Statistics for Social and Behavioral Sciences

The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science. Successful applications of technology require comparable measurements. While the applications here focus on educational tests, score linking issues are directly applicable to medicine and many branches of behavioral science.

Since the 1980s, the fields of educational and psychological measurement have enhanced and widely applied techniques for producing linked scores that are comparable. The interpretation attached to a linkage depends on how the conditions of the linkage differ from the ideal. In this book, experts in statistics and psychometrics describe classes of linkages, the history of score linkings, data collection designs, and methods used to achieve sound score linkages. They describe and critically discuss applications to a variety of domains including equating of achievement exams, linkages between computer-delivered exams and paper-and-pencil exams, concordances between the current version of the SAT® and its predecessor, concordances between the ACT® and the SAT®, vertical linkages of exams that span grade levels, and linkages of scales from high-stakes state assessments to the scales of the National Assessment of Educational Progress.

Dr. Neil J. Dorans is a Distinguished Presidential Appointee at Educational Testing Service. During his 27 years at ETS, he has had primary responsibility for the statistical work associated with the AP®, PSAT/NMSQT®, and SAT® exams. He was the architect for the recentered SAT scales. He has guest edited special issues on score linking for *Applied Measurement in Education*, *Applied Psychological Measurement*, and the *Journal of Educational Measurement*.

Dr. Mary Pommerich is a psychometrician in the Personnel Testing Division of the Defense Manpower Data Center, where she works with the ASVAB (Armed Services Vocational Aptitude Battery) testing program. She guest edited a special issue on concordance for *Applied Psychological Measurement*. Her research is typically generated by practical testing problems and has focused on a wide variety of issues, including linking and concordance.

Dr. Paul W. Holland is the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service and before that professor in the School of Education and the department of Statistics at the University of California, Berkeley. His books include *Discrete Multivariate Analysis*, *Differential Item Functioning*, *Perspectives on Social Network Research*, and two books on test score equating. He is a fellow of the American Statistical Association and the Institute for Mathematical Statistics, was designated a National Associate of the National Academies, was awarded for his career contributions by the National Council on Measurement in Education, and was elected to the National Academy of Education.

EDUCATION

ISBN 978-0-387-49770-9

9 780387 497709

springer.com

Dorans · Pommerich · Holland

Linking and Aligning Scores and Scales

Statistics for Social and Behavioral Sciences

Neil J. Dorans · Mary Pommerich
Paul W. Holland (Editors)

Linking and Aligning Scores and Scales

Springer

Statistics for Social and Behavioral Sciences

Neil J. Dorans · Mary Pommerich
Paul W. Holland (Editors)

Linking and Aligning Scores and Scales

Springer

Statistics for Social and Behavioral Sciences

Advisors:
S.E. Fienberg
W.J. van der Linden

Statistics for Social and Behavioral Sciences

Brennan: Generalizability Theory
DeBoeck/Wilson: Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach
Devlin/Fienberg/Resnick/Roeder: Intelligence, Genes, and Success: Scientists Respond to The Bell Curve
Dorans/Pommerich/Holland: Linking and Aligning Scores and Scales
Finkelstein/Levin: Statistical for Lawyers, 2nd ed.
Gastwirth: Statistical Science in the Courtroom
Handcock/Morris: Relative Distribution Methods in the Social Sciences
Johnson/Albert: Ordinal Data Modeling
Kolen/Brennan: Test Equating, Scaling, and Linking: Methods and Practices, 2nd ed.
Longford: Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician
Morton/Rolph: Public Policy and Statistics: Case Studies from RAND
van der Linden: Linear Models for Optimal Test Design
von Davier/Carstensen: Multivariate and Mixture Distribution Rasch Models
von Davier/Holland/Thayer: The Kernel Method of Test Equating
Zeisel/Kaye: Prove It with Figures: Empirical Methods in Law and Litigation

Neil J. Dorans
Mary Pommerich
Paul W. Holland
(Editors)

Linking and Aligning Scores and Scales

Foreword by Ida M. Lawrence

Springer

Neil J. Dorans
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
ndorans@ets.org

Paul W. Holland
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
pholland@ets.org

Mary Pommerich
Monterey Bay Defense Manpower Data Center
400 Gigling Rd.
Seaside, CA 93955
USA
mary.pommerich@osd.pentagon.mil

Disclaimer: The opinions expressed by Mary Pommerich are her own and not necessarily those of the Department of Defense or the United States Government.

Library of Congress Control Number: 2007920470

ISBN-10: 0-387-49770-6 e-ISBN-10: 0-387-49771-4
ISBN-13: 978-0-387-49770-9 e-ISBN-13: 978-0-387-49771-6

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Dedication

To Kirsten.
—N. J. D.

To Bob and Mellita
—M. P.

To Martha
—P. W. H.

Foreword

In their preface to the second edition of *Test Equating, Scaling, and Linking,* Mike Kolen and Bob Brennan (2004) made the following observation: “Prior to 1980, the subject of equating was ignored by most people in the measurement community except for psychometricians, who had responsibility for equating” (p. vii). The authors went on to say that considerably more attention is now paid to equating, indeed to all forms of linkages between tests, and that this increased attention can be attributed to several factors:

1. An increase in the number and variety of testing programs that use multiple forms and the recognition among professionals that these multiple forms need to be linked.
2. Test developers and publishers, in response to critics, often refer to the role of linking in reporting scores.
3. The accountability movement and fairness issues related to assessment have become much more visible.

Those of us who work in this field know that ensuring comparability of scores is not an easy thing to do. Nonetheless, our customers—the test-takers and score users—either assume that scores on different forms of an assessment can be used interchangeably or, like the critics above, ask us to justify our comparability assumptions. And they are right to do this. After all, the test scores that we provide have an impact on decisions that affect people’s choices and their future plans. From an ethical point of view, we are obligated to get it right.

With the increased spotlight on linking, we have also seen interest in providing more sophisticated and complex kinds of assessment for tests designed for making high-stakes decisions. As we introduce more constructed response questions into our assessments, the challenge of linking increases. For example, when constructed response items are used as linking items, we are making the implicit claim that the raters scored the question the same way both times. How to control for differences in scoring at different administrations is a tricky business but is essential to successful linking. When test questions are scored by humans, instead of by machines, what mechanisms are needed to ensure that scores on reused test forms can be reported without a check on the stability of the scoring of the constructed response portions?

The No Child Left Behind Act of 2001 has spawned a strong market interest in formative assessments and assessments for other low-stakes decisions. We need to remind ourselves, and others, that linking issues need to play a role in assessment for lower stakes decisions. Without attention to score comparability on these formative assessments, we run the risk of giving bad instructional advice. The challenge lies in determining what kinds of standards need to apply to scores on these kinds of test.

A final challenge relates to improved communication about the practical consequences of addressing linking issues at the design phase for a testing program and as an ongoing activity in order to ensure fair and meaningful scores. We need to do a better job of helping decision-makers and policy folks understand the issues around equating and linking. We need to explain the limitations of the methods and the cost of being able to make truthful claims about score comparability.

This volume takes important steps in preparing us for these challenges. It examines foundational issues that cut across different types of linking. It delves into issues that are particularly germane to different classes of linking.

Ida M. Lawrence
ETS Senior Vice-President of Research & Development
January 2007

Preface

In 1980, an Educational Testing Service (ETS) equating conference led to a book (Holland & Rubin, 1982) that was one of first to bring professional attention to the critical statistical practice of equating. At that time, equating was a trade practiced by a small group of applied psychometricians, and equating practices were passed down from experts to novices.

Shortly after that book was published, both Neil Dorans and Paul Holland became intrigued by a simple question: When is an equating a good equating? Put another way, how do we evaluate the quality of an equating?

About 15 years later, Holland chaired a National Research Council committee that produced a report, *Uncommon Measures* (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), giving an accessible summary of informed, professional judgment about the issues involved in linking scores on different educational tests. Congressional requests to provide advice on how to link scores on tests that cover similar material was the impetus for the profession's response delivered in *Uncommon Measures*.

Around the same time, Neil Dorans and Mary Pommerich collaborated to produce a concordance between scores on the ACT® and SAT®, the two major college admissions tests in the United States (Dorans, Lyu, Pommerich, & Houston, 1997). This work led to an interest in better understanding how equating differs from other types of linkage between scores and when different types of linkage should be conducted. In time, a special *Applied Psychological Measurement* issue on concordance was co-edited by Pommerich and Dorans (2004a). Drawing distinctions among types of linkage was an important theme in that special issue.

Returning to the question of what constitutes an equating, Dorans and Holland (2000) introduced indexes for quantifying how much an equating depends on the subpopulation in which it is conducted. The importance of population invariance as a check on equatability has developed rapidly since 2001, as evidenced by a special issue on the topic in the *Journal of Educational Measurement*, edited by Dorans (2004a).

In June, 2005, Dorans and Holland organized another ETS-sponsored conference.[1] Demonstrating a shift in focus from the seminal conference held 25 years earlier, the 2005 conference focused on the more general issue of linking, of which equating was but one topic of discussion. The conference was dedicated to Professor Ledyard R Tucker,[2] one of the early theorists and practitioners of equating. The conference provided raw material for this volume.

During the 25 years between the two ETS conferences, several books addressed issues in score linking. The volume by Kolen and Brennan (2004), in its second edition, is an encyclopedic treatment of the field of equating, scaling, and linking. von Davier, Holland, and Thayer (2004b) focused on kernel equating as a unified approach that introduces several new ideas of general use in equating. In addition to *Uncommon Measures*, another report on score linking from the National Research Council is *Embedding Questions* (Koretz, Bertenthal, & Green, 1999). Finally, the work of Livingston (2004) is a user-friendly account of many of the major issues and techniques.

Where does this volume fit into the array of books that have been written about equating and linking? Simply, it is more about score linking than score equating. We place a strong emphasis on distinguishing between different kinds of linking and the inferences that can be associated with each type of linking. This volume examines the different types of linking from both theoretical and practical perspectives. Theory that ignores reality is doomed to be irrelevant. Practice that occurs without an appreciation of the theory of linking is likely to be influenced by the biases of the practitioner. This volume emphasizes the importance of both theory and practice.

Several ETS staff provided essential support. Martha Thompson organized the linking conference that was attended by 200 assessment professionals. She and Liz Brophy turned a concept into a reality. John Mazzeo, Associate Vice-President for Statistical Analysis and Research, and Ida Lawrence, Senior Vice-President of Research and Development at ETS, supported the conference. As experienced linkers, they readily endorsed production of this volume as well. The volume benefited from the administrative skills of Liz Brophy and the editorial skills of Kim Fryer.

[1] Linking and Aligning Scores and Scales, a conference in honor of Ledyard R Tucker's approach to theory and practice, was held at Princeton University on June 24–25, 2005.

[2] A brief history of Ledyard R Tucker's professional life can be found in Dorans (2004b).

Contents