

Statistics for Social and
Behavioral Sciences

Statistics for Social and
Behavioral Sciences

The comparability of measurements made in differing circumstances for different methods and instruments has long been a pre-occupation for all of us. Various theoretical applications of linking major comparative measurements, while the applications here focus on educational measurements, bring these two closely together in a unified and comprehensive treatment of educational issues.

Over the years, the field of educational and psychological measurement has advanced and widely applied techniques for producing related scores that are comparable. The comprehensive and state-of-the-art approach in this book, the conditions of the linkage effort, take for what is the most logical to consider and fundamental to the theory of linkage, the history of such linkages, how different linkages and methods relate to each other, how linkages may differ, and how to choose applications to a variety of domains including, reporting of information, testing, linkage between computer-administered tests and paper-and-pencil tests, cross-validation between the various modes of the SAT and its predecessor, comparisons between the SAT and SAT-2, modes of linkage of scores for year-grade levels, and linkage of scores from high school tests administered in the states of the National Assessment of Educational Progress.

Dr Neil J. Dorans is a distinguished professional specialist in Educational Testing Service. During his 27 years at ETS he has had primary responsibility for the national level operations and all the SAT, SAT-2, SAT-3, and SAT-9 exams. He was the architect of the national SAT scale linkage program which provided more test score linkage for applied measurement, education, applied psychology, management, and the journal of Educational Measurement.

Dr Mary Pommerich has professionalized in the Educational Testing Service of the United States since 1976. She has published widely with the journal, Journal of Educational Measurement. Her research interests include cross-national test score linkage, the joint effect of a national test on construction of the applied psychology research area. She currently is largely concerned with providing testing professionals the theoretical and practical aspects of cross-testing testing and measurement.

Dr Paul W. Holland is the Professor of Educational and Psychological Measurement for the Center and Center that provides a theoretical foundation and for applications of statistics to the measurement of education. He has published in Educational Measurement, Journal of Educational Measurement, Psychometrics in Social Science Research, and other journals. He has been awarded the Excellence in the Education of Measurement and the Institute for Educational Statistics has designated a National Institute of the National Academies has awarded for his work contributions to the National Institute for Measurement in Education, and was awarded the National Institute of Education.

© Neil J. Dorans



Springer.com

Psychology Research Review



Linking and Aligning Scores and Scales

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

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.

ETS, GRE, Praxis and TOEFL are registered trademarks of Educational Testing Service (ETS).
CLEP, COLLEGE BOARD, PSAT/NMSQT, and SAT are registered trademarks of College Board.
SAT REASONING TEST is a trademark of College Board.
ACT is a registered trademark of ACT, Inc. in the U.S.A. and other countries.

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.

© 2007 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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.

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

Contents

1 Overview	1
Part 1: Foundations	3
2 A Framework and History for Score Linking.....	5
<i>Paul W. Holland</i>	
2.1 Introduction	5
2.2 Predicting	6
2.2.1 Predicting Observed Scores.....	7
2.2.2 Projecting Distributions of Observed Scores.....	9
2.2.3 Predicting True Scores.....	10
2.2.4 Summary.....	11
2.3 Scale Aligning	11
2.3.1 Battery Scaling: Different Constructs and a Common Population of Examinees.....	12
2.3.2 Anchor Scaling: Different Constructs and Different Populations of Examinees	14
2.3.3 Vertical Scaling: Similar Constructs and Similar Reliability, But Different Difficulty and Different Populations of Examinees	17
2.3.4 Calibration: Same Construct, Different Reliability, and the Same Population of Examinees.....	18
2.3.5 Concordances: Similar Constructs, Difficulty, and Reliability	19
2.4 Equating: Same Construct and the Same Intended Difficulty and Reliability	20
2.4.1 What Makes a Linking an Equating?	22
2.4.2 A Crucial Consideration for Scale Aligning and Equating.....	25
2.4.3 A Brief Outline of Equating Methods	26
2.5 A Brief Note on the Theory of Equating	29
3 Data Collection Designs and Linking Procedures.....	31
<i>Michael J. Kolen</i>	
3.1 Introduction	31
3.2 Features of Testing Situations	32

3.2.1 Test Content.....	32
3.2.2 Conditions of Measurement.....	33
3.2.3 Examinee Population.....	33
3.2.4 Construct Measured.....	33
3.3 Types of Linking Considered	33
3.4 Linking Functions and Features of Testing Situations	35
3.5 Linking Designs.....	37
3.5.1 Random Groups Design for Equating	38
3.5.2 Random Groups Design and Variations for Linking.....	39
3.5.3 Single Group Design with Counterbalancing for Equating	41
3.5.4 Single Group Design with Counterbalancing and Variations for Linking	43
3.5.5 Common-Item Nonequivalent Groups Design for Equating ...	45
3.5.6 Anchor-Test Nonequivalent Groups Design for Linking	47
3.6 Linking Procedures.....	49
3.6.1 Traditional Statistical Methods for Equating.....	49
3.6.2 IRT Statistical Methods for Equating.....	52
3.6.3 Methods for Linking Tests Intended to Measure Similar Constructs.....	53
3.7 Summary and Conclusions	54
Part 2: Equating.....	57
4 Equating: Best Practices and Challenges to Best Practices	59
<i>Nancy S. Petersen</i>	
4.1 Equating.....	59
4.2 Best Practices.....	61
4.2.1 Data Collection.....	61
4.2.2 Total Tests and Anchor Tests	62
4.2.3 Equating Process	62
4.3 Challenges to Best Practices.....	63
4.3.1 Choice of Data Collection Design.....	64
4.3.2 Psychometric Properties of the Total Tests and Anchor Test..	66
4.3.3 Samples.....	67
4.3.4 Problems in Implementation.....	67
4.4 Discussion.....	70
4.5 Summary.....	71
5 Practical Problems in Equating Test Scores: A Practitioner’s Perspective.....	73
<i>Linda L. Cook</i>	
5.1 Introduction	73
5.2 The Nonequivalent Groups Anchor Test Design.....	75

5.3 Characteristics of the New and Old Forms	76
5.4 Characteristics of the Groups Used for Equating	79
5.5 Characteristics of the Anchor Test (Common Items)	84
5.6 Conclusions	87
6 Potential Solutions to Practical Equating Issues.....	89
<i>Alina A. von Davier</i>	
6.1 Introduction	89
6.2 Observed-Score Equating Methods	91
6.3 Addressing the Fairness Issue: Population Invariance of Equating Functions.....	92
6.3.1 Definitions and Measures of Population Differences in Equating	92
6.3.2 Criteria for Detecting Subpopulation Differences in Equating Functions.....	93
6.3.3 Implications of Population Sensitivity of Equating Functions	94
6.3.4 Discussion and Future Research Directions	96
6.4 Addressing the Small-Samples Issue: Synthetic Linking Functions	96
6.5 Addressing Differences in Ability in the Two Populations of the NEAT Design.	99
6.6 Addressing the Stability of Equating Results: Kernel Equating and Applications.....	101
6.6.1 The Gaussian Kernel Method.....	101
6.6.2 Applications of the KE Framework.....	102
6.6.3 Discussion and Future Research Directions	103
6.7 Discussion.....	105
Part 3: Tests in Transition	107
7 Score Linking Issues Related to Test Content Changes	109
<i>Jinghua Liu and Michael E. Walker</i>	
7.1 Introduction	109
7.2 Major Linking Issues for Tests in Transition	111
7.2.1 Changes to the New SAT Critical Reading in 2005	112
7.2.2 Requirements for Equating	114
7.2.3 Criteria to Determine Equatability of the New SAT	116
7.3 Considerations for Data Collection Design	117
7.3.1 Linking Designs.....	117
7.3.2 Sample Size Considerations	120
7.4 Equatability Analyses	122

7.4.1 Comparison of Test Specifications Between the New SAT and the Old SAT	122
7.4.2 Empirical Relationship Between the Old and New Tests.....	123
7.4.3 Comparison of Measurement Precision on the Old and New Tests	124
7.4.4 Score Equity Assessment	126
7.5 Discussion.....	133
8 Linking Scores Derived Under Different Modes of Test Administration	135
<i>Daniel R. Eignor</i>	
8.1 Introduction	135
8.2 Background.....	136
8.3 Comparability Issues Involving Scores from Computer-Based and Paper-and-Pencil Tests	138
8.4 Mode of Presentation Linking Designs	141
8.5 Random Groups Design.....	142
8.5.1 General Discussion.....	142
8.5.2 Equating Studies Done with the Random Groups Design.....	145
8.5.3 Calibration Studies Done with the Random Groups Design	145
8.6 Single Group Counterbalanced Test Design.....	147
8.6.1 General Discussion.....	147
8.6.2 Equating Studies Done with the Single Group Counterbalanced Design	149
8.6.3 Calibration Studies Done with the Single Group Counterbalanced Design	150
8.6.4 Concordance Studies Done with the Single Group Counterbalanced Design	152
8.7 Anchor Test: Nonequivalent Groups Design.....	154
8.7.1 General Discussion.....	154
8.7.2 Calibration Studies Done with the Anchor Test Design.....	156
8.8 Summary.....	157
9 Tests in Transition: Discussion and Synthesis.....	161
<i>Robert L. Brennan</i>	
9.1 The Liu and Walker Chapter on Test Content Changes	162
9.1.1 Content Specifications and Item Characteristics	163
9.1.2 Empirical Relationships.....	164
9.1.3 Reliability and CSEMs	165
9.1.4 Subpopulation Invariance for Males and Females.....	165
9.1.5 Other Comments.....	167
9.2 Eignor Chapter on Mode of Administration.....	168
9.2.1 Types of Linking	168