Preface

Prior to 1980, the subject of equating was ignored by most people in the measurement community except for psychometricians, who had responsibility for equating. Beginning in the early 1980s, the importance of equating was recognized by a broader spectrum of people associated with testing. This increased attention to equating is attributable to at least three developments. First, there continues to be an increase in the number and variety of testing programs that use multiple forms of tests, and the testing professionals responsible for such programs have recognized that scores on multiple forms should be equated. Second, test developers and publishers often have referenced the role of equating in arriving at reported scores to address a number of issues raised by testing critics. Third, the accountability movement in education and issues of fairness in testing have become much more visible. These developments have given equating an increased emphasis among measurement professionals and test users.

In addition to statistical procedures, successful equating involves many aspects of testing, including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests. Of course, psychometricians who conduct equating need to become knowledgeable about all aspects of equating. The prominence of equating, along with its interdependence with so many aspects of the testing process, also suggests that test developers and all other testing professionals should be familiar with the concepts, statistical procedures, and practical issues associated with equating.

Before we published the first edition in 1995, the need for a book on equating became evident to us from our experiences in equating hundreds of test forms in many testing programs, in training psychometricians to conduct equating, in conducting seminars and courses on equating, and in publishing on equating and other areas of psychometrics. Our experience suggested that relatively few measurement professionals had sufficient knowledge to conduct equating. Also, many did not fully appreciate the practical consequences of various changes in testing procedures on equating, such as the consequences of many test-legislation initiatives, the use of constructed-response items in assessments, and the introduction of computer-based test administration. Consequently, we believed that measurement professionals needed to be educated in equating methods and practices; the 1995 book was intended to help fulfill this need. Although several general published references on equating existed at the time (e.g., Angoff 1971;

Harris and Crouse 1993; Holland and Rubin 1982; Petersen et al. 1989), none of them provided the broad, integrated, in-depth, and up-to-date coverage of the first edition of this book.

After the publication of the first edition in 1995, a large body of new research was published. Much of this work was in technical areas that include smoothing in equipercentile equating, estimation of standard errors of equating, and the use of polytomous item response theory methods in equating. In addition, the use of constructed-response items and computer-based tests became more prominent. These applications create complexities for equating beyond what is typically encountered with paper-and-pencil multiple-choice tests. Thus, updating the material in the first edition was one of the reasons for publishing a second edition.

The first edition briefly considered score scales and test linking. The second edition devoted whole chapters to each of these topics. The development of score scales is an important component of the scaling and equating process. Linking of tests has been of much recent interest, due to various investigations of how to link tests from different test publishers or constructed for different purposes (e.g., Feuer et al. 1999). Because both scaling and linking are closely related to test equating, it seemed natural to extend coverage along these lines.

Following the publication of the second edition in 2004, a considerable amount of research was conducted on equating, scaling, and linking. In addition to a substantial number of journal articles, Dorans, Pommerich, and Holland (2007) and von Davier (2011) published edited books on equating, scaling, and linking. In addition, a substantial chapter by Holland and Dorans (2006) provides a conceptual framework for classifying equating and linking methodology that focuses on the properties of scores that are linked and on the requirements of different types of linking. A chapter by Kolen (2006) provides a updated discussion of score scales. The third edition updates all chapters to incorporate this recent literature. Following is a brief overview of the chapters of the third edition.

In Chap. 1, a general introduction is provided, primarily in terms of a conceptual overview. In this chapter, we define equating, describe its relationship to test development, and distinguish equating from scaling and linking. We also present equating designs, properties of equating, and introduce the concept of equating error.

In Chap. 2, using the random groups design, we illustrate traditional equating methods, such as equipercentile and linear methods. We also discuss here many of the key concepts of equating, such as properties of converted scores and the influence of the resulting scale scores on the choice of an equating result.

In Chap. 3, we cover smoothing methods in equipercentile equating. We show that the purpose of smoothing is the reduction of random error in estimating equating relationships in the population. We describe methods based on log-linear models, cubic splines, and strong true score models.

In Chap. 4, we treat linear equating with nonequivalent groups of examinees. We derive statistical methods and stress the need to disconfound examinee-group and test-form differences. Also, we distinguish observed score equating from true score equating.

In Chap. 5, we continue our discussion of equating with nonequivalent groups with a presentation of equipercentile methods.

In Chap. 6, we describe item response theory (IRT) equating methods under various designs. This chapter covers issues that include scaling person and item parameters, IRT true and observed score equating methods, equating using item pools, and equating using polytomous IRT models.

Chapter 7 focuses on standard errors of equating; both bootstrap and analytic procedures are described. We illustrate the use of standard errors to choose sample sizes for equating and to compare the precision in estimating equating relationships for different designs and methods.

In Chap. 8, we describe many practical issues in equating, including the importance of test development procedures, test standardization conditions, and quality control procedures. We stress conditions that are conducive to adequate equating. Also, we discuss comparability issues for mixed-format assessments and computer-based tests.

Chapter 9 is devoted to score scales for tests. We discuss different scaling perspectives. We describe linear and nonlinear transformations that are used to construct score scales, and we consider procedures for enhancing the meaning of scale scores that include incorporating normative, content, and score precision information. We discuss procedures for maintaining score scales and scales for batteries and composites. We conclude with a section on vertical scaling that includes consideration of scaling designs and psychometric methods and a review of research on vertical scaling.

In Chap. 10, we describe linking categorization schemes and criteria and consider equating, vertical scaling, and other related methodologies as a part of these categorization schemes. An extensive example is used to illustrate how the lack of group invariance in concordance relationships can be examined and used as a means for demonstrating some of the limitations of linking methods.

We use a random groups illustrative equating example in Chaps. 2, 3, and 7; a nonequivalent groups illustrative example in Chaps. 4–6; a second random groups illustrative example in Chaps. 6 and 9; and a single-group illustrative example in Chap. 10. We use data from the administration of a test battery in multiple grades for an illustrative example in Chap. 9, and data from the administration of two different tests for an illustrative example in Chap. 10. Chapters 1–10 each have a set of exercises that are intended to reinforce the concepts and procedures in the chapter. The answers to the exercises are in Appendix A. We describe computer programs and how to obtain them in Appendix B.

In addition to updating the review of literature for all of the chapters, the third edition incorporates substantial new material as follows:

- Chapter 3 includes additional procedures to choose models in log-linear pre-smoothing and includes a new brief section on the kernel method of equating.
- Chapter 4 includes a new section on chained linear equating and incorporates chained linear equating in the illustrative example. In addition, it includes a new

discussion of the relationships among linear methods in the common-item nonequivalent groups design.
- Chapter 5 includes new descriptions of modified frequency estimation equating and chained equipercentile equating, and incorporates these methods in the illustrative example.
- Chapter 8 includes a new extensive section on equating criteria in research studies. Material on equating mixed-format tests containing multiple-choice and constructed-response items is significantly updated.
- Chapter 9 includes a new section on unit scores, item scores, and raw scores. A new section on scores for mixed-format tests, including issues in weighting scores for different item types, is added. In addition, a new section on score scales and growth is added.
- Chapter 10 includes a new summary of the Holland and Dorans (2006) linking framework.

In addition, each chapter contains a reference list, rather than having a single reference list at the end of the volume as in the first two editions.

We anticipate that many readers of this book will be advanced graduate students, entry-level professionals, or persons preparing to conduct equating, scaling, or linking for the first time. Other readers likely will be experienced professionals in measurement and related fields who will want to use this book as a reference. To address these varied audiences, we make frequent use of examples and stress conceptual issues. This book is not a traditional statistics text. Instead, it is meant for instructional use and as a reference for practical use that is intended to address both statistical and applied issues. The most frequently used methodologies are treated, as well as many practical issues. Although we are unable to cover all of the literature on equating, scaling, and linking, we provide many references so that the interested reader may pursue topics of particular interest.

The principal goals of this book are for the reader to understand the principles of equating, scaling, and linking; to be able to conduct equating, scaling, and linking; and to interpret the results in reasonable ways. After studying this book, the reader should be able to

- Understand the purposes of equating, scaling, and linking and the context in which they are conducted.
- Distinguish between equating, scaling, and linking methodologies and procedures.
- Appreciate the importance to equating of test development and quality control procedures.
- Understand the distinctions among equating properties, equating designs, and equating methods.
- Understand fundamental concepts—including designs, methods, errors, and statistical assumptions.
- Compute equating, scaling, and linking functions and choose among methods.
- Interpret results from equating, scaling, and linking analyses.

- Design reasonable and useful equating, scaling, and linking studies.
- Conduct equating, scaling, and linking in realistic testing situations.
- Identify appropriate and inappropriate uses and interpretations of equating, scaling, and linking results.

We cover nearly all of the material in this book in a three semester-hour graduate seminar at The University of Iowa. In our course, we supplement the materials here with general references (Angoff 1971; Holland and Dorans 2006; Holland and Rubin 1982; Petersen et al. 1989) so that the students become familiar with other perspectives and notational schemes.

We have used much of the material in this book in various training sessions, including those at the annual meetings of the National Council on Measurement in Education, the American Educational Research Association, and the American Psychological Association, and in workshops given in Israel, Japan, South Korea, Spain, Taiwan, and The University of Iowa.

We acknowledge the generous contributions that others made to the first edition of this book. We benefitted from interactions with very knowledgeable psychometricians at ACT and elsewhere, and many of the ideas in this book came from conversations and interactions with these people. Specifically, Bradley Hanson reviewed the entire manuscript and made valuable contributions, especially to the statistical presentations. He conducted the bootstrap analyses that are presented in Chapter 7 and, along with Lingjia Zeng, developed much of the computer software used in the examples. Deborah Harris reviewed the entire manuscript, and we thank her especially for her insights on practical issues in equating. Chapters 1 and 8 benefitted considerably from her ideas and counsel. Lingjia Zeng also reviewed the entire manuscript and provided us with many ideas on statistical methodology, particularly in the areas of standard errors and IRT equating. Thanks to Dean Colton for his thorough reading of the entire manuscript, Xiaohong Gao for her review and for working through the exercises, and Ronald Cope and Tianqi Han for reading portions of the manuscript. We are grateful to Nancy Petersen for her in-depth review of a draft of the first edition, her insights, and her encouragement. Bruce Bloxom provided valuable feedback, as did Barbara Plake and her graduate class at the University of Nebraska–Lincoln. We thank an anonymous reviewer, and the reviewer's graduate student, for providing us with their valuable critique. We are indebted to all who have taken our equating courses and training sessions.

For the second edition, we are grateful to Ye Tong for the many hours she spent on electronic typesetting, for all of the errors she found, and for helping with many of the examples and the exercises. We thank Amy Hendrickson for helping to develop the polytomous IRT examples in Chapter 6, Seonghoon Kim for reviewing the additions to Chapter 6 on polytomous IRT and for developing the computer program POLYST, and Ping Yin for her work on Chapters 4 and 10. We acknowledge the work of Zhongmin Cui and Yueh-Mei Chien on the computer programs, and the work of Noo Ree Huh on checking references. We thank the students in our equating and scaling classes at The University of Iowa who discovered many errors and for helping us clarify some confusing portions of earlier drafts. We are grateful to Neil

Dorans, Samuel Livingston, and Paul Holland for reviewing portions of the new material in the second edition. We express our appreciation to the Iowa Measurement Research Foundation for providing support to the graduate students who worked with us on the second edition. For the third edition, we thank Wei Wang for her many hours spent on electronic typesetting. We also thank many graduate students at The University of Iowa for helping us correct errors that appeared in the second edition. Amy Kolen deserves thanks for her superb editorial advice for all three editions.

Iowa City, IA November, 2013

Michael J. Kolen
Robert L. Brennan

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Research Council.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education and Praeger.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. New York: Springer.

http://www.springer.com/978-1-4939-0316-0

Test Equating, Scaling, and Linking
Methods and Practices
Kolen, M.J.; Brennan, R.L.
2014, XXVI, 566 p. 68 illus., Hardcover
ISBN: 978-1-4939-0316-0