EXPLORATIONS IN MENTAL HEALTH

Quantitative and Qualitative Methods in Psychotherapy Research

Edited by
Wolfgang Lutz and Sarah Knox

Quantitative and Qualitative Methods in Psychotherapy Research

In this collection, international contributors come together to discuss how quantitative and qualitative methods can be used in psychotherapy research. The book considers the advantages and disadvantages of each approach, and recognizes how each method can enhance our understanding of psychotherapy.

Divided into two parts, the book begins with an examination of quantitative research and discusses how we can transfer observations into numbers and statistical findings. Chapters on quantitative methods cover the development of new findings and the improvement of existing findings, identifying and analyzing change, and using meta-analysis.

The second half of the book comprises chapters considering how qualitative and mixed methods can be used in psychotherapy research. Chapters on qualitative and mixed methods identify various ways to strengthen the trustworthiness of qualitative findings via rigorous data collection and analysis techniques. Adapted from a special issue of *Psychotherapy Research*, this volume will be key reading for researchers, academics, and professionals who want a greater understanding of how a particular area of research methods can be used in psychotherapy.

Wolfgang Lutz is Full Professor and Chair of Clinical Psychology and Psychotherapy at the Department of Psychology as well as Director of the Clinical Training Program and the Outpatient Research Clinic at the University of Trier, Germany. He is Editor of *Psychotherapy Research* and on the editorial board of several journals in the field, such as *Cognitive Therapy and Research*. He has published widely on outcome management in psychotherapy, therapist effects and the prediction of treatment progress for individual patients.

Sarah Knox is Professor and Director of Training for the Counseling Psychology Ph.D. program in the College of Education at Marquette University, Milwaukee, USA. She is Co-Editor-in-Chief for *Counselling Psychology Quarterly*, and has served on the editorial board of several journals. She also publishes extensively on the psychotherapy process, as well as on training and supervision.

Explorations in Mental Health series

Books in this series:

New Law and Ethics in Mental Health Advance Directives
The convention on the rights of persons with disabilities and the right to choose
Penelope Weller

The Clinician, the Brain, and I
Neuroscientific findings and the subjective self in clinical practice
Tony Schneider

A Psychological Perspective on Joy and Emotional Fulfillment
Chris M. Meadows

Brain Evolution, Language and Psychopathology in Schizophrenia
Edited by Paolo Brambilla and Andrea Marini

Quantitative and Qualitative Methods in Psychotherapy Research
Edited by Wolfgang Lutz and Sarah Knox

Quantitative and Qualitative Methods in Psychotherapy Research

Edited by Wolfgang Lutz and Sarah Knox

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2014
by Routledge
27 Church Road, Hove, East Sussex BN3 2FA

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Quantitative and qualitative methods in psychotherapy research/
edited by Wolfgang Lutz and Sarah Knox.
pages cm
1. Psychotherapy – Research. 2. Evidence-based psychotherapy.
I. Lutz, Wolfgang, 1966–. II. Knox, Sarah (Psychologist).
RC337.Q358 2014
616.89′14 – dc23
2013025209

ISBN: 978-0-415-82070-7 (hbk)
ISBN: 978-0-203-38607-1 (ebk)

Typeset in Baskerville
by Florence Production Ltd, Stoodleigh, Devon, UK

Contents

Illustrations

Figures

Tables

Contributors

Rebecca Alperin, Ph.D.
Department of Psychology
University of Toledo
Toledo, OH, USA

Scott A. Baldwin, Ph.D.
Department of Psychology
Brigham Young University
Provo, UT, USA

Arjan Berkeljon, Ph.D.
Department of Psychology
Brigham Young University
Provo, UT, USA

Alan W. Burkard, Ph.D.
Department of Counselor Education and Counseling Psychology
Marquette University
Milwaukee, WI, USA

Madeline M. Carrig, Ph.D.
Department of Psychology and Neuroscience
Duke University
Durham, NC, USA

Louis G. Castonguay, Ph.D.
Department of Psychology
Pennsylvania State University
University Park, PA, USA

Marc J. Diener, Ph.D.
Clinical Psychology Doctoral Program
Long Island University, Post
Brookville, NY, USA

Robert Dobrenski, Ph.D.
Department of Psychology
University of Toledo
Toledo, OH, USA

Ann M. Doucette, Ph.D.
Midge Smith Center for Evaluation Effectiveness
The George Washington University
Washington, DC, USA

Michael Eid, Ph.D.
Department of Psychology and Educational Science
Freie Universität Berlin
Berlin, Germany

Robert Elliott, Ph.D.
School of Psychological Sciences and Health
University of Strathclyde
Glasgow, United Kingdom

Constance T. Fischer, Ph.D.
University of Kentucky
Pittsburgh, PA, USA

Jonathan L. Fishman, Ph.D.
Department of Psychology
Miami University
Oxford, OH, USA

Robert Gallop, Ph.D.
Department of Mathematics, Applied Statistics Program
West Chester University
West Chester, PA, USA

Nicola Gazzola, Ph.D.
Faculty of Education
University of Ottawa
Ottawa, ON, Canada

Christian Geiser, Ph.D.
Department of Psychology
Utah State University
Logan, UT, USA

Leslie S. Greenberg, Ph.D.
Department of Psychology
York University
Toronto, ON, Canada

Mark J. Hilsenroth, Ph.D.
Derner Institute of Advanced Psychological Studies
Adelphi University
Garden City, NY, USA

William T. Hoyt, Ph.D.
Department of Counseling Psychology
University of Wisconsin-Madison
Madison, WI, USA

Vanessa Illing, Ph.D.
Department of Psychology
University of Ottawa
Ottawa, ON, Canada

Shigeru Iwakabe, Ph.D.
Graduate School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan

Anthony S. Joyce, Ph.D.
Department of Psychiatry
University of Alberta
Edmonton, Canada

Alan E. Kazdin, Ph.D.
Department of Psychology
Yale University
New Haven, CT, USA

David A. Kenny, Ph.D.
Department of Psychology
University of Connecticut
Storrs, CT, USA

Sarah Knox, Ph.D.
Department of Counselor Education and Counseling Psychology
College of Education
Marquette University
Milwaukee, WI, USA

Gregory G. Kolden, Ph.D.
Department of Psychiatry
University of Wisconsin Hospital and Clinics
Madison, WI, USA

Michael J. Lambert, Ph.D.
Department of Psychology
Brigham Young University
Provo, UT, USA

Scott Leon, Ph.D.
Department of Psychology
Loyola University
Chicago, IL, USA

Kenneth N. Levy, Ph.D.
Department of Psychology
Pennsylvania State University
University Park, PA, USA

Eric G. Loken, Ph.D.
Department of Human Development and Family Studies
Pennsylvania State University
University Park, PA, USA

Wolfgang Lutz, Ph.D.
Department of Psychology
University of Trier
Trier, Germany

Zoran Martinovich, Ph.D.
Department of Psychiatry and Behavioral Sciences
Northwestern University
Evanston, IL, USA

Stanley B. Messer, Ph.D.
Graduate School of Applied and Professional Psychology
Rutgers University
Piscataway, NJ, USA

Joel M. Monroe, Ph.D.
American School of Professional Psychology
Argosy University
Washington DC
and Broughton Hospital, Morganton, NC, USA

Susan L. Morrow Ph.D.
Department of Educational Psychology
University of Utah
Salt Lake City, UT, USA

Fridtjof W. Nussbeck, Ph.D.
Department of Psychology
University of Bielefeld
Bielefeld, Germany

Benjamin M. Ogles, Ph.D.
College of Family, Home, and Social Sciences
Brigham Young University
Provo, UT, USA

John S. Ogrodniczuk, Ph.D.
Department of Psychiatry
University of British Columbia
Vancouver, BC, Canada

Katerine Osatuke, Ph.D.
Department of Psychology
Miami University
Oxford, OH, USA

Rhea Partyka, Ph.D.
Department of Psychology
University of Toledo
Toledo, OH, USA

Antonio Pascual-Leone, Ph.D.
Department of Psychology
University of Windsor
Windsor, ON, Canada

Juan Pascual-Leone, Ph.D.
Department of Psychology
York University
Toronto, ON, Canada

Fabian Ramseyer, Ph.D.
University Hospital of Psychiatry
University of Bern
Bern, Switzerland

Stephen M. Saunders, Ph.D.
Department of Psychology
Marquette University
Milwaukee, WI, USA

Hugo J. Schielke, Ph.D.
Department of Psychology
Miami University
Oxford, OH, USA

William B. Stiles, Ph.D.
Department of Psychology
Miami University
Oxford, OH, USA

Timothy J. Strauman, Ph.D.
Department of Psychology and Neuroscience
Duke University
Durham, NC, USA

Niklaus Stulz, Ph.D.
Psychiatric Services Aargau and Department of Psychology
University of Bern
Bern, Switzerland

Giorgio A. Tasca, Ph.D.
Institute of Mental Health Research
University of Ottawa
Ottawa, ON, Canada

Ladislav Timulak, Ph.D.
School of Psychology
Trinity College
Dublin, Ireland

Wolfgang Tschacher, Ph.D.
University Hospital of Psychiatry
University of Bern
Bern, Switzerland

John Wagner, Ph.D.
Dialectical Behaviour Therapy Centre of Vancouver, Inc.
Vancouver, BC, Canada

Rachel H. Wasserman, Ph.D.
Cambridge Health Alliance
Department of Psychiatry
Division of Psychology
Cambridge, MA, USA

Jeanne C. Watson, Ph.D.
Ontario Institute for Studies in Education
Toronto, ON, Canada

Joel Weinberger, Ph.D.
Department of Psychology
Adelphi University
Garden City, NY, USA

Elizabeth Nutt Williams, Ph.D.
St. Mary's College of Maryland
St. Mary's City, MD, USA

Abraham W. Wolf, Ph.D.
Center for Marital and Sexual Health
Case Western Reserve University
Cleveland, OH, USA

Preface

The content of this book consists of two parts, paralleling the structure of the 2009 issue of *Psychotherapy Research* for which this volume serves as an expansion and update. First, we focus on issues related to quantitative research, which relies on the transfer of observations into numbers and statistical analyses to specify findings. Quantitative methods reflect the predominant paradigm used for psychotherapy research since its start approximately 100 years ago. Quantitative methods have become increasingly complex over the last ten years, however, and hence there are many new issues of which a psychotherapy researcher should be aware; such issues are covered within this book. Chapters on quantitative methods are further organized into three subsections: (a) developing new and improving existing measures in psychotherapy research, (b) identifying and analyzing change in psychotherapy, and (c) aggregating research findings via meta-analysis. In the second part of the book, we discuss qualitative and mixed methods, which arose in part as a juxtaposition to quantitative methods, and have recently gained increasing presence in the field. Qualitative methods rely on words, narratives, and clinical judgment, thereby restoring a humanistic quality to research. Currently, both quantitative and qualitative methods are regarded positively within psychotherapy research. Although individual researchers often prefer one approach over the other, they nevertheless should know as much as possible about both methods.

We hope that this book will stimulate not only better research, but also more research on research methods themselves. In addition, we hope that this book stimulates thoughtful discussion regarding the advantages and disadvantages of both approaches to research, and a recognition of how each method may enhance our understanding of psychotherapy. Producers, consumers, and instructors of psychotherapy research will thus find in one volume a wide but related range of content that addresses current topics in the field. Much of this content speaks to cutting-edge areas in psychotherapy research (e.g., item response theory and multimethod analyses to assess and improve measurement validity; use of fMRIs to assess therapy process and outcome; sophisticated analyses that enable examination of longitudinal data while also integrating different levels of change; advances in meta-analyses techniques for both quantitative and qualitative data; a growing number of rigorous qualitative research methods; increased attention to the

trustworthiness of qualitative data). The book also provides many links to websites of program syntax in R, HLM, Mplus, SPSS, or SAS, with additional material for interested readers to get a start for their own analysis. Some of this additional material, especially for the quantitative chapters, is only referenced in the text body and can be found on the following website: www.methodsbook.uni-trier.de (User: book; Password: methods). Therefore, perhaps the volume's primary contribution and the need it best addresses is assembling in one place the most recent and most "hot" topics of interest to producers, consumers, and instructors of psychotherapy research.

We would like to thank all contributing authors for their participation, careful work, and inspiring chapters. Their expertise, acquired over many years in the profession, provides the combination of breadth and depth that makes this volume useful for both beginning and advanced researchers. We are likewise pleased that Jane Madeley, of Taylor & Francis, asked us to edit this book on *Quantitative and Qualitative Research Methods in Psychotherapy Research*, and want to thank her for her support. We would also like to thank Dipl.-Psych. Antje Welscher, Dipl.-Psych. Julian Rubel, B.Sc. Hanna Epping, B.Sc. David Gerhard, B.Sc. Julia Maria Kaspar, B.Sc. Annegret Knape, B.Sc. Svea Susan Schmidt, and B.Sc. Andreas Martin Siegbert for their help in some of the technical elements of bringing this book to press.

Wolfgang Lutz and Sarah Knox
May, 2013

Abbreviations

AAS	Adult Attachment Scale
AHQ	Attachment History Questionnaire
AIC	Akaike Information Criterion
AICC	Akaike Information Corrected Criterion
ANCOVA	analysis of covariance
ANOVA	analysis of variance
APA	American Psychological Association
APES	Assimilation of Problematic Experiences Scale
AT	Affirmative Team
BDI	Beck Depression Inventory
BED	binge eating disorder
BEDS	Binomial Effect Size Display
BDNF	brain-derived neurotrophic factor
BIC	Bayesian Information Criterion
BOLD	blood-oxygen-level dependent
BORTTI	Bell Object Relations and Reality Testing Inventory
BSI	Brief Symptom Inventory
CALPAS-G	California Psychotherapy Alliance Scale—Group patient version
CAMS	Classification of Affective-Meaning States
CAQ	Components of Attachment Questionnaire
CATS	Client Attachment to Therapist Scale
CBT	cognitive–behavioral therapy
CCT	Cognitive Control Training
CFA	confirmatory factor analysis
CGM	centering at the grand mean
CQR	Consensual Qualitative Research
CT	cognitive therapy
CTT	classical test theory
CWC	centering within cluster
DBT	Dialectical Behavior Therapy
ECRS	Experiences in Close Relationships Scale
EEG	electroencephalography

EFT	emotion-focused therapy
EN	Edwards–Nunnally method
ES	effect size
EST	empirically supported treatment
ETR	expected treatment response
fMRI	functional magnetic resonance imaging
GAS	goal attainment scale
GCBT	Group Cognitive Behavioral Psychotherapy
GEE	Generalized Estimating Equations
GLM	general linear model
GLN	Gulliksen–Lord–Novick method
GMM	Growth Mixture Models
GPIP	Group Psychodynamic Interpersonal Psychotherapy
GSI	General Severity Index
GT	Grounded Theory
HA	Hageman–Arrindell method
HAT	Helpful Aspects of Therapy
HLM	hierarchical linear model
HRSD	Hamilton rating depression scale
HSCED	Hermeneutic Single Case Efficacy Design
ICC	intraclass correlation coefficient
ICC	item character curve
ICT	integrative cognitive therapy
IIP	Inventory for Interpersonal Problems
IPT	interpersonal therapy
IQ	intelligence quotient
IRC	item response curve
IRF	item response function
IRT	item response theory
ITP	interpersonal psychotherapy
LL	log likelihood
LOCF	last observation carried forward
LSQ	Life Status Questionnaire
LST	latent state–trait
MAO-A	monoamine oxidase A
MAR	missing at random
MCAR	missing completely at random
MHI	Mental Health Index
MINQUE	minimum norm quadratic unbiased estimation
MIRT	multidimensional item response theory
MIVQUE	minimum variance quadratic unbiased estimation
ML	maximum likelihood
MLM	multilevel models/multilevel modeling
MQL	marginal-quasi likelihood
MRI	magnetic resonance imaging

MSE	mean square error
MTMM	multitrait–multimethod
MTMM–MO	multitrait–multimethod–multioccasion
NIMH	National Institute for Mental Health
NN	nearest neighbors
NNT	number needed to treat
OLS	ordinary least squares
OQ	Outcome Questionnaire
PCA	principal–components analysis
PET	positron emission tomography
PFC	prefrontal cortex
PPRS-BPD	Psychotherapy Process Rating Scale for Borderline Personality Disorder
PQ	Personal Questionnaire
PQL	penalized quasi-likelihood
QOR	quality of object relations
QOROM	Quality of Reporting of Meta-Analyses
rANOVA	repeated measurement ANOVA
RC	reliable change
RCI	reliable change index
RCT	randomized clinical trial
RCT	randomized controlled trial
REML	restricted maximum likelihood
RQ	Relationship Questionnaire
RSQ	Relationship Styles Questionnaire
SEM	standard error of measurement
SL	Schmid–Leiman method
SPT	Supportive Psychotherapy
SRC	scale response curve
ST	Skeptic Team
STAI	State–Trait Anxiety Inventory
STG	short-term group therapy
TCC	test characteristic curve
TDCRP	Treatment of Depression Collaborative Research Program
TESF	Therapist Experiential Session Form
TFP	Transference Focused Psychotherapy
TIF	test information function
TR	repetition time
TSPA	time-series panel analysis
VAR	vector autoregression
VEV	Veränderungsfragebogen des Erlebens und Verhaltens
WAI	Working Alliance Inventory

This page intentionally left blank

1 Quantitative and qualitative methods for psychotherapy research

Introduction

Wolfgang Lutz and Sarah Knox

We are pleased that, because the Special Issue on "Quantitative and Qualitative Methods" that appeared in 2009 in the journal *Psychotherapy Research* was such a success, the publisher (Taylor & Francis) has asked us to develop an update in the form of a book. Many psychotherapy researchers throughout the world have used that special issue to teach classes on psychotherapy research methods and to expose students to the most exciting developments in the area. Now with this volume, such vital content is available as a user-friendly book, thereby providing three advantages: First, the material from the Special Issue has been updated to reflect recent developments in the field since 2009; second, presenting the material in book format is ideally suited for classroom teaching; and third, additional material for some of the quantitative chapters that provides support for researchers seeking to begin their own analysis in a specific area, or that gives helpful links to additional material, can be found on the following website: www.methodsbook.uni-trier.de (User: book; Password: methods).

We are thus excited to offer this volume, and hope that it will be helpful for the field of psychotherapy research. The book can be used as a resource for planning and designing new studies, for analyzing the results of existing data, and as a basis for teaching advanced methods in psychotherapy research.

The volume can be roughly divided into two parts. In the first part, we focus on issues related to quantitative research, which relies on numbers and statistical analyses. Quantitative methods reflect the predominant paradigm used for psychotherapy research since its inception over a hundred years ago. Quantitative methods have become increasingly complex, and hence there are many issues about which the average psychotherapy researcher needs to be aware. In the second part, we focus on qualitative and mixed methods, which have arisen more recently, and offer additional ways of increasing our understanding of psychotherapy. Qualitative methods rely more on words, narratives, and clinical judgment, bringing back some of the humanistic quality to research. At this point, both quantitative and qualitative methods are regarded positively within psychotherapy research, although individual researchers often prefer one approach over the other.

Quantitative methods

The goal of psychotherapy research is to advance our knowledge about the process, as well as the course and the outcome, of psychotherapy. Researchers try to identify the best treatment options possible for patients with a given problem, disorder, or set of problems or disorders. Ideally, then, we can select optimal treatments for individual patients. Quantitative research methods are helpful tools for achieving these goals because they enable us to study the complex relations between the patient, the therapist, the process of therapy, external events in the lives of patients, and in-session progress, post-session progress, and therapy outcome at the end of treatment as well as at follow-up. Such methods also help us aggregate and integrate findings about psychotherapy (e.g., via meta-analysis).

The quality and scope of the 14 chapters on quantitative methods depict the progress in the areas just described. These articles are grouped into three major categories: (a) developing new and improving existing measures in psychotherapy research, (b) identifying and analyzing change in psychotherapy, and (c) aggregating research findings via meta-analysis.

Measure development

The clinical and scientific value of the psychotherapy research enterprise depends on the validity of our measures. One trend that can be seen in the field is the development of new research tools (e.g., IRT or multitrait–multimethod analysis) to improve the validity of our measures.

The papers on measurement issues start with Doucette and Wolf (Chapter 2), who discuss advances in latent trait and item response theory (IRT) and their advantages over classical test theory. In the next paper, Eid, Geiser, and Nussbeck (Chapter 3) discuss multitrait–multimethod procedures and their implications for test validity, multimethod assessment, and psychotherapy research in general. The chapter by Wasserman, Levy, and Loken (Chapter 4) then introduces generalizability theory as a framework within which multiple sources of error can be simultaneously evaluated; generalizability theory allows researchers to improve the accuracy of reliability estimates and provides critical information for the modification of coding procedures in psychotherapy research. Carrig, Kolden, and Strauman (Chapter 5) then discuss functional magnetic resonance imaging (fMRI), a new methodological tool to assess outcome and process in psychotherapy, one with the potential to provide new insights for psychotherapy research.

Identifying and analyzing change

A decade or so ago, researchers were pleased when they were able to demonstrate the average difference between two groups using only pre-post change as put forward by Beutler and Howard (1998) and Newman and Howard (1991). Since that time, several advancements have been made in quantitative methods that allow us to complete more sophisticated analyses, such as examining longitudinal data

on the course of psychotherapy, as well as integrating several levels of change (e.g., sessions, patients, treatments) and differences across therapists. These new developments are particularly important because we may soon be fortunate enough to have data for thousands of patients, arising from the new developments in scientist-practitioner networks or patient-focused research (e.g., Castonguay, Barkham, Lutz, and McAleavey, 2013; Lambert, 2001, 2007, 2013; Lutz, 2002). To be able to analyze such vast data, we need new tools to aggregate the information without neglecting interindividual differences.

In the first paper on how to identify and to analyze change, Kazdin (Chapter 6) discusses moderators, mediators, and mechanisms of change, and shows promising lines of work to better identify these components of the change process. This paper pinpoints the conceptual and research difficulties in studying change mechanisms and presents recommendations for future research on how and why therapy works. Anecdotally, this paper was one of the most downloaded papers of the 2009 special series and addresses the pressing questions related to the identification of central change agents in psychotherapy. The set of papers by Tasca and Gallop (Chapter 7), Gallop and Tasca (Chapter 8), and Tasca, Illing, Joyce, and Ogrodniczuk (Chapter 9) introduce fundamentals and complexities of multilevel models (MLM), as well as a three-level growth MLM approach for the analysis of longitudinal data and nested data in general. These modern sophisticated statistical tools allow researchers to model individual change and group change, and provide new opportunities for handling missing data in longitudinal designs and nested designs. Kenny and Hoyt (Chapter 10) extend the multilevel approach to the analysis of group as well as rolling group data. They also give practical guidelines on how to conduct multilevel analyses within a statistical package (SPSS), and include a link to the website that provides the R-syntax on how to conduct such analyses. Tschacher and Ramseyer (Chapter 11) then introduce the methodology of aggregated time-series analysis (time-series panel analysis, TSPA), which allows for the identification of prototypical and fine-grained process patterns to approximate causal dynamic structures.

The two final papers in this section focus on the evaluation of progress. We start with a debate about a common method to assess and evaluate clinically significant change (Lambert and Ogles, Chapter 12). These authors suggest using Jacobson and Truax's (1991) method as a standard way of defining clinical significance in psychotherapy research. The use of a definition and classification of clinically meaningful change on an individual basis is essential in reporting and communicating results in efficacy and effectiveness studies, as well as in patient-focused research.

Lutz, Stulz, Martinovich, Leon, and Saunders (Chapter 13) then present different kinds of decision rules as the basis for the evaluation of progress and the application of feedback tools. Rational as well as empirical approaches are discussed, and examples as well as material for the application of Growth Mixture Models (GMM) using Mplus are provided.

The aggregation of research findings via meta-analysis

Finally, the field now benefits from new developments in empirically aggregating information over many studies via meta-analysis. In comparison to the original meta-analysis by Smith, Glass, and Miller (1980), several advancements have been made that allow us to more precisely define weighted effect size, the statistical significance of effects, as well as tests of homogeneity (including file-drawer analysis and moderator variables). Meta-analysis is an important tool, one that not only allows researchers to aggregate information over hundreds of studies, but also allows them to demonstrate the efficacy and effectiveness of psychotherapy in comparison to other treatments (e.g., medical, psychopharmacological).

Berkeljon and Baldwin (Chapter 14) provide an introductory tutorial about conducting meta-analysis in psychotherapy outcome research. Their central topics involve identifying and collecting studies, coding effect sizes, coding substantive and methodological information, combining effect sizes, and interpreting effect sizes; moderator analyses are also introduced. Finally, Diener, Hilsenroth, Weinberger, and Monroe (Chapter 15) provide a primer on using meta-analysis for correlation coefficients. Based on an example of the relationship between patient-reported therapeutic alliance and adult-attachment style, they demonstrate aspects and calculations of the weighted average effect size, the statistical significance of effects, a test of homogeneity, confidence intervals, and file-drawer analysis.

Qualitative and mixed methods

Despite the many strengths of rigorous quantitative methods, some researchers have long expressed dissatisfaction with such empirical approaches, particularly for investigating psychotherapy process (e.g., Goldman, 1976, 1979). The standard for such research in those days was that you could only publish something if you observed it and measured it reliably and validly. Unfortunately, this demand often forced researchers to study relatively trivial things because they could easily be seen and coded (e.g., head nods), and left unexamined the vast amount of clinically rich data from psychotherapy. Kiesler (1973) aptly summarized this dilemma: "If you can't count it, it doesn't count; if you can count it, that ain't it" (p. 16). This rift led many clinicians to bemoan the gap between science and practice (see Morrow-Bradley and Elliott, 1986).

As a result, some in the field of psychotherapy research became excited when learning of qualitative methods used by colleagues in education and anthropology. Indeed, many people wrote about the promise of qualitative research to enhance our understanding of psychotherapy (Borgen, 1992; Hill and Gronsky, 1984; Hoshmand, 1989; Howard, 1983; Neimeyer and Resnikoff, 1982; Polkinghorne, 1984). Although it took some time before people developed rigorous approaches suitable for psychotherapy research, we now have a number of sound methods: comprehensive process analysis (Elliott, 1989), consensual qualitative research (Hill et al., 2005; Hill, Thompson, and Williams, 1997), grounded theory (Rennie, Phillips, and Quartaro, 1988; Strauss and Corbin, 1990, 1998), and phenomenological approaches (Giorgi, 1985).

Methodological developments in qualitative research

Qualitative psychotherapy research methods continue to evolve and be refined. We highlight here new developments within task analysis (Pascual-Leone, Greenberg, and Pascual-Leone, Chapter 16), a method that "pushes the envelope" in combining aspects of qualitative and quantitative approaches to explain therapy processes. Task analysis procedures have existed for some time (see Greenberg, 2007), and these authors describe how this approach can be used in a programmatic way to study client change processes. In addition, Elliott et al. (Chapter 17) present a mixed-method approach that involves hermeneutic case studies, borrowing from legal processes to use arguments to determine the weight of the evidence about whether change has occurred in psychotherapy. Next, Schielke, Fishman, Osatuke, and Stiles (Chapter 18) present an intriguing new approach borrowed from architecture (called the Ward method) to help researchers more effectively integrate the multiple voices inherent on a research team as they seek to understand phenomena.

A major complaint about qualitative research has been the difficulty of aggregating findings across projects. For example, how do researchers compare words to discern whether they have the same meaning across studies? Fortunately, researchers have been developing methods of qualitative meta-analysis or metasynthesis to address this concern. Timulak (Chapter 19), for instance, discusses how to perform a rigorous secondary qualitative analysis of primary qualitative findings. In addition, he also proposes a means to examine, again via a qualitative meta-analysis, the effects of the method on the findings themselves. Iwakabe and Gazzola (Chapter 20) likewise present ideas about combining results across single case studies.

Methodological issues in qualitative research

As qualitative psychotherapy research matures, we must also consider important methodological questions. First, we include a chapter by Knox and Burkard (Chapter 21) on the interviewing process, given that many of the qualitative methods rely on interviews as their means of collecting data. Williams and Morrow (Chapter 22) then examine the construct of trustworthiness in qualitative data, another crucial element worthy of consideration. Finally, yet another vital concept in qualitative research is bracketing, or becoming aware of biases and expectations and setting them aside so that they do not unduly influence data collection or analysis. Fischer (Chapter 23) addresses the philosophical and methodological issues related to bracketing.

We are indeed excited to witness both the evolution of new quantitative methods, as well as the evolution from sole reliance on quantitative methods to inclusion of qualitative methods as viable means of conducting psychotherapy research (see Chapter 24 by Knox and Lutz). We are also honored to continue the thoughtful debate about research methods. We hope that this book will stimulate not only better research, but also more research on research methods themselves.

This page intentionally left blank

Part I

Quantitative methods

Developing new and improving existing measures

This page intentionally left blank

2 Increasing measurement precision in psychotherapy research

Item response theory and bifactor models

Ann M. Doucette and Abraham W. Wolf[1]

Questioning the measurement precision of psychotherapy research

The value of psychotherapy has been characterized in terms of the beneficial change clients experience as a result of therapeutic intervention. Although there is now general agreement that psychotherapy yields favorable effects for those seeking treatment for emotional and psychological distress (Campbell, Norcross, Vasquez, and Kaslow, 2013; Lambert and Bergin, 1994; Lipsey and Wilson, 1993, Miller, Hubble, Chow, and Siedel, 2013; Smith, Glass, and Miller, 1980), there is much debate on how to build an adequate evidence base on which to investigate *what works for whom*, *under what conditions*, and *why* (Howard, Orlinsky, and Lueger, 1995; Margison et al. 2000; Roth and Fonagy, 1996; Stiles, 2013). This debate is even more pronounced as those paying for psychotherapy treatment, private insurance and public support (Medicaid and Medicare) rely on outcome measures as primary indicators of treatment need, and evidence of treatment effectiveness, begging the question of whether treatment need and effectiveness can be adequately reduced to a single or parsimonious set of numbers (McElvaney and Timulak, 2013). Rarely is the set of items yielding an outcome score questioned in terms of its sufficiency in capturing patient status or the changes that occur as attributable to treatment. Seldom is the set of items yielding an outcome score, examined in terms of its appropriateness in reflecting changes in the precipitating problem that brought the patient to treatment. Instead, these considerations are taken for granted and accepted as being satisfactory.

We can no longer simply assume that the measures used in psychotherapy research are adequate markers of treatment need and its effectiveness. Measures used in sophisticated analytic models allowing us to parcel out the variance attributed to client and therapist characteristics and the contribution of specific treatment approaches must be scrutinized in terms of their precision. The decisions made based on outcome measures must be in accordance with that precision—a precision that is not likely known or even pondered at the time decisions are made.

The chapter addresses the critical importance of measurement in psychotherapy research and presents an overview of item response theory (IRT: Hambleton, Swaminathan, and Rogers, 1990; Wright and Stone, 1979) and bifactor models (Reise, 2013) as a more comprehensive measurement model approach for psychotherapy outcome research. Basic principles and assumptions of the single and multi-parameter IRT and bifactor models are described. The advantages of IRT and bifactor models will be illustrated in a reanalysis of a psychotherapy outcome measure that was originally developed using classical test theory measurement (CTT: Gulliksen, 1950; Lord, 1980; Novick, 1966). The CTT measurement model is not addressed in detail, given its prominent use and coverage in the measurement literature (Crocker and Algina, 1986; Lord and Novick, 1968).

Measurement models

> Measurement, as a scientific method, is a way of finding out (more or less reliably) what level of an attribute is possessed by the object or objects under investigation . . . the magnitude of a level of an attribute via its numerical relation (ratio) to another level of the same attribute.
>
> (Michell, 2001, p. 212)

Classical test theory (CTT) and latent trait models

Classical test theory

There are two distinct approaches to measurement, CTT and latent trait models (Lazarsfeld and Henry, 1968). Essentially CTT, also known as *true score theory*, posits that measurement is an additive composition of the respondent's true ability plus random error ($X = T + E$). Error is assumed to be random, uncorrelated with the true score, equivalent across all sample respondents, and as a consequence, follows a normal distribution. This assumption is the axiomatic foundation of CTT. The emphasis of CTT is not on individual item or measure scores, but rather on the properties of measure scores relative to samples of individual respondents. Fan (1988) characterizes CTT as having a *circular dependency*, where the quality of the measure is dependent on the response sample, and the respondent scores are dependent on the quality of the items making up the measure.

Latent trait

In contrast, latent trait models, also known as *strong true score theory*, are the foundation of IRT measurement models. Latent traits are measured through the direct assessment of observed values of categorical indicators (e.g., level of sadness, hopelessness endorsed by the respondent) believed to represent the unobserved construct. From an IRT perspective, each item is assumed to measure a specific and unique facet of the latent construct of interest at varying levels (e.g., mild to

severe impairment, low to high service need, etc.). Although psychotherapy research customarily utilizes summative scores, IRT item-level estimates allow us to determine the contribution of each item to the overall measurement of the latent trait. For example, an item asking about suicidal intent on a depression scale would reflect a more substantive contribution to severe depression than would an item indicating feeling blue once in a while, offering a more comprehensive opportunity to investigate and interpret psychotherapy outcomes.

One of the most important advantages of the IRT model is the transformation of item responses to a scale-free metric of the latent trait. Item-level estimates are modeled in terms of logistic probability distributions based on *ability*, an individual's location on the latent construct (e.g., mild, moderate, severe psychological distress) and the *difficulty* of the item in terms of its location on the measured construct—mild to severe. For example, individuals with moderate distress would be expected to positively endorse items reflecting mild and moderate distress, but would have low probability of positively endorsing items reflecting profound distress (such as suicidality) independent of the instrument used and measurement occasion. IRT measures are considered *sample invariant* based on this logistic probability (Wright and Douglas, 1977). IRT measures yield item and latent trait estimates that do not vary as a result of respondent samples, as do CCT measures. Individuals with moderately high levels of depression would be expected to have similar response profiles—trait levels (moderately high) linked to specific item content.

In addition to examining item-level properties, IRT offers the advantage of comprehensively examining the adequacy of the response scale properties of a measure (Wright, 1977; Yen, 2005). Statistical methods are impervious to the level of measurement (nominal, ordinal, interval, and ratio) and response scale imprecision (Baker, Hardyck, and Petrinovich, 1966). In many instances, if not most, response scale data are treated as interval, assumed to meet statistical assumptions; ordered monotonically; and, are adequate in segmenting the sample into distinct groupings along the measured attribute (e.g., low, moderate, and high; strongly agree to strongly disagree, etc.).

While the advantages of the IRT approach seem obvious (Embretson and Reise, 2000), it is important to remember that the advantages of IRT models are only realized if the data (measure scores) fit the measurement model assumptions (Reise and Haviland, 2005).

Item response theory fundamentals

Although the use of IRT models is relatively new to the psychotherapy research literature, IRT has a longer history than is evident in the psychotherapy research literature (Loevinger, 1957; Lord, 1953; Rasch 1960/1980). The slow uptake of IRT as a prominent set of measurement models is largely due to its computational complexity and until the last decade a limited availability of user-friendly analytic software (e.g., Winsteps, Parscale, Bilog, Multilog). In the last decade there has been substantial growth in the use of IRT measurement models in behavioral and physical healthcare investigations (Becker et al., 2007; Burlew, Feaster, Brecht, and

Hubbard, 2009; Cella et al., 2007; Sotsky et al., 2006; Streiner, 2013; Liu and Verkuilen, 2013).

There are several IRT models, which differ in terms of approach and parameterization (Rasch, 1PL, 2PL, 3PL, and 4PL). In addition, recent developments in IRT measurement approaches have resulted in nonparametric in addition to the more traditional parametric approaches (Junker and Sijtsma, 2001; Sijtsma and Meijer, 2007; Thissen and Steinberg, 1986). The essential difference between these approaches is as follows. The parametric IRT approach tests the null hypothesis in terms of model assumptions. For example, a parametric approach would accept or reject the assumption of unidimensionality. In contrast, the nonparametric IRT approach does not assume model assumptions such as unidimensionality, but instead, observes the data as a point of departure and examines the data to determine its true dimensionality. The distinction between the two approaches could be described in terms of a confirmatory versus an exploratory approach. The focus of the chapter is on parametric IRT models; however, it is well recognized that nonparametric IRT models offer increased flexibility in terms of data analyses, based on less restrictive assumptions that typically lead to the inclusion of more items in the scale. While many IRT models exist, there are assumptions that are shared across these models: unidimensionality, local independence, and monotonicity.

IRT model assumptions

Unidimensionality

A core assumption of IRT is that the measured construct is unidimensional, measuring a single measured trait/ability, and that this trait/ability accounts for all item intercorrelations (McDonald, 1981). If a measure is considered unidimensional, the scores across all of the items (total score) can be used to characterize an individual on the measured attribute. If the measure is multidimensional, it is more accurate to use subsets of item scores (e.g., domain, subscales, etc.) to reflect the measured attribute, as a single score would not be reflective of a specific trait/ability, for example, depression, anxiety and the like. It is important to note that unidimensionality is never perfect; it is always approximate. The most important question is whether deviation from unidimensionality is substantial enough to warrant the construction of two or more subscales representing distinctions within the measure. Although multidimensional IRT models (MIRT) have been developed (Bartolucci, 2007; Embretson and Yang, 2013; Christensen, Bjorner, Kreiner, and Petersen, 2002; Reckase, 1997; Zwick, 1987) over the past two decades, IRT approaches to multidimensional data are complex, and a detailed explanation is beyond the scope of this chapter.

While confirmatory factor analysis (CFA) can be conducted to examine the extent to which the items measure a dominant factor, newer methods are available

to examine measurement data that are thought to be multidimensional. These include the bifactor model (Bock et al., 2002; Gibbons et al. 2007; Gibbons and Hedeker, 1992; Immekus, 2013; Immekus and Imbrie, 2008; Reise, 2013; Reise and Haviland, 2005). The bifactor structure constrains the items to have a non-zero loading on the dominant factor, and at most one group factor. There may be several group factors in the model. Another alternative is the *testlet* approach which parcels items together into testlets that have common content (Wainer and Kiely, 1987; Wilson and Adams, 1995). The testlets then become the measure. There is an underlying assumption that the item/testlet composition is known in advance of IRT analysis.

Local independence

Local independence assumes that the dominant factor is the sole source of influence in how a person responds to an item. The response to an item is independent (there is no significant association) of responses to other items in a scale after controlling for the latent trait (dominant factor) measured by the scale (Wainer and Thissen, 1996; Yen, 1993). In other words, the probability of a person endorsing an item is determined by his/her ability level, and is not influenced by item content. By default, when the assumption of unidimensionality is achieved, local independence is obtained (McDonald, 1981). A solution to resolving violations of unidimensionality and local independence is removal of the offending item(s) from the scale. As noted previously, unidimensionality is never perfect. The question is always, is the deviation from unidimensionality substantive enough to degrade the measurement model (addressed more specifically in chapter section on the bifactor model).

Monotonicity

As noted previously, the probability of endorsing an item response is directly related to the person's ability. A person experiencing a lessening of depression should accordingly select item responses indicative of improved psychological function. Monotonicity is easily examined graphically; plotting item mean scores conditional on rest-scores (total raw score minus the item score). In the graph of a well-fitting item, the conditional item mean systematically increases with one unit change in rest-score level (Reeve et al., 2007). For example, an individual having severe depression will respond to items located at or around that trait level using the same response option. In other words, an individual consistently selects response options (e.g., strongly agree to strongly disagree, none of the time to almost always, etc.) that reflect his/her level on the measured trait—a person selecting *strongly agree* to feeling worthless would select accordingly to an item asking about having no sense of purpose.

IRT basics

Person–item map

The fundamental purpose of measurement is to objectively determine how much of a latent attribute a person has; to locate a person on a *measurement ruler* in terms of how much of the attribute of interest they possess, using responses to the set of items representing the latent attribute (see Figure 2.1). The person's location on this ruler is an estimate of their *ability* and the items are additive, meaning that the measure score is an additive function of the items in the measure. Items are also arrayed on this *ruler* in terms of their *difficulty*. To satisfy the additive property of measurement, the IRT *ruler* is standardized using a logit scale having equivalent units that could theoretically range from negative infinity to positive infinity, with the differences between adjacent values being equal. This scale is referred to using the Greek letter theta (θ). Items are also positioned in relationship to the same ruler on which persons are located in terms of their ability, a distinct advantage of the IRT model. Figure 2.2 depicts a person–item map, where the respondent sample is to the left of the axis, and the items are to the right (see Figure 2.2). In this example, both persons and items are on the same scale ranging from minimal/mild distress to greater distress.

Item characteristic curve (ICC)

In addition to the assumptions that IRT models share (unidimensionality, local independence and monotonicity), another element shared by all IRT models is the *item characteristic curve* (ICC), also referred to as the *item response function* (IRF) or *item response curve* (IRC). In IRT, item function is modeled using a cumulative form of the logistic function, an S shaped ogive. The items representing the latent attribute are located along the ruler in terms of *difficulty* (easy to hard, mild to severe). In the case of dichotomous responses, item difficulty is defined as the point on the ruler where the probability of a correct response is 0.5. In the case of rating scales using polytomous data, where the respondent identifies a level of response to the item stem (e.g., agree versus disagree, etc.), the IRT model establishes the relative

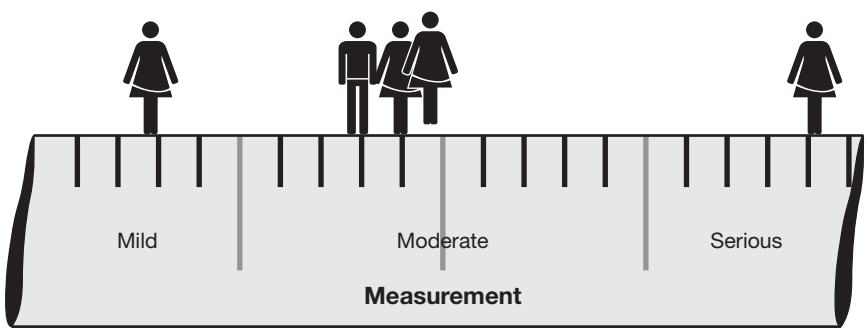

Figure 2.1 Measurement ruler

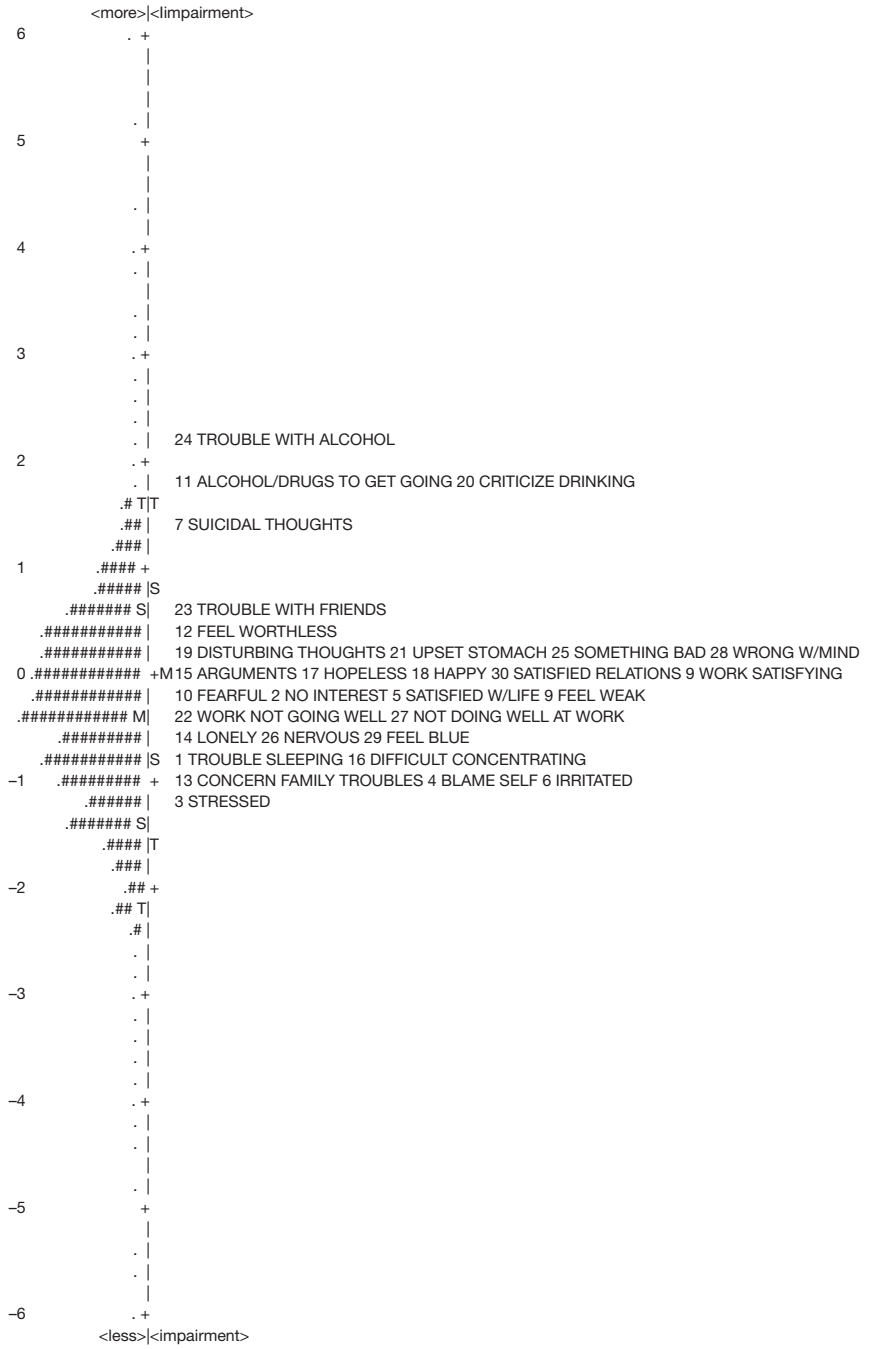

Figure 2.2 The IRT ruler: person–item map

Source: Reprinted from "Questioning the measurement precision of psychotherapy research," by A. Doucette and A. W. Wolf, 2009, *Psychotherapy Research*, *19* (4), p. 377. Copyright 2009 Routledge.

difficulty of the item, and models the pattern of use of the scale response categories, yielding a rating scale structure that is shared across all items in the measure. Persons with low ability (e.g., depression) would have high probability of correct or positive endorsement of easy items (e.g., feeling blue occasionally), but a low probability of correct or positive endorsement of more difficult items (e.g., suicidal intent). In the case of rating scale data, a person selecting "agree" on a 5-point Likert scale ranging from strongly disagree to strongly agree would be interpreted to imply that the respondent went beyond the response thresholds below agree (strongly disagree, disagree, neither disagree/agree), but failed to surpass the remaining threshold, strongly agree. The slope of the S shaped item curve changes as a function of ability, achieving a maximum when the item difficulty and ability levels are equal.

Item difficulty is located on the horizontal axis (theta θ) and found by dropping a vertical line from the inflection point (vertical axis, probability = 0.50). Item curves shift from left (easiest items) to right (more difficult items). Figure 2.3 provides ICCs for two hypothetical items. Item one has a difficulty estimate of –1.19 and item two has an estimate of 1.36. Item two is considered a more difficult item. An individual having a 50 percent probability of endorsing item two correctly/positively would have a 90 percent probability of endorsing item one, the easier item, correctly/positively (follow the dotted line from the inflection point of item two up to meet the item one ICC).

Item information function (IIF)

The item information function measures item precision and is one of the more important aspects of IRT. The IIF indicates how sensitive an item is in determining where a respondent is located on the IRT ruler (theta θ). At its maximum, the IIF locates the point on the latent trait θ at which an item provides the most information, and the level (or magnitude) of information that is provided at that point. Person ability (theta θ) is more precisely measured at the ability level that corresponds to the item's difficulty parameter. Item information decreases as the person ability level departs from the item difficulty. Information function approaches zero at the extremes of the ability scale. For example, little information is gained about the level of impairment that is experienced by an individual with mild depression who responds negatively to an item assessing suicidality, other than the individual is not suicidal. The negative response says nothing descriptive about the level distress experienced as a result of mild depression.

IIF is estimated using item parameters (the parameters used vary across the IRT models—these models will be described in the next section of this chapter), and is essential in informing measurement development. In developing measures (e.g. psychotherapy outcome measures) a measure developer could select items that will match the purposes of the test. For example, if a clinical cut-off score is used, the developer would purposefully select items that provide high information around the cut-off score area of the measured construct, in order to make a precise determination about whether responses satisfy the criteria set for a specific clinical

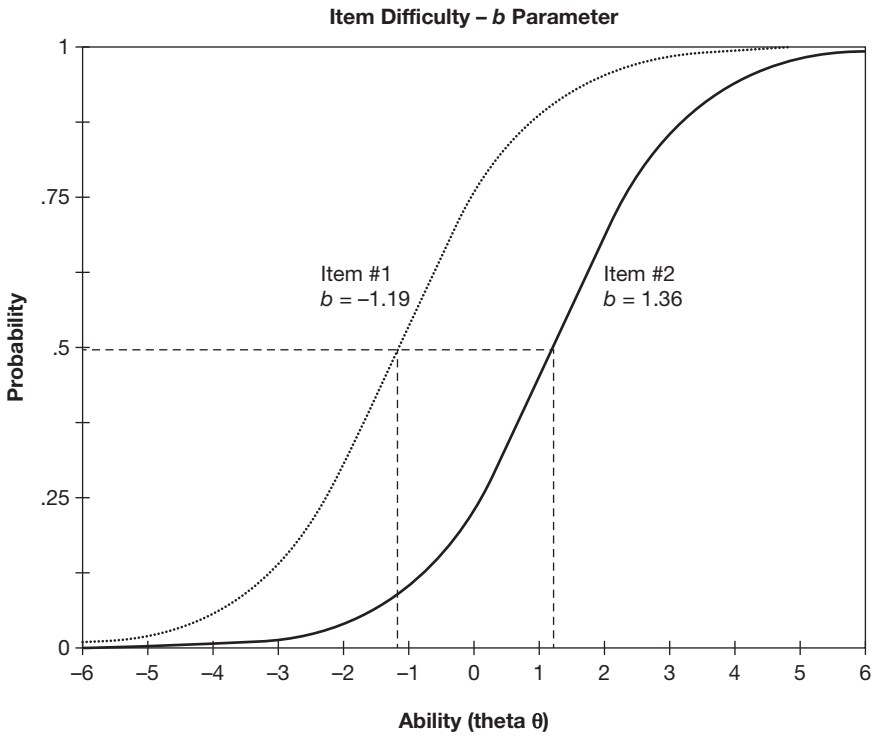

Figure 2.3 Example: item character curves (ICC)

Source: Reprinted from "Questioning the measurement precision of psychotherapy research," by A. Doucette and A. W. Wolf, 2009, *Psychotherapy Research, 19* (4), p. 378. Copyright 2009 Routledge.

cut-off score. Although the measurement ideal is to have a set of items with maximum bandwidth (i.e., items that are maximally effective at differentiating persons along a wide range of the latent trait), in most cases scale items vary considerably in terms of the amount of information they contain and where they cluster along the latent trait. (For a more extensive discussion of IRT parameters and information functions, see Reise and Henson, 2003.)

Test information function (TIF)

The information function at a specified ability level for a measure is the sum of the information across the items at that ability level. Test information will consequently be higher than the information function for a single item; the greater the number of items in a measure the greater the information for the test. Test information illustrates how well a measure estimates ability across the score range on ability. TIF curves may be peaked, indicating an ability to better distinguish groups in terms of selection or cut-off criteria at an identified point on the ability continuum (area under the peak). Precision at points along the attribute continuum,

more distant from the peak would be considerably less precise. Flatter TIF curves characterize measures having systematic levels of precision across a broad ability score range (e.g., mild to severe levels of impairment). Peaked TIF measures are precise in terms of identifying clinical versus nonclinical populations; however, they are far less precise in terms of information toward the tails (e.g., mild and severe impairment, etc.). If peaked TIF measures are used in longitudinal studies assessing change in emotional distress over time, the measures will be less sensitive, as individuals improve and/or deteriorate. As individuals move from one end of the continuum, improving as a result of psychotherapy, they would eventually move to a score range where the peaked test was less precise and subject to increased measurement error. It is critically important to determine the intent of the measure. Measures to determine treatment need more precision around established cut-off scores. Measures assessing change over time need broad bandwidth, measures that have sufficient items having sensitivity in detecting change across the measured construct (e.g., from mile to severe distress). Without sufficient measurement bandwidth, there may be insufficient items to yield precise estimates at the tails of the measured construct resulting in substantive inaccuracy regarding decisions made about the need and benefits of psychotherapy treatment.

IRT models

As noted earlier, there are essentially four unidimensional IRT models, characterized by the number of parameters that are included in the model. The Rasch and 1PL models include an item difficulty parameter that is allowed to vary as a function of ability (theta θ). The 2PL models measure data in terms of item difficulty and item discrimination, a second parameter. The 3PL and 4PL models incorporate guessing and carelessness parameters, respectively, and are seldom used in psychotherapy research. These models are most often used in educational assessment, in which individuals are encouraged to complete all items and to guess if they are unsure, and in personality testing to detect intentional response distortion (Chernyshenko, Oleksandr, Chan, Lee, and Drasgow, 2001; Ellis, Becker, and Kimmel, 1993; Rouse, Butcher, and Miller, 1999; Zumbo, Pope, Watson, and Hubley, 1997). Measurement models that include multiple parameters create challenges in interpreting measurement data. A distinct advantage of measures meeting Rasch and 1Pl model requirements is the fact that summative scale scores are sufficient in reflecting an individual's status along the measured attribute continuum.

Rasch and 1 parameter (1PL) models

There is much debate regarding the use of single parameter versus multiparameter models. Essentially, the Rasch and 1PL models question how well empirical data (measure scores/response patterns) fit in terms of the measurement model constraints. Multiparameter models ask an opposite question: How can additional

parameters be manipulated to increase model fit to the available data? These models attempt to explain deviation from the measurement model through the introduction of additional parameters, which are described later. Proponents of the 1PL approach argue that "the researcher's task is to work toward a better fit of the data to the model's requirements until the match is sufficient for practical measurement purposes in that field" and not to explain the variance from the measurement model (Bond and Fox, 2001, p. 191).

The Rasch/1PL IRT model focuses on ability, the capacity of a person to positively endorse (e.g., agree) or answer correctly an item at varying degrees of item difficulty. For example, items on a satisfaction with services scale would be identified in terms of levels of approval/contentment and so on and accordingly matched to persons having more or less of the attribute, in this case favorable and unfavorable perceptions of care. The difficulty parameter sets the location of the ICC on the horizontal axis (θ/ability). The item difficulty parameter, identified as b, is the only parameter that is allowed to vary in the Rasch or 1PL model. All other parameters (item discrimination, guessing, and carelessness) are constrained and set to one. Estimates for the second, third, and fourth parameters are nonetheless reported in Rasch and 1PL models. Substantive deviation from a value of 1 is an indication of deviation from the measurement model.

Item discrimination is expected to be uniform and is indicated by parallel ICC curves. ICCs that cross indicate that items change in their difficulty relative to the placement of persons on the attribute level (θ). If items contribute differentially along the ability continuum (crossed ICCs), it becomes challenging to objectively interpret the construct represented by the items across all respondents. The intent of measurement in psychotherapy research is to identify the amount of distress experienced by a client relative to some specified quantity (e.g., clinical cut-off, change from intake/baseline). The differential behavior of items at some placements on the ability continuum calls into question the construct validity of the measure, because the function of the items is not stable and ordered across the ability continuum. A measure is objective if the data it yields are not dependent on which items are used in assessing the trait or on the differential function of items at various trait (θ) levels (Rasch, 1977).

Rasch versus 1PL models

The Rasch and 1PL models are often discussed interchangeably. There are, however distinctions between the two (Hambleton, 1989; Wright and Masters, 1982; Wright and Stone, 1979). The intention of IRT is to seek a model to fit the data to understand and explain item response data patterns, while the intention of the Rasch model is to develop a measurement system based on measurement principles of objectivity, sufficiency and concatenation. Table 2.1 lists some of the characteristic distinctions between these two approaches. While both the Rasch and the 1PL models question how well empirical data (measure scores/response patterns) fit in terms of the measurement model constraints, the 1PL response is to consider

Table 2.1 Distinctions between the Rasch and 1PL models

Rasch	*1PL*
• Local *fit* of the data to the model, one parameter at a time	• Global *fit of the model to the data*
• Parameterizes each member of the respondent sample individually	• Summarizes the respondent sample as a normal distribution
• Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)	• ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
• Item data violating models assumptions (not supporting parameter separability) in a linear framework is examined for possible deletion from the model	• Model reconsidered in the event of data misfit in terms of adding additional parameters (discrimination, lower asymptote, etc.)

Source: Reprinted from "Questioning the measurement precision of psychotherapy research," by A. M. Doucette and A. W. Wolf, 2009, *Psychotherapy Research*, *19* (4), p. 380. Copyright 2009 Routledge.

additional parameters to improve the model fit, as opposed to the Rasch response to examine items violating model assumptions for possible deletion from the measure (Blais et al., 1999; Pastor and Beretvas, 2006; Simon et al., 2006).

Two-parameter model

As opposed to a 1PL model, in which the individual's response is a function of his or her level on the measured attribute and the difficulty of the item, a 2PL model allows the individual's response to vary in terms of the individual's level on the measured attribute, the item difficulty, plus item discrimination, the degree to which the item can differentiate attribute levels (e.g., individuals with high vs. low emotional distress). Figure 2.4 illustrates items having variable item discrimination. All three of the curves have the same location of 0, but the shapes are different (interception of probability = 0.5 and $\theta = 0$). Some of the curves are steeper than the others. The discrimination parameter, a, refers to the slope of the item curves. Curves with lower values of a are flatter, whereas those with higher values are steeper. The slope of an item characterizes how well an item is able to discriminate between respondents having different levels on the latent trait.

In 2PL models, item information is greatest at the point of item difficulty, but since the slope of an item is allowed to vary by the addition of a new parameter, the amount of information for an item also varies. While 2PL models offer the advantage of more precision in describing measurement data, they challenge us in terms of uniformly describing what is actually being measured. In other words, an item may have differential meaning for individuals relative to their response to other items in the same measure. For example, the expectation of high scores across all items in a measure of emotional distress for individuals experiencing severe psychological impairment may not be apparent in the measurement model and may indicate multidimensionality (Balsis, Gleason, Woods, and Oltmanns, 2007).

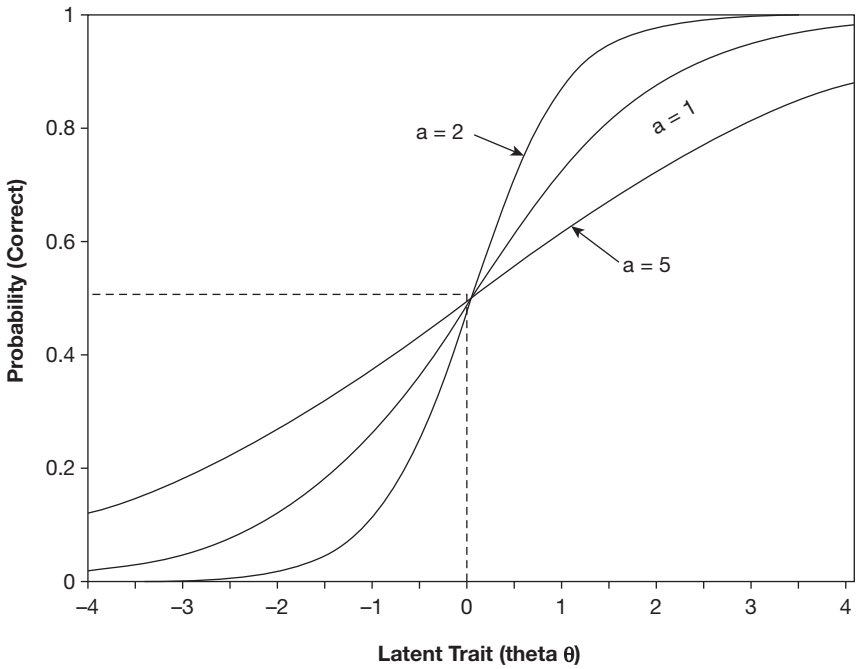

Figure 2.4 Example: item discrimination (*b*)

Source: Reprinted from "Questioning the measurement precision of psychotherapy research," by A. M. Doucette and A. W. Wolf, 2009, *Psychotherapy Research*, *19* (4), p. 380. Copyright 2009 Routledge.

Three- and four-parameter models

The IRT framework accommodates additional parameters to estimate the function of guessing (3PLmodels) and carelessness (4PL models). While most often used in educational assessment where individuals are encouraged to complete all items and to guess if they are unsure, the 3PL models can be useful in personality testing to detect intentional response distortion, inconsistent responses or responses not fitting a profile (Chernyshenko et al., 2001; Ellis et al., 1993; Rouse et al., 1999; Zumbo et al., 1997). A four-parameter model has been developed to account for carelessness responses, where an individual with a high level of a specific trait endorses items reflecting that trait, but fails to endorse an item or two in the expected manner. Although the four-parameter model is not cited in the psychotherapy research literature, it could be used to examine aberrant response patterns that might result from carelessness that is associated with measure familiarity in longitudinal studies or to identify respondent inattention to items that are phrased differentially to detect social desirability (e.g., negatively worded items imbedded among positively worded items, used as a lie scale). The three- and four-parameter models have received minimal attention in the psychotherapy literature.

As noted earlier, the addition of multiple parameters creates challenges in interpreting measurement data. A distinct advantage of measures meeting Rasch

and 1Pl model requirements is the fact that summative scale scores are sufficient in reflecting an individual's status along the measured attribute continuum.

Applying IRT to psychotherapy research: example

Data from a large U.S. commercial health plan using a measure of global distress as a psychotherapy outcome measure is reanalyzed using IRT. As mentioned above, the Rasch model is the most parsimonious of the IRT models and is considered the strongest measurement model for latent trait variables (Fischer and Molenaar, 1995) and, thus is used to illustrate issues of concern regarding measurement precision in psychotherapy research.

A specific advantage of the Rasch and other IRT models is the graphical output that accompanies analyses, allowing individuals with less mathematical and statistical knowledge to readily grasp the measurement concepts. The analyses of the measure presented next illustrate the item-level information provided by the Rasch measurement model. The intent of these analyses is to inform psychotherapy and other researchers using self-report outcome measurement of the merits of this approach.

Example using US commercial health plan data

Sample

Data used in this example are from a measure used by a large commercial health plan in the United States to assess outcome for the treatment of behavioral health disorders. The health plan dataset included 258,393 unique participants, ranging in age from 18 to 65 years of age (9 percent—18–24 years, 26 percent—25–35 years, 31 percent—36–45 years, 30 percent—45–60 years, 4 percent—61–65 year). Thirty-two percent of the sample is male. Depression (42 percent) and adjustment disorder (20 percent) were the two most prominent diagnoses, followed by anxiety (9 percent), bipolar (7 percent), and alcohol and drug (6 percent) disorders. On average, the health plans clients in this sample received 12 sessions of treatment. The instrument was administered at intake and multiple points during treatment.

Measurement

A 30-item instrument assessing general distress, the Life Status Questionnaire (LSQ—Lambert, Hatfield, and Vermeersch, 2001) was used as the outcome measure. The measure, developed using CTT, asks about symptomatology, work/school problems, interpersonal relationships, and alcohol/drug problems using a 5-point Likert scale (0 = *never* to 4 = *almost always*; scale score range 0 to 120). The health plan uses the total score across the 30 items as their outcome index. Reliability for this sample is 0.93.

Item characteristics, sufficiency (bandwidth), and scale dimensionality

The LSQ measure dataset was reanalyzed using the Rasch model with Winsteps (Linacre, 2001).[2] The Rasch Winsteps software provides *infit* and *outfit* statistics for each scale item. Infit and outfit estimates are based on chi-square distributions, weighting each item observation, by scale information (variance). Infit is sensitive to response patterns across items that are aligned to the person's ability estimate, and outfit is sensitive to item responses far away from the person's ability, typically item response on items that are at the ends of the trait continuum (theta θ, very easy or very difficult). Recommended infit and outfit estimate boundaries are expected to range between 0.7 and 1.4 (Bond and Fox, 2007, p. 243; Wright, 1995). Estimates less than 0.05 are overly predictive and deceptive, falsely indicating that the measure is more precise than it likely is. Estimates exceeding 2.0 indicate the *noise* associated with the item is greater than the information provided by the item. In addition to infit and outfit estimates, P-value (proportion of correct and point biserial correlations (Pearson product-moment correlation between the item score and the total item score with the item excluded) were also examined as initial estimates of item quality. Initial analyses indicated that six items exceeded acceptable infit/outfit ranges. The infit/outfit estimates for three of these items (suicidality, trouble with friends, and concern with family trouble) were closer to the recommended range, and had acceptable point biserial and *p*-value estimates. The three alcohol and drug items had the most serious infit and outfit estimates, as well as low *p*-value and point biserial estimates indicating that these items had substantively poorer fit than did the 27 other items in the scale (see Table 2.2).

An important aspect of an IRT approach is the ability to determine whether the scale items sufficiently cover the trait range. The Rasch model illustrates this with the person–item map (see Figure 2.2). The distribution of persons (respondents) is depicted on the left of the vertical axis; with the item distribution in terms of difficulty (theta θ, IRT *b* parameter) shown to the right of the same axis. Items are arrayed from top to bottom in terms of difficulty. The alcohol and drug (items identified as misfitting) and suicide items are the most difficult for respondents to endorse (indicating higher levels of distress), while feeling stressed, self-blame, and concerns about family troubles are among the easiest items to endorse (indicating lower levels of distress) for this treatment sample. Items appearing on the same line or in close proximity to each other, tap the same trait range. These items, especially if the content is similar, provide redundant information. For example, items 22 and 27 (*work/school is not going too well*, *not doing well in work/school*) have difficulty estimates of .34 and .37 respectively. The residual correlation for these two items is 0.33, indicating that these items share a third of their random unexplained variance. Residual correlations yielded from a principal–components analysis (PCA) are examined to detect item dependency, the potential for duplicating some feature or content across items. In the aforementioned case, if an individual reports that *work is not going too well*, it is likely that he/she will also report that he/she is *not doing well at work*. This suggests that