on factorial invariance, see Meredith (1993) and Vandenberg and Lance (2000).

2 Information-theoretic measures of fit (e.g., Akaike information criterion, Bayesian information criterion) are also suitable for evaluating factorial invariance, but they have not been supported in the literature as being informative beyond the CFI and the other fit measures discussed here.

3 The constrained-baseline strategy is similar to the “top-down” approach for assessing scale-level invariance in that it starts with a model that imposed the most restrictive (or full) metric or scalar invariance. In contrast, the free-baseline strategy has similarities with the “bottom-up” approach in that it starts with the least restrictive (or partial) metric or scalar invariance model. For more details on these two approaches, see Welkenhuyzen-Gybel and van de Vijver (2001).

4 When evaluating scale-level invariance, the variance and/or mean of a construct are freely estimated in one group. In contrast, when evaluating item-level invariance, they are constrained to equality across groups. In other words, the configural invariance model is used as the baseline model when examining nonuniform DIF and uniform DIF.

5 For example, in Figure 3.3, two nested-model comparisons are possible for testing nonuniform or uniform DIF. Thus, the corrected critical p value equals .025 (= .05/2).

6 In supplementary analyses, we found that the critical values of RMSEA and SRMR suggested by Chen (2007) are not suitable for DIF analysis as well. When used to detect nonuniform and uniform DIF under our simulated conditions, generally they inflated Type I error above the nominal alpha level and/or provided very low power.

7 In fact, WLS and RWLS are not recommended in some cases. For example, Flora and Curran (2004) noted that the WLS $\chi^2$ is inflated, as are the parameter estimates, whereas their standard errors are negatively biased. Also, French and Finch (2006) found that the RWLS LR test provides very low power for testing scale-level metric invariance.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


NOTES

1 See a contrasting view on partial invariance by De Beuckelaer (2005).
2 An exception is Model D testing for partial scalar invariance that relaxes two constraints imposed by Model C.
3 A freeware program to calculate the Satorra-Bentler corrected chi-square difference can be downloaded from the Web site http://www.abdn.ac.uk/~psy086/dept/psychom.htm

REFERENCES


SUGGESTIONS FOR FURTHER READING


APPENDIX 4.A: LISREL SYNTAX

*************** GREAT BRITAIN ***************

DA NI = 4 NO = 2378 NG = 2 MA = CM
CM FU
0.6654  0.2030  0.3353  0.1536
0.2030  0.5382  0.1605  0.1050
0.3353  0.1605  1.0788  0.1497
0.1536  0.1050  0.1497  1.0491
ME
3.7642  3.8756  3.5008  2.8158
LA
pstmzs_r dngval_r flrms nhpftr
MO NX = 4 NK = 1 LX = FU,FI TD = FU,FI PH = SY,FR TX = FU,FI
KA = FU,FI !In addition to usual matrices, TX and KA are added that contain intercepts and means
even more recent paper by Weijters, Schillewaert, and Geuens (2008) introduces a sound but sophisticated approach to correct for different types of response styles. Their approach is based on the inclusion of a separate, heterogeneous set of response style indicators drawn from a wide universe of multi-item survey measures. This approach enables a valid and reliable assessment of response styles (i.e., due to the heterogeneity of response style indicators), while avoiding a possible confound between questionnaire content and response style of the respondent (i.e., as response style indicators are not used for substantive purposes; see De Beuckelaer, Weijters, & Rutten, 2010).

**ACKNOWLEDGMENTS**

We would like to thank Albert Satorra, Steffen Kühnel, and Gerrit K. Janssens for their valuable comments on an earlier version of this chapter. In addition, we are also indebted to the critical reviewers Eldad Davidov and Holger Steinmetz, who provided us with valuable suggestions on how to improve this chapter.

**SUMMARY**

De Beuckelaer and Swinnen raise the question to what extent latent means difference tests are robust against noninvariant measurement parameters. An extensive simulation study shows that especially differences in item intercepts can seriously threaten the validity of latent means comparisons.

**NOTES**

1 In this study, we deal only with the comparison of LV means across groups. Cross-group comparisons of structural relationships between observed or LVs are not considered. Such comparisons require only the factor loadings of LV indicators to be identical across the groups (i.e., metric invariance across groups) involved in the comparison (see, for instance, De Beuckelaer, Lievens, & Swinnen, 2007; Van de Vijver & Leung, 1997).

2 More precisely, it is the mean- and variance-adjusted chi-square statistic with robust standard errors (see Muthén & Muthén, 1999), which is used in this simulation study.
Especially because of these scales, the present simulation study is to be conceived as more realistic than the study by Kaplan and George (1995).

The reliability of the ith LV indicator is calculated as follows: $1 - \frac{\text{error variance}}{\lambda_i^2 + \text{error variance}}$. The error variance is always fixed to 0.51 in the simulation study.

Related noninvariance conditions are characterized by an identical LV mean difference between both groups and a noninvariant indicator having an unequal factor loading and/or indicator intercept across groups.

The specification of a confidence interval (CI) is always a somewhat arbitrary decision. Changing from a 99% CI to a 95% CI would not have had a substantial impact on the decisions regarding robustness/nonrobustness of the noninvariance condition (i.e., on average across all noninvariance conditions, less than one replication [i.e., 0.80; SD = 0.60] would have been classified differently).

These additional tables were not included in this chapter to save space. These tables do, however, show that the test of equality of LV means in both groups does provide reasonable control over the type I error rate (i.e., as assessed in scalar invariance conditions). As type I error rates occur when one falsely rejects the hypothesis of equal LV means at population level (i.e., the null-hypothesis), the test's control over the type I error rate is evaluated by examining statistical results obtained in “no discrepancy conditions” (i.e., those conditions in which the null-hypothesis truly holds at population level). In addition, such tables provide useful information on the power of the statistical test (i.e., as assessed in noninvariance conditions).

REFERENCES


Vandenbult, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158.


### SUGGESTIONS FOR FURTHER READING


### APPENDIX 5.A

**Classification and Regression Trees (C&Rtrees)**

Percentage of correct statistical conclusions (latent variable [LV] mean difference test)

ILLUSTRATIVE EXAMPLE of a C&RTree (see also Figure 5.A.1) Example_C&RTree

(4 indicators; N=198) [[Overall: 22.3%];

[Split 1=F4_D3 [upper=root]: nil=7.9%; one=63.9%];

[Split 1.1=F5_D2 [upper=nil]: nil=1.1%; one=25.0%];

[Split 1.2=F5_D4 [upper=one]: nil=79.6%; one=16.7%];

[Split 1.1.1=F2_D1 [upper=nil]: nil=0.0%; one=6.7%];

[Split 1.1.2=F4_D1 [upper=one]: nil=50.0%; one=0.0%];

[Split 1.2.1=F4_D3*F5_D3 [upper=nil]: nil=88.9%; one=61.1%]]

Notes: This example C&RTree is identical to RT_02 (see below); clarification of the notation used: F4_D3=1 means the factor loading of indicator 1, i.e. λ2, equals 0.8 in Group 2 (see F4 in Table 5.1); F5_D2=1 means the intercept of indicator 2 is 0.15 higher for Group 2 (see F5 in Table 5.1); F5_D4 means the intercept of indicator 2 is 0.45 higher for Group 2 (see F5 in Table 5.1).
ordinal data yields very similar results compared to when the indicators are treated as continuous.

NOTES

1 These studies report simulations that examine whether assuming normality and continuity of measurement scales when using ordinal categorical scales yields different conclusions in a cross-cultural invariance test. Comparisons of several estimation methods based on different assumptions for other types of models have also been conducted. They generally conclude that the maximum likelihood parameter estimates and standard errors are rather robust for small violations of normality (see e.g., Coenders & Saris, 1995; Coenders, Satorra, & Saris, 1997).

2 The negative cross loadings indicate that the association (covariance) between the opposing latent value constructs did not capture all of the opposition for these items. The positive cross loadings indicate that these associations overestimated the opposition for two items. The need for these cross loadings may be due to the reduction from 10 original values to 7. Without introducing them, the model fit was not acceptable. From a measurement point of view, cross loadings are not elegant. Cross loadings contaminate correlations between factors, a problem if one is interested in the correlations. However, our main interest was not to evaluate the strengths of relationships between values but to examine whether measurement properties, such as factor loadings and intercepts, are invariant across countries.

3 A threshold captures transitions from one category to another. Thus, if there are $K$ response categories for an indicator, there are $K-1$ threshold parameters for the latent response variable.

4 In principle, the equality constraint of the intercepts may be released in Mplus by introducing a perfectly measured factor behind the latent response variable (Muthén & Asparouhov, 2002, p. 15).

5 Two parameterizations are possible for running the model: Theta and Delta (Muthén & Muthén, 2007). The Theta parameterization includes residual variances for the continuous latent response variables (see Muthén & Muthén, 2007, pp. 485–486). This has the advantage of also permitting a test of the invariance of the residual variances (Millsap & Yun-Tein, 2004, Muthén & Asparouhov, 2002). We applied both the Theta and Delta parameterizations and obtained essentially the same results. See Appendix 6.A for the final model.

6 Before testing for measurement invariance, we examined the level of skewness and kurtosis of the values across the countries. Skewness (both left and right, depending on the item) was significant for all 21 items in the three countries. Kurtosis was significant for 20 items.

7 Though the power of MGCFA is lower than proportional odds modeling (see Welkenhuysen-Gybel’s & Billiet, 2002, p. 216).

REFERENCES


Vandenberg, R. J. (2002). Towards a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158.


**SUGGESTIONS FOR FURTHER READING**


**APPENDIX 6.A: MPLUS SYNTAX FOR THE FINAL MODEL WITH INVARIANCE OF THREE VALUES**

```plaintext
1 title: MGCFA Values ESS Round 2 - Belgium, Luxembourg, Netherlands;
2 data: file is benelux_values.dat;
3 variable: names are ipcrtiv imprich ipeqopt ipshabt impsafe impdiff ipfrule ipudrst
4 ipmodst ipgdtim impfree iphlppl ipsuces ipstrgv ipadvnt ipbhprp iprspot iplylfr
5 impenv imptrad impfun country;
6 categorical are all;
7 grouping is country (1=BE 2=LU 3=NL);
8 missing = all (7-9);
9 Analysis: PARAM = THETA;
10 Model:
11 SD by impfree* ipcrtiv@1;
12 UNBE by ipeqopt* ipudrst@1 impenv iphlppl iplylfr ipadvnt imprich;
13 COTR by ipmodst* imptrad* ipfrule@1 ipbhprp iprspot imprich;
14 SEC by impsafe@1 ipstrgv;
15 POAC by imprich* iprspot* ipshabt@1 ipsuces ipmodst;
16 HE by ipgdtim@1 impfun;
```
feature of MGSEM is the estimation of more complex path models with direct and indirect effects, a feature that is very useful when analyzing a variable, such as religious involvement, which takes an intermediary position. Another advantage is that MGSEM, contrary to multilevel modelling, does not require a minimum number of countries to guarantee an accurate estimate—even with as little as two countries, meaningful comparisons can be made. At the same time, the MGSEM approach naturally has some drawbacks. First, it is far from straightforward as to how context-level variables could be added in the analysis, although one could think of a two-step approach in which certain parameters of the MGSEM model (e.g., means or effects) are regressed on context-level variables (cf. Achen, 2005). And secondly, MGSEM, at least when polychoric correlations are used, is computationally quite demanding, which puts a limit on the complexity of the model. Including too many variables in the model can lead to very long runtimes and/or estimation difficulties.

SUMMARY

The authors study the interplay between social structure, religious involvement, values, and social attitudes using ESS data from 25 countries. This analysis illustrates how the similarity of measurement and structural relations can be tested using a MGSEM model.

NOTES

1 This was also the case, however, for a-religious humanists (“free thinkers”).
2 This operationalization of religious involvement does not admit a distinction between people who adhere strongly to a nonreligious way of life, people who only marginally relate to their religion, and people who belong to a religion but vary in terms of degree of involvement, in both the past and in present situations. There is only a differentiation going from actually not involved to strongly involved in the present situation.
3 The equivalence of the error covariances over the countries has neither been specified nor tested in any of the models.
4 All models are estimated with LISREL 8.7 (Jöreskog & Sörbom, 1993). Because the three items are measured on ordinal scales, we decided to use a weighted least squares
(WLS) estimation procedure in which polychoric correlations and asymptotic covariance matrices are used as input rather than regular covariance matrices (Jöreskog, 1990).

5 Detailed results can be obtained from the authors.

6 Secularization is conceived of here as a process at the individual level, which is measured by the decline in individual participation in religious associations and services. This is only one aspect of the concept of secularization (Dobbelaere, 2002).

7 We should be cautious of comparing the proportion of explained variance across countries. After all, $R^2$ is a standardized measure and therefore assumes equivalence of the variances of latent concepts over countries. Variance equivalence was not tested in this study (we would like to thank Eldad Davidov for his comment on this).

8 We would like to thank Hermann Dülmer for suggesting to us a regional treatment of Germany.

9 Due to lack of space, it is not possible to present these analyses in great detail. However, full results can be obtained from the first author.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


assumptions. Second, the geometric representation of items allows for a direct and explicit examination of the complete pattern of relationships among items and dimensions. The SSA maps offer a clear visual depiction of these relationships, which when combined with information from CFAs, provide for a more complete and concrete understanding of a scale’s structure. Results from the two analyses serve to validate one another and together provide strong support for the RTC scale’s structure across cultural contexts. We recommend that future work take on similar procedures, employing distinct analytic techniques, for the validation of other constructs as well.²

SUMMARY

This chapter compares two approaches to test measurement equivalence of the Resistance to Change (RTC) scale, namely MGCFA and Smallest Space Analysis (SSA). While there is considerable overlap in the conclusions of both methods, each technique highlights distinct insights in the data.

NOTES

1 Additional analyses were conducted, comparing the interrelated four-factor model to a model with a higher-order, overarching, latent variable. Results of these additional analyses are reported in Oreg et al. (2008).

2 The authors wish to thank Shmuel Shye, Adi Amit, and Liat Levontin for their assistance in running the FSSA software. We also thank Shalom Schwartz and Holger Steinmetz for helpful comments on a previous version of this manuscript. A number of paragraphs in this article are © 2008 by the American Psychological Association and reproduced with permission. The use of this information does not imply endorsement by the publisher.

The Czech participation was supported by grant IAA700250702 from the Grant Agency of the Academy of Sciences of the Czech Republic and is related to research plan AV0Z0250504 of the Institute of Psychology, Academy of Sciences of the Czech Republic. Funds for the Croatian participation have been provided by Grant 194-1941558-1530 from the Ministry of Science, Education, and Sports in Croatia. Slovak participation was supported by grant no. 2/6185/26 approved by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, VEGA. Data collection in China was supported by the Chinese National Science Foundation.
REFERENCES


**SUGGESTIONS FOR FURTHER READING**


ACKNOWLEDGMENTS

We would like to thank Jaak Billiet for his valuable suggestions on how to improve this chapter.

NOTES

1 In the first edition of this book we suggested measurement equivalence testing 1.0. This was a rather complicated procedure that introduced the same two aspects as the current chapter. The current chapter—Measurement Equivalence Testing 2.0—is, however, a lot easier to comprehend.

2 JRule is freeware and can be requested by sending an email to William van der Veld (w.vanderveld@ru.nl). More information can be found on William M. van der Veld’s ResearchGate pages.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**

SUMMARY

In this chapter, Meuleman illustrates how multiple group SEM can be employed as solution to overcome some of the problems associated with using classical multilevel analysis for cross-national data. Using a two-step procedure, the chapter shows how contextual characteristics directly and indirectly affect anti-immigration attitudes.

NOTES

1 The study by Sides and Citrin (2007) is an exception here.
2 Some studies, for example, Coenders (2001), report tests for the cross-cultural equivalence of measurements. However, in these studies only metric equivalence is assessed (i.e., equality of factor loadings), while the use of multilevel models presupposes scalar equivalence (i.e., additional equality of intercepts).
3 The different forms of ethnic threat can be expected to be triggered by different antecedents, and therefore imply different theoretical models. In this chapter, however, I confine myself to perceptions of economic threat.
4 One could argue that this relationship also operates in the other direction, and that perceived immigrant presence is caused by, rather than causing, threat perceptions. Unfortunately, the causality of this relationship cannot be tested with the available dataset. The theoretical framework used in this chapter (group conflict theory), however, predicts that this relationship flows predominantly from perceived immigrant presence to perceived threat.
5 Israel also participated in round 1 of the ESS. However, I decided to remove the results of this country from the analyses because of large differences in the immigration context.
6 If sufficient information was available (i.e., missings on less than one-third of all items used), missing values were imputed by means of the EM-algorithm available in LISREL 8.7.
7 Strictly speaking, scalar equivalence is also required for the other endogenous individual-level variables that are directly linked to contextual variables (i.e., perceived presence of immigrants and contact). However, as these are single-indicator concepts, this assumption cannot be tested. (I would like to thank Eldad Davidov for this useful addition.)
8 In Section 10.4.2.1, more information is given on how I treat the countries for which partial scalar equivalence is lacking.
9 Technical specifications: The estimated model is a multigroup model with 21 groups. The mean structure of the data is not taken into account, so that no intercepts or latent means are estimated. All factor loadings are set equally across groups, except for the significant deviations from metric equivalence that were detected during equivalence testing (Meuleman, 2009). Correlations were estimated between all exogenous
variables of the model. Because most items are measured on ordinal scales, and various items have a strongly skewed distribution, I used a weighted least squares (WLS) estimation procedure, in which polychoric correlations and asymptotic covariance matrices are used as input rather than regular covariance matrices (Jöreskog, 1990). All models are estimated with LISREL 8.7 (Jöreskog & Sörbom, 1993). In some countries, the direct paths from gender (DK, FI, GB, NL, PT, SI), age (DK, IT), or education (FI, IT, NL, PT) to REJECT were deleted, because they are virtually identically to zero and cause difficulties in the estimation (i.e., large standard errors for some parameters).

For reasons of conciseness, contact is operationalized here as a single dummy variable indicating contact with immigrant friends and/or colleagues. Exploratory analyses with separate indicators for both modalities of contact show that having immigrant friends has a substantially stronger effect than having immigrant colleagues (see also Schneider, 2008). Probably, this finding can be explained by self-selection mechanisms: especially individuals with proimmigration attitudes will select immigrants as friends. However, even contact with colleagues, which is far less a matter of free choice, turns out to temper threat perceptions considerably. I would like to thank Peter Schmidt for drawing my attention to this issue.

Complete results can be obtained from the author.

Concretely, I compared latent country-mean rankings of: (1) the model with partial scalar equivalence for all countries and (2) the best-fitting model, where partial scalar equivalence is violated for seven countries. Although the latent mean scores are somewhat different, both models yield almost identical country rankings. Consequently, it can be concluded that the present measurement for inequivalence only biases the results to a limited extent, and that country rankings are largely unaffected.

The author would like to thank Jaak Billiet, Marc Swyngedouw, Hideko Matsuo, Eldad Davidov, and Peter Schmidt for comments on earlier versions of this chapter.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


but rather assumed that the capitalist order, once established in a society, would force the individual to work like the Puritan who wanted to work in a calling.

**SUMMARY**

Dülmer illustrates the disadvantages of using OLS regression for international comparisons by analyzing work ethics in 53 countries. A comparison with multilevel regression models—which are more appropriate to analyze this hierarchical data—shows remarkable differences.

**NOTES**

1. Unfortunately, the b-coefficient for religiosity that was controlled for (cf. Norris & Inglehart, 2004, p. 166) is missing in the Table A7.1 with the complete regression results (p. 179).
2. In this article I will report robust standard errors (cf. Hox, 2002, p. 200). Using robust standard errors for MLSEM requires for reasons of power a minimum number of 40 countries (cf. Meuleman & Billiet, 2009). Inferences based on robust standard errors sacrificing some statistical power in order to be less dependent on the assumption of normality (cf. Hox, 2002, p. 201). For less than 40 countries remains the option to use conventional standard errors.
3. Appendix 11.A includes the complete information about the type of religious culture and the HDI 1998.
4. Together with Norris and Inglehart (2004) measurement invariance has been assumed but not tested (for testing measurement invariance see Meuleman and Billiet in this book). Testing measurement invariance with more than 30 countries is "unwieldy at best" (Selig et al., 2008, p. 105). However, the very good fit of the SRMR within and the SRMR between of Model 1b might be a hint that measurement invariance is given for the vast majority of countries. In any case, research is needed in order to know more about the relationship between fit measures and measurement invariance in Multilevel CFA.
5. Mplus uses the \( z \)- instead of the \( t \)-distribution. Hence, in cases with comparably few macro-level units and a relatively low number of degrees of freedom for estimating a b-coefficient (say, \( df < 120 \)) the acceptance region for the null hypothesis might be slightly too small; that is, the null hypothesis might be rejected too easily. This affects in general (a) macro-level main effects if the intercept is estimated with a random component, (b) slopes of micro-level indicators that are estimated with their own random component, and (c) macro-level indicators that interact with a micro-level indicator that is estimated with a random component. The predictor variables of macro-level latent factors are also affected. Substantial differences between both tests affect only borderline effects.
6 The first partial derivative for the cohort born from 1946 to 1965 is, for instance, \( \Delta \bar{Y} / \Delta \text{Born 1946–1965} = 0.194 – 0.485 \times (\text{HDI 1998}) \).
7 The formula for calculating conditional standard errors is given by Friedrich (1982, p. 810). Conditional z-values can be calculated by dividing the conditional b-coefficient by its conditional standard error.
8 One might try to save H1a by assuming that at least the majority of traditional countries would have not developed or even have suffered from economic decline during the second half of the twentieth century. Hence, modernization would have taken place, at best, in a minority of traditional countries. This assumption, however, clearly contradicts Inglehart’s (1997, pp. 229–230, cf. also p. 332) own finding according to that poorer countries showed in the past higher growth rates than wealthier countries.
9 The first partial derivative for HDI 1998 is \( \Delta \bar{Y} / \Delta \text{HDI 1998} = 4.178 – 2 \times 3.969 \times (\text{HDI 1998}) – 0.485 \times (\text{Born 1946–1965}) – 0.375 \times (\text{Born 1966–1987}) – 0.273 \times (\text{Education 2nd Level}) – 0.737 \times (\text{Education 3rd Level}) \).
11 There are no substantial changes in the other effects (including the conditional effects) if the squared HDI for 1998 would be dropped from Model 2c. The AIC and the adjusted BIC would, however, slightly increase (by less than 10).

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


**APPENDIX 11.A**

**Included Countries (and Year of Fieldwork)**

*European Values Study (Wave III):* Belarus 2000 (0.781, IN, O), Belgium 1999 (0.925, PI, C), Bulgaria 1999 (0.772, IN, O), Croatia 1999 (0.795, IN, C), Czech Republic 1999 (0.843, IN, C), Denmark 1999 (0.911, PI, P), Estonia 1999 (0.801, IN, P), Finland 2000 (0.917, PI, P), France 1999 (0.917, PI, C), Germany 1999 (East and West separately; 0.911, PI, P), Great Britain 1999 (0.918, PI, P), Hungary 1999 (0.817, IN, C), Iceland 1999 (0.927, PI, P), Ireland 1999 (0.907, PI, C), Italy 1999 (0.903, PI, C), Latvia 1999 (0.771, IN, P), Lithuania 1999 (0.789, IN, C), Luxembourg 1999 (0.908, PI, C), Malta 1999 (0.865, IN, C), Netherlands 1999 (0.925, PI, P), Northern Ireland 1999 (0.918, PI, P), Poland 1999 (0.814, IN, C), Romania 1999 (0.770, IN, O), Russia 1999 (0.771, IN, O), Slovakia 1999 (0.825, IN, C), Slovenia 1999 (0.861, IN, C), Spain 1999 (0.899, PI, C), Sweden 1999 (0.926, PI, P), Turkey 2001 (0.732, IN, M), Ukraine 1999 (0.744, IN, O).
ACKNOWLEDGMENT

The authors thank Peter Lugtig for his comments on an earlier version of the chapter.

NOTES

1 The following countries participated in the first round: Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Luxemburg, the Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. The same countries, except Israel, participated in the second round of the ESS, and Estonia, Iceland, Slovakia, Ukraine, and Turkey participated for the first time. In the third round Czech Republic, Greece, Iceland, Israel, Italy, Ireland, Luxemburg, and Turkey did not participate, where Bulgaria, Cyprus, Estonia, Latvia. Romania, Russia participated in this round for the first time.

2 The exact wording of the questions was: “To what extent do you think [country] should allow people of the same race or ethnic group as most [country’s] people to come and live here,” “How about people of a different race or ethnic group from most [country] people?” and “How about people from the poorer countries outside Europe?” for items allow 1, 2, 3, respectively. All answer categories are coded as 1 = Allow many to come and live here; 2 = Allow some; 3 = Allow a few; 4 = Allow none. Age is measured in years, divided by 10.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


parameters are different across countries. For that purpose, the authors propose a model with random factor loadings and a Bayesian estimation procedure.

NOTES

1 Equations 13.2 to 13.6 are written in matrix/vector notation. By consequence, no subscripts are included to refer to the distinct items.

2 In this particular case, however, the inclusion of main effect $\gamma_1$ is not strictly necessary for the interaction term to be meaningful. In this MLSEM, the random slope variance (i.e., the component that is explained by the cross-level interaction) is completely separated from the random intercept variances (i.e., the component relevant for the main effect). This makes it possible, for example, to reduce model complexity by removing all random intercepts from the data (by group-mean centering) and retain the random slope variation only—this procedure yields the same estimate for the cross-level interaction effect.

3 The rationale for focusing on the $\textit{born}$ item is twofold. First, we estimated similar models with random slopes for the other items as well. Although all four items contain significant random slope variation, the largest variance component is found for $\textit{born}$. Second, and perhaps even more importantly, there are convincing theoretical arguments explaining why this item functions differentially across countries.

REFERENCES


Meuleman, B. (2012). When are item intercept differences substantively relevant in measurement invariance testing? In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences. Festschrift for Peter Schmidt* (pp. 97–104). Wiesbaden, Germany: Springer VS.


**SUGGESTIONS FOR FURTHER READING**


attractive option for studying measurement invariance in any situation in which the indicators are discrete variables.¹

SUMMARY

The authors show how measurement invariance can be tested using LCA instead of techniques that rely on stronger distributional assumptions, such as MGCFA or IRT. Multigroup latent class models for nominal as well as ordinal indicators are discussed and illustrated by means of empirical examples.

NOTE

¹ This work was partially supported by a grant No. BFR06/040 from the “Fonds National de la Recherche” (Luxembourg).

REFERENCES


SUGGESTIONS FOR FURTHER READING


classifying individuals that are spiritual but not religious is inappropriate because both terms have huge intersections from the perspective of the spiritual movement.

Further research is needed to validate the findings from the MGLCA model. As a first step, it is necessary to improve the indicators for different forms of religion. Particularly, the emotional aspects of religion are underdeveloped in survey research but will gain increasing attention. Second, it has to be tested if alternative spirituals develop in place of church religiosity or in place of unbelief. Only if this question is answered it is possible to say whether the existence of alternative spiritualities challenges secularization theory. The study shows that in each case, alternative spiritualities have to be considered by empirical research on religion in Europe.

SUMMARY

Siegers applies multiple group LCA to construct a typology of religious orientations in Europe. A typology consisting of five classes—church religiosity, moderate religiosity, alternative spiritualities, religious indifferences, and atheism—is found to be measured in a partially homogenous way.

NOTES

1 The items frequently included in surveys are: frequency of church attendance, the importance of God for respondentís life, religious self-assessment, denominational membership, and frequency of prayer.
2 RAMP was coordinated by Wolfgang Jagodzinski (University of Cologne) and Karel Dobbleare (Catholic University of Leuven). I am grateful to Wolfgang Jagodzinski for providing me the dataset.
3 All models reported here are estimated with Mplus Version 5.2. Missing cases were excluded listwise. Overall, 10,809 cases are included in the analysis.

REFERENCES


SUGGESTIONS FOR FURTHER READING
of constraints to this model in order to test full and different kinds of partial invariance. Correlation of full class profiles were suggested as a way to assess configural invariance. We demonstrated this approach using the data on basic human values in two groups of countries—West & North Europe and Eastern Europe. The formal criteria pointed to the existence of only configural invariance of the four latent classes across the two groups. However, these criteria might be too strict and, given the substantial considerations, it might be concluded that the three latent classes are invariant across the groups. Using this result, the class probabilities were legitimately compared, and analyses revealed Eastern Europeans were much more prone to membership in the Weak Social values class than members of the West & North European population.

Overall, further research is needed to guide the model selection process for deciding between partial and configural measurement invariance of latent classes as well as to develop model fit statistics that are more flexible and robust to a sample size, alternative to the overly conservative likelihood ratio test. The overly strict LRTs and uncertainty of information criteria such as BIC need to be amended with newer approaches, such as newly introduced approximate (Bayesian) measurement invariance in factor analysis. Nonetheless, there are several tools currently available to test for measurement invariance of latent classes across groups, some of which were demonstrated in this chapter.

**SUMMARY**

Rudnev describes how a group-as-covariate approach can be used to test the invariance of unordered latent class models. The chapter highlights the links between LCA and CFA models, and provides an illustration using data on basic human values from two groups of countries (West & North European and Eastern European countries).

**NOTES**

1 This chapter is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE). The author is grateful to the two anonymous reviewers and to the editors for their useful comments, which helped to significantly improve the chapter.
Testing for Invariance of Latent Classes

2 It is not related to latent profile analysis, which is a generalization of the LCA model to continuous indicators (Vermunt, 2004). Class profile is a full set of conditional probabilities for the specific class.

3 The label “metric” might be misleading in the context of LCA, as classes do not have a metric; still, we keep it to compare with the levels of measurement invariance in factor analysis.

4 The formula is a simplified expression 13.5 given in the Kankaraš et al. chapter, this volume.

5 Burnham and Anderson (2002) suggested rules of thumb for model selection for nested models using difference in AIC: 0–2 being negligible, 4–7 moderate, and more than 10 substantial. As models with different number of classes are not truly nested, it is not clear whether these rules are applicable here.

6 Technically, such a model is called a factor mixture model (see Muthén, 2008).

7 This problem is sometimes referred as a “big data problem” in computer science and is often addressed by sampling of the data. Sampling procedure draws a random sample from each group of respondents, and then all the models required are estimated in a usual way using this smaller subsample. From our experience, application of sampling to LCA gives inaccurate results as compared to a full-sample estimation, which is due to the sensitivity of LCA to heterogeneity in the data.

8 Here, we use scaled LRT (Satorra & Bentler, 1999), as the models were estimated with the maximum likelihood robust algorithm which adjusts the likelihood for nonnormality.

REFERENCES


**SUGGESTIONS FOR FURTHER READING**


and a group parameter (Meulders & Xie, 2004), which is called differential feature functioning (DFF).

DIF research has a long tradition in educational measurement where matters of fairness and equity were paramount. In this context, DIF was commonly referred to by the notion of item bias. Concerns about test bias typically centers around the differential performance by groups based on gender or race. Currently, the scope of DIF research is much wider. Apart from issues on equity and fairness and apart from investigating lack of invariance, Zumbo (2007) discerned other general uses of DIF, such as the investigation of the comparability of translated or adapted measures, or the use of DIF as a method to help understand the cognitive and/or psychosocial process of item responding and test performance.

SUMMARY

This chapter discusses how various item response theory techniques (the Rash model as well as the two and the three parameter logistic model) can be applied to detect differential item functioning. The author illustrates the procedure with an application using different modes of data collection (paper-and-pencil vs. computerized modes).

REFERENCES


Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233.

**SUGGESTIONS FOR FURTHER READING**


SUMMARY

Quandt explores the advantages and limitations of using polytomous Rasch models for identifying potentially heterogeneous populations. Basic strategies to test equivalence are discussed, and the model is applied to ISSP data on national pride.

NOTES

1 IRT modeling has rapidly followed the statistical advances of the last decades, among them Bayesian methods and hierarchical or multilevel modeling. Chapter 19 demonstrates an elaborate hierarchical IRT approach with PISA data.
2 In turn, restricting the threshold distances of the same item to be equal can serve to actually test the interval scale quality of items (Andrich, 1978; Rost, 2004).
3 One consequence of this is that the concept of differential discriminatory power of different items indirectly appears in the Rasch model, although the individual logistic item characteristic curves (ICC) are still estimated with constant slope, and a single location parameter per ICC. This often goes unnoticed, because no explicit discrimination parameter is used as, for example, in the two parameter logistic model.
4 See Stark, Chernyshenko, and Drasgow (2006) and Ewing, Salzberger, and Sinkovics (2005) for overviews of commonalities and differences of IRT and MGCFA.
5 A number of related model variations have been proposed before and after that, see for an overview of recent developments the volume edited by von Davier and Carstensen (2007).
6 However, it is possible to compute confidence intervals around threshold parameter estimates per class, and check whether these overlap across classes.
7 The battery has seven items in ISSP 2003. One item was not included because it was not a replication from a previous application of this instrument in ISSP 1995 (ZA No. 2880), although comparison of both ISSP modules over time is not in the scope of this paper. Another item was dropped after a principal component analysis. Both items appear to have somewhat lower face validity in the judgment of this author, and incidentally, they are the only negatively worded items of the battery.
8 Although we have not yet shown scalar invariance for the battery, we know enough about the quality of the item battery that it would be a surprise if the country ranks changed by more than a few positions when using latent trait estimates instead of raw scores.
9 One reason for selecting this subset of countries is that all used self-administered questionnaires, with many other countries using face-to-face interviews for the ISSP survey. Interview mode effects could be plausible sources of DIF, which would however mandate a more systematic investigation than can be provided here.
10 Other options for handling polytomous responses available in the software would be the rating scale, dispersion, and equidistance models, which pose different additional restrictions on the distance of the category thresholds (von Davier, 2001). Being more
restrictive, they are less likely to fit the data and also less prone to uncover aberrant item category profiles. The relatively poorer fit was confirmed by according model runs (not reported).

11 A 5-class model did not converge after 1200 iterations and its preliminary estimates showed no improvement in fit over the 4-class model.

12 The number of parameters is determined by the number of observed response patterns for the saturated model. For the Rasch models, every item requires one parameter each for estimating its category thresholds, plus two parameters for a logistic function that approximates the distribution of response scores. For each additional class, one parameter for the class size has to be included.

13 To avoid some complexity, the explanation and reporting of certain Rasch-specific measures computed by WinMira, such as Andrich’s index of person separability, which is largely equivalent to coefficient α (Andrich, 1988, p. 84), is omitted here. Results are available from the author on request.

14 Together with the distribution of classes over countries, the item location profiles provide one bit of information that could be pursued for further interpretation: just by the size of Class 2 in the U.S. sample, the higher relative difficulty of items citz and ctrs in that class appears to be almost a U.S. specialty.

15 Analyses with education and (high) age as proxies did not prove useful. One problem is that these proxies are also correlated with the substantive scale. Presentation has been omitted for want of space. An application of hierarchical IRT methods focused on explicitly estimating respondents’ propensity to extreme response styles can be found in De Jong, Steenkamp, Fox, and Baumgartner (2008).

16 I am indebted to a reviewer for helpful hints and comments. All remaining errors are my own.

REFERENCES


Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

**SUGGESTIONS FOR FURTHER READING**


REFERENCES


**SUGGESTIONS FOR FURTHER READING**


(rather than frequentist) estimation and imposes less restrictive equality constraints on measurement parameters.

NOTES

1 For an overview of recent developments in the analysis of measurement invariance, see Van de Schoot, Schmidt, De Beuckelaer, Lek, and Zondervan-Zwijnenburg (2015).
2 For an introduction to Bayesian approximate measurement invariance using a simulated data example, see Lek et al. (in press).
3 The percentages of foreign-born immigrants in a country are averaged between 1995 and 1999 according to data from the World Bank Group (2016).
4 Currently, only Mplus and the R package Blavaan (Merkle & Rosseel, 2016) allow testing for approximate measurement invariance.
5 We used the Mplus default criterion `bconvergence = .05`.
6 Other studies (Cieciuch et al., 2014; Muthén & Asparouhov, 2013; Van de Schoot et al., 2013) have also considered a prior variance of 0.010 as a reasonable choice and compromise between model fit and precision.

REFERENCES


for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*. Advance online publication. doi:10.1037/met0000113


**SUGGESTIONS FOR FURTHER READING**


NOTES

1 Studies of yet another type try to assess the sources of noninvariance (see, e.g., Davidov, Dülmer, Cieciuch, Kunz, Seddig, & Schmidt, 2016; Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012; Meuleman, 2016; Meuleman & Schlüter, this volume; as well as Schlüter & Meuleman, 2009).

2 It should be noted that another strength of the alignment procedure in Mplus is its option to conduct simulations to explore the stability in the ranking of group means (see Munck et al., in press). Furthermore, the alignment procedure can be expanded to analyze the invariance of uniqueness and of factor variances or covariances and to conduct multiple-group multiple indicators multiple causes (MIMIC) models (see Marsh et al., 2017).

REFERENCES


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158. doi:10.1177/1094428102005002001


**SUGGESTIONS FOR FURTHER READING**


the fact that several countries’ estimates were close together. This means that it will be difficult to find a misspecified model in which the mis-
specifications do not affect the rank order at all. In a sense, this observa-
tion merely reflects the law that, as more precise conclusions are drawn from the data, more stringent requirements on those data are needed. From that perspective, this aspect of sensitivity analysis simply reveals a truth about the limits to the conclusions that can be drawn from a given analysis.

**SUMMARY**

This chapter argues that comparative researchers should not focus exclusively on cross-group differences in measurement parameters, but rather assess the impact that such differences have on substantive conclusions. The chapter shows how the expected change in the parameter of interest (EPC-interest) can be used for that purpose.

**ACKNOWLEDGMENTS**

Thanks are due to the anonymous reviewer for various suggestions that improved this chapter, especially the suggestion to consider the effect of loadings and intercepts jointly. Thanks are also due to Jeroen Vermunt and Yves Rosseel for their software implementations of the EPC-interest.

**NOTES**

1. This follows the logic that a regression with the constant one as predictor will give as a regression coefficient “of” this constant the dependent variable’s mean (when there are no other predictors, i.e., simple regression) or intercept (when there are, i.e., multiple regression).
2. This could be remedied by specifying missing = ‘fiml’ as an argument to the lavaan function call.
REFERENCES


SUGGESTIONS FOR FURTHER READING


Second, some of the open answers the respondents provided contained connotations that were difficult to code. This is a general disadvantage of web probing compared to conventional cognitive interviewing where the researcher can, by asking additional questions, contribute to the clarification of meaning.

Finally, we found some considerable differences between countries but no consistent relationship between (groups of) countries and (types of) argumentation patterns. Most country differences seem to be based on local peculiarities.

In spite of these limitations, however, the mixed methods approach we employed allowed us to gain insights which we could not have arrived at by using quantitative methods alone.¹

**SUMMARY**

This chapter proposes a mixed methods approach to test for measurement invariance. Using a qualitative approach alongside quantitative procedures yields better understanding of the possible causes of noninvariance. The invariance of a scale measuring attitudes towards immigration is tested to illustrate the approach.

**NOTE**

¹ This research was supported by grants from the German Research Foundation (DFG) to the first author (BR 908/3–1 and BR 908/5–1). We would also like to thank Dorothée Behr and Alexandra Kireew for their support.

**REFERENCES**


Michael Braun and Timothy P. Johnson

(ALLBUS) [The ALLBUS guest worker items. On the history of a standard instrument in the general surveys in social sciences (ALLBUS)]. Soziale Welt, 58, 145–161.


SUGGESTIONS FOR FURTHER READING

