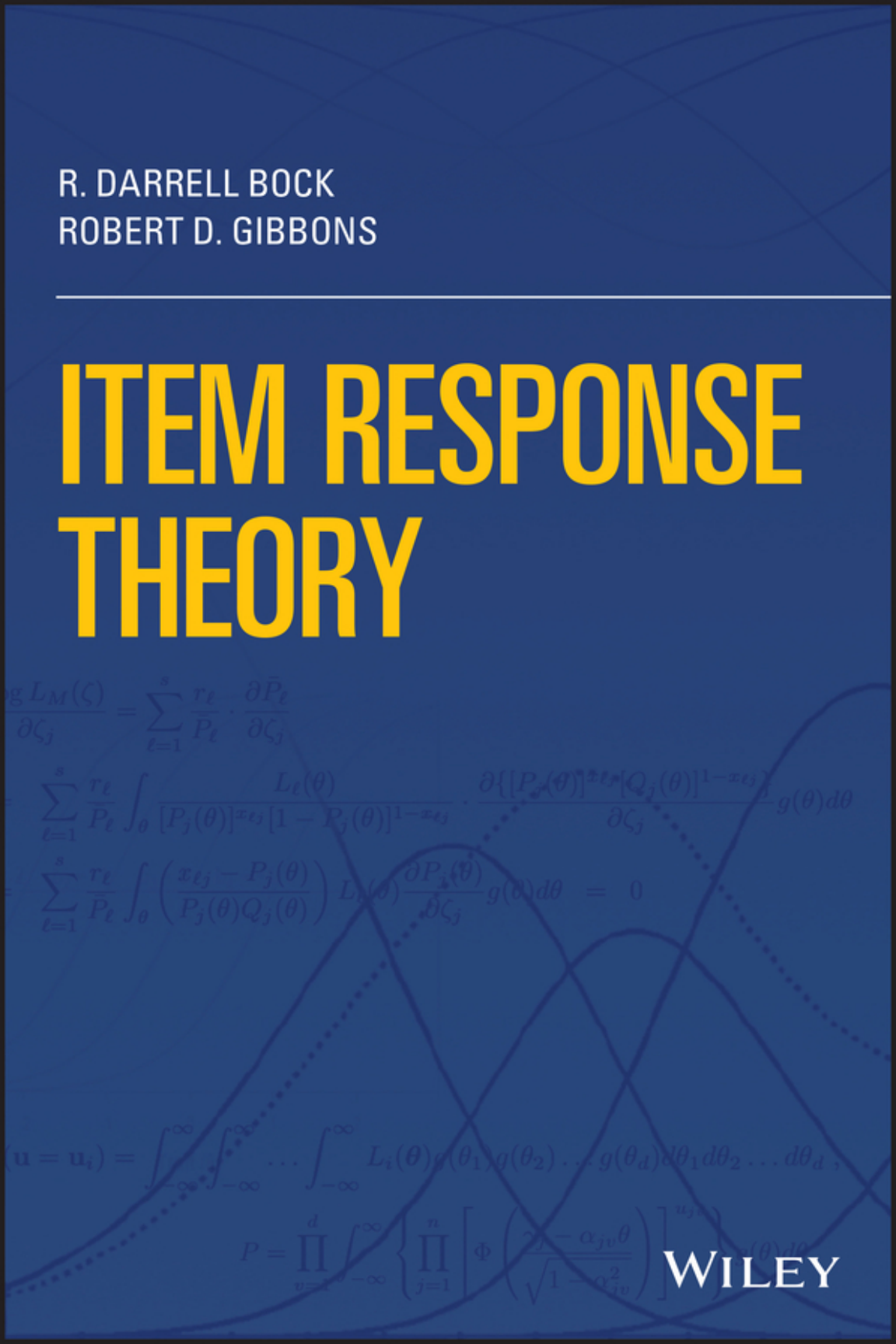
R. DARRELL BOCK
ROBERT D. GIBBONS

ITEM RESPONSE THEORY



WILEY

Item Response Theory

Item Response Theory

R. Darrell Bock
Robert D. Gibbons
University of Chicago

WILEY

This first edition first published 2021

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

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

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

Library of Congress Cataloging-in-Publication Data

Names: Bock, R. Darrell, author. | Gibbons, Robert D., 1955- author.
Title: Item response theory / Richard Darrell Bock, Robert David Gibbons, University of Chicago.
Description: First edition. | Hoboken : Wiley, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020055709 (print) | LCCN 2020055710 (ebook) | ISBN 9781119716686 (hardback) | ISBN 9781119716679 (adobe pdf) | ISBN 9781119716716 (epub)
Subjects: LCSH: Item response theory. | Psychology–Mathematical models.
Classification: LCC BF39.2.I84 B63 2021 (print) | LCC BF39.2.I84 (ebook) | DDC 150.28/7–dc23
LC record available at https://lccn.loc.gov/2020055709
LC ebook record available at https://lccn.loc.gov/2020055710

Cover Design: Wiley
Cover Image: © Image by Robert Gibbons

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

10 9 8 7 6 5 4 3 2 1

To Renee, Monica, Paul, and Conrad

R.D.B.

To Carol, Julie, Jason, Ethan, and Michael

R.D.G.

Contents

Preface *xv*
Acknowledgments *xvii*

1 Foundations *1*
1.1 The Logic of Item Response Theory *3*
1.2 Model-Based Data Analysis *4*
1.3 Origins *5*
1.3.1 Psychometric Scaling *6*
1.3.2 Classical Test Theory *9*
1.3.3 Contributions from Statistics *10*
1.4 The Population Concept in IRT *12*
1.5 Generalizability Theory *14*

2 Selected Mathematical and Statistical Results *23*
2.1 Points, Point Sets, and Set Operations *23*
2.2 Probability *25*
2.3 Sampling *27*
2.4 Joint, Conditional, and Marginal Probability *27*
2.5 Probability Distributions and Densities *29*
2.6 Describing Distributions *34*
2.7 Functions of Random Variables *36*
2.7.1 Linear Functions *36*
2.7.2 Nonlinear Functions *39*
2.8 Elements of Matrix Algebra *40*
2.8.1 Partitioned Matrices *43*
2.8.2 The Kronecker Product *44*
2.8.3 Row and Column Matrices *45*
2.8.3.1 Rank and Nullity *45*
2.8.4 Matrix Inversion *45*

2.9 Determinants *47*
2.10 Matrix Differentiation *48*
2.10.1 Scalar Functions of Vector Variables *48*
2.10.2 Vector Functions of a Vector Variable *50*
2.10.3 Scalar Functions of a Matrix Variable *50*
2.10.4 Chain Rule for Scalar Functions of a Matrix Variable *51*
2.10.5 Matrix Functions of a Matrix Variable *52*
2.10.6 Derivatives of a Scalar Function with Respect to a Symmetric Matrix *53*
2.10.7 Second-Order Differentiation *54*
2.11 Theory of Estimation *55*
2.11.1 Analysis of Variance *58*
2.11.2 Estimating Variance Components *59*
2.12 Maximum Likelihood Estimation *62*
2.12.1 Likelihood Functions *63*
2.12.2 The Likelihood Equations *63*
2.12.3 Examples of Maximum Likelihood Estimation *64*
2.12.4 Sampling Distribution of the Estimator *65*
2.12.5 The Fisher-Scoring Solution of the Likelihood Equations *66*
2.12.6 Properties of the Maximum Likelihood Estimator (MLE) *67*
2.12.7 Constrained Estimation *67*
2.12.8 Admissibility *67*
2.13 Bayes Estimation *68*
2.14 The Maximum A Posteriori (MAP) Estimator *71*
2.15 Marginal Maximum Likelihood Estimation (MMLE) *72*
2.15.1 The Marginal Likelihood Equations *73*
2.15.2 Application in the "Normal–Normal" Case *75*
2.15.2.1 First-Stage Estimation *75*
2.15.2.2 Second-Stage Estimation *76*
2.15.3 The EM Solution *78*
2.15.4 The Fisher-coring Solution *78*
2.16 Probit and Logit Analysis *80*
2.16.1 Probit Analysis *80*
2.16.2 Logit Analysis *82*
2.16.3 Logit-Linear Analysis *83*
2.16.4 Extension of Logit-Linear Analysis to Multinomial Data *85*
2.16.4.1 Graded Categories *86*
2.16.4.2 Nominal Categories *88*
2.17 Some Results from Classical Test Theory *91*
2.17.1 Test Reliability *93*

2.17.2 Estimating Reliability *94*
2.17.2.1 Bayes Estimation of True Scores *99*
2.17.3 When are the Assumptions of Classical Test Theory Reasonable? *100*

3 Unidimensional IRT Models *103*
3.1 The General IRT Framework *105*
3.2 Item Response Models *107*
3.2.1 Dichotomous Categories *107*
3.2.1.1 Normal-Ogive Model *107*
3.2.1.2 2PL Model *111*
3.2.1.3 3PL Model *113*
3.2.1.4 1PL Model *115*
3.2.1.5 Illustration *116*
3.2.2 Polytomous Categories *118*
3.2.2.1 Graded Categories *119*
3.2.2.2 Illustration *121*
3.2.2.3 The Nominal Categories Model *121*
3.2.2.4 Nominal Multiple-Choice Model *130*
3.2.2.5 Illustration *131*
3.2.2.6 Partial Credit Model *131*
3.2.2.7 Generalized Partial Credit Model *135*
3.2.2.8 Illustration *135*
3.2.2.9 Rating Scale Models *136*
3.2.3 Ranking Model *138*

4 Item Parameter Estimation – Binary Data *141*
4.1 Estimation of Item Parameters Assuming Known Attribute Values of the Respondents *142*
4.1.1 Estimation *143*
4.1.1.1 The One-Parameter Model *143*
4.1.1.2 The Two-Parameter Model *145*
4.1.1.3 The Three-Parameter Model *145*
4.2 Estimation of Item Parameters Assuming Unknown Attribute Values of the Respondents *146*
4.2.1 Joint Maximum Likelihood Estimation (JML) *147*
4.2.1.1 The One-Parameter Logistic Model *147*
4.2.1.2 Logit-Linear Analysis *149*
4.2.1.3 Proportional Marginal Adjustments *152*
4.2.2 Marginal Maximum Likelihood Estimation (MML) *155*
4.2.2.1 The two-parameter Model *161*

5 Item Parameter Estimation – Polytomous Data *175*
5.1 General Results *175*
5.2 The Normal Ogive Model *180*
5.3 The Nominal Categories Model *181*
5.4 The Graded Categories Model *183*
5.5 The Generalized Partial Credit Model *186*
5.5.1 The Unrestricted Version *187*
5.5.2 The EM Solution *188*
5.5.2.1 The GPCM Newton–Gauss Joint Solution *189*
5.5.3 Rating Scale Models *190*
5.5.3.1 The EM Solution for the RSM *190*
5.5.3.2 The Newton–Gauss Solution for the RSM *191*
5.6 Boundary Problems *192*
5.7 Multiple Group Models *194*
5.8 Discussion *196*
5.9 Conclusions *199*

6 Multidimensional IRT Models *201*
6.1 Classical Multiple Factor Analysis of Test Scores *202*
6.2 Classical Item Factor Analysis *203*
6.3 Item Factor Analysis Based on Item Response Theory *205*
6.4 maximum Likelihood Estimation of Item Slopes and Intercepts *207*
6.4.1 Estimating Parameters of the Item Response Model *208*
6.5 Indeterminacies of Item Factor Analysis *212*
6.5.1 Direction of Response *212*
6.5.2 Indeterminacy of Location and Scale *212*
6.5.3 Rotational Indeterminacy of Factor Loadings in Exploratory Factor Analysis *213*
6.5.3.1 Varimax Factor Pattern *214*
6.5.3.2 Promax Factor Pattern *214*
6.5.3.3 General and Group Factors *214*
6.5.3.4 Confirmatory Item Factor Analysis and the Bifactor Pattern *215*
6.6 Estimation of Item Parameters and Respondent Scores in Item Bifactor Analysis *217*
6.7 Estimating Factor Scores *219*
6.8 Example *219*
6.8.1 Exploratory Item Factor Analysis *220*
6.8.2 Confirmatory Item Bifactor Analysis *225*
6.9 Two-Tier Model *229*
6.10 Summary *230*

7 Analysis of Dimensionality *231*
7.1 Unidimensional Models and Multidimensional Data *232*
7.2 Limited-Information Goodness of Fit Tests *236*
7.3 Example *238*
7.3.1 Exploratory Item Factor Analysis *238*
7.3.2 Confirmatory Item Bifactor Analysis *239*
7.4 Discussion *240*

8 Computerized Adaptive Testing *243*
8.1 What Is Computerized Adaptive Testing? *243*
8.2 Computerized Adaptive Testing – An Overview *244*
8.3 Item Selection *245*
8.3.1 Unidimensional Computerized Adaptive Testing (UCAT) *246*
8.3.1.1 Fisher Information in IRT Model *246*
8.3.1.2 Maximizing Fisher Information (MFI) and Its Limitations *248*
8.3.1.3 Modifications to MFI *249*
8.3.2 Multidimensional Computerized Adaptive Testing (MCAT) *251*
8.3.2.1 Two Conceptualizations of the Information Function in Multidimensional Space *252*
8.3.2.2 Selection Methods in MCAT *253*
8.3.3 Bifactor IRT *256*
8.4 Terminating an Adaptive Test *257*
8.5 Additional Considerations *258*
8.6 An Example from Mental Health Measurement *260*
8.6.1 The CAT-Mental Health *261*
8.6.2 Discussion *264*

9 Differential Item Functioning *267*
9.1 Introduction *267*
9.2 Types of DIF *268*
9.3 The Mantel–Haenszel Procedure *269*
9.4 Lord's Wald Test *271*
9.5 Lagrange Multiplier Test *271*
9.6 Logistic Regression *273*
9.7 Assessing DIF for the Bifactor Model *274*
9.8 Assessing DIF from CAT Data *275*

10 Estimating Respondent Attributes *279*
10.1 Introduction *279*
10.2 Ability Estimation *280*
10.2.1 Maximum Likelihood *280*

10.2.2 Bayes MAP *281*
10.2.3 Bayes EAP *281*
10.2.4 Ability Estimation for Polytomous Data *282*
10.2.5 Ability Estimation for Multidimensional IRT Models *283*
10.2.6 Ability Estimation for the Bifactor Model *284*
10.2.7 Estimation of the Ability Distribution *284*
10.2.8 Domain Scores *285*

11 Multiple Group Item Response Models *287*
11.1 Introduction *287*
11.2 IRT Estimation When the Grouping Structure Is Known: Traditional Multiple Group IRT *288*
11.2.1 Example *290*
11.3 IRT Estimation When the Grouping Structure Is Unknown: Mixtures of Gaussian Components *292*
11.3.1 The Mixture Distribution *293*
11.3.2 The Likelihood Component *295*
11.3.3 Algorithm *297*
11.3.4 Unequal Variances *297*
11.4 Multivariate Probit Analysis *297*
11.4.1 The Model *298*
11.4.2 Identification *300*
11.4.3 Estimation *300*
11.4.4 Tests of Fit *301*
11.4.5 Illustration *302*
11.5 Multilevel IRT Models *305*
11.5.1 The Rasch Model *306*
11.5.2 The Two-Parameter Logistic Model *307*
11.5.3 Estimation *308*
11.5.4 Illustration *309*

12 Test and Scale Development and Maintenance *311*
12.1 Introduction *311*
12.2 Item Banking *311*
12.3 Item Calibration *314*
12.3.1 The OEM Method *315*
12.3.2 The MEM Method *315*
12.3.3 Stocking's Method A *315*
12.3.4 Stocking's Method B *316*
12.4 IRT Equating *318*
12.4.1 Linking, Scale Aligning and Equating *318*

12.4.2 Experimental Designs for Equating *319*
12.4.2.1 Single Group (SG) Design *319*
12.4.2.2 Equivalent Groups (EG) Design *319*
12.4.2.3 Counterbalanced (CB) Design *319*
12.4.2.4 The Anchor Test or Nonequivalent Groups with Anchor Test (NEAT) Design *320*
12.5 Harmonization *320*
12.6 Item Parameter Drift *322*
12.7 Summary *323*

13 Some Interesting Applications *325*
13.1 Introduction *325*
13.2 Biobehavioral Synthesis *325*
13.3 Mental Health Measurement *329*
13.3.1 The CAT-Depression Inventory *329*
13.3.2 The CAT-Anxiety Scale *331*
13.3.3 The Measurement of Suicidality and the Prediction of Future Suicidal Attempt *331*
13.3.4 Clinician and Self-Rated Psychosis Measurement *333*
13.3.5 Substance Use Disorder *334*
13.3.6 Special Populations and Differential Item Functioning *335*
13.3.6.1 Perinatal *336*
13.3.6.2 Emergency Medicine *336*
13.3.6.3 Latinos Taking Tests in Spanish *337*
13.3.6.4 Criminal Justice *339*
13.3.7 Intensive Longitudinal Data *339*
13.4 IRT in Machine Learning *340*

Bibliography *343*
Index *361*

Preface

Not everything that can be counted counts, and not everything that counts can be counted.

(Albert Einstein)

If we date the origin of modern item response theory from Derrick Lawley's pioneering 1943 paper, "On problems connected with item selection and test construction," or Frederic Lord's 1950 Psychometric Monograph, "A theory of test scores," the field has now enjoyed nearly 75 years of vigorous development. It appears to have reached a level of maturity sufficient to warrant a comprehensive review of the accomplishments up to this point. The previous effort, Lord and Novick's 1968 monograph "Statistical theories of mental test scores," while incorporating the innovative contributions of Allan Birnbaum, was necessarily a report of work in progress in a young field. Only the models for binary-scored items were available at that time, and the estimation theory required to implement them, although intimated, was not yet well developed.

Results in the field are now much richer. Currently, there are models to fit many forms of item response data, and the statistical methods for estimating the parameters of these models exist and are implemented. Procedures for assigning scale scores to respondents are more varied and include efficient adaptive algorithms. Entirely new methods exist for estimating latent distributions of populations without computing scores for individual sample members. Better solutions have been found for the classic problems of test maintenance and forms equating. Connections between item response theory and multilevel sampling models have been clarified. Perhaps most important, the computing facilities now exist to make large-scale applications of these developments practical.

Our aim in the present text is to present a reasonably complete account of this progress with special emphasis on the computer applications. Our discussion therefore includes details on numerical procedures suitable for practical

applications of IRT. The most difficult aspect of writing this text has been finding the right level at which to present these topics, which are inherently mathematical and statistical. To make the discussion accessible as possible, without violating the spirit of the subject, we have adopted a level of presentation that assumes a first-year graduate background in the behavioral or social sciences, together with mathematics preparation through calculus and courses in statistics through generalized linear models. To supplement that preparation, we offer in Chapter 2 a review of some of the mathematical and statistical foundations required in the sequel. To motivate the reader and to fix ideas, we have everywhere tried to find real and interesting data with which to illustrate the theory.

Acknowledgments

We thank Bob Mislevy and David Thissen for their contributions to early work on this project; Don Hedeker, Li Cai, and Yanyan Sheng for their contributions, review, and helpful comments; and Cody Brannan for help in preparing the manuscript. We also acknowledge support of the Office of Naval Research, N00014-85-K-0586, and the National Institute of Mental Health R01 MH100155 and R01 MH66302 which funded our work on multidimensional item response theory and computerized adaptive testing. The content is solely the responsibility of the authors.

1

Foundations

To measure is to know.

(Source: Lord Kelvin (William Thompson) 1824–1907)

Most observations of behavior are recorded as distinct qualitative events. For example:

> a student responds correctly to certain specified questions, responds incorrectly to others, and declines to respond to still others;
> on the fifth trial of a learning experiment, the subject recalls six abstract and ten concrete words from a list of thirty;
> a reader rates each paragraph of an essay exercise on a scale of rhetorical effectiveness graded from 1 to 7.
> a participant in a class discussion group speaks up three times on issue A, once on issue C, but not at all on issues B, D or E;
> in response to a social survey, a head of household endorses five out of ten statements concerning a public issue, but disagrees with the others;
> an applicant for a secretarial position makes two spelling errors in transcribing 300 words of dictation;
> a patient in a primary care clinic reports specific problems with mood, cognition and somatic symptoms of depression during the past two weeks.

These types of data have in common the fact that each respondent is reacting qualitatively to multiple stimuli in a specified set. In the present context, we call all such stimuli *items* and define item response theory, or "IRT," as the statistical study of data that arise in this way. That each respondent is responding to more than one item is essential to the definition: if each respondent were presented only one item, an enumeration of the observed qualitative responses would result in a simple contingency table that could be analyzed by conventional chi-square or

Item Response Theory, First Edition. R. Darrell Bock and Robert D. Gibbons.

log-linear methods. Such methods can be extended to perhaps three or four items by assigning respondents to distinct categories generated by all possible combinations of the repeated qualitative response, but they quickly become unworkable as the numbers of items increase. When there are repeated qualitative responses to relatively large numbers of items, the data are the special province of IRT. This form of data must be regarded to arise from two stages of sampling – the sampling of responses within each respondent, and the sampling of respondents from some population.

There are three main uses of IRT methods of data analysis. The first is to summarize information in the responses in a way that is suitable for some practical decision about a given respondent. The IRT reduction of the data either classifies the respondents qualitatively or assigns each a quantitative measure that supports such a classification. This is a traditional treatment of data from multiple-item tests. A score on an educational test, for example, may support the decision to admit a student to a college or university; a profile of performance in a battery of vocational tests may influence the choice of job applicants or military recruits; a self-report on a personality inventory may suggest the best approach to counseling or psychotherapy. Under favorable conditions, these kinds of uses of item response data can substantially improve the chances of successful outcomes of the decision compared to subjective or more arbitrary methods of selection or classification.

The second important use of these methods is to describe various groups to which the respondents may belong. The paradigm of this use is the randomized experiment, in which subjects are assigned with equal probability to control or treatment groups and a multiple-item test is administered in order to evaluate the effects of the treatments. The object of the IRT analysis of the test data is to estimate the distribution of response tendencies in the several groups. Similarly, in survey studies, respondents may be randomly selected from defined subpopulations of some larger population and administered an attitude or opinion questionnaire. The responses to the questionnaire items could in principle be used to classify the individual respondents, as in an employment interview, but this is not the purpose of the typical survey. The aim is rather to compare the subpopulations with respect to the distribution of response tendencies among their members.

The classical approach to analysis of data from either of these sources is to make comparisons among the groups or subpopulations by estimating scores for the respondents and analyzing them as if they were the primary data. One of the important contributions of IRT has been to show that this is not necessarily the best way to proceed. We present methods by which population characteristics can be directly estimated from the original item responses without computing intermediate respondent-level scores.

The third use of IRT analysis is to characterize the items. In some types of study, the items themselves are the objects of interest. For example, the aim

may be to construct items with levels of difficulty and discriminating power suitable for a particular population of respondents. These activities will inevitably involve the analysis of empirical item data in order to verify that the methods of item construction are succeeding. Some items may be far wide of the mark and will have to be made easier or harder. Others, especially among multiple choice items, may contain hidden ambiguities that weaken their discriminating power; they can usually be corrected by reworking the response alternatives. In either case, IRT methods can identify and estimate characteristics of the items that are diagnostic of these problems. These methods now extend beyond the traditional multiple-choice item formats to rating scales, nominal categories, and item clusters. They also encompass the empirical study of the cognitive processes involved in the item response. In these studies, interest centers on classes of items distinguished by common stimulus or task features. The objective is to identify such features, connect them to other cognitive theory, and predict their effect on statistical characteristics of the items (e.g., item difficulty or discriminating power). Chapters of this text devoted to this aspect of IRT are Chapter 4 on parameter estimation for binary items, Chapter 5 on multiple-category items, and Chapter 6 on item factor analysis.

1.1 The Logic of Item Response Theory

Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. The difference is clarified by the following analogy. Imagine a track and field meet in which 10 athletes participate in men's 110-m hurdles race and also in men's high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined not only by the runner's time but also by the number of hurdles successfully cleared, i.e. not tipped over. On the other hand, the high jump is conducted in the conventional way: The crossbar is raised by, say, 2-cm increments on the uprights, and the athletes try to jump over the bar without dislodging it. The first of these two events is like a traditionally scored objective test: Runners attempting to clear hurdles of varying heights is analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions. On the high jump, ability is measured by a scale in millimeters and centimeters at the highest scale position of the crossbar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until she can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly.

In IRT, ability is measured by a scale point, not a numerical count. These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: If hurdles are arbitrarily added or removed, the number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures noncomparable. The same is true of traditional number-right scores of objective tests: Scores lose their comparability if item composition is changed. The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2-cm positions, or if one of those positions were omitted, height cleared is unchanged, and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted, or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected. This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of educational and psychological measurement.

1.2 Model-Based Data Analysis

The IRT discussed in this text is aptly described as "model-based." There are cogent reasons for taking this approach to item response data rather than relying on enumerative summaries or nonparametric statistical methods. Perhaps most important is the economy of thought and discussion that results from substituting quantitative complexity for voluminous descriptive detail. In this, IRT emulates modern physical science, which attempts to account for a wide range of observable phenomena by a possibly complicated mathematical function depending on relatively few free parameters. IRT achieves this kind of economy by expressing the probability of an observed response to a stimulus in terms of a limited number of characteristics of the stimulus and of the respondent. The mathematical functions used for this purpose, the most important of which we discuss in Chapters 3–6, are called *item response models*. They are a central feature of IRT. They are capable of accounting succinctly for the kinds of data exemplified above, and their parameters concisely describe the operating characteristics of the items.

Another merit of the model-based approach is that, when it leads us to a restricted class of parsimonious models that fit a wide range of data, our confidence in the theory behind the models is strengthened. We are then encouraged to extend the theory to new situations and further test its generality. Apart from

fortuitous discoveries, this is the main avenue of progress in scientific work. Purely descriptive methods of data analysis do not give us the same reassurance that we have the right conception of the phenomenon. Nonparametric curve fitting procedures, for example, merely produce smoothed representations of the data, possibly under continuity restrictions. They have no definite limit on the number of free parameters implicitly fitted in the construction of the curve. Having no definite form, they are difficult to compare, discuss, or extend to other domains. Admittedly, they are useful for limited purposes, such as interpolating values between observations when no suitable functional forms can be found. But absence of suitable functions is not generally the case in item response data: most of the response models that have been proposed for both binary and multiple category data account for the observations within the limits of sampling error, and they do so with comparative few free parameters. It is unlikely that worthwhile improvement in fit could be expected by a model-free approach to item response data in the domains typically analyzed.

A main strength of existing IRT is that it provides a coherent and rigorous methodology for the analysis of a very wide range of multiresponse qualitative data. The familiar statistical tools for measured, quantitative variables are not generally suitable for such data. The most widely used procedures for such variables, including linear least-squares regression, univariate and multivariate analysis of variance, discriminant analysis, linear structural analysis, etc., all model the distribution of the observations on a continuous interval scale and assume homogeneous error variation. Except in limiting cases, qualitative response data do not even remotely satisfy these assumptions. They are discrete events: They do not refer to any continuum, do not have interval scale properties or have homogeneous error from one stimulus to another. Even the familiar population descriptors in the classical statistical analysis – means, standard deviations, product-moment correlations, etc. – do not serve these forms of data well.

1.3 Origins

IRT is not primarily a theory in the sense of a putative explanation of some phenomenon. Rather, it is a coherent methodological system, similar to estimation theory or least-squares theory in the field of statistics. The exception in IRT is the concept of an observed qualitative responses arising from underlying quantitative variation through the action of an intervening threshold process. This conception, especially as it applies to sensory discriminations or preference judgments, is implicit in the more psychologically oriented applications of the theory. In other areas of application, such as educational measurement, IRT is viewed merely as a means of relating the response probabilities to a much smaller number

of underlying parameters in terms of which respondents can be characterized, populations compared, or items described.

The psychological orientation in IRT had its origins in the nineteenth and early twentieth-century work on scaling of stimuli; the educational measurement orientation is associated with the development of educational tests during the twentieth century and is now referred to as "classical test theory." Running through both of these approaches is a common thread of concepts and methods borrowed from mathematics and mathematical statistics. An understanding of these sources of present theory is a good foundation for study of the topic. In Section 1.3.1, we review briefly the contributions from each and discuss their relationships to the theory in its present form.

1.3.1 Psychometric Scaling

The first attempt to estimate scale values from discrete data was by Fechner (1966) in connection with his study of Weber's Law. Weber had found in a careful series of experiments that the magnitude of errors made by human observers in judging the size or intensity of a physical stimulus tends to be proportional to the intensity of the stimulus. Fechner reasoned that this indicated the existence of a sensory continuum on which the intensities of the stimuli are perceived as the logarithm of their physical measures. In a typical experiment demonstrating the Weber effect, the investigator requires the observer, by a pulley arrangement, to adjust the length of a variable line to match that of a displayed line of fixed length. This procedure is called the "method of adjustment." The general finding by this method, that the average absolute error in reproducing the stimulus is a constant proportion of the stimulus size (usually about 10%), is now known as Weber's Law. To establish the generality of this law and gain support for his theory relating stimulation to sensation, Fechner wanted to extend these studies to stimuli that could not easily be adjusted continuously, such as flavors, odors, or weights. For this purpose, he developed what he called the "method of right-and-wrong cases," but which is now called the "method of constant stimuli," or the "constant method."

In the modern version of the "lifted-weight" experiment discussed by Fechner in 1860 (Guilford 1954), the observer is presented a trial weight, and an identically appearing standard weight, and asked to lift them and state whether the former is heavier than the latter. The trial weights, which are set at several different levels smaller than, equal to, and larger than the standard, are sufficiently close to the standard that the observer makes a certain proportion of errors in repeated attempts at this task. The data from the experiment consist of the number of times the observer chooses the trial weight as heavier in a fixed number of attempts.

To infer a continuous measure of the average absolute error from these frequencies, Fechner invoked the same assumption that Gauss had made earlier

for errors in astronomical observations – namely, that the errors are normally distributed with a mean and standard deviation typical of the observer. On this assumption, the expected size of the error associated with each observed proportion of "greater-than" judgments is the deviate at the corresponding percentage point of the normal distribution. If the assumption of normally distributed errors is tenable and Fechner's theory is correct, the plot of the corresponding normal deviates versus the log stimulus intensity (or in this case the difference between the logs of the test weights and the standard weight) should form a straight line, apart from sampling error in the observed proportions. With properly counterbalanced orders of stimulus presentation, the 50 percent point (or zero deviate) should occur at the point of stimulus equality. In that case, the probable error in judgments involving the standard stimulus can be defined as difference between the 75 and 25 percent points read from the fitted line – the so-called "difference limen" or "difference threshold" (Bock and Jones 1968).

Moreover, if Weber's Law holds over a wide range of physical intensities, the difference limen is constant on the sensory continuum and can serve as the unit of the scale that measures the psychological construct "sensation." The origin of the continuum can be set at the log of that value of the stimulus that can be correctly distinguished from the null stimulus 50% of the time (the "absolute threshold"). The logarithmic relationship between stimulation and sensation is now referred to as "Fechner's Psychophysical Law."[1]

In 1928, Louis Leon Thurstone formalized the concept of a sensory scale by introducing the discriminal process construct and a threshold mechanism. We discuss his model in more detail in Chapter 3, but briefly his assumption was that the stimulus gives rise in the observer to an unobservable random variable consisting of a fixed component attributable to the stimulus and a random component due to temporal instability of the sensory system. He called this unobservable variable a "discriminal process." To explain the observed response, he posited the existence of a point, or threshold, on the continuum such that the observer responds in one category if the process is above the threshold, and in another category if not.[2]

In Thurstone's model, it is the *difference* of the discriminal processes that is the relevant variable. If the process for the test stimulus momentarily exceeds that of the standard, the difference is positive and the observer responds that the test

1 In the 1950s, Stevens (1961) disputed the validity of Fechner's Law and proposed in its place a class of power functions relating physical to sensory intensity. But Stevens was discussing the relationship between magnitude judgments and stimulus intensity, which is a quite different phenomenon than the stimulus confusions on which Fechner's logarithmic psychophysical law is based. Although the power-law includes the logarithmic function in the limiting case as the exponent goes to zero, it makes no special contribution to the modeling of sensory discrimination phenomena.

2 Fechner (1860) attributed a similar account of the constant method results to his colleague, Möbius.

stimulus appears greater or more intense; otherwise, not. Thus, for the difference process, the threshold is at zero. Like Fechner, Thurstone assumed that the discriminal processes, and thus the difference process, are normally distributed. The continuous value attributed to the difference between the two stimuli is therefore consistently estimated by the normal deviate corresponding to the proportion of times the trial stimulus is judged greater than the standard.

Thurstone's most important contribution to scaling was the demonstration that a purely psychological scale not dependent on any form of physical measurement can be constructed from a suitable set of interlocking comparisons. If the deviates corresponding to these comparisons determine, or overdetermine, the locations of the stimuli on the sensory continuum, it then becomes possible to assign to the stimuli quantitative values in a well-defined metric. In particular, Thurstone (1928) showed that in the method of paired comparisons, where the observer compares all $n(n-1)/2$ distinct pairs of n stimuli, the equations implied by the discriminal process model overdetermine the locations of the stimuli on the psychological continuum. From the observed qualitative judgments, the locations can be estimated on a continuous scale with arbitrary origin and unit, and $(n-1)(n-2)/2$ degrees of freedom remain to test the fit of the model (see Chapter 5 in Bock and Jones 1968). In principle, the dispersions of the random processes could also vary and require estimation, but in the classical "Case V" analysis of paired-comparisons, they are assumed constant. To emphasize that such a scale could be constructed for any sort of pairwise orderable objects, Thurstone referred to it as a "psychological" continuum rather than a sensory continuum.

IRT borrows heavily from these earlier conceptions of psychological scaling. It accepts the idea that a discrete behavioral response to a set task, object, or proposition (in short, to an *item*) is the expression of a stochastic mechanism that can be modeled by an unobservable random variable and a threshold. The variable is assumed to have some form of distribution on an infinite latent continuum. The items are characterized by their locations, or thresholds, on this continuum and by the dispersion of the corresponding random variable. If an external variable or criterion is correlated with the unobservable random variable, the IRT scaling can proceed on the same basis as psychophysical scaling (see Section 4.1). But in the more typical case where no such external variable exists, the scale depends only on relationships internal to the data, which in IRT are direct responses to multiple items rather than the multiple comparisons of Thurstonian scaling.

As we elaborate in Chapter 3, when Lawley (1943) and Lord (1952) formulated the IRT model, they added a random component associated with the observer (or in our terminology the *respondent*). The paired comparison model does not require

this component because, if present, it would subtract out of the difference process (Andrich 1978).[3]

The main objective of IRT is to estimate from the item responses the attribute values of the respondents. We discuss statistical methods for this purpose in Chapters 4–6 and 10. Because the values are on a scale determined by the error process implicit in the item response models, they are often referred to as "scale scores" to distinguish them from the traditional "test score," which is just a count of the number of correct item responses.

There is a precise formal sense in which the psychological scaling model, which does not include an individual difference component, stands in the same relationship to the IRT model, as the analysis-of-variance fixed-effects model stands to the mixed-effects model with one random dimension. In the IRT model, the parameters that characterize items are the fixed effects and the attribute components associated with the respondents are the random effects. The only difference between the models is that the estimators of these quantities are linear in the analysis of variance and nonlinear in IRT. This is a very big difference methodologically; however, the simple noniterative calculations of the mixed-effects analysis of variance give way to the more complex iterative procedures of nonlinear estimation that characterize much of IRT analysis (see Chapters 4–6).

In IRT, the item response model expresses the probability of a specified response to a test item as a function of the quantitative attribute of the respondent and one or more parameters of the items. Such models now exist for many types of item responses. Those for binary (right–wrong) scored items were the first to be developed and are still the most common, but models for ratings and graded or multiple-category item responses are now available and enjoying increasing application. We discuss binary item response models in Chapter 3 and multiple-category items in Chapter 5. (See Thissen and Steinberg (1986) for a taxonomy of item response models.)

1.3.2 Classical Test Theory

Classical test theory is essentially an extension of the Gaussian theory of errors (Gauss 1809) to the measurement of individual differences. Originating in the work of Spearman (1907), Brown, E.L. Thorndike, and others, the classical theory was first applied to scores from cognitive tests in which item responses were

3 If each observer judges all stimuli in a paired comparison experiment, random interactions between observers and stimuli need to be included in the *sampling* model for the Case V analysis. Otherwise, the correlation between deviates for comparisons involving common objects are correlated, and the test of fit is grossly biased (Chapman and Bock 1958).

scored "right" or "wrong." The test score of a respondent was the number of right responses. Later the theory was extended to any multiple-item psychological test in which items can be meaningfully scored in a consistent direction. Sometimes the direction cannot be specified *a priori*, and empirical evidence must be invoked. In the personality tests – scales of the Minnesota Multiphasic Personality Inventory (MMPI), for example – the direction for scoring each item is chosen so that the scale will best discriminate between normal and pathological groups identified by expert judgments. A similar method is available in IRT whenever an external variable correlated with the attribute dimension is available (see Section 4.1).

Classical test theory assumes that the test score obtained by counting "right" responses is an additive linear model consisting of two random components – one due to the individual differences in the population of respondents – the other due to error, defined as the item-by-respondent interaction. Both components are assumed normally distributed, although this assumption is not required in many of the results. A main motivation of the theory is to obtain a scale-free index of the precision with which test scores estimate attributes of the respondents. Because the origin and range of the test score depends arbitrarily on the number of items in the test, the scale-dependent measure of precision used by Gauss (the reciprocal of the mean square error) is not applicable in classical test theory. It is replaced by the *reliability coefficient*, defined as the ratio of the variance of the individual difference component (true score variance in classical test theory terms) to the sum of that variance plus the variance of the error component. In statistical terms, reliability is the intraclass correlation of within-to between-individual variation in the population of respondents. The reliability coefficient therefore ranges from 1, indicating error-free measurement, and 0, indicating no variation other than error.

One of the main results of reliability theory, based on the assumption of independent responses within individuals, is that the size of the error component in the test score is constant as the number of items increases, whereas the variance of the individual-difference component increases proportionately with the number of items. The reliability coefficient therefore tends to unity as the number of items increases indefinitely, assuming that the items that are added have parallel content. The effect on reliability of increasing the test length by some arbitrary factor is given in the *Spearman–Brown formula*. In practice, the between-respondent variance component and the item-by-respondent interaction are estimated by analysis of variance methods or related formulas.

1.3.3 Contributions from Statistics

Simultaneous with, but largely independent of, the developments in psychological scaling and reliability theory, similar concepts and methods were elaborated in the field of statistics. Fechner's constant method reappeared in the 1920s