Lecture Notes in Statistics 202
Proceedings

Looking Back

Proceedings of a Conference in Honor of Paul W. Holland

Springer

Lecture Notes in Statistics – Proceedings **202**

Edited by P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

For further volumes:
http://www.springer.com/series/694

Neil J. Dorans • Sandip Sinharay
Editors

Looking Back

Proceedings of a Conference in Honor of Paul W. Holland

Springer

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

Sandip Sinharay
Educational Testing Service
Rosedale Road
Princeton, NJ 08541, USA
ssinharay@ets.org

ISSN 0930-0325
ISBN 978-1-4419-9388-5 e-ISBN 978-1-4419-9389-2
DOI 10.1007/978-1-4419-9389-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011931535

To Paul and Roberta

Foreword

It is honor and privilege to be asked to provide the foreword to *Looking Back*. As an academic statistician, as a director of a research department at Educational Testing Service (ETS), as a colleague, as a mentor, and as a consultant inside ETS as well as to various external statistical and scientific agencies, Paul Holland throughout his illustrious career has made significant contributions to theory and practice in the fields of psychometrics, statistics, and social science research. On a more personal note, I have been fortunate to have spent most of my career at ETS during a period in which Paul was also employed there. Although I did not collaborate as directly with Paul as did the authors of the various chapters of this book, it is not difficult to discern Paul's influence on my own professional career in terms of what I know about statistics and psychometrics, the kind of activities I engaged in as a practicing psychometrician, and the stewardship of testing programs I was required to provide as an ETS technical leader. What is true for me is, I believe, true for many of the statistics and psychometric staff of my vintage – at ETS as well as elsewhere.

I attended graduate school at the University of Arizona in the late 1970s and early 1980s and, as part of my degree program, took an applied statistics course in the sociology department. The course was in the area of analysis of contingency tables using log-linear models. The primary text for the course was a book by Stephen Fienberg (one of the contributors to this volume) called *The Analysis of Cross-Classified Data* (2nd edition). But looming in the background as highly recommended supplementary material was a more imposing tome, *Discrete Multivariate Analysis: Theory and Practice* by Yvonne Bishop, Stephen Fienberg, and one Paul W. Holland. Throughout the course, we were assigned sections of this tome as supplementary reading and, for someone like me with relatively modest mathematical training, I found the material enlightening, though challenging and intimidating as well. As a result of this experience, I was very familiar with the name Paul Holland and had learned at least some of what I know about log-linear models and their applications from him well before I ever set foot on the ETS campus. I viewed Paul as a sort of rock star in the area of discrete data analysis,

and one of the things that made it exciting and desirable to come to ETS after I completed graduate school was the opportunity to work for an organization that employed the great man himself.

I joined ETS in 1984, as what we called then an associate measurement statistician. I was responsible for overseeing statistical and psychometric support activities for several ongoing ETS testing programs. While I had some measurement and applied statistics background, like many freshly minted graduate students, I had very limited experience with score equating – the statistical process testing companies use to ensure that scores from different forms of the same test (e.g., different administrations of the SAT) are expressed on a common scale. Then, as well as today, equating tests constituted a large portion of the activities of ETS psychometricians. So as part of my early on-the-job education, I tried to learn as much as I could, and as quickly as I could, about equating. Of course, I read various ETS memos and orientation materials that were given to me as a new employee. However, I also read what was then a relatively new book, *Test Equating*, edited by Paul Holland and Don Rubin. In it was a chapter by Paul and Henry Braun titled "Observed-Score Test Equating: A Mathematical Analysis of Some ETS Procedures." In that chapter, Paul and Henry laid out a formal statistical framework for describing equating procedures in widespread use at ETS. This chapter helped me greatly to organize and make sense of the various documents about equating that I was reading and to better understand the nature of what I was seeking to accomplish in my day-to-day work as an ETS measurement statistician. I am certain that Paul and Henry's chapter accelerated my development and made me a more effective measurement professional than I otherwise would have been.

Of course, throughout the 1980s and early 1990s, like most of my ETS colleagues I had the pleasure to see Paul's work on differential item functioning (DIF) develop and contribute directly to a substantial research program and, more importantly, to improved statistical procedures for ensuring fairness. The resulting methodologies and rules of thumb that Paul and his colleagues developed became standard operating procedure at ETS and continue to this day. So, once again, my understanding of statistical approaches to assessing fairness issues and the day-to-day activities of testing professionals at ETS, and I would guess other companies as well, were in no small part shaped by Paul's contributions to psychometric theory and practice.

Paul, much to our chagrin, left ETS in 1993, taking an academic position at the University of California at Berkeley. Near the end of last century, Paul Ramsey and Drew Gitomer, both ETS vice presidents at that time, initiated a concerted effort to strengthen ETS's statistical and psychometric foundation. Paul Ramsey asked Steve Lazer and me to speak with colleagues and to prepare A and B lists of statisticians/psychometricians we should try to hire. After a number of colleagues were consulted, it was clear that at the top of everyone's A list was Paul Holland. Fortunately, Paul was ready to consider coming back to ETS, as he notes in *Returning to ETS From Berkeley* in this volume, and Paul Ramsey and Drew Gitomer were able to make that happen. The impact of Paul's return to the ETS was immediate and profound. He re-established his program of research on

equating, presaged in the Braun and Holland chapter, which resulted in the publication of the book *The Kernel Method of Test Equating* with Alina von Davier and Dorothy Thayer. This work also led to the creation and deployment of software for implementing the approach operationally.

Paul began attending National Assessment of Educational Progress technical advisory committee meetings – contributing to discussions surrounding technical matters associated with this important testing program. He produced several white papers on issues associated with the impact on NAEP of the newly passed No Child Left Behind Act, and, generally, through his wisdom and guidance, helped those of us charged with directing NAEP psychometric activities better manage the NAEP program through a period of rapid change. Through his activities he demonstrated to the NAEP sponsors (the National Center for Education Statistics and the National Assessment Governing Board) what we all knew from working with him over the years – that he is not only a world-class researcher, but one who is willing to use those gifts in tackling problems of real practical importance.

But the impact of Paul's return on ETS went beyond his contributions to NAEP. Drew Gitomer recounted to me how he had sent a company-wide announcement of Paul's return to ETS and was amazed at the sheer number of positive responses he received from not just the technical areas but from all parts of ETS, indicating how happy people were that he was returning and how they were looking forward to working with him. The conference proceedings that are captured here in *Looking Back* are a fitting recognition and celebration of Paul's substantial impact on ETS and the profession.

John Mazzeo
Vice President
Statistical Analysis &
Psychometric Research
Educational Testing Service

Preface

In 2006, Paul W. Holland retired from Educational Testing Service (ETS) after a career spanning five decades. In 2008, ETS sponsored a conference, *Looking Back*, honoring Paul's contributions to applied and theoretical psychometrics and statistics. *Looking Back* attracted a large audience that came to pay homage to Paul and to hear presentations by colleagues who worked with Paul in special ways over those 40+ years. This book contains papers based on these presentations, as well as vignettes provided by Paul before each section.

Shelby Haberman, the eminent statistician who is a long-time contemporary of Paul's, was attracted to ETS by Paul in 2002. Shelby is very conversant about the history of statistics. In *The Contributions of Paul Holland*, Shelby provides a history with commentary on some of Paul's major contributions.

The first collection of papers appears under the heading *Holland the Young Scholar*. Two well-known statisticians, who worked closely with Paul in the 1970s when they all were young, contributed papers in this collection. Stephen Feinberg, co-author with Paul and Yvonne Bishop of the classic *Discrete Multivariate Analysis: Theory and Practice*, contributes *Algebraic Statistics for p_1 Random Graph Models: Markov Bases and Their Uses* with Sonja Petrović and Alessandro Rinaldo. In *Mr. Holland's Networks*, Stanley Wasserman, who was a doctoral student when Paul taught at Harvard, reports on work in social network theory that has evolved since Paul's seminal work with Sam Leinhardt.

As the title *Holland Shaping ETS* states for the next collection of papers, Paul applied statistical thinking to a broad range of ETS activities in test development, statistical analysis, test security, and operations. Donald Rubin attracted Paul to ETS in 1975 and co-edited with Paul the book *Test Equating*, which was one of first to bring professional attention to the critical statistical practice of score equating. Donald's *Bayesian Analysis of a Two-Group Randomized Encouragement Design* addresses a practical problem in causal inference, an area to which he and Paul made significant contributions. The development and implementation of procedures for differential item functioning (DIF) was one major application. Michael Zieky, who was at ETS when DIF was introduced, provides a valuable history of DIF in the 1980s in *The Origins of Procedures for Using Differential Item Functioning Statistics at*

Educational Testing Service. Brian Junker, who was a summer intern under Paul in the 1980s, contributes *The Role of Nonparametric Analysis in Assessment Modeling: Then and Now*. Paul Rosenbaum, an expert on statistical treatment of data from observational designs, contributes *What Aspects of the Design of an Observational Study Affect Its Sensitivity to Bias From Covariates That Were not Observed?*

Holland left ETS in the early 1990s to become a professor. The next section, *Holland the Berkeley Professor*, contains papers from two of his former students. Derek Briggs addresses a very current topic in *Cause or Effect? Validating the Use of Tests for High-Stakes Inferences in Education*. Ben Hansen assesses coaching effects in *Propensity Score Matching to Extract Latent Experiments From Nonexperimental Data: A Case Study*.

While Paul was at Berkeley, the productive group he left behind at ETS missed his guidance and leadership. Paul returned to ETS in 2000 and began to mentor a new set of young ETS professionals. Three of those lucky individuals contributed to *Holland Rebuilding ETS*. Tim Moses worked closely with Paul on several topics, including, as the title of his paper states, *Log-Linear Models as Smooth Operators: Holland's Statistical Applications and Their Practical Uses*. Sandip Sinharay, who worked with Paul on several topics, contributed *Chain Equipercentile Equating and Frequency Estimation Equipercentile Equating: Comparisons Based on Real and Simulated Data*. Alina von Davier discusses her work with Paul on his kernel-equating model and its extensions in *An Observed-Score Equating Framework*.

When Paul returned to ETS, he asked two ETS employees whom he had mentored to join his group. Henry Braun currently of Boston College and a former ETS Vice-President for Research and Neil Dorans of ETS made contributions to *Holland: From Mentor to Colleague*. Henry, an expert in the application of statistics to issues in educational policy, contributes *An Exploratory Analysis of Charter Schools*. Neil, who focuses on fairness assessment topics including DIF and equating, builds upon Paul's historical review of testing in *Holland's Advice for the Fourth Generation of Test Theory: Blood Tests Can Be Contests*.

The papers in this book attest to how Paul's pioneering ideas influenced and continue to influence several fields such as social networks, causal inference, item response theory, equating, and DIF.

Through *Looking Back* and this book, we thank Paul for service to our field and years of generous and wise advice to us and to his many students and colleagues. Anyone who has met and talked with Paul will share our gratitude to a man who inspired with his intelligence and encouraged with his enthusiasm for life.

Our deepest thanks go to all contributors for their generosity, help, and patience and also to the participants in *Looking Back*. Several ETS staff provided essential support. Liz Brophy and Jazzme Blackwell organized the conference, which was attended by 100 scholars. The book benefited from the editorial acumen of Kim Fryer. The conference and book were supported by a research allocation from the ETS Research & Development division led by Senior Vice President Ida Lawrence.

Princeton, NJ

Neil J. Dorans
Sandip Sinharay

Contents

Contributors

Henry I. Braun Lynch School of Education, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

Derek C. Briggs School of Education, University of Colorado at Boulder, 249 UCB, Boulder, CO 80309, USA

Neil J. Dorans Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Stephen E. Fienberg Department of Statistics, Carnegie Mellon University, 132G Baker Hall, Pittsburgh, PA 15213, USA

Shelby J. Haberman Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Ben B. Hansen Statistics Department, 439 West Hall, University of Michigan, Ann Arbor, MI 48109–1107, USA

Paul W. Holland 703 Sayre Dr., Princeton, NJ 08540, USA

Brian W. Junker Department of Statistics, Carnegie Mellon University, 132E Baker Hall, Pittsburgh, PA 15213, USA

John Mazzeo Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Tim P. Moses Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Sonja Petrović Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 322 Science and Engineering Offices (M/C 249), 851 S. Morgan Street, Chicago, IL 60607–7045, USA

Alessandro Rinaldo Department of Statistics, Carnegie Mellon University, 229I Baker Hall, Pittsburgh, PA 15213, USA

Paul R. Rosenbaum Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104–6340, USA

Donald B. Rubin Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, USA

Kathleen M. Sheehan Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Sandip Sinharay Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Christina Tang Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Alina A. von Davier Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Stanley Wasserman Department of Statistics, Indiana University, 309 North Park Street, Bloomington, IN 47408, USA

Michael J. Zieky Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Part I
Paul Holland's Contributions

Chapter 1
The Contributions of Paul Holland

Shelby J. Haberman

1.1 Introduction

Paul Holland's work over his long and varied career has shown both breadth and depth. He has made major contributions to the analysis of discrete data, to the study of social networks, to equating, to differential item functioning (DIF), to item response theory (IRT), and to causal inference. He has worked on a wide variety of applied problems ranging from scanner accuracy to test security to summarization of data on candidates. Any review of his contributions will necessarily provide a rather limited indication of his achievements. Nonetheless, several instructive themes can be found in his work. One is the long-standing connection with the analysis of discrete data. A second is a longstanding connection to the social and behavioral sciences. A third is an emphasis on the observed over the unobserved in the analysis of data. These themes interact and have been demonstrated in Paul's work at least since graduate school. Paul's doctoral dissertation concerned a new minimum chi-square test. His involvement in research in the social sciences reflects both his family background and his early association with his dissertation advisor Patrick Suppes (Robinson, 2005). The emphasis on the observed can be seen in his emphasis on observed-score equating and log-linear models rather than on latent-structure models, although Paul has made major contributions to IRT.

This overview of Paul's work is necessarily selective and biased. For example, Paul is a coauthor of a highly influential work on discrete multivariate analysis (Bishop, Fienberg, & Holland, 1975); however, I will concentrate here on contributions that are more specifically connected to Paul himself. In addition, due to my own limited knowledge, causal inference will be less examined than is appropriate given its significance in Paul's work. This review will emphasize DIF,

S.J. Haberman (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: shaberman@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_1,

IRT, social networks, and kernel equating. Briefer consideration will be given to other contributions to equating, causal inference, and the analysis of empirical data.

1.2 Differential Item Functioning

A good example of the application of methodology for analysis of contingency tables to educational measurement arises in testing for DIF by use of the Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959). In Bishop et al. (1975, pp. 147–148), this statistic is described in terms of a test of conditional independence of two dichotomous random variables given a polytomous variable. No connection to psychometrics is contemplated. The data are independent and identically distributed triples (A_h, B_h, C_h), $1 \le h \le N$, with A_h and B_h equal 0 or 1 and C_h an integer from 0 to $k-1$ for some integer $k \ge 2$. The probability p_{abc} that $A_h = a$, $B_h = b$, and $C_h = c$, $0 \le a \le 1$, $0 \le b \le 1$, and $0 \le c \le k-1$, is assumed to be positive. The null hypothesis under study is that A_h and B_h are conditionally independent given C_h. To test this hypothesis, one considers the counts n_{abc} of h such that $A_h = a$, $B_h = b$, and $C_h = c$. Let n_{a+c} be the number of h with $A_h = a$ and $C_h = c$, let n_{+bc} be the number of h with $B_h = b$ and $C_h = c$, and let n_{++c} be the number of h with $C_h = c$. Under the null hypothesis, the expected value $m_{abc} = Np_{abc}$ of n_{abc} has maximum-likelihood estimate $\hat{m}_{abc} = n_{a+c}n_{+bc}/n_{++c}$, at least if n_{++c} is positive. Mantel and Haenszel considered the marginal total n_{11+}, the number of h with $A_h = B_h = 1$. Under the null hypothesis, the estimated expected value of n_{11+} is $\hat{m}_{11+}$, the sum of the expected values $\hat{m}_{11c}$ for $1 \le c \le k$. If $n_{++c} > 1$ for each c, then conditional on the observed values of n_{a+c} and n_{+bc}, the difference $n_{11+} - \hat{m}_{11+}$ has variance

$$V = \sum_{c=1}^{k} \hat{m}_{11c} n_{2+c} n_{+2c} / [n_{++c}(n_{++c} - 1)]$$

Mantel and Haenszel (1959) suggested use of $Z = (n_{11+} - \hat{m}_{11+})/V^{1/2}$ to test the hypothesis of conditional independence. If the null hypothesis holds, then Z converges in distribution to a standard normal random variable.

As noted in Bishop et al. (1975), the MH statistic has an important optimality property. Consider the log-linear model of no three-factor interaction in which it is assumed that each log cross-product ratio

$$\log\left(\frac{m_{11c}m_{22c}}{m_{21c}m_{12c}}\right) = \log m_{11c} - \log m_{21c} - \log m_{12c} + \log m_{22c}$$

has a common value τ. If $\tau = 0$, then A_h and B_h are conditionally independent given C_h. The uniformly most powerful unbiased test of the null hypothesis of conditional independence of A_h and B_h given C_h against the alternative hypothesis of no three-factor interaction depends on the MH statistic Z (Birch, 1964).

In a typical application to DIF, $A_h = 1$ if h is an examinee with a correct response to an item, $A_h = 0$ otherwise, $B_h = 1$ if h belongs to some group of interest, say female examinees, $B_h = 0$ if h belongs to a reference group, say male examinees, and C_h is a polytomous variable typically determined by the total score of h on the examination. The null hypothesis is that the relationship of the item response A_h to the score variable C_h is unaffected by the group B_h (Holland & Thayer, 1988), so that A_h and B_h are conditionally independent given C_h. This application of this familiar statistic had a remarkable effect on an entire field, as is evident from an edited volume on DIF that soon appeared (Holland & Wainer, 1993).

An interesting aspect of the development of DIF is the decision to use the MH estimate of the common cross-product ratio $q = \exp(\tau)$ (Mantel & Haenszel, 1959). Let

$$d_c = (n_{11c} + n_{22c})/n_{++c},$$

$$e_c = (n_{12c} + n_{21c})/n_{++c},$$

$$f_c = n_{11c}n_{22c}/n_{++c},$$

$$g_c = n_{12c}n_{21c}/n_{++c},$$

$$f_+ = \sum_{c=1}^{k} f_c,$$

$$g_+ = \sum_{c=1}^{k} g_c,$$

and

$$v_c = \frac{1}{n_{11c}} + \frac{1}{n_{12c}} + \frac{1}{n_{21c}} + \frac{1}{n_{22c}}.$$

Then q has MH estimate $O = f_+/g_+$ and τ has estimate $T = \log O$. The considerations that entered into this decision reflected the computational environment in existence at the time. The MH estimate is easily computed, and has a normal approximation. Let

$$s^2(T) = \frac{1}{2}\sum_{c=1}^{k}(d_c/f_+ + e_c/g_+)(f_c/f_+ + g_c/g_+)$$

and

$$s(O) = Os(T).$$

As the sample size N becomes large, $(Q-q)/s(O)$ and $(T-\tau)/s(T)$ both converge in distribution to a standard normal random variable (Phillips & Holland, 1987), so that approximate confidence intervals are readily derived. A variety of alternatives to $s(T)$ and $s(O)$ are also available.

Nonetheless, alternatives to the MH estimate have been available since before the MH statistic was ever introduced (Woolf, 1955). The estimate

$$O_W = \exp(T_W)$$

can be used with

$$T_W = \frac{\sum_{c=1}^{k} \hat{\tau}_c/v_c}{\sum_{c=1}^{k} 1/v_c}$$

and

$$\hat{\tau}_c = \log(n_{11c}) - \log(n_{21c}) - \log(n_{12c}) + \log(n_{22c}).$$

As the sample size N becomes large, $(O_W-q)/s(O_W)$ and $(T_W-\tau)/s(T_W)$ converge in distribution to a standard normal random variable, where

$$s^2(T_W) = \frac{1}{\sum_{c=1}^{k} v_c^{-1}}$$

and

$$s(O_W) = Os(T_W).$$

To improve the accuracy of large-sample approximations and to avoid problems that arise if some count n_{abc} is 0, it is helpful to replace n_{abc} by $n_{abc}+0.5$ in the formulas for T_W and $s(T_W)$ (Haldane, 1955). Unless τ is 0, so that conditional independence holds, the probability is 1 that $s(T_W) < s(T)$ for sufficiently large N. If τ is 0, then $s(O)/s(O_W)$ converges to 1 with probability 1. It is not clear that the MH estimate O should be used rather than the Woolf estimate O_W, although study of O for use in DIF did yield results that suggested that s_O and s_W should be rather similar for the small values of τ of primary interest.

The common cross-product ratio q can also be obtained by maximum likelihood, but iterative computation is needed. Iterative proportional fitting was well known at the time, as evident in Paul's publications (Bishop et al., 1975, chap. 3), and Newton-Raphson algorithms were also available (Haberman, 1978, chap. 3); however, iterative computation was unattractive at the time. Similarly, use of conditional maximum likelihood to alleviate problems of small frequency counts was not practical given computational constraints (Birch, 1964). The question now is whether improvements in the computational environment warrant revisiting the methodology for DIF.

1.3 Item Response Theory

A somewhat more complex application of contingency tables has been to IRT. Here the basic observation is that in a right-scored test with $k \geq 2$ items and $n \geq 1$ examinees, the item responses X_{ij} of examinee i, $1 \leq i \leq n$, on item j, $1 \leq j \leq k$, can be used to develop a 2^k contingency table. Let X_{ij} be 1 if the response is correct, and let X_{ij} be 0 otherwise. Let $\mathbf{X}_i$ be the vector with coordinates X_{ij}, $1 \leq j \leq k$, and assume that the $\mathbf{X}_i$ are independent and identically distributed. For simplicity, assume that each response X_{ij} is 1 with positive probability and is 0 with positive probability. For each k-dimensional vector $\mathbf{x}$ with coordinates x_j equal to 0 or 1, let $p(\mathbf{x})$ be the probability that $\mathbf{X}_i = \mathbf{x}$, and let $f(\mathbf{x})$ be the number of examinees i with $\mathbf{X}_i = \mathbf{x}$, so that $f(\mathbf{x})$ has expected value $m(\mathbf{x}) = Np(\mathbf{x})$. Then the array of $f(\mathbf{x})$ forms a 2^k contingency table with a multinomial distribution. To be sure, the number of cells in the table will be extremely large for an assessment with 100 items; however, techniques associated with the analysis of contingency tables remain applicable when IRT is introduced.

In typical item-response models, a d-dimensional latent random vector θ_i is assumed to exist, and it is assumed that the X_{ij}, $1 \leq j \leq k$, are conditionally independent given θ_i. The conditional probability that $X_{ij} = 1$ given $\theta_i = \omega$ is the item characteristic curve $P_j(\omega)$. Item response models restrict the distribution of θ_i and the item characteristic curves $P_j(\omega)$ in a variety of ways. In typical cases, one has the monotonicity condition that $P_j(\omega) \geq P_j(\omega')$ if each coordinate of ω is at least as large as the corresponding coordinate of ω'. In such case, one may exploit the mathematical concept of total positivity (Karlin, 1968).

In an early example of this approach (Cressie & Holland, 1983), the one-dimensional Rasch model is considered. Here the dimension d is 1 and

$$P_j(\omega) = \exp(\omega - b_j)/[1 + \exp(\omega - b_j)]$$

for real b_j. The Rasch model implies that the log-linear model

$$\log p(\mathbf{x}) = c_m - \sum_{j=1}^{k} x_j b_j, \sum_{j=1}^{k} x_j = m,$$

holds (Tjur, 1982). On the other hand, the log-linear model does not imply the Rasch model. Indeed, the Rasch model holds if, and only if, a positive random variable T exists such that $\exp(c_m - c_0)$ is the mth moment of T for $1 \leq m \leq k$ (Cressie & Holland, 1983). Under the Rasch model, $\exp(c_m - c_0)$ is the mth moment of a random variable with density $uv(\omega)$ relative to the ability distribution, where u is a positive constant and

$$1/v(\omega) = \prod_{j=1}^{k} [1 + \exp(\omega - b_j)]$$

for ω real. The well-known result that k moments do not specify a distribution implies that the ability distribution cannot be determined from the k items even if a linear constraint is imposed on the item parameters b_j in order to determine them. In practice, the identification problem is much less significant if a parametric model is employed for the distribution of θ_i. For example, if θ_i is assumed to have a normal distribution with mean 0 and positive variance σ^2, then the item parameters b_j and the variance σ^2 can be estimated (Bock & Aitkin, 1981). A variety of cases can also be considered in which θ_i is assumed to be polytomous (Heinen, 1996).

Although initial results were obtained without explicit use of total positivity (Holland, 1981), total positivity provides a number of generalizations (Holland & Rosenbaum, 1986). A few simple illustrations of findings are instructive. Any pair of item responses X_{ij} and $X_{ij'}$, $j \neq j'$, must have a nonnegative correlation. If T_i is the sum of the $X_{ij''}$ for j'' for 1 to k, then the conditional correlation of X_{ij} and $X_{ij'}$ given $T_i - X_{ij} - X_{ij'}$ must be nonnegative. One learns that negative point-biserial correlations are fundamentally incompatible with item-response models, for X_{ij} and $T_i - X_{ij}$ must have a nonnegative correlation and X_{ij} and T_i must have a positive correlation.

Work on the Dutch identity (Holland, 1990) considered the relationship between item-response models and log-linear models with only main effects and two-factor interactions. A rather striking result is that the log-linear model holds if, for some possible value $\mathbf{x}$ of $\mathbf{X}_i$, the conditional distribution of θ_i given $\mathbf{X}_i = \mathbf{x}$ is multivariate normal with positive covariance matrix and if the item logit function $\log\{P_j(\omega)/[1 - P_j(\omega)]\}$ is a linear function of ω for each item j. This result leads to an even more striking series of conjectures based on Bayes' theorem and on Taylor's theorem. The suggestion is that, for an item-response model with a large number of items, the item characteristic curves can only be estimated without problems of parameter identification if each curve is determined by no more than two parameters. This claim suggests difficulties can be anticipated with the three-parameter logistic model. The influence of the Dutch identity in IRT has continued. For example, when the Rasch model is applied and the θ_i have normal distributions, then bounds can be obtained on the log cross-product ratios for responses X_{ij} and $X_{ij'}$ (Haberman, Holland, & Sinharay, 2008). Similar results can also be obtained with the two-parameter logistic model.

1.4 Social Networks

The use of techniques associated with the analysis of contingency tables is also quite evident in Paul's joint work with Samuel Leinhardt on analysis of social networks. From a statistical point of view, an inherent challenge in the study of social networks is that observations are usually dependent in complex ways. The techniques used often come from the analysis of contingency tables, but treatment

of dependence complicates analysis. For a basic case to explore, consider nodes (individuals) 1 to g, and let X_{ij} describe the relationship of node i to node j, say whether individual i regards individual j as a friend. The sociomatrix $\mathbf{X}$ is the g by g matrix with row i and column j equal to X_{ij}. Various descriptive terms can be used for relationships. The essential feature is that X_{ij} is equal to 1 if i relates to j and X_{ij} is 0 otherwise. Relationships need not be reciprocal, so that X_{ji} and X_{ij} need not be the same. The convention is adopted that $X_{ii} = 0$, so that nodes are not related to themselves. Analysis of data can involve both descriptive statistics and probability models. For instance, the sum X_{i+} of the X_{ij}, $1 \leq j \leq g$, measures the tendency of node i to relate to other nodes, the sum X_{+j} of the X_{ij}, $1 \leq i \leq g$, measures the tendency of other nodes to relate to node j, the sum X_{++} of the X_{ij} for $1 \leq i \leq g$ and $1 \leq j \leq q$ measures the overall level of relationship in the group, and the sum $M = \sum_{i=2}^{g} \sum_{j=1}^{i-1} X_{ij}X_{ji}$ measures the extent to which relationships are mutual (Holland & Leinhardt, 1970). Far more complex analysis may be based on results for all combinations of three nodes (triads) i, j, and k for $1 \leq i<j<k \leq g$, and analysis can consider changes in networks over time (Holland & Leinhardt, 1977). The descriptive statistics X_{++}, X_{i+}, X_{+j}, and M form the basis of the log-linear model in which, for each $\mathbf{x}$ in the set G of possible sociomatrices for g nodes, the probability $p(\mathbf{x})$ that $\mathbf{X} = \mathbf{x}$ satisfies

$$\log p(\mathbf{x}) = \kappa + \rho m + \theta x_{++} + \sum_{i=1}^{g} \alpha x_{i+} + \sum_{j=1}^{g} \beta_j x_{+j}, \tag{1.1}$$

where x_{i+} is the sum of x_{ij} over j, x_{+j} is the sum of x_{ij} over i, x_{++} is the sum of x_{ij} over i and j, and m is the sum of $x_{ij}x_{ij}$ for $1 \leq i<j \leq g$. The model parameters ρ, θ, α_i, and β_j determine the constant κ due to the constraint that the sum of the $p(\mathbf{x})$, $\mathbf{x}$ in G, must be 1. To identify model parameters, the constraints are imposed that the sum of the α_i is 0 and the sum of the β_j is also 0 (Holland & Leinhardt, 1981a). The model implies that the pairs (X_{ij}, X_{ji}) are independent for $1 \leq i<j \leq g$, and each pair (X_{ij}, X_{ji}) has common log cross-product ratio ρ. The conditional log odds

$$\log \left[P(X_{ij} = 1|X_{ji} = 0)/P(X_{ij} = 0|X_{ji} = 0)\right] = \theta + \alpha_i + \beta_j$$

then satisfies an additive model.

Numerous special cases of (1.1) appear in the literature (Holland & Leinhardt, 1979). If $\rho = \theta = \alpha_i = \beta_j = 0$, then $\mathbf{X}$ is uniformly distributed on G. Consider the following cases:

1. $\rho = \alpha_i = \beta_j = 0$, so that the X_{ij} are independent and identically distributed with θ the logit of the probability that $X_{ij} = 1$.
2. $\alpha_i = \beta_j = 0$, so that all pairs (X_{ij}, X_{ji}), $i \neq j$, are identically distributed.
3. $\rho = \beta_j = 0$, so that for each node i, the X_{ij} are independent and identically distributed for $j \neq i$.

4. $\rho = \alpha_i = 0$, so that for each node j, the X_{ij} are independent and identically distributed for $i \neq j$.
5. $\rho = 0$, so that the Rasch model holds in which node i and node j, both i and j integers between 1 and g and $i \neq j$, are in effect regarded as examinee i and item j (Haberman, 1981).

Statistical inferences can be straightforward or remarkably challenging in (1.1). Straightforward cases involve strong parameter restrictions. If Case 1, 2, 3, or 4 is assumed, then conventional use of maximum likelihood is satisfactory for g large. Case 5 is challenging, for use of maximum likelihood leads to the customary problems associated with joint estimation in the Rasch model. The case in which no parameter is restricted in (1.1) is even more difficult (Haberman, 1981; Holland & Leinhardt, 1981b). The challenges of the model specified by (1.1) can be treated by linear restrictions on the α_i and β_j or by use of random effects models as in item-response theory. Statistical analysis of social networks continues; however, Paul has not been involved for some time.

1.4.1 Log-Linear Smoothing and Kernel Equating

In work on kernel equating with Dorothy Thayer and later Alina von Davier, Paul used log-linear models to improve efficiency of estimation of probabilities prior to application of kernel smoothing (von Davier, Holland, & Thayer, 2004). The log-linear models, typically polynomial models for one-dimensional or two-dimensional contingency tables, are employed to estimate probabilities for specific scores or pairs of scores. In equating applications, these estimated probabilities are then added together to estimate distribution functions of individual variables. The kernel part of kernel equating is a traditional approach to estimation in applications far removed from psychometrics such as density estimation and estimation of the power spectrum associated with a stochastic process. The notable feature of kernel equating is the combination of statistical concepts that have little relationship to each other.

The kernel part of kernel equating is more essential in equating than is the application of log-linear models. Consider any two real random variables X and Y. Suppose that X has distribution function F, and Y has distribution function G. Let $F_{1/2}$ be the percentile rank function defined for real x to be $F_{1/2}(x) = P(X < x) + \frac{1}{2}P(X = x)$. Similarly, let $G_{1/2}$ be the percentile rank function of Y. Note that $F_{1/2}(x) = F(x)$ if F is continuous at x, a condition equivalent to the condition that $X = x$ with probability 0. A similar remark applies to $G_{1/2}$ and G.

Equipercentile methods of equating seek monotone real conversion functions $e_{Y\cdot X}$ and $e_{X\cdot Y}$ such that $e_{Y\cdot X}$ is the inverse of $e_{X\cdot Y}$, $G(e_{Y\cdot X}) = F$, and $F(e_{X\cdot Y}) = G$. The function $e_{Y\cdot X}$ is used to convert X to Y in the sense that $e_{Y\cdot X}(X)$ and Y have the same distribution. The function $e_{X\cdot Y}$ is used to convert Y to X in the sense that $e_{X\cdot Y}(Y)$ and X have the same distribution. If F and G are both strictly increasing and

continuous, then F has an inverse F^{-1}, G has an inverse G^{-1}, $e_{Y \cdot X} = G^{-1}(F)$, and $e_{X \cdot Y} = F^{-1}(G)$. If X has a normal distribution with mean μ_X and with positive variance σ_X^2, and Y has a normal distribution with mean μ_Y and positive variance σ_Y^2, then $e_{Y \cdot X}(x) = \mu_Y + (\sigma_Y/\sigma_X)(x - \mu_X)$ for real x and $e_{X \cdot Y}(y) = \mu_X + (\sigma_X/\sigma_Y) \times (y - \mu_Y)$ for real y, so that the conversion functions are linear.

If X is discrete, then F is not continuous, so that the inverse F^{-1} does not exist. A similar comment applies if Y is discrete. The functions $e_{Y \cdot X}$ and $e_{X \cdot Y}$ may still exist if X and Y are discrete. For example, if X and Y have the same distribution, then $e_{X \cdot Y}$ and $e_{Y \cdot X}$ can be chosen to be the identity function. Nonetheless, in typical cases in which X and Y are discrete, no functions $e_{X \cdot Y}$ and $e_{Y \cdot X}$ can satisfy all requirements. This problem has two consequences in equipercentile equating. The first consequence involves discrete test scores. In virtually all applications of observed-score equating, the test scores of each test are discrete variables. As a consequence, the desired conversion functions $e_{X \cdot Y}$ and $e_{Y \cdot X}$ do not generally exist. The second consequence involves use of empirical distribution functions. For positive integers m and n, consider independent and identically distributed random variables X_i, $1 \leq i \leq m$, with common distribution function F and independent and identically distributed random variables Y_i, $1 \leq i \leq n$, with common distribution function G. In equating, equivalent-groups designs have sampling with the X_i, $1 \leq i \leq m$, and the Y_i, $1 \leq i \leq n$, independent. In single-groups designs, the pairs (X_i, Y_i) are independent and identically distributed as (X, Y) and $m = n$. For either case, let χ_S be the indicator function of a set S of the real line. The empirical distribution function $\hat{F}$ is defined for real x by the equation

$$\hat{F}(x) = m^{-1} \sum_{i=1}^{m} \chi_{(-\infty, x]}(X_i),$$

so that $\hat{F}(x)$ is the fraction of the X_i that do not exceed x. Similarly,

$$\hat{G}(y) = n^{-1} \sum_{i=1}^{n} \chi_{(-\infty, y]}(Y_i).$$

For each x, $\hat{F}(x)$ converges almost surely to $F(x)$ as m approaches ∞. For each y, $\hat{G}(y)$ converges almost surely to $G(y)$ as n approaches ∞. Nonetheless, $\hat{F}$ and $\hat{G}$ are not continuous functions, so that they do not lead directly to estimates of the conversion functions $e_{Y \cdot X}$ and $e_{X \cdot Y}$.

It is possible to consider imperfect conversion functions. Kernel equating provides one source of such functions. In general, strictly increasing continuous functions $d_{X \cdot Y}$ and $d_{Y \cdot X}$ are considered such that $d_{X \cdot Y}$ is the inverse of $d_{Y \cdot X}$ and the expectation

$$K = E([G_{1/2}(Y) - F_{1/2}(d_{X \cdot Y}(Y))]^2) + E([F_{1/2}(X) - G_{1/2}(d_{Y \cdot X}(X))]^2)$$

is small. If the conversion functions $e_{X\cdot Y}$ and $e_{Y\cdot X}$ are defined, then $K = 0$ when $d_{Y\cdot X} = e_{Y\cdot X}$ and $d_{X\cdot Y} = e_{X\cdot Y}$. If Y and $c(X)$ have the same distribution for a continuous and strictly increasing function c, then $K = 0$ if $d_{Y\cdot X} = c$ and $d_{X\cdot Y} = c^{-1}$.

For a general version of kernel equating, consider a continuous symmetric random variable Z with mean 0 and variance 1. Let Z have a distribution function W and a continuously differentiable and positive density w. For example, Z might have a standard normal distribution with $W = \Phi$ and $w = \phi$. Let h_X and h_Y be positive real constants. Let X have finite variance $\sigma^2(X) > 0$, and let Y have finite variance $\sigma^2(Y) > 0$. Let Z be independent of X and Y. Let $a_X = \sigma_X/(\sigma_X^2 + h_X^2)^{1/2}$ and $a_Y = \sigma_Y/(\sigma_Y^2 + h_Y^2)^{1/2}$. Then $U_X = a_X(X + h_X Z)$ and $U_Y = a_Y(Y + h_Y Z)$ are both random variables, U_X has the same mean and variance as X, U_Y has the same mean and variance as Y, the distribution function D_X of U_X is twice continuously differentiable and strictly increasing, the distribution function D_Y of U_Y is twice continuously differentiable and strictly increasing, and one may consider the imperfect conversion functions $d_{Y\cdot X} = D_Y^{-1}(D_X)$ and $d_{X\cdot Y} = D_X^{-1}(D_Y)$. Note that D_X converges to $F_{1/2}$ as h_X approaches 0, and D_Y converges to $G_{1/2}$ as h_Y approaches 0. If X and Y are distinct, then K does not normally approach 0 as h_X and h_Y approach 0. If X has a normal distribution with mean μ_X and variance $\sigma_X^2 > 0$, if Y has a normal distribution with mean μ_Y, and variance $\sigma_Y^2 > 0$, and if Z has a standard normal distribution, then $d_{Y\cdot X}(x) = e_{Y\cdot X}(x) = \mu_Y + (\sigma_Y/\sigma_X)(x - \mu_X)$ for real x, $d_{X\cdot Y}(y) = e_{X\cdot Y}(y) = \mu_X + (\sigma_X/\sigma_Y)(y - \mu_Y)$ for real y, and $K = 0$, so that conversion functions are linear. If F and G are continuous and strictly increasing, then $d_{X\cdot Y}$ converges to $e_{X\cdot Y}$, $d_{Y\cdot X}$ converges to $e_{Y\cdot X}$, and K converges to 0 as h_X and h_Y approach 0. If X and Y are discrete, then K does not typically converge to 0 as h_X and h_Y approach 0.

The conversion functions $d_{Y\cdot X}$ and $d_{X\cdot Y}$ are readily estimated from the empirical distribution functions $\hat{F}$ and $\hat{G}$. Assume that the samples sizes m and n are both at least 2. Let s_X^2 be the standard estimated sample variance of the X_i, and let s_Y^2 be the estimated sample variance of the Y_i. Let

$$\hat{a}_X = s_X/(s_X^2 + h_X^2)^{\frac{1}{2}}$$

and

$$\hat{a}_Y = s_Y/(s_Y^2 + h_Y^2)^{\frac{1}{2}}.$$

One estimates $D_X(x)$ by $\hat{D}_X(x) = E(\hat{F}(x/\hat{a}_X - h_X Z))$ and $D_Y(y)$ by $\hat{D}_Y(y) = E(\hat{G}(y/\hat{a}_Y - h_Y Z))$. It is easily verified that

$$\hat{D}_X(x) = m^{-1}\sum_{i=1}^{m} W((x/\hat{a}_X - X_i)/h_X)$$

and

$$\hat{D}_Y(y) = n^{-1} \sum_{i=1}^{n} W((y/\hat{a}_Y - Y_i)/h_Y).$$

It is easily seen that $\hat{D}_X$ and $\hat{D}_Y$ are continuous and strictly increasing. For real x, $\hat{D}_X(x)$ converges to $D_X(x)$ with probability 1 as the sample size m increases. For real y, $\hat{D}_Y(y)$ converges to $D_Y(y)$ with probability 1 as the sample size n increases. The estimated conversion functions $\hat{d}_{Y\cdot X} = \hat{D}_Y^{-1}(\hat{D}_X)$ and $\hat{d}_{X\cdot Y} = \hat{D}_X^{-1}(\hat{D}_Y)$ have the consistency properties that, for real y, $\hat{d}_{X\cdot Y}(y)$ converges to $d_{X\cdot Y}(y)$ with probability 1 as m and n increase and, for real x, $\hat{d}_{Y\cdot X}(x)$ converges to $d_{Y\cdot X}(x)$ with probability 1 as m and n increase. In addition, normal approximations are available for both $\hat{d}_{Y\cdot X}(x)$ and $\hat{d}_{X\cdot Y}(y)$, and asymptotic confidence intervals may be derived. Appropriate formulas depend on the equating design.

The selection of the constants h_X and h_Y does depend on whether X and Y are continuous. Let X and Y both have positive continuous density functions, so that F and G are both continuous and strictly increasing. As previously noted for this case, $d_{Y\cdot X}$ converges to $e_{Y\cdot X}$ and $d_{X\cdot Y}$ converges to $e_{X\cdot Y}$ as h_X and h_Y become small. In addition, $\hat{D}_X(x)$ converges to $F(x)$ for x real as m becomes large and h_X approaches 0, and $\hat{D}_Y(y)$ converges to $G(y)$ for y real as n becomes large and h_Y approaches 0. The bias $E(\hat{D}_Y(y) - G_{1/2}(y))$ is of order h_Y, and the bias $E(\hat{D}_X(x) - F_{1/2}(x))$ is of order h_X. The normal approximations for $\hat{d}_{Y\cdot X}(x)$ and $\hat{d}_{X\cdot Y}(y)$ also require that $h_X m$ and $h_Y n$ approach ∞ as the sample sizes become large. For x and y real, the differences $\hat{d}_{Y\cdot X}(x) - e_{Y\cdot X}(x)$ and $\hat{d}_{X\cdot Y}(y) - e_{X\cdot Y}(y)$ are of order $(m^{-1} + n^{-1})^{1/2}$ if $h_X^2 m$ and $h_Y^2 n$ approach 0 and $h_X m$ and $h_Y n$ approach ∞ as the sample sizes m and n increase. Thus in the continuous case, estimation results are rather satisfactory, at least with sufficient sample size.

The discrete case customarily encountered in equating applications is far less satisfactory. If X and Y are discrete, then the conversion functions $e_{Y\cdot X}$ and $e_{X\cdot Y}$ are not available, and selection of h_X and h_Y involves a compromise between the desire for relatively smooth estimated functions $\hat{d}_{X\cdot Y}$ and $\hat{d}_{Y\cdot X}$ and a desire that the distribution function $\hat{D}_X$ be close to the percentile rank function $F_{1/2}$ and $\hat{D}_Y$ be close to $G_{1/2}$. This compromise is not influenced very strongly by considerations of sample size. The constants h_X and h_Y should not approach 0 as the sample sizes m and n become large. Very small h_X and h_Y result in very high variances of $\hat{d}_{Y\cdot X}(x)$ and $\hat{d}_{X\cdot Y}(y)$ in typical cases. Suggestions on selection of h_X and h_Y are examined in von Davier et al. (2004).

Use of log-linear models in kernel equating provides an opportunity to reduce the asymptotic variances of equating functions in the typical case in which X and Y are polytomous. A compromise between bias and variance is typically involved when distribution functions F and G are estimated by use of log-linear models for the univariate probabilities $P(X = x)$, x in the range of X, or $P(Y = y)$, y in the range of Y. In single-groups designs, one may also consider log-linear models for estimation of the joint probability $P(X = x, Y = y)$ for x in the range of X and y in

the range of Y. Variance estimates for conversion functions derived from kernel equating with log-linear models have been derived for a variety of equating designs (von Davier et al., 2004).

1.5 Applications to the Social Sciences

A consistent theme in Paul's work has been application of statistical methods to the social sciences. Many of these applications have already been apparent in the review of work related to the analysis of discrete data. Social networks, DIF, and kernel equating are normally applied to sociology and educational measurement. Further work related to the social sciences has been less connected to the analysis of discrete data. His work on causal inference, although not inherently confined to the social sciences, is especially important in fields in which randomized experiments are not readily conducted. Paul's work on equating is by no means confined to kernel equating. In addition, Paul has often collaborated with other researchers to analyze data in the social sciences.

1.5.1 Causal Inference

Paul's work on causal inference is a significant foray into an area of substantial interest both to philosophers and to scientists. It is of interest to note that Paul's doctoral thesis at Stanford, *A Variation on the Minimum Chi-Square Test*, was supervised by Patrick Suppes. Despite his advisor's distinguished work in philosophy that includes work on causality (Suppes, 1970), Paul has indicated that his interest in causal inference comes from interaction with Don Rubin (Robinson, 2005), and joint papers on causal inference appeared in the 1980s (Holland & Rubin, 1983, 1988). For a discussion by Paul on causality, approaches of philosophers, approaches of statisticians, and approaches in medicine and social science, see Holland (1986). My treatment of this area is limited for several reasons. My professional competence is much more limited here than in other areas in which Paul has worked. In cases in my life in which decisions of consequence have required reliance on scientific data, I have been struck by the extraordinary difficulty in applying the data even when studies have been properly randomized.

1.5.2 Equating

Paul's work on equating dates back to his first career at Educational Testing Service (ETS). Much of his work was collaborative. The names Rubin, Braun, Kingston, and Thayer appear in joint papers and edited volumes on equating. The edited volume

with Don Rubin is quite helpful as an indication of the state of equating in the early 1980s (Holland & Rubin, 1982). A major aspect of the work is the introduction of more formal statistical theory into the theory and practice of equating. For example, standard statistical theory is used to introduce rigorously derived approximate standard deviations, and careful consideration is given to the meaning of equipercentile equating from a population perspective. The work informs research into kernel equating but applies much more broadly. His second career at ETS has also had a strong emphasis on equating, although, with the exception of Dorothy Thayer, the collaborators have changed. In addition, to the collaboration with Alina von Davier and Dorothy Thayer on kernel equating, there has been collaboration with Neil Dorans and with Mei Liu on population invariance, with Tim Moses on log-linear smoothing, and with Sandip Sinharay on design of anchor tests and on missing-data assumptions in the NEAT (nonequivalent group anchor test) design. The edited volume with Neil Dorans and Mary Pommerich provides some indication of the status of equating and linking during Paul's second career at ETS (Dorans, Pommerich, & Holland, 2007). The work on equating has included applications to GRE® subject tests, the LSAT, and the SAT®.

1.5.3 Analysis of Data

Paul has collaborated with many investigators at ETS and elsewhere on analysis of data. As is quite often true of statisticians, data and statistical methods involved have been quite varied. For example, Paul provided early assistance in the study of scanner accuracy at ETS and in the development of methods to detect possible student collaboration on examinations. He also collaborated on analysis of items for TOEFL®, study of examinee-selected responses for AP®, development of methods to improve accuracy of transcriptions of test items, evaluation of measures to prevent driving under the influence of alcohol, measurement of food insecurity, and ranking graduate programs.

1.6 Concluding Remarks

Paul has been active as a statistician for more than four decades, so an overview of accomplishments is necessarily quite selective. His influence also extends well beyond his published work. He has served on editorial boards of major journals in statistics and psychometrics, he has made major contributions to the National Research Council's work, he has served as president of the Psychometric Society, and he has repeatedly been a member of advisory boards for testing organizations and for research organizations. Paul has also had a major influence on ETS and on the professions of statistics and psychometrics through his teaching and mentoring of students and junior colleagues. Such mentoring has a cascading effect, for the

statisticians and psychometricians whom Paul has mentored have in turn taught and mentored other statisticians and psychometricians.

Acknowledgement The study was funded by ETS Research allocation. The author thanks the reviewers of this paper, Dan Eignor and Sandip Sinharay, and the editors of this volume. Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Birch, M. W. (1964). The detection of partial association, I: The 2x2 case. *Journal of the Royal Statistical Society, Series B, 26*, 313–324.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48*, 129–141.

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

Haberman, S. J. (1978). *Analysis of qualitative data, Volume I: Introductory topics*. New York, NY: Academic.

Haberman, S. J. (1981). An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association, 76*, 60–61.

Haberman, S. J., Holland, P. W., & Sinharay, S. (2008). Limits on log odds ratios for unidimensional item response theory models. *Psychometrika, 72*, 551–561.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics (London), 20*, 309–311.

Heinen, T. (1996). *Latent class and discrete latent trait models*. Thousand Oaks, CA: Sage.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79–92.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55*, 5–18.

Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology, 70*, 492–513.

Holland, P. W., & Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology, 5*, 5–20.

Holland, P. W., & Leinhardt, S. (1979). Structural sociometry. In P. W. Holland & S. Leinhardt (Eds.), *Perspectives on social network research* (pp. 63–83). New York, NY: Academic.

Holland, P. W., & Leinhardt, S. (1981a). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association, 76*, 33–65.

Holland, P. W., & Leinhardt, S. (1981b). An exponential family of probability distributions for directed graphs: Rejoinder. *Journal of the American Statistical Association, 76*, 62–65.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.

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

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick Lord* (pp. 3–25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review, 12*, 203–231.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Karlin, S. (1968). *Total positivity*. Stanford, CA: Stanford University Press.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425–431.

Robinson, D. (2005). Profiles in research: Paul W. Holland. *Journal of Educational and Behavioral Statistics, 30*, 343–350.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam, The Netherlands: North-Holland.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics, 9*, 23–30.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Annals of Human Genetics (London), 19*, 251–253.

Part II
Holland the Young Scholar

Comments on My Social Network Research

Paul W. Holland

A year or so after I had started teaching in the Statistics Department at Harvard University, Fred Mosteller asked me if I would mind substituting for him as a discussant for a session at the American Sociological Association that was to be held in Boston, Massachusetts, a few months hence. It was a session in which Nathan Keyfitz, Leo Goodman, and Jim Davis, along with his student Sam Leinhardt, would be giving methodological papers. I agreed to do this, even though up to that point I had never had the experience of being a discussant, because the paper by Davis and Leinhardt had really caught my attention. They used graph theory and some statistical models to analyze social network data – including a data bank of over 800 sociograms (i.e., the 0/1 adjacency matrices that describe the directed graphs [i.e., digraphs] of various forms of friendship).

While I had heard about the mathematical study of graphs and digraphs in my undergrad and graduate education, it had not been systematic nor very complete. But I found the Davis and Leinhardt paper fascinating, and I figured out how to calculate a variance that went along with an expected value that they had used in their paper. It was for the number of triads within the graph that formed certain patterns under the assumption that graph was constructed at random but with the same number of reciprocated and unreciprocated friendships as in the observed digraph. Davis and Leinhardt had focused on the distribution of the number of certain specific triads (triples of nodes) in a digraph to see if there was evidence in these social network data that showed "interesting" structure that went beyond the nonrandom amounts of reciprocated friendship that pervaded the social networks in their data base. I think I surprised Davis and Leinhardt by making a small contribution in my discussion that allowed them to move their work forward. And it led, for at least 10 years thereafter, to joint work with Leinhardt as a steady and important part of my own research.

One of the major contributions of that work, in my opinion, was the p_1-distribution for stochastic digraphs. Unlike earlier random graph distributions that provided tests of structure in network data, the p_1-distribution had parameters in it for the differential attractiveness and expansiveness of the nodes, as well as for tendencies toward friendship reciprocation, which could all be estimated from

data. The p_1-distribution moved us, in a small step, from testing to estimation for directed graphs. We called it the p_1 distribution because we thought that it was the *first interesting* distribution for digraphs and that it would be followed by other distributions that had even more interesting parameters. This evolution required some serious extensions of our work on p_1, and has yet to take place, as far as I know.

Also during my early years at Harvard, I met Steve Fienberg who was a graduate student in the department. Eventually, Steve, Yvonne Bishop, and I wrote the discrete data book for which we are known. During this period Steve became especially adept at, among many other things, figuring out how to use arrays that included structural zeros to fit interesting models to data without these structural zeros. After Sam and I had introduced the p_1-distribution, Steve figured out how to fit it using his structural zero trick. Steve's paper in this volume is an outgrowth of this observation that uses much more modern mathematics than any of us were using back in those days.

Stanley Wasserman was also a graduate student at Harvard Statistics. He got there toward the end of my time, and he became interested in the social network research that Sam and I had done. He continues his interest in social networks to this day, as his paper in this volume shows. It is gratifying to see good research done as a consequence of some work of your own. Social network research has become an industry in the social sciences with contributors from many fields and from all over the world.

Chapter 2
Algebraic Statistics for p_1 Random Graph Models: Markov Bases and Their Uses

Stephen E. Fienberg, Sonja Petrović, and Alessandro Rinaldo

2.1 Introduction

The term *social network* connotes a social structure composed of individuals (or organizations), typically labeled as *nodes*, linked by one or more relations, such as friendship, information sharing, financial transaction, and so on. Social network analysis uses a graphical representation where the individuals correspond to nodes in the graph and the presence of a relationship to *edges*. The term social network now also describes specific social structures on the World Wide Web and many authors have examined the Web itself in various forms as a social network.

One can find several strains of probabilistic/statistical modeling in modern literature on networks. Some models attempt to capture empirically observed characteristics in a descriptive integrated form. Others look at large sample properties for *randomly* generated graphs possessing properties of different sorts and then attempt to discover such properties empirically in large-scale networks. One example of a property of widespread interest is tightly connected blocks of individuals, also referred to sometimes as *communities*. Indeed, the study of *blockmodels* has a long history in sociology. Yet other more recent approaches to partition networks into blocks or clusters have imported tools from machine learning, such as mixed-membership models that allow individuals to belong to more than one cluster simultaneously, thereby characterizing properties of networks and relationships among nodes (Airoldi, Blei, Fienberg, & Xing, 2008).

S.E. Fienberg (✉)
Department of Statistics, Carnegie Mellon University,
132G Baker Hall, Pittsburgh, PA 15213, USA
e-mail: fienberg@stat.cmu.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_2,

Almost all of the "statistically" oriented literature on networks derives from a handful of seminal papers. In sociology there is the early work of Jacob Moreno (1934)and the empirical studies of Stanley Milgram (1967)and Travers and Milgram (1969). In mathematics/probability there are early papers by Leo Katz and collaborators, e.g., see Katz (1951) and Katz and Powell (1957) and especially the Erdös-Rényi (1960) paper on random graph models.

Moreno (1951) invented the *sociogram* – a diagram of points and lines used to represent relations among persons, a precursor to the graph representation for social networks. Milgram (1967) gave the name to what is referred to as the *small world phenomena* – short paths of connections linking most people in social spheres. His experiments had provocative results – a median length of completed chains on the order of six, the famous *six degrees of separation*.

The network research community that arose in the 1970s was composed of mathematicians, sociologists, and statisticians. It built upon all of these earlier efforts, and we can see the direct impact of the Erdös-Rényi model on the most relevant of them, the work of Holland and Leinhardt (1970, 1971, 1976, 1978). Their work culminated in a paper (Holland & Leinhardt, 1981) in which they described their p_1 model for the analysis of networks, which models dyadic pairs of nodes independently. The p_1 model built on work in a number of their earlier papers on the topic of network modeling, and it allowed for differential attractiveness (popularity) – incoming links – and expansiveness – outgoing links – as well as an additional effect associated with mutual links due to reciprocation. Holland and Leinhardt's p_1 model was in fact log-linear in form, and this allowed for easy computation of maximum likelihood estimates using a special contingency table representation of the data (Fienberg & Wasserman, 1981a, 1981b), various generalizations to multidimensional network structures (Fienberg, Meyer, & Wasserman, 1985), and stochastic block models. These quickly evolved into the class of p^* models (now referred to as exponential random graph models – ERGMs) due to Frank and Strauss (1986) and expanded upon by Strauss and Ikeda (1990) and Wasserman and Pattison (1996). Over the past 15–20 years, ERGMs have been widely used in a descriptive form for cross sectional network structures or cumulative links for networks.

In this paper, we reconsider the Holland-Leinhardt p_1 model using the tools of algebraic geometry now embodied in the area of research referred to as algebraic statistics (see Diaconis & Sturmfels, 1998; Drton, Sturmfels, & Sullivant, 2009; Pistone, Riccomagno, & Wynn, 2001; Gibilisco et al., 2009). In particular, we derive the basic algebraic generators for the mathematical structure of p_1, known as Markov bases. We also expect to link these results to those on Markov bases for working with log-linear models for contingency tables (e.g., as described in Diaconis & Sturmfels, 1998; Dobra, Fienberg, Rinaldo, Slavkovic, & Zhou, 2008; Fienberg, 2007) because of the contingency table representation of Fienberg and Wasserman (1981a, 1981b), but this is still work in progress.

In a concluding discussion, we describe potential uses of the Markov bases and mention some possible generalizations to the class of ERGMs.

2.2 Notation and Structure of the Holland-Leinhardt p_1 Model

In this section, we describe in detail the p_1 model of Holland-Leinhardt and provide a short summary of the algebraic statistics tools and language used in our analysis.

We are concerned with describing probability distributions over a directed graph on the set of n nodes. The nodes correspond to units in a network, such as individuals, and the edges correspond to links or relationships between two units. We focus on dyadic pairings (i.e., pairs of nodes in the graph) and keep track of whether node i sends an edge to j, or vice versa, or none, or both. Let us define four probabilities: let $p_{ij}(1,0)$ be the probability of node i sending an edge toward j (1 denotes the outgoing side of the edge); $p_{ij}(0,1)$ the probability of node j sending an edge toward i; $p_{ij}(0,0)$ the probability that there is no edge between i and j; and $p_{ij}(1,1)$ the probability of i sending an edge to j *and* j sending an edge to i. For each dyad (e.g., a pair of nodes (i,j)), these four probabilities sum to 1, and we assume that the $\binom{n}{2}$ dyads are mutually independent.

The Holland-Leinhardt p_1 model of interest postulates that for each dyad (i,j), the probability of observing the four possible configurations satisfies the following equations (see Holland & Leinhardt, 1981):

$$\begin{aligned}
\log p_{ij}(0,0) &= \lambda_{ij} \\
\log p_{ij}(1,0) &= \lambda_{ij} + \alpha_i + \beta_j + \theta \\
\log p_{ij}(0,1) &= \lambda_{ij} + \alpha_j + \beta_i + \theta \\
\log p_{ij}(1,1) &= \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}.
\end{aligned}$$

where

$$\sum_i \alpha_i = \sum_j \beta_j = 0.$$

For each dyad, (i,j), the parameter α_i describes the effect of an outgoing edge from i, and β_j the effect of an incoming edge pointed towards j, while ρ_{ij} corresponds to the added effect of reciprocated edges. The parameter θ quantifies the average *density* of the network (i.e., the tendency of having edges), and λ_{ij} is simply a normalizing parameter to ensure that the probabilities for each dyad (i,j) add to 1.

Clearly p_1 reduces to the Erdös-Rényi model (Erdös & Rényi, 1960) when $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\rho\}$ are all set equal to 0. For the present purposes, we assume that the dyad is in one and only one of the possible states. As pointed out in Fienberg and Wasserman (1981a, 1981b), the p_1 model can be represented as a log-linear model over a $2 \times 2 \times n \times n$ table. In particular, the reciprocation parameter is a simple log-odds ratio for a 2×2 table associated with the dyad (i,j):

$$\rho_{ij} = \log\left[\frac{p_{ij}(0,0)p_{ij}(1,1)}{p_{ij}(1,0)p_{ij}(0,1)}\right]$$

and the other parameters of interest have related representations.

In this work, we study the following special cases of the general p_1 structure:

1. $\rho_{ij} = 0$, no reciprocal effect.
2. $\rho_{ij} = \rho$, constant reciprocation.
3. $\rho_{ij} = \rho + \rho_i + \rho_j$, edge-dependent reciprocation.

The first two of these cases were studied originally by Holland and Leinhardt (1981), and the third was introduced in Fienberg and Wasserman (1981a, 1981b).

2.2.1 Algebraic Statistics of the p_1 Models

In this section we present the essential algebraic statistics background necessary for our analysis of p_1 models.

Any log-linear model consists of probability distributions whose logarithms lie in the linear span of the rows of a matrix A. This matrix, whose entries can be typically chosen to be integers, is also called the *design matrix* of the model. In algebraic geometry, this integer matrix determines two objects: algebraically, an *ideal* of *polynomials*, and geometrically, a *variety* representing the set of solutions to the system of these polynomial equations. (For basic background on algebraic geometry and its applications, see Cox, Little, & O'Shea, 2005, 2007.)

Let us formally construct the design matrix for each of the three versions of the p_1 model described above. To each probability $p_{ij}(\bullet,\bullet)$ we can associate a monomial. Recall that a monomial is a product of powers of indeterminates. Here, indeterminates are the parameters λ_{ij}, α_i, β_i, ρ_{ij}, and θ for $i,j \in \{1,\ldots,n\}$. The correspondence is established as follows:

$$\lambda_{ij}\alpha_i^a\alpha_j^b\beta_i^b\beta_j^a\theta^{a+b}\rho_{ij}^{\min(a,b)} \mapsto p_{ij}(a,b) \tag{2.1}$$

where $a,b \in \{0,1\}$.

The design matrix A is a matrix that encodes this map in the following way: the columns of A are indexed by $p_{ij}(\bullet,\bullet)$'s and its rows by the model parameters. The entries of the design matrix are either 0 or 1; there is a 1 in the (r,c)-entry of the matrix if the parameter indexing row r appears in the monomial corresponding to the probability $p_{ij}(\bullet,\bullet)$ indexing the column c. For example, in the case $\rho_{ij} = 0$, the matrix is of size $\left(4 \times \binom{n}{2}\right) \times \left(\binom{n}{2} + 2n\right)$. For $n = 2$, the graph on two nodes consisting of a single edge, the no-reciprocation case design matrix is:

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & \lambda_{12} \\ 0 & 1 & 1 & 2 & \theta \\ 0 & 1 & 0 & 1 & \alpha_1 \\ 0 & 0 & 1 & 1 & \alpha_2 \\ 0 & 0 & 1 & 1 & \beta_1 \\ 0 & 1 & 0 & 1 & \beta_2 \end{array}\right]$$

The rows of Z_2 are indexed by parameters as indicated, while the columns are indexed by $p_{12}(0,0)$, $p_{12}(1,0)$, $p_{12}(0,1)$, $p_{12}(1,1)$. This means that, for example, $p_{12}(1,0)$, encoded by the second column of the matrix, corresponds to the monomial $\lambda_{12}\theta\alpha_1\beta_2$. Notice that this monomial simply consists of those parameters (indeterminates) that we see in the definition of the p_1 model. We have just taken the logarithms of the probabilities here and encoded the result in matrix form.

We will create the design matrices in a systematic way: The rows will always be indexed by λ_{ij}, θ, $\alpha_1,\ldots,\alpha_n$, $\beta_1,\ldots,\beta_n$, ρ_{ij}, lexicographically in that order. The columns will be ordered in the following way: First fix i and j in the natural lexicographic ordering; then, within each set, vary the edge directions in this order: $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$.

It is easy to see that the design matrix for the network on n nodes will consist of several copies of the two-node matrix, placed as a submatrix in rows and columns corresponding to all two-node subnetworks.

Let p be the $4\times\binom{n}{2}$-dimensional vector containing the probabilities $p_{ij}(\bullet,\bullet)$'s. The vector p being in the p_1 model means that the probability vector is described by the design matrix. Geometrically, this is very similar to saying that the *point* p lies in the *toric variety* parametrized by the design matrix. To be a little more precise, the probability distributions representing the p_1 model equal the *real positive part* of the corresponding toric variety. Equivalently, they are the points in the intersection of the variety and the probability simplex. (Geometers do not require points to be positive and sum to 1, while in probability theory, points are required to be positive and sum to 1 – hence the discrepancy.) The beauty of algebraic geometry is that the geometric object we just described can be described *implicitly*: The points of the variety form a set of solutions to a system of *polynomial equations*. Note that the solution set (or, the variety) is a subset of $R^{4\binom{n}{2}}$. The polynomial equations for a toric variety have very special form: They are binomials, and they can be read off from the matrix A. These equations are collected into an *ideal* of polynomials, called the *toric ideal of* A:

$$I_A \equiv (p^{b^+} - p^{b^-} : b^+ - b^- \in \ker A), \tag{2.2}$$

where $\ker A$ denotes the kernel of the design matrix: the set of all vectors $b = b^+ - b^-$ such that $Ab = 0$ (equivalently, all solutions of the homogeneous system of polynomials defined by A). Note that the notation p^u is standard multi-index notation for a monomial, where we think of p as the vector of probabilities, and the vectors b^+ and b^- are restricted to have integer coordinates. For example, for the two-node design matrix above, there is one vector in the kernel of the matrix: $[1,-1,-1,1]$, representing the following relation on the columns of the matrix: $p_{12}(1,0)p_{12}(0,1) - p_{12}(1,1)p_{12}(0,0)$. Every other vector in the kernel is a multiple of this one, so this binomial is enough to encode the whole toric ideal.

We represent the observed network as a $4\binom{n}{2}$-dimensional vector X with $0/1$ entries, so that the cardinality of the sample space is $2^{n(n-1)}$. Indeed, the observed edge configuration for a given pair of nodes (i,j) is a draw from a multinomial

distribution with size 1 and class probabilities $p_{ij}(0,0), p_{ij}(1,0), p_{ij}(0,1), p_{ij}(1,1)$. Thus, X is composed of $\binom{n}{2}$ independent multinomial draws, one for each pair of nodes. Note that the expected value of X is p. By standard theory of exponential families (see Barndorff-Nielsen, 1978; Brown, 1986), the vector $T = AX$ contains the sufficient statistics for the model parameters. (See also Geiger, Meek, & Sturmfels, 2006.)

Unlike Holland and Leinhardt (1981), who encode the observed network using the $n(n-1)$ off-diagonal elements of the incidence matrix, we choose to represent the observed network using a vector of dimension $4\binom{n}{2} = 2n(n-1)$. The significant advantage of using this different redundant representation is that the sufficient statistics are the image of a *linear* mapping specified by the design matrix A. Thus, under our parametrization, p_1 models are linear exponential families supported over a polyhedral set (that is, log-linear models), a well-understood class of statistical models that enjoy remarkable algebraic and geometric properties.

For a given vector t of observed sufficient statistics, the *fiber* of t is defined to be the set of all possible observable networks with the same sufficient statistic t. In statistical jargon, the probability distribution over the networks in the fiber at t is called the exact distribution corresponding to t. The exact distribution is used to perform model selection and goodness-of-fit testing in cases in which standard asymptotic methods, such as χ^2 approximations, are deemed unreliable. Thus, practitioners may want to resort to this form of conditional inference when the observed data are sparse or when the asymptotic validity of standard procedures has not been theoretically established. Both situations apply to network data in general and to p_1 models in particular (cf., Haberman, 1981). Unfortunately, the size and combinatorial complexity of the fiber is typically very large, even in small problems, and complete fiber enumeration, which is required for determining the exact distribution, is often unfeasible.

The theory of Markov bases provides a possible solution to the problem of finding the exact distribution corresponding to a given observed sufficient statistics. A Markov basis (see Diaconis & Sturmfels, 1998) consists of a set of *moves* that, starting from any point in the fiber, allows one to perform a random walk over any fiber in such a way that any point in the fiber has a positive probability of being visited. Markov bases thus provide a way to compute approximately the exact distribution of the model for goodness-of-fit purposes, and thus provide alternative means for model selection and goodness-of-fit testing to standard χ^2 asymptotic approximations. The fundamental theorem of Markov bases (see Diaconis & Sturmfels) states that a set B of vectors is a Markov basis for the log-linear model associated to the design matrix A, *if and only if* the corresponding set of binomials $\{p^{b^+} - p^{b^-} : b^+ - b^- \in B\}$ generates the toric ideal I_A. Thus one of our main goals is to explore these toric ideals, use algebraic geometry to derive Markov bases, and then further our analysis by considering which of the elements will satisfy the requirement that the probability distributions lie in the probability simplex. Our analysis is particularly challenging as many of the Markov bases we obtain from the fundamental theorem of Markov bases violate the p_1 constraints

that each dyad is associated to a multinomial with size 1 (i.e., the constraint that for each dyad we observe one and only one of the possible four dyadic configurations).

It is worth noting that along with Markov bases, useful for sampling the fibers, researchers are often interested in other binomial bases for the toric ideal, in particular the *Gröbner bases*. A Gröbner basis is a special set of generators that is equivalent to a row-echelon form for a linear system of equations. The set contains a Markov basis, and generally it is strictly larger, but it has some special properties desirable for certain computations (for example, one can do polynomial multi-indeterminate long division only if one has a Gröbner basis of a system of polynomials). For background see Cox et al. (2005, 2007), and for a standard reference on Gröbner bases for toric ideals, see Sturmfels (1996).

The primary goal of this paper is to understand the structure of these Markov bases for the three cases of the p_1 model. Our secondary goal is to provide a geometric characterization of the conditions for the existence of the maximum likelihood estimates (MLEs) of the model parameters; see Haberman (1977) for more details and general results on existence of the MLEs in models exponential response models.

For both our goals, we rely on the following software to perform algebraic and computational geometry calculations:

- 4ti2 (4ti2 Team, 2008) – a software package for algebraic, geometric, and combinatorial problems on linear spaces – generates basis elements (perhaps redundant) for Markov bases for specific values of n. In turn, we can use these to compute exact distribution given the minimal sufficient statistics using Monte Carlo Markov chain methods.
- Polymake (Gawrilow & Joswig, 2008) – a software package for analyzing convex polytopes. In Sect. 2.4 we use Polymake to explore conditions for the existence of (nonzero) MLEs when $\rho_{ij} = 0$. Nonexistence of MLEs effects both the computation and the assessment of fit.

The strategy we follow is one that has proved successful in other categorical data problems. We investigate the form of our problem (determining the Markov basis or investigating the fiber) in low dimensions (e.g., $n = 3, 4, 5$). Then we posit features of the general structure that will hold for arbitrary n.

2.3 Markov Bases Associated with p_1 for Small Networks

In this section, we will describe the toric varieties corresponding to the variation on p_1 models. Unfortunately, the algebraic geometry machinery generates *universal* basis elements that can take all possible values. Because we are dealing with probabilities that are non-negative and add to 1, some basis elements ignore the fact that we get to observe the dyad in one and only one of the four possible states. Thus once we find the Markov bases, we still need to be careful in identifying those elements that are useful for our statistical enterprise. The good news is that we are

able to decompose every Markov basis element using certain basic moves. We can carry out this construction sequentially in a way that creates only statistically meaningful moves. The key idea is to decompose the toric ideal of the p_1 model using ideals which are known and easier to understand by ignoring the dyadic constraints represented by the normalizing parameters λ_{ij}. This approach reveals a connection between p_1 models and toric varieties, which are associated to certain graphs that have been studied by the commutative algebra and algebraic geometry community, specifically Villarreal (2001) and Ohsugi and Hibi (1999).

In terms of ideal generators for the p_1 model, reintroducing the normalizing constants adds another level of algebraic difficulty, and we defer the details of how to accomplish this to a later more mathematically technical paper' with 'we refer the reader to Petrović, Rinaldo, & Fienberg (2010) for details. In terms of moves on the network, however, we can avoid this difficulty by exhibiting the decomposition of the moves (although inapplicable in terms of ideal generators) using well-understood binomials arising from graphs. This approach reduces the complexity and size of Markov moves, and in addition, allows us to bypass the study of those basis elements that are not applicable due to the constraints described above. Finally, we point out that even though the size of the Markov bases grows rapidly as we increase the number of nodes, there is quite a lot of structure in these generating sets. In what follows, we will first illustrate this structure on some small networks. Because there are three cases of the p_1 model, we need three different labels for the design matrices of the n-node network. The design matrix depends on the choice of n and ρ_{ij}:

1. For the case $\rho_{ij} = 0$, when the reciprocal effect is *zero*, we denote the design matrix for the n-node network by Z_n.
2. For the case of *constant* reciprocation (i.e., $\rho_{ij} = \rho$), we denote the n-node network matrix by C_n.
3. When reciprocation is *edge-dependent* (i.e., $\rho_{ij} = \rho + \rho_i + \rho_j$), we denote the design matrix by E_n.

2.3.1 Case I: No Reciprocation ($\rho_{ij} = 0$)

While this is clearly a special case of $\rho_{ij} = \rho$, we treat it separately as it is algebraically interesting in its own right.

We start with the simplest nontrivial example: $n = 2$. The design matrix Z_2, which encodes the parametrization φ_2 of the variety was given in Sect. 2.2.1. The toric ideal I_{Z_2} is the principal ideal generated by one quadric:

$$I_{Z_2} = (p_{12}(1,0)p_{12}(0,1) - p_{12}(1,1)p_{12}(0,0))$$

and thus this single binomial is a Markov basis and also a Gröbner basis with respect to any term order. We can verify this by hand, or by using software (4ti2 Team, 2008).

In general, we can translate binomials into moves in the following way: We will remove all edges that are represented by the p_{ij}'s in the negative monomial and add all edges represented by the p_{ij}'s in the positive monomial. Note that if $p_{ij}(0,0)$ occurs in either, it has no effect: It says to remove or add the "no-edge," so we do nothing. The reason why the terms $p_{ij}(0,0)$ are kept is to ensure that the binomial is homogeneous with respect to the pair $\{i,j\}$. Here, for example, since the positive monomial is of degree two, the negative monomial has $p_{12}(0,0)$ attached to it to ensure it also is of degree two.

Thus, the generator of I_{Z_2} represents the following Markov move: Delete the bidirected edge between 1 and 2. Replace it by an edge from 1 to 2 and an edge from 2 to 1. If we need to allow only one edge per dyad, however, this binomial is meaningless and there are not really any allowable Markov moves. Philosophically, the case of no reciprocation somehow contradicts this assumption, since if $\rho_{ij} = 0$, a bidirected edge between two nodes is always valued the same as two edges between them. For this reason, the requirement of only one edge per dyad makes this problem so much more complicated – relations such as this one for any dyad in an n-node network will appear in the generating sets of the ideal I_{Z_n}, but we will *never* want to use them.

Next, let $n = 3$. The design matrix Z_3 encodes the parametrization φ_3 of the variety as follows

$$Z_3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{matrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \\ \theta \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{matrix}$$

where the columns of Z_3 are indexed by $p_{12}(0,0)$, $p_{12}(1,0)$, $p_{12}(0,1)$, $p_{12}(1,1)$, $p_{13}(0,0)$, $p_{13}(1,0)$, $p_{13}(0,1)$, $p_{13}(1,1)$, $p_{23}(0,0)$, $p_{23}(1,0)$, $p_{23}(0,1)$, $p_{23}(1,1)$.

The toric ideal I_{Z_3} is minimally generated by the following set of binomials:

$$\begin{gathered} p_{23}(0,1)p_{23}(1,0) - p_{23}(1,1)p_{23}(0,0), \\ p_{13}(0,1)p_{13}(1,0) - p_{13}(1,1)p_{13}(0,0), \\ p_{12}(0,1)p_{12}(1,0) - p_{12}(1,1)p_{12}(0,0), \\ p_{12}(0,1)p_{13}(1,0)p_{23}(0,1) - p_{12}(1,0)p_{13}(0,1)p_{23}(1,0). \end{gathered}$$

The first three generators are precisely the binomials from I_{Z_2} for the three dyads $\{1,2\}$, $\{1,3\}$, $\{2,3\}$. The only statistically meaningful generator is the cubic. It represents the following move: Replace the edge from 1 to 2 by the edge from 2 to 1; replace the edge from 2 to 3 by the edge from 3 to 2; replace the edge from 3 to 1 by the edge from 1 to 3. Graphically, it represents the *three-cycle* oriented two different ways: the positive monomial represents the cycle $1 \to 3 \to 2 \to 1$, while the negative monomial represents the cycle $1 \to 2 \to 3 \to 1$. It also corresponds to the contrast associated with test involving the fit of either the quasi-independence or the quasi-symmetry model in a 3×3 table with structural zeros down the diagonal:

0	$p_{12}(1,0)$	$p_{13}(1,0)$
$p_{12}(0,1)$	0	$p_{23}(1,0)$
$p_{13}(0,1)$	$p_{23}(0,1)$	0

Quasi-independence posits a model for the nonzero cell probabilities that involves a product of a parameter for the row and one for the column, whereas quasi-symmetry allows for symmetry in the interactions but different marginal totals. The one degree of freedom contrast needed to examine both of these models is the same when $n = 3$, for example,

$$\log p_{12}(0,1) + \log p_{13}(1,0) + \log p_{23}(0,1) - \log p_{12}(1,0) - \log p_{13}(0,1) - \log p_{23}(1,0)$$

(cf., the related discussion of these models in Bishop, Fienberg, & Holland, 1975).

Suppose now that $n = 4$. A minimal generating set for the ideal I_{Z_4} consists of 151 binomials:

- 6 quadrics
- 4 cubics
- 93 quartics
- 48 quintics

Some of these violate the requirement that each dyad can be observed in only one state. As it is impractical to write all of these binomials down, we will list just a few of those that are statistically meaningful (i.e., respect the requirement of at most one edge per dyad at any time). As expected, the quadrics and the cubics are simply the generators of I_{Z_3} for the 4 three-node subnetworks of the four-node network. The quadrics are not of interest. The cubics represent the three-cycles. Here is a list of sample quartics, written in binomial form as it is most appropriate at the moment:

$$\begin{aligned}
&p_{12}(1,1)p_{34}(1,1)p_{23}(0,0)p_{14}(0,0) - p_{12}(0,0)p_{34}(0,0)p_{23}(1,1)p_{14}(1,1),\\
&p_{23}(1,1)p_{14}(1,1)p_{13}(0,0)p_{24}(0,0) - p_{23}(1,0)p_{14}(1,0)p_{13}(0,1)p_{24}(0,1),\\
&p_{23}(1,1)p_{14}(1,1)p_{12}(0,0)p_{34}(0,0) - p_{12}(1,0)p_{23}(1,0)p_{34}(1,0)p_{14}(0,1),\\
&p_{12}(0,0)p_{23}(1,1)p_{34}(0,1)p_{14}(1,0) - p_{12}(1,0)p_{23}(1,0)p_{34}(1,1)p_{14}(0,0).
\end{aligned}$$

Finally, we consider a pair of representative quintics:

$$p_{12}(0,0)p_{23}(1,1)p_{34}(0,1)p_{14}(0,1)p_{24}(1,0) - p_{12}(0,1)p_{23}(1,0)p_{34}(1,1)p_{14}(0,0)p_{24}(0,1),$$
$$p_{12}(1,0)p_{23}(1,0)p_{14}(0,0)p_{13}(1,1)p_{24}(1,0) - p_{12}(0,1)p_{23}(1,1)p_{14}(1,0)p_{13}(1,0)p_{24}(0,0).$$

This set of Markov moves is much more complex then the 10 Markov moves resulting from the simpler parametrization of the p_1 model on four nodes described by Holland and Leinhardt (1981). We postpone further analysis of these binomials to another paper (see Petrović, Rinaldo, & Fienberg, 2010), where we develop a general characterization of Markov moves that go beyond the simple case of $\rho = 0$. For now, we simply note that all of them preserve the in- and out-degree distributions of the nodes in the network. After we study the other two cases for ρ_{ij}, we will see a recurring underlying set of moves that can be used to understand these ideals.

2.3.2 Case II: Constant Reciprocation ($\rho_{ij} = \rho$)

Now we introduce one more row to the zero-ρ design matrix Z_n to obtain the constant-ρ matrix C_n. Namely, this row represents the constant ρ added to those columns indexed $p_{ij}(1,1)$ for all $i,j \in [n]$. It is filled with the pattern $0,0,0,1$ repeated as many times as necessary. For example, the design matrix for the two-node network is as follows:

$$C_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \lambda_{12} \\ \theta \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \rho \end{matrix}$$

In this case, the ideal is empty (there is nothing in the kernel of C_2). We should have expected this result since the case of $\rho_{ij} = 0$ requires no reciprocation effect. Here, the bidirected edge is valued differently than the two single edges in a dyad; this is the meaning of the last row of the design matrix.

For the three-node network, we have the design matrix,

$$C_3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \\ \theta \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \rho \end{matrix}$$

and the Markov basis consists only of the cubic move from the case $\rho_{ij} = 0$, for example,

$$p_{12}(0,1)p_{13}(1,0)p_{23}(0,1) - p_{12}(1,0)p_{13}(0,1)p_{23}(1,0),$$

used to extract *all possible* Gröbner bases of the ideal. Graver bases are not known for many families of toric ideals. A precise definition of the Graver basis can be found, for example, in Sturmfels (1996); the motivation for it comes from integer programming.

Let $n = 4$. The software 4ti2 (n.d.) outputs a minimal generating set of the ideal I_{C_4} consisting of:

- 4 cubics
- 57 binomials of degree 4
- 72 of degree 5
- 336 of degree 6
- 48 of degree 7
- 18 of degree 8

Out of this large set, the applicable Markov moves are the same as in the case $\rho_{ij} = 0$ with a few degree-six binomials added, such as:

$$\begin{aligned} &p_{12}(0,0)p_{13}(1,1)p_{14}(1,1)p_{23}(0,1)p_{24}(1,0)p_{34}(0,0) \\ &\quad - p_{12}(1,1)p_{13}(0,1)p_{14}(1,0)p_{23}(0,0)p_{24}(0,0)p_{34}(1,1) \end{aligned}$$

2.3.3 Case III: Edge-Dependent Reciprocation ($\rho_{ij} = \rho + \rho_i + \rho_j$)

To construct the design matrix E_n for this case, we start with the matrix C_n from the case $\rho_{ij} = \rho$, and introduce n more rows indexed by $\rho_1, \ldots, \rho_n$. Every fourth column of the new matrix, indexed by $p_{ij}(1,1)$, has two nonzero entries: a 1 in the rows corresponding to ρ_i and ρ_j. For example, when $n = 2$, the matrix looks like this:

$$E_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \lambda_{12} \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \theta \\ \rho \\ \rho_1 \\ \rho_2 \end{matrix}$$

Because this is a full-rank matrix, the ideal for the two-node network is empty.

With $n = 3$ we get the expected result; the ideal I_{E_3} is the principal ideal

$$I_{E_3} = (p_{12}(1,0)p_{23}(1,0)p_{13}(0,1) - p_{12}(0,1)p_{23}(0,1)p_{13}(1,0)).$$

With $n = 4$ we get the first interesting Markov moves for the edge-dependent case. The software 4ti2 (4ti2 Team, 2008) outputs a minimal generating set of the ideal I_{E_4} consisting of:

- 4 cubics
- 18 binomials of degree 4
- 24 of degree 5

The cubics, as usual, represent reorienting a three-cycle. Similarly some of the quartics represent four-cycles. And then we get a few more binomials, of the following types:

$$p_{13}(0,0)p_{24}(0,0)p_{14}(0,1)p_{23}(0,1) - p_{13}(0,1)p_{24}(0,1)p_{14}(0,0)p_{23}(0,0)$$

of degree four, and

$$\begin{aligned} &p_{13}(0,0)p_{24}(0,0)p_{14}(0,1)p_{12}(1,0)p_{23}(1,0) \\ &\quad - p_{13}(1,0)p_{24}(0,1)p_{14}(0,0)p_{12}(0,1)p_{23}(0,0) \end{aligned}$$

of degree five. Note that these two are just representatives; we may, for example, replace every $p_{ij}(0,0)$ in each of them by $p_{ij}(1,1)$ and get other Markov moves that are minimal generators of the toric ideal I_{E_4}.

2.4 The Marginal Polytope and the Maximum Likelihood Estimate

In this final section, we turn our attention to the problem of nonexistence of the MLE of the model parameters for p_1 models. See also Haberman (1977) and Fischer (1981) for existing results in the literature. In particular, we briefly derive a geometric characterization of the necessary and sufficient conditions for existence of the MLE. We refer the reader to Petrović et al. (2010) for more details and more general statements.

Let X be the sample space (i.e., the set of all observable networks), and for a given design matrix A, consider the set

$$S = convhull(\{t = Ax, x \in X\}),$$

consisting of the convex combinations of all the possible observable sufficient statistics for the p_1 model specified by A. Since X is finite, the set S is a polytope

(a bounded and closed convex set cut out by a finite number of hyperplanes). Borrowing the terminology from Eriksson, Fienberg, Rinaldo, and Sullivant (2006), we call the polytope S the *marginal polytope* of the model. From the theory of exponential families (see Barndorff-Nielsen, 1978; Brown, 1986), the marginal polytope is the convex support of the family of distributions representing the p_1 model whose design matrix is A. Furthermore, the MLE $\hat{p}$ of the vector of probabilities p exists if and only if the vector t of observed sufficient statistics is in the interior of S. Here, by existence of the MLE we mean that the vector $\hat{p}$ has strictly positive coordinates or, equivalently, that the model parameters are strictly positive.

Thus, in order to decide if the MLE corresponding to an observed sufficient statistics t exists, it is necessary to decide whether t belongs to the interior of S. When the MLE does not exist, it then becomes important to identify which coordinates of $\hat{p}$ are estimated to be zero, or equivalently, which of the model parameters are estimated to be zero. Both tasks require dealing directly on the geometric and combinatorial properties of the marginal polytope and of its boundary. Unfortunately, due to the product-multinomial constraints, the set S is fairly difficult to describe, even with the full knowledge of A. Both problems, however, can be solved by looking at the simpler set

$$C = cone(\{t = Ax, x \geq 0\})$$

which is the polyhedral cone generated by the sufficient statistics t. Unlike S, C does not encode the multinomial constraints and, therefore, is much easier to handle. In particular, there already exists an algorithm to check whether t belongs to the interior of C (which can be decided by solving a feasibility problem via linear programming) and, more importantly, to decide which coordinates of $\hat{p}$ are zero (see Eriksson et al., 2006; Rinaldo, 2006; Rinaldo, Fienberg, & Zhou, 2009). We refer the reader to Petrović et al. (2010) for rigorous statements of these results. We remark that these geometric results reflect the well known fact that, under appropriate conditions satisfied by p_1 models, in more general log-linear model settings, the MLE under product-multinomial sampling scheme (captured by the set S) exists if and only if it exists under the Poisson sampling scheme (captured by the set C). See, for instance, Haberman (1974).

We conclude this section with a more detailed analysis of the p_1 model with $\rho = 0$. Unlike the other p_1 models we consider, for this special case, explicit and simple conditions for the existence of the MLE can be directly derived. To this end, for notational convenience, we no longer represent the network X as a vector. Instead, we adopt the original notation of Holland and Leinhardt (1981) and define X is a $n \times n$ matrix whose (i,j) entry is 1 if there is an edge leaving i into j and 0 otherwise. Note that since self-loops are not allowed, $X_{ii} = 0$ for all i. Using the matrix notation, for the p_1 model with $\rho = 0$, the $2n$-dimensional sufficient statistics are the row and column sums of X. Indeed, when $\rho = 0$, the p_1 model in the parametrization of Holland and Leinhardt is a linear exponential family.

There are three cases where the MLE does not exist. The first two cases are clear.

- If a row or column sum is equal to $n-1$.
- If a row or column sum is equal to 0.

The third case is subtler, as it corresponds to situations in which the minimal sufficient statistics can be positive and less than $n-1$. From the theory of exponential families, the MLE $\hat{p}$ satisfies the moment equations, namely the row and column sums of $\hat{p}$ match the corresponding row and column sums of the observed network. Thus, the MLE does not exist whenever this constraint cannot be satisfied by any strictly positive vector. As a result, for $n=3$, besides the two obvious cases indicated above, the MLE does not exist if the following three patterns of zeros are observed:

$$\begin{bmatrix} \times & 0 & \\ 0 & \times & \\ & & \times \end{bmatrix}, \quad \begin{bmatrix} \times & & 0 \\ & \times & \\ 0 & & \times \end{bmatrix}, \quad \begin{bmatrix} \times & & \\ & \times & 0 \\ 0 & & \times \end{bmatrix}.$$

When $n=4$, there are four patterns of zeros leading to a nonexistent MLE, even though the margins can be positive and smaller than three:

$$\begin{bmatrix} \times & 0 & & 0 \\ 0 & \times & & 0 \\ & & \times & \\ 0 & 0 & & \times \end{bmatrix}, \quad \begin{bmatrix} \times & 0 & 0 & \\ 0 & \times & 0 & \\ 0 & 0 & \times & \\ & & & \times \end{bmatrix}, \quad \begin{bmatrix} \times & & 0 & 0 \\ & \times & & \\ 0 & & \times & 0 \\ 0 & & 0 & \times \end{bmatrix}, \quad \begin{bmatrix} \times & & & \\ & \times & 0 & 0 \\ & 0 & \times & 0 \\ & 0 & 0 & \times \end{bmatrix}.$$

We found these patterns using polymake. Based on our computations, we conjecture that for a network on n nodes, the number of patterns of zeros that lead to nonexistence of the MLE and that fall in this third category is always $2n$.

2.5 Discussion

In this paper we begin a reconsideration of the Holland-Leinhardt p_1 model using the tools of algebraic statistics. In particular, we attempt to derive Markov bases for p_1. We have yet to link these to those on Markov bases for log-linear models for contingency tables, (e.g., as described by Diaconis & Sturmfels, 1998; Dobra et al., 2008; Fienberg, 2007). But because of the contingency table representation of Fienberg and Wasserman (1981a, 1981b), we expect some form of congruence.

One of the interesting aspects of the 1981 Holland-Leinhardt paper was its focus on assessing the fit of the model to actual network data, although they made clear that there was little or no theory to rely upon. As a reviewer of an earlier draft of this paper noted: "The asymptotic challenge [for assessing the fit of p_1] is considerable, and a comparison with related item-response theory suggests that some aspects of the problem are deeply impossible. The asymptotic challenge derives from the very

high dimension relative to the number of observations and from the considerable variation in the accuracy of different model parameters."

Haberman (1981) made precisely this argument in his discussion of the 1981 Holland and Leinhardt paper. What is apparent from a reading of the current network literature for exponential random graphs is that, more than 35 years later, this problem has not been solved, either asymptotically or for small samples (e.g., Hunter, Goodreau, & Handcock, 2008), although the use of *random effects* or hierarchical Bayesian versions of p_1 is one way that has been used to reign in the difficulty of high dimensionality. Because of the inherent sparseness of the p_1 model, we expect that the Markov bases and related algebraic geometry notions discussed in this paper will ultimately be useful for exploring two statistical problems: (a) determining condition for the existence of MLEs, and (b) using them to traverse conditional (given minimal sufficient statistics) sample spaces, generating exact distributions useful for assessing goodness of fit.

The p_1 model has been generalized in a variety of ways and is usually now discussed in the context of exponential random graph models (ERGMs, also known as p^* models see Frank & Strauss, 1986; Strauss & Ikeda, 1990; Wasserman & Pattison, 1996; see Rinaldo et al., 2009, for recent results on the geometric properties of ERGMs). Laying out a full algebraic statistics framework for ERGMs, such as that introduced in this paper for p_1, appears to be quite difficult. We believe this difficulty is a consequence of the fact that the likelihood function does not decompose into independent components in the way that the p_1 likelihood decomposes into independent dyadic components.

We believe that the algebraic statistics framework for the Holland-Leinhardt p_1 model, which we have introduced in this paper, is not only mathematically elegant, but that it also offers a statistically interesting complement to their pioneering work on network modeling.

Acknowledgements To Paul Holland, whose work on contingency tables and network models continues to provide us with research ideas. This paper continues the exploration of their connections.

This research was supported in part by NSF grant DMS-0631589 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program. We have received valuable comment from a number of colleagues when we presented preliminary versions of this paper at a series of workshops. We are especially grateful to two reviewers for helpful comments on an earlier draft of the paper.

References

4ti2 Team. (2008). 4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces [Computer software]. Available from http://www.4ti2.de.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9*, 1823–1856.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. New York, NY: Wiley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. New York, NY: Springer-Verlag (Original work published 1975).

Brown, L. D. (1986). *Fundamentals of statistical exponential families*. Hayward, CA: Institute of Mathematical Statistics.

Cox, D., Little, J., & O'Shea, D. (2005). *Using algebraic geometry*. Berlin, Germany: Springer-Verlag.

Cox, D., Little, J., & O'Shea, D. (2007). *Ideals, varieties, and algorithms*. Berlin, Germany: Springer-Verlag.

Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distribution. *Annals of Statistics, 26*, 363–397.

Dobra, A., Fienberg, S. E., Rinaldo, A., Slavkovic, A. B., & Zhou, Y. (2008). Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In M. Putinar & S. Sullivant (Eds.), *Emerging applications of algebraic geometry, pp. 63–88*. New York, NY: Springer-Verlag.

Drton, M., Sturmfels, B., & Sullivant, S. (2009). *Lectures on algebraic statistics*. Basel, Switzerland: Springer Basel AG.

Erdös, P., & Rényi, A. (1960). The evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl., 5*, 17–61.

Eriksson, N., Fienberg, S. E., Rinaldo, A., & Sullivant, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *Journal of Symbolic Computation, 41*, 222–233.

Fienberg, S. E. (2007). Editorial, expanding the statistical toolkit with algebraic statistics. *Statistica Sinica, 17*, 1261–1272.

Fienberg, S. E., Meyer, M. M., & Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association, 80*, 51–67.

Fienberg, S. E., & Wasserman, S. S. (1981a). Categorical data analysis of single sociometric relations. *Sociological Methodology, 1981*, 156–192.

Fienberg, S. E., & Wasserman, S. S. (1981b). Discussion of Holland, P. W. and Leinhardt, S., An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association, 76*, 54–57.

Fischer, G. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59–77.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81*, 832–842.

Gawrilow, E., & Joswig, M. (2008). Polymake – A software package for analyzing convex polytopes [Computer software]. Available from http://www.math.tu-berlin.de/polymake/.

Geiger, D., Meek, C., & Sturmfels, B. (2006). On the toric algebra of graphical models. *Annals of Statistics, 34*(3), 1462–1492.

Gibilisco, P., Riccomagno, E., Rogantin, M. P., & Wynn, H. P. (Eds.). (2009). *Algebraic and geometric methods in statistics*. Cambridge, England: Cambridge University Press.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago, IL: University of Chicago Press.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5*(5), 815–841.

Haberman, S. J. (1981). An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association, 76*, 60–61.

Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology, 70*, 492–513.

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Small Group Research, 2*, 107–124.

Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology, 7*, 1–45.

Holland, P. W., & Leinhardt, S. (1978). An omnibus test for social structure using triads. *Sociological Methods Research, 7*, 227–256.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association, 76*, 33–65.

Hunter, D. R., Goodreau, S., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association, 103*, 248–258.

Katz, L. (1951). The distribution of the number of isolates in a social group. *Annals of Mathematical Statistics, 23*, 271–276.

Katz, L., & Powell, J. H. (1957). Probability distributions of random variables associated with a structure of the sample space of sociometric investigations. *Annals of Mathematical Statistics, 28*, 442–448.

Milgram, S. (1967). The small world problem. *Psychology Today, 1*(1), 60–67.

Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease Publishing Company.

Moreno, J. L. (1951). *Sociometry, Experimental method and the science of society. An approach to a new political orientation*. New York, NY: Beacon House.

Ohsugi, H., & Hibi, T. (1999). Toric ideals generated by quadratic binomials. *Journal of Algebra, 218*(2), 509–527.

Petrović, S., Rinaldo, A., & Fienberg, S. E. (2010). Algebraic statistics for a directed random graph model with reciprocation. In M. Viana & H. Wynn (Eds.), *Algebraic methods in statistics and probability, Volume II*. Contemporary Mathematics, Providence, RI: American Mathematical Society, *516*, 261–283.

Petrović, S., Rinaldo, A., & Fienberg, S. E. (2011). *Maximum likelihood estimation in network models*. Manuscript submitted for publication.

Pistone, G., Riccomagno, E., & Wynn, H. P. (2001). *Algebraic statistics: Computational commutative algebra in statistics*. New York, NY: Chapman & Hall.

Rinaldo, A. (2006). *Computing maximum likelihood estimates in log-linear models* (Technical Rep. No. 835). Pittsburgh, PA: Carnegie Mellon University, Department of Statistics.

Rinaldo, A., Fienberg, S. E., & Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics, 3*, 446–484.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association, 85*, 204–212.

Sturmfels, B. (1996). *Gröbner bases and convex polytopes* (*University Lecture Series* (Vol. 8)). Providence, RI: American Mathematical Society.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32*, 425–443.

Villarreal, R. (2001). *Monographs and textbooks in pure and applied mathematics: Vol. 8. Monomial algebras*. New York, NY: Marcel Dekker.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika, 61*, 401–425.

Chapter 3
Mr. Holland's Networks: A Brief Review of the Importance of Statistical Studies of Local Subgraphs or One Small Tune in a Large Opus

Stanley Wasserman

3.1 Introduction

I enrolled in the Graduate School of Arts and Sciences at Harvard in fall 1973 to do graduate work in statistics. I had six classmates in my cohort, four of whom eventually received PhDs. One, Richard Hill, a recent graduate of MIT, was also an administrator at the Computer Research Center (CRC) of the National Bureau Economic Research (NBER). He worked there with his former classmate Mark Eisner, technical director of the CRC.

Paul Holland was full time at CRC, as a senior research associate, from 1972 to 1975, while maintaining a lectureship in the Department of Statistics up the river. After my first year of studies, Richard and Paul were looking for a research assistant to begin full-time for summer 1974 and to continue part-time during the academic year. I was interested, and very pleased to be hired. So, my work with Paul began that summer and continued for the next 2 years while Paul and I were both in Cambridge. Paul moved to Educational Testing Service (ETS) late in 1975, so I moved on as well, spending my last year in graduate school (1976–1977) at Carnegie-Mellon University, working with his close collaborator at the time, Sam Leinhardt.

Paul directed my thesis and was my mentor throughout the 1970s. I owe much to him: I valued his enthusiasm, enjoyed his humor, and was very grateful that much of what I did back in those early years was regarded by him as a "thing of beauty" (a standard which I never met again). I do regret that we were not more in contact over the last two decades.

I am grateful to be able to write a short note for his *Festschrift*, commenting on the importance of his research to the burgeoning field of network science.

S. Wasserman (✉)
Department of Statistics, Indiana University, 309 North Park Street,
Bloomington, IN 47408, USA
e-mail: stanwass@indiana.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_3,

3.2 Notation

We begin with a graph (or a directed graph), a single set of nodes N, and a set of lines or arcs L. It is common to use this mathematical concept to represent a *network*. We use the notation of Wasserman and Faust (1994), especially Chaps. 13 and 14. There are extensions of these ideas to a wide range of networks, including multiple relations, affiliation relations, valued relations, and social influence and selection situations (in which information on attributes of the nodes is available); see the chapters of Carrington, Scott, and Wasserman (2005).

The purpose of this short exposition is to discuss the developments in statistical models for networks that have occurred over the past 10 years and relate them to Paul's early statistical network research. Background for much of this paper is summarized in the statistical chapters (Chaps. 8–11) of Carrington, et al. (2005) (which were written in 2002). More of it can be found in the statistical physics literature, for example, the review paper of Newman (2003) or the edited volume of Newman, Barabasi, and Watts (2006). The statistical modeling of social networks is advancing quite quickly. The many exciting new developments include, for instance, longitudinal models for the coevolution of networks and behavior (Snijders, Steglich, & Schweinberger, 2007) and latent space models for social networks (Handcock, Raftery, & Tantrum, 2007; Hoff, Raftery, & Handcock, 2002). Here, we review a few developments that are relevant to Paul's work in the early 1970s.

3.3 The Importance of Mr. Holland to Network Science

3.3.1 Some Past History

One of the most important structural theories in network analysis is *structural balance*, and its many derivatives. The history of structural balance, clusterability, and ranked clusterability began in network science in the 1940s when a variety of mathematicians invaded the structural space occupied by the early sociometricians. The forefront of this research yielded a variety of theorems, rooted in graph theory, that allowed for checks on whether a particular graph was structurally balanced or clusterable. With these clusterability theorems in hand, a number of researchers embarked on empirical investigations. Questions such as how common clusterable signed (di)graphs are, and whether such signed (di)graphs were balanced, needed answers. These investigations required surveying many sociomatrices obtained from diverse sources. Further, the empirical studies had to be accompanied by statistical models that allowed those interested to study whether departures from theoretical models such as clusterability were *statistically large*.

The necessary statistical techniques are a bit too long and tedious for the scope of the current chapter. A few details appear below with a reference to Chap. 14 of

Wasserman and Faust (1994) for lots more information. But we can report here how the theorems of clusterability were generalized due to unexpected empirical evidence.

The standard Holland-Leinhardt index for clusterability or transitivity starts with the triad census, a vector of isomorphic triad counts, either of length 4 (for graphs) or 16 (for directed graphs). As usual, let $\mathbf{T}$ denote the triad census vector. Mathematically, let $\boldsymbol{l}$ be a weighting vector, designed to count the frequency of a particular structural tendency. Then, $\boldsymbol{l}'\,\mathbf{T}$ is a linear combination of the triad census, using one of the weighting vectors derived from the substantive hypothesis under study.

This linear combination is the number of times that the specific configuration, associated with the chosen weighting vector, occurs in the observed sociomatrix. Under one of the random directed graph distributions, we can calculate the expected value and covariance matrix of $\mathbf{T}$, and hence the expected number for this configuration and its variance. This expected number is $\boldsymbol{l}'\,\boldsymbol{\mu}_T$, and the standard error is $\sqrt{\boldsymbol{l}'\,\Sigma_T\boldsymbol{l}}$, where $\boldsymbol{\mu}_T$ is the mean triad census vector, and Σ_T is the 16×16 (or 4×4) covariance matrix of the counts of the triad census.

This standardized index is then used as a test statistic for a variety of substantive null hypotheses. The first two moments of the linear combination of raw counts are calculated under these null hypotheses, which invariably assume some particular random digraph distribution. From substantive hypothesis, to weighting vector, to test statistic, to a statistical evaluation... very good science, popularized in the network science methodological toolkit by Holland and Leinhardt (1970). And they did the data analysis necessary to prove it was good science, as well.

3.3.2 Empirical Evidence

Davis and Leinhardt (1968, 1972), and Davis (1970) gathered nearly 800 sociomatrices from many different sources and discovered a few interesting facts. First, they found that many relations measured were directional. The old, recommended strategy of focusing on semicycles in such structures was difficult to implement. Secondly, asymmetric dyads, in which one actor chooses another actor but the choice is not reciprocated, were very common. The ideas of balance and clusterability needed to be modified to take such situations into account (rather than ignoring the directionality of these arcs, which was the current practice when attention is focused on semicycles). Thirdly, they found that signed relations were rather rare. Thus, they decided to modify the theories of balance and clusterability so that the theories could be applied to signed directional relations. When even these new theories were later found lacking, Holland and Leinhardt (1971) revised them to unsigned directional relations.

Davis and Leinhardt (1972) also found that in some digraphs one subset of actors chose a second, while actors in this second subset chose members of a third subset. The clusters of actors appeared to be *ranked*, or hierarchical in nature, with the actors on the bottom choosing those at the top (but not vice versa). Figure 3.1 shows the triads for a signed, directed relation.

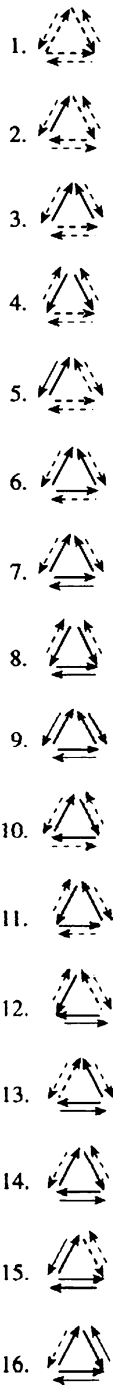

Fig. 3.1 Triads for a signed, directed relation

Holland and Leinhardt (1970) were the first to suggest the extension of these ideas to nonsigned directional relations. To turn ranked clusterability for complete signed digraphs into an equivalent idea for digraphs without signs is quite simple. Take the idea of ranked clusters for complete signed digraphs and do not consider arcs with negative signs. Then, any arc with a sign of " – " is removed from the signed digraph. Drop the positive signs from the remaining arcs. The assumption is that the relation under study is the *positive* part of the signed relation – for example, we study only "like," "not like," and "dislike." Figure 3.1 shows the triples of Fig. 3.2, without the negative arcs. The triples arising from directional relations are commonly referred to as *triads*, since we consider the threesome of nodes and all the arcs between them.

Note that the two problematic triads from ranked clusterability found empirically to be quite common have one and five arcs. These triads are numbered 2 and 16 in Fig. 3.1. Holland and Leinhardt (1971) showed that ranked clusterability is a special case of a more general set of theorems that naturally blend balance, clusterability, and ranked clusterability. Their *partially ordered clusterability* leads naturally to a consideration of the concept of *transitivity*.

Holland and Leinhardt (1971) reviewed the postulates of balance theory, clusterability, and ranked clusterability, as well as transitive tournaments (Hempel, 1952; Landau, 1951a, 1951b, 1953), and proposed the very general concept of transitivity to explain social structures. Transitivity includes all the earlier ideas as special cases. From a transitive digraph, one can obtain balanced, clusterable, and ranked clusterable graphs by making various assumptions about reciprocity and asymmetry of choices. During the past two decades, evidence has accumulated that transitivity is indeed a compelling force in the organization of social groups. What is even more remarkable, is that the idea was discovered anew, by the physicists invading the network science world 10 years ago. And now, transitivity and clusterability are very *hot*.

3.4 Some Current History

Early work on distributions for graphs was quite limited, forcing researchers to adopt independence assumptions that were not terribly realistic (see Chaps. 13–16 of Wasserman & Faust, 1994). It is hard to accept the standard assumption common in much of the literature, especially in physics, of complete independence and then to adopt the misnamed and overly simplistic *random graph* distribution (there are, of course, an infinite number of random graph distributions). *The* random graph distribution to the physicists, usually referred to as a *Bernoulli graph* (Wasserman & Faust, 1994, Chap. 13), assumes no dependencies at all among the random components of a graph. Equally hard to believe as a true representation of social

Fig. 3.2 Triads for a directed relation

behavior are the many conditional uniform distributions and p_1, which assumes independent dyads (Holland & Leinhardt, 1977, 1981).

The breakthrough in statistical modeling of networks was first exposited by Frank and Strauss (1986), who termed their model a *Markov random graph*. Further developments, especially commentary on estimation of distribution parameters, were given by Strauss and Ikeda (1990). Wasserman and Pattison (1996) elaborated upon the model, describing a more general family of distributions. Pattison and Wasserman (1999), Robins, Pattison, and Wasserman (1999), and Anderson, Wasserman, and Crouch (1999) further developed this family of models, showing how a Markov parametric assumption gives just one, of many, possible sets of parameters. This family, with its variety and extensions, was named p^*, a label by which it has come to be known. The parameters (which are determined by the hypothesized dependence structure) reflect structural concerns, which are assumed to be governing the probabilistic nature of the underlying social and/or behavioral process.

Work continues on this family, pointing out generalizations (Pattison & Robins, 2002; Robins, Elliot, & Pattison, 2001; Robins, Pattison, & Elliott, 2001; Snijders, Pattison, Robins, & Handcock, 2006), degeneracies (Handcock, 2002), and new estimation strategies (Hunter, 2007; Hunter & Handcock, 2006; Snijders, 2002).

The early work by the first researchers extended p^* in a variety of ways and laid the foundation for work in this decade on the estimation problems inherent in the early formulations. This research also was an important forerunner of the new parametric specifications that promise wider usage of the family. A more thorough history of this family of distributions, including a discussion of its roots in spatial modeling and statistical physics, can be found in Borner, Sanyal, and Vespignani (2007). Wasserman and Robins (2005) offered a review of p^* circa 2003, while Robins, Pattison, Kalish, and Lusher (2007) reviewed the 2003–2006 period. Other recent thoughts can be found in the May 2007 issue of *Social Networks*, a special issue devoted, in part, to p^*.

The work of Frank and Strauss (1986) did indeed begin a new era for statistical modeling of networks, although it took 10 years for Markov random graphs to be discussed at more length by network methodologists. What is remarkable is how the wheel keeps getting reinvented. Witness the rebirth of Holland and Leinhardt's transitivity index, as we describe below.

3.5 Clustering Coefficients

The clustering coefficient of a vertex in a graph quantifies how close the vertex and its neighbors are to being a clique (complete graph). Duncan Watts and Steven Strogatz introduced the measure in 1998 to determine whether a graph is a small-world network (hubs, which are part of cliques, but also adjacent to other hubs).

3.5.1 Formal Definition

A graph $G = (V, E)$ formally consists of a set of vertices V and a set of edges E between them. An edge e_{ij} connects vertex i with vertex j. The neighborhood N for a vertex v_i is defined as its immediately connected neighbors as follows:

$$N_i = \{v_j\} : e_{ij} \in E \text{ or } e_{ji} \in E. \tag{3.1}$$

The degree k_i of a vertex is defined as the number of vertices, $||N_i||$, in its neighborhood N_i. The clustering coefficient C_i for a vertex v_i is then given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. For a directed graph, e_{ij} is distinct from e_{ji}, and therefore for each neighborhood N_i there are $k_i(k_i - 1)$ links that could exist among the vertices within the neighborhood (k_i is the total (in + out) degree of the vertex). Thus, the clustering coefficient for directed graphs is given as

$$C_i = \frac{||\{e_{jk}\}||}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E. \tag{3.2}$$

An undirected graph has the property that e_{ij} and e_{ji} are equal by definition. Therefore, if a vertex v_i has k_i neighbors, edges could exist among the vertices within the neighborhood. Thus, the clustering coefficient for undirected graphs can be defined as

$$C_i = \frac{2||\{e_{jk}\}||}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E. \tag{3.3}$$

Let $\lambda_{G(v)}$ be the number of triangles on $v \in V(G)$ for undirected graph G. That is, $\lambda_{G(v)}$ is the number of subgraphs of G with three edges and three vertices, one of which is v. Let $\tau_{G(v)}$ be the number of triples on $v \in V(G)$. That is, $\tau_{G(v)}$ is the number of subgraphs (not necessarily induced) with two edges and three vertices, one of which is v and such that v is incident to both edges. Then we can also define the clustering coefficient as

$$C_i = \frac{\lambda_G(v)}{\tau_G(v)}. \tag{3.4}$$

It is simple to show that the two preceding definitions are the same, since

$$\tau_G(v) = \frac{1}{2}k_i(k_i - 1). \tag{3.5}$$

These measures are 1 if every neighbor connected to v_i is also connected to every other vertex within the neighborhood and 0 if no vertex that is connected to v_i connects to any other vertex that is connected to v_i.

The clustering coefficient for the whole system is usually defined as the average of the clustering coefficient for each vertex:

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} C_i. \tag{3.6}$$

3.5.2 Clustering Coefficients in Usage and Practice

In words... the clustering coefficient $\overline{C}$ is defined as follows: Suppose that a vertex v has k_v neighbors (or *alters*, a term usually used in the social network literature); then at most $k_v(k_v - 1)/2$ edges can exist between the alters (this occurs when every neighbor of v is connected to every other neighbor of v). Let C_v denote the fraction of the allowable edges that actually exist. Then the clustering coefficient $\overline{C}$ is simply the average of C_v over all v.

This definition, introduced by Watts and Strogatz (1998), has been very important empirically, as featured in Duncan Watts' books (1999, 2003), and much research (see, for example, Robins, Pattison, & Woolcock, 2005).

Besides being used over the past decade to check a nondirected graph for transitivity, it has been used extensively to study the small-world nature of a graph. From Matt Jackson's (2008) nice text, a graph exhibiting the *small world property* has a small diameter and small average path length (as well illustrated in Watts, 1999). Quantitatively, a graph is considered small-world if its average clustering coefficient is significantly larger than that for a random graph constructed on the same vertex set, and if the graph has a small mean-shortest path length.

The small-world paradigm, introduced by Stan Milgram in the mid-1960s has stormed into our culture. Milgram's (1967) study, published in *Psychology Today*, showed that people in the United States seemed to be connected by approximately six acquaintanceship links, on average. From this finding, the notion of *six degrees of separation* was born. Milgram actually never used this now very popular phrase; the most likely popularizer of the term *six degrees of separation* would be John Guare, whose Pulitzer Prize–winning and Tony Award play with the same name was published in 1990 (Guare, 1990).

A generalization of the clustering coefficient to directed graphs is obvious and straightforward, thus bringing this idea directly in line with Holland and Leinhardt's (1971) τ index for transitivity. The only difference, of course, is the normalization for $\overline{C}$ and the standardization (z-scoring) for τ. The similarity was not noticed by Watts (1999, 2003) or Newman (2003). Holland and Leinhardt's (1971) research on this is not even mentioned by Jackson (2008).

The moral: Paul Holland's work with Sam Leinhardt on indexes for triads was replicated, with very little attribution, by the current generation of poorly educated physicists doing network science.

Science, and posterity, will note that Paul and Sam were the first to quantify the notion and importance of transitivity and clusterability.

Acknowledgement This research was supported by a grant from the U.S. Office of Naval Research (#N00014-02-1-0877). Ann McCranie graciously provided research and data analysis assistance.

References

Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p^* primer: Logit models for social networks. *Social Networks, 21*, 37–66.

Borner, K., Sanyal, S., & Vespignani, A. (2007). Network science: A theoretical and practical framework. In B. Cronin (Ed.), *Annual review of information science & technology* (Vol. 4, pp. 537–607). Medford, NJ: Information Today/American Society for Information Science and Technology.

Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2005). *Models and methods in social network analysis*. New York, NY: Cambridge University Press.

Davis, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two theoretical models on 742 sociograms. *American Sociological Review, 35*, 843–852.

Davis, J. A., & Leinhardt, S. (1968). *The structure of positive interpersonal relations in small groups*. Paper presented at the annual meeting of the American Sociological Association, Boston, MA.

Davis, J. A., & Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. In M. J. Berger, M. Zelditch Jr., & B. Anderson (Eds.), *Sociological theories in progress* ((pp, Vol. 2, pp. 218–251). New York, NY: Houghton-Mifflin.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81*, 832–842.

Guare, J. (1990). *Six degrees of separation*. New York, NY: Random House.

Handcock, M. S. (2002). Statistical models for social networks: Degeneracy and inference. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis* (pp. 229–240). Washington DC: National Academies.

Handcock, M. S., Raftery, A. E., & Tantrum, J. (2007). Model-based clustering for social networks (with discussion). *Journal of the Royal Statistical Society, Series A, 170*, 301–354.

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Chicago, IL: University of Chicago Press.

Hoff, P., Raftery, A., & Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97*, 1090–1098.

Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology, 70*, 492–513.

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies, 2*, 107–124.

Holland, P. W., & Leinhardt, S. (1977). *Notes on the statistical analysis of social network data*. Unpublished manuscript.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association, 76*, 33–65.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks, 29*, 216–230.

Hunter, D., & Handcock, M. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics, 15*, 565–583.

Jackson, M. O. (2008). *Social and economic networks*. Princeton, NJ: Princeton University Press.

Landau, H. G. (1951a). On dominance relations and the structure of animal societies: I. Effect of inherent characteristics. *Bulletin of Mathematical Biophysics, 13*, 1–19.

Landau, H. G. (1951b). On dominance relations and the structure of animal societies: II. Some effects of possible social factors. *Bulletin of Mathematical Biophysics, 13*, 245–262.

Landau, H. G. (1953). On dominance relations and the structure of animal societies: III. The condition for a score structure. *Bulletin of Mathematical Biophysics, 15*, 143–148.

Milgram, S. (1967). The small world problem. *Psychology Today, 1*(1), 60–67.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167–256. doi:10.1137/S003614450342480 DOI:dx.doi.org.

Newman, M., Barabasi, A.-L., & Watts, D.J. (Eds.). (2006). *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.

Pattison, P. E., & Robins, G. L. (2002). Neighbourhood-based models for social networks. *Sociological Methodology, 32*, 301–337.

Pattison, P. E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology, 52*, 169–193.

Robins, G. L., Elliot, P., & Pattison, P. E. (2001). Network models for social selection processes. *Social Networks, 23*, 1–30.

Robins, G. L., Pattison, P. E., & Elliott, P. (2001). Network models for social influence processes. *Psychometrika, 66*, 161–190.

Robins, G. L., Pattison, P. E., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks, 29*, 169–172.

Robins, G. L., Pattison, P. E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks, III. Valued relations. *Psychometrika, 64*, 371–394.

Robins, G. L., Pattison, P. E., & Woolcock, J. (2005). Social networks and small worlds. *American Journal of Sociology, 110*, 894–936.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure, 3*, 2.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology, 36*, 99–153.

Snijders, T. A. B., Steglich, C., & Schweinberger, M. (2007). Modeling the co-evolution of networks and behavior. In K. van Monfort, H. Oud, & A. Satoraa (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 41–71). Mahwah, NJ: Lawrence Erlbaum Associates.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association, 85*, 204–212.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York, NY: Cambridge University Press.

Wasserman, S., & Pattison, P. E. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p^*. *Psychometrika, 60*, 401–426.

Wasserman, S., & Robins, G. L. (2005). An introduction to random graphs, dependence graphs, and p^*. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 148–161). New York, NY: Cambridge University Press.

Watts, D. J. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton, NJ: Princeton University Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York, NY: W.W. Norton.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.

Part III
Holland Shaping ETS

Some of My Favorite Things About Working at ETS

Paul W. Holland

Of course it was the great colleagues and interesting problems that I had at ETS that kept me there for so many years. It is impossible to list all those with whom I have worked, but several are represented in this volume, and I will try to say something about them, as well as a few others.

I first met Dorothy Thayer a few days after I joined the group headed by Albert Beaton, the Office of Data Analysis Research (ODAR). She was what was called a *scientific programmer* elsewhere, but at ETS she was one of several data analysts. Many of my publications from my years at ETS are jointly authored with her because we had a good division of labor; I did the theory and formulas, and she did the programming and data analysis. I was not the only statistician with whom she worked: Mel Novick, Don Rubin, and currently, Charles Lewis were also blessed by her careful and thoughtful work. Her care was not always fully appreciated, however, when it required me (and her other colleagues) to rework the theory in which Dotty had found an error through her calculations. But more often than not, if Dottie could not understand what was supposed to be going on in the theory, I usually concluded that neither did I.

I worked with many other data analysts over my many years at ETS, too many to list, but I would be remiss if I did not mention a few. John Barone, Jim Ferris, Bruce Kaplan, Dave Saxe, and Judy Pollack all did amazing things for me with computers in the days when doing it yourself on a laptop was simply not the option it is today. I remember when Sam Leinhardt and I got personal computers for our work on social networks. Mine was the first personal computer at ETS, and now it is impossible to imagine how work got done without them. Well, in the 1970s and 1980s it did get done, and in those days, ODAR and its data analysts were the key for getting useful things out of computers.

I first met Don as a graduate student at Harvard when he took a course that Fred Mosteller and I ran on mathematical models in the social sciences. His did his thesis work with Bill Cochran and then came to ETS in Beaton's ODAR group. There is no question that Don is the main reason I eventually came to work at ETS. He and I were the two mathematical/consulting statisticians in ODAR and worked closely with the data analysts on projects that covered a vast array of

education-related research – from evaluating children's educational TV to computer-aided instruction. A few years after I had come on board, Don got the ETS management to fund the Program Statistics Research Project, which supported statistical research that had relevance to ETS programs. It just so happened that ETS had suffered some client push-back due to two different types of statistical problems – the bouncing beta problem in the Law School Validity Study Service and a significant equating error for one of the graduate school programs. Don argued that if some resources were put into the study of more modern statistical ways to doing these tasks that it would help ETS. So, he took on the problem of stabilizing the regression coefficients used in validity studies, while I began my foray into test equating that eventually resulted in two books and many papers. Of course, my greatest debt to Don is his introducing me to the statistical issues in causal inference. Surprising to some, his initial work on this topic grew out of his early interest in missing data. The "counterfactual" way of looking at causal inference problems posits a great deal of missing data (e.g., the response to the treatment condition of those in the control group) that these responses are always missing. We wrote several papers together on various subtopics within causal inference, but his ideas are his own and my contributions were more of explication in terms that I found easier to understand. Don left ETS and eventually ended up back at Harvard University's Department of Statistics where he has had a long and distinguished career and many distinguished students. I was disappointed when he left ETS but happy with his many successes.

I first met Brian Junker when he was a graduate student of Bill Stout at the University of Illinois. Bill and I had a common interest in nonparametric evaluations of the fit of item response theory (IRT) models in psychometrics, and we had similar approaches to this. Brian was one of a very solid group of Bill's students, and I was able to convince him to be a summer intern at ETS. During that visit we made a curious discovery about ourselves. For some reason we discussed the fact that my mother lived in Santa Maria, California. He said, "Oh, where does she live?" I gave the street name, and he then said, "What is the address on that street?" I gave it, and from the address he believed that his parents-in-law lived exactly across the street from my mother. We checked it out and found that it is truly a small world. Brian's paper for this volume has, to me, a very clear connection to the type of foundational IRT work that interested me when I first met him. He puts a variety of models into a common framework that makes their common features clearer.

Paul Rosenbaum got his degree with Don at Harvard and then joined the ETS staff in my group as a young scholar who had wide interests and a willingness to do serious consulting for the rest of ETS. He and I did some interesting work on the foundations of IRT, the basic workhorse of modern psychometrics. He too, was destined to be a great teacher and left ETS for a distinguished career at Penn. His contribution to this volume is on a topic of great interest to me: the proper design of good observational studies for drawing causal inferences.

At ETS there are a considerable number of psychometricians and statisticians, but their numbers are dwarfed by the test development staff that is drawn from

many academic fields – literature, mathematics, political science, and so on, depending on the subject matter required. Mike Zieky is one of the first test developers that I met at ETS, and we were both heavily involved when ETS was setting up its system for identifying potentially biased items using the DIF measures that I and others at ETS had developed. The process of setting up DIF analyses for operational use took many meetings and many brains working together to make it both practical and effective. This effort apparently paid off because the approach we took to making DIF operational has been copied in various ways by testing organizations all over the world. Mike's contribution to this volume describes the result of this effort and reflects his many years of experience with it.

Chapter 4
Bayesian Analysis of a Two-Group Randomized Encouragement Design

Donald B. Rubin

4.1 Randomized Encouragement Designs

Randomized *encouragement designs*, terminology established in the seminal article by Holland (1988) although used earlier (e.g., Swinton, 1975), are the norm when dealing with human populations. At least in much of the world today, thankfully, we cannot force anyone to take a randomly assigned treatment; rather, we can only encourage them to do so, typically after describing some details of what to expect under each of the treatment conditions prior to participation, so-called *informed consent*. As a consequence, human experiments often face the complication of noncompliance with assigned treatment. For example, the treatment group may be randomly assigned to be encouraged to study more, whereas the control group receives no extra encouragement. In this example, hours of studying will be measured in both groups of the study – treatment and control – as will the primary outcome variable, final achievement on a test. Not only is it of interest to study the effects of the encouragement on the amount of studying and on achievement, but it is also of interest to investigate the relationship between the amount of studying, an *intermediate* outcome variable, on the *primary* outcome variable in the treatment and control groups.

Holland (1988) was an early statistically coherent and principled attack on this problem of intermediate outcomes. Yet, it was partially ignored by a student, colleague, and old friend of his when recently writing with his student about the problems of noncompliance. Who is the scoundrel and in what publication? The answer is: the author of this contribution in Jin and Rubin (2008) – JR, henceforth – that reanalyzed data from Efron and Feldman (1991) – EF, henceforth. EF concerned a randomized double-blind trial of an active drug (cholestyramine) versus a placebo for cholesterol reduction – the primary outcome variable, where compliance, or dose, was measured by the proportion of assigned pills, either active or placebo, taken.

D.B. Rubin (✉)
Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, USA
e-mail: rubin@stat.harvard.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_4,

EF tried to estimate a dose-response relation in the active treatment group even though the dose taken was self-selected and not randomized; EF described the true dose-response relation as the one that would be observed in a large study with dose randomly assigned and strictly enforced. As published, the EF article included written discussion, and the consensus seemed to be that EF's analysis was not convincing. I was among the discussants reaching this conclusion, and in my discussion, I referred to Holland (1988), calling it a "particularly relevant" article. But for some reason, JR did not cite this important paper. I redress that oversight here and clarify and simplify the analysis of the EF data presented in JR.

4.2 The Efron and Feldman Data and Some Initial Descriptive Analyses

The patients in the trial were all men who had been told that their cholesterol levels were elevated, and that they should exercise and diet more, as well as possibly take a cholesterol-reducing drug; 164 were randomized to take the active drug, indicated by $Z_i = T$ for the i^{th} man, and 171 were randomized to the control group and assigned to take the placebo, indicated by $Z_i = C$. For each patient, cholesterol levels were measured before and after taking the drug or placebo. The potential outcome variables for the i^{th} man, $Y_i(T)$, $Y_i(C)$, were the decrease in cholesterol level when assigned T or when assigned C, and the observed cholesterol reduction is the only variable used by EF or JR besides the treatment indicator, Z_i, and the dose taken (notation for doses later). The observed cholesterol reduction for the i^{th} man is denoted $Y_i.obs$, where $Y_i.obs = Y_i(T)$ when $Z_i = T$, and $Y_i.obs = Y_i(C)$ when $Z_i = C$. The notation used here relatively closely follows that used in JR.

The dose taken suffers from *partial compliance* complications: Most patients in the treatment group took only a fraction of their assigned dose, and most patients in the control group took only a fraction of their assigned dose. EF opined that this complication may have a hidden benefit in that it may allow us to estimate a dose-response relation. Let $D_i(T)$ be the proportion of assigned dose actually taken (as estimated by a pill count) when patient i is assigned treatment, which is observed when patient i is assigned drug, but missing when assigned control. By design, $D_i(C)$ is zero because the men have no access to the drug except within the experiment. Analogously, let $d_i(C)$ be the proportion of placebo dose taken when patient i is assigned to control, which is observed when patient i is assigned control and missing when assigned treatment. By design, $d_i(T)$ is zero because there is no access to the placebo when assigned treatment. The need for both $D_i(T)$ and $d_i(C)$ is called "extended noncompliance" in JR, extended from the more usual situation where we could use just $D_i(T)$ and $D_i(C)$. The observed values $\{D_i.obs\}$ are $\{D_i(T)\}$ in the treated group and all zero in the control group, and similarly, the observed values $\{d_i.obs\}$ are $\{d_i(C)\}$ in the control group and all zero in the treated group.

Figure 4.1, based on data in Efron and Feldman (1991), reveals an apparent dose-response relationship in the treated group, with a quadratically increasing trend,

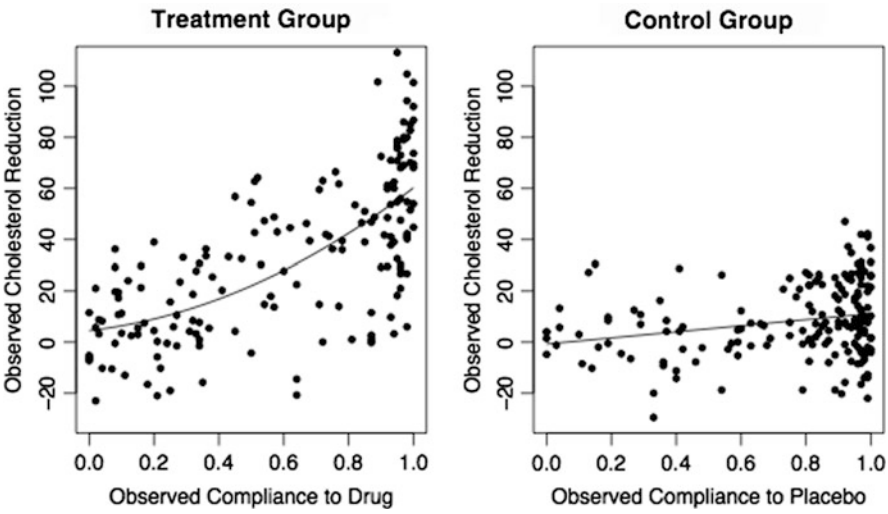

Fig. 4.1 Relationship between observed cholesterol reduction and observed compliance. From "Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data," by H. Jin and D. B. Rubin, 2008, *Journal of the American Statistical Association, 103*(481), p. 102. Copyright 2008 by the American Statistical Association. Reprinted with permission

as estimated by EF, which is not surprising given the nature of the drug. More surprising, however, is EF's estimation of a linearly increasing dose-response relationship in the control group. The placebo is thought to be totally inert, with no possible causal connection to cholesterol reduction. What is going on in the control group? Recall that both doses are self-selected, so that the plot suggests those who are more compliant in the control group (i.e., take more of their assigned dose) have more cholesterol reduction. This makes sense because all of these men know that they have high cholesterol, so that those of them who take more of their daily dose of placebo are also more likely to exercise regularly, watch their intake of fatty foods, and so on, and therefore would be expected to have more cholesterol reduction in time than the more noncompliant patients. Placebo compliance is really a descriptor of the patients, a covariate, which is only observed in the control group. In some sense, what we want to do is subtract the control group's apparent dose-response from the treatment group's apparent dose-response and be left with a true apparent dose-response relationship. EF attempted this, but the discussion indicated that at least several readers were unconvinced. The objective in JR was to do this subtraction correctly under explicit assumptions, based on the framework of *principal stratification* (Frangakis & Rubin, 2002).

A very important consideration is that the observed compliance rates in the two groups are very different. Figure 4.2, based on data from JR, reveals that compliance is much better in the control group than in the treatment group, which is not surprising because cholestyramine works by inhibiting the absorption of fat,

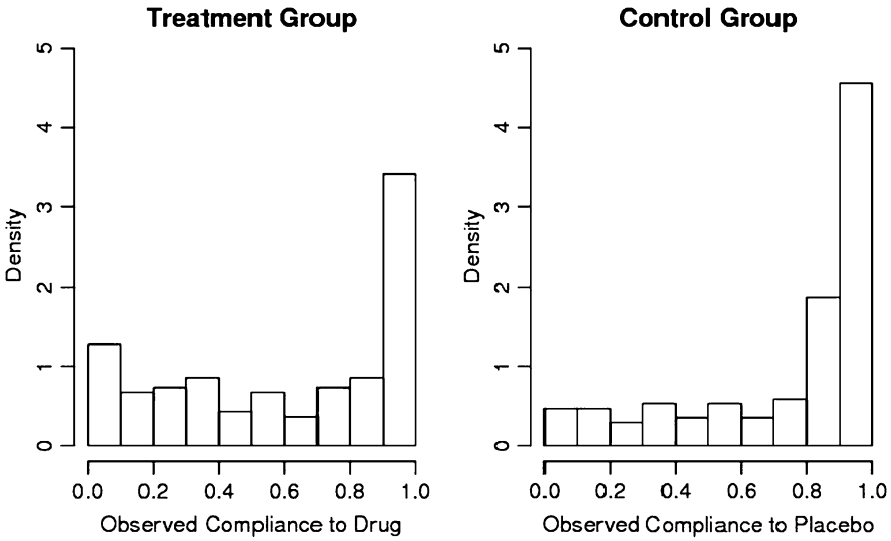

Fig. 4.2 Histograms of observed compliance. From "Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data," by H. Jin and D. B. Rubin, 2008, *Journal of the American Statistical Association, 103*(481), p. 102. Copyright 2008 by the American Statistical Association. Reprinted with permission

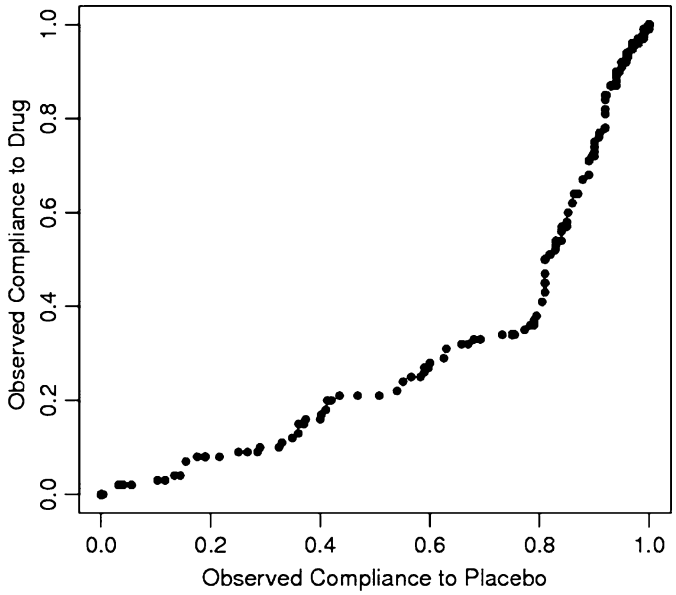

Fig. 4.3 Q-Q plot of observed drug and observed placebo compliance. From "Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data," by H. Jin and D. B. Rubin, 2008, *Journal of the American Statistical Association, 103*(481), p. 102. Copyright 2008 by the American Statistical Association. Reprinted with permission

and as a consequence, it can produce excessive colonic gas in some people. Because of the randomization, the men in the treatment group would be expected to have the same distribution of placebo compliance as observed in the placebo group, and analogously, the men in the placebo group would be expected to have the same distribution of drug compliance as observed in the treatment group. Figure 4.3 displays the differences in the observed compliance behaviors based on a Q-Q plot from JR; if the placebo were a perfect blind, we would expect to see roughly a 45-degree line, but this is clearly not the case.

4.3 Efron and Feldman's Approach to Estimating Dose-Response

The way EF handled the difference between the distributions was to *equipercentile equate* (Holland & Rubin, 1983) D_i.obs and d_i.obs, thereby making $D_i(T)$ known for everyone and making $d_i(C)$ known for everyone. Effectively then, $d_i(C)$ becomes a known covariate, and $D_i(T)$ becomes the dose in the active treatment group, which is then essentially considered ignorably assigned (Rubin, 1978) by EF given $d_i(C)$, which is the only *covariate* used in either the EF analysis or the JR analysis. But a few issues occur with the EF analysis. First, although the equipercentile equating is correct in expectation because of the randomization, there is no reflection of any uncertainty of this imputation of all the missing dose data. Second, why should we accept the ignorability of the assignment of active dose given placebo dose, especially when placebo dose is, in fact, fully missing in the group being assigned active doses? And, third, how can active dose simultaneously be assumed to be a one-one function of placebo dose and stochastically assigned given it? A fourth issue concerns the possibility of dose-response changing with the type of placebo-complier; for example, maybe the better placebo compliers will benefit more from the same dose of drug because they are also doing other things to improve their cholesterol, or maybe the opposite is true because there is less room for improvement in cholesterol due to the drug because of their other activities. The EF analysis does not allow any possibility to study this issue because placebo compliance and dose of the active drug taken are one-one functions of each other.

The analysis in JR addressed these issues using a Bayesian model that explicated all needed assumptions, and the presentation here is a crisper and more direct one than in JR. This perspective is basically the Rubin Causal Model (Holland, 1986) as expanded to include principal stratification (Frangakis & Rubin, 2002).

4.4 JR's Assumptions and Hypothetical Experiment

We begin by stating two standard assumptions made by both EF and JR. First, we accept the stable unit treatment value assumption (SUTVA; Rubin, 1980), whereby there is no interference between units and no hidden versions of either treatment or

control for any unit; SUTVA allows us to write the matrix of all values that could be observed in this experiment as a matrix with 164 + 171 rows, and with a column for each variable (defined more precisely shortly). Second, both EF and JR assume the ignorability of treatment assignment of active drug versus placebo, which is justified by the randomization.

JR also, like EF, considered the active dose taken to be part of the experimental assignment, but JR made this assumption explicit by describing a hypothetical experiment that could have led to the EF observed data and that allowed placebo compliance to be missing in the active treatment group. Thus, JR eliminated EF's equipercentile equating of compliances and replaced it with the explicit description of the assignment mechanism for the dose of the active drug given placebo compliance, which was assumed latently ignorable (Frangakis & Rubin, 1999), that is, probabilistic as a function of placebo compliance, rather than just ignorable.

More precisely, in this hypothetical experiment, there was a run-in period prior to randomization where each man's baseline placebo compliance was measured using the same assigned dose as in the actual control group; call this d_i^*. Then at randomization, 164 men were randomly selected for treatment with the active drug, and 171 were randomly selected for control and given placebo, just as with the actual study. For the 164 men selected for active drug treatment, for a man with baseline placebo compliance, d_i^*, the active dose was assigned according to a drawn value of a Beta random variable (on [0,1]), which gave the active dose that was to be assigned and enforced, as a fraction of d_i^* (e.g., if the drawn Beta was 0.9, the man was assigned $0.9 \times d_i^*$ for his active dose). Because of the known negative side effects of the drug, no effort was made to assign and enforce a larger dose of the active drug than the man would take of an inert placebo.

Continuing with this hypothetical experiment, when the study was complete, it was noticed that observed placebo compliance in the control group, $d_i(C)$, was identical to placebo compliance at baseline, d_i^*, for all men. As a result, the investigators naively threw away d_i^* in both groups of the experiment, and also, they forgot to record the parameters of the Beta distribution used to draw values of active dose, thinking that these values were irrelevant once the actual dose of the active drug was known (naïve Bayesians, no doubt). Thus, the assignment of active drug versus control is, as it actually was, strongly ignorable (Rosenbaum & Rubin, 1983), but the assignment of dose of active drug in the treated group was latently ignorable (Frangakis & Rubin, 1999) – that is, it would be ignorable if d_i^* were observed, and it would be precisely known if the parameters of the Beta distribution had been remembered.

4.5 JR'S Formal Notation and Model

Figure 4.4, based on data from JR, displays all the variables described in this hypothetical experiment. One potential Y outcome exists for each possible dose, which are labeled T_0 for zero dose, …, and T_1 for full dose. The objective is to estimate true dose-response for this hypothetical experiment from its observed data.

i	d_i^*	Z_i	Z_{Di}	$d_i(T)$	$d_i(C)$	$Y_i(T_0)$	...	$Y_i(T_D)$	...	$Y_i(T_1)$	$Y_i(C)$
1	?	T	T_0	0	?	⋆	?	?	?	?	?
...	?	T	...	0	?	...	...	...	...	...	?
...	?	T	T_D	0	?	?	?	⋆	?	?	?
...	?	T	...	0	?	...	...	...	...	...	?
n_T	?	T	T_1	0	?	?	?	?	?	⋆	?
n_T+1	?	C	?	0	⋆	?	?	?	?	?	⋆
...	?	C	?	0	⋆	?	?	?	?	?	⋆
...	?	C	?	0	⋆	?	?	?	?	?	⋆
n	?	C	?	0	⋆	?	?	?	?	?	⋆

"⋆" represents observed data, "?" represents missing data.

Fig. 4.4 Principal stratification framework for dose-response with $d_i(C)$ defining strata and $Z_{Di}(T)$ defining dose. From "Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data," by H. Jin and D. B. Rubin, 2008, *Journal of the American Statistical Association, 103*(481), p. 108. Copyright 2008 by the American Statistical Association. Reprinted with permission

Let θ equal all parameters in the model. Also, assume $d_i(C) = d_i^*$ for everyone, denoted d_i in the following expressions. Formally, the treatment assignment mechanism has two parts. First, the actual randomization of $Z_i = T$ versus $Z_i = C$: $\Pr[Z|d_i, Y_i(C), \{Y_i(Z_{Di})\}, \theta] = \Pr[Z|\theta] \sim$ constant. And second the hypothetical randomization of dose Z_{Di} given $Z = T$: $\Pr[Z_{Di}|d_i, Y_i(C), \{Y_i(Z_{Di})\}, Z = T, \theta] = \Pr[Z_{Di}|d_i, Z = T, \theta] = d_i \times \text{Beta}(\alpha_1, \alpha_2)$. This model has active dose latently ignorable given the partially observed variable d_i, which is, actually, fully missing for those assigned treatment.

Next, the model for the covariate distribution is $\Pr[d_i|\theta] = \text{Beta}(\alpha_3, \alpha_4)$. Diagnostics presented later suggest that these Beta assumptions are reasonable. Next, we summarize JR's parametric model for the potential outcomes joint distribution given d_i and θ. Here is where the science of dose-response enters. First, $\Pr[Y_i(C)|\, d_i, \theta] = N(\beta_0 + \beta\, d_i, \sigma_C^2)$; that is, response under control is linearly related to placebo compliance. Also $\Pr[Y_i(Z_{Di})|\ Y_i(C),\ d_i,\ \theta] \sim N[Y_i(C) + \gamma_1\, Z_{Di} + \gamma_2\, Z_{Di}^2 + \gamma_3\, Z_{Di}\, d_i, \sigma_{T.C}^2]$, mutually conditionally independent across the Z_{Di}, and $\gamma_1 \geq 0$, $\gamma_2 \geq 0$, $\gamma_1 + \gamma_3 \geq 0$; when $Z_{Di} = 0$, the expectation of $Y_i(Z_{Di}) - Y_i(C)$ is zero; thus, the causal effect of a zero dose of the drug is constrained to be zero in expectation; moreover, dose-response is constrained to be monotonally and quadratically increasing for this range of doses.

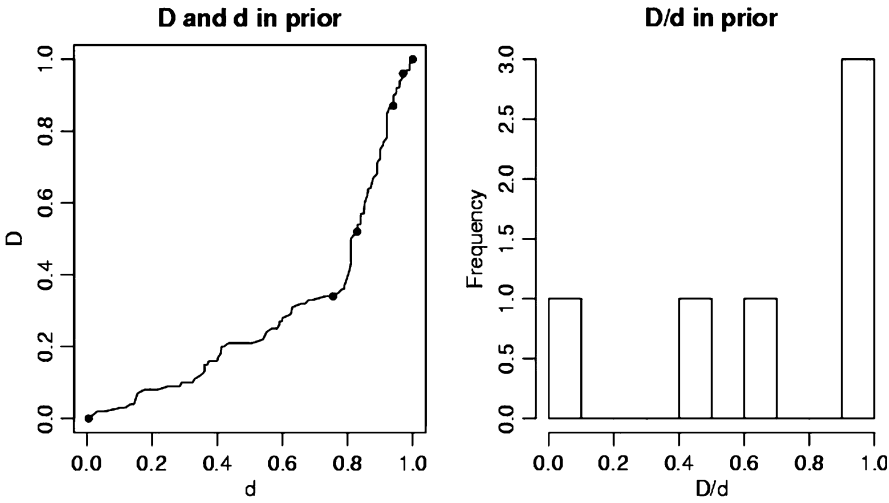

Fig. 4.5 Six prior data points for $\pi(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$

Finally, we summarize the prior distribution on θ. The prior distribution on the parameters of the Betas is specified by adding six "fake" men with both Z_{Di} and d_i observed on the equal percentile equating line, where these six men have nothing else observed. These are the minimum, 25th percentile, median, 75th percentile, maximum on the equipercentile equating line, as displayed in Fig. 4.5. The purpose of these fake men is simply to stabilize computation and has little influence on inference because there are only six fake men and 335 real ones, and the fake values are accurate in expectation because of the randomization. The prior distribution on the rest of θ is independent and is the standard "noninformative" prior proportional to $1/(\sigma_C\ \sigma_{T.C})$.

4.6 JR's Computation and Diagnostic Checks

This missing data problem is addressed by JR using MCMC to draw Bayesian inferences by iterative simulation. The parameters are θ, and the key missing data are d_i for those assigned treatment and Z_{Di} for those assigned control. The steps of the simulation are as follows: given θ, draw the key missing data; given the key missing data, draw θ; iterate until approximate convergence. A large number of such draws approximates the posterior distribution of dose-response as a function of principal strata defined by $d_i(C)$. The details are found in JR.

The propriety of the Beta-Beta part of their model was addressed by JR using diagnostic plots given in Fig. 4.6, which display one representative draw of the key missing data. The upper left Q-Q plot reveals that the drawn doses of the active drug in the control group (the values on the vertical axis) have nearly the same

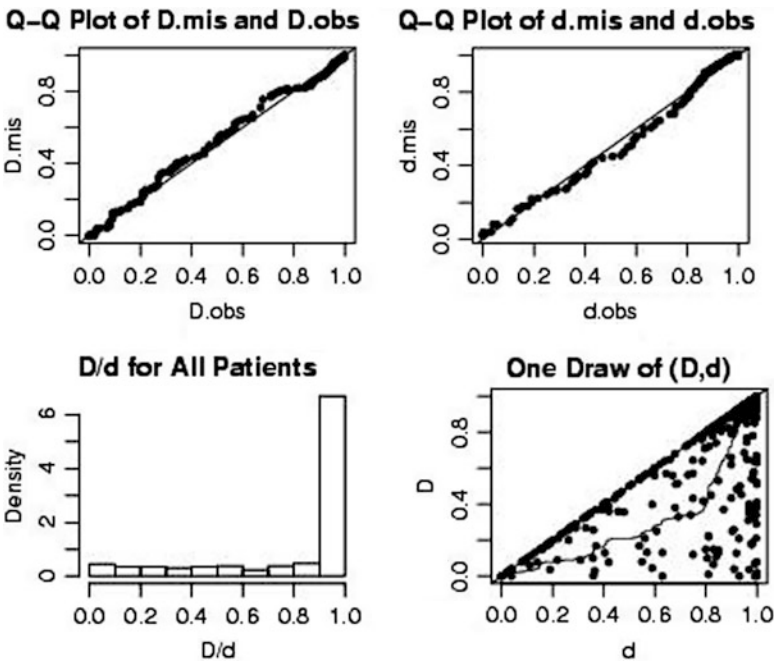

Fig. 4.6 Diagnostic checks for JR's model, one posterior draw of key missing data

distribution as the actual doses of the active drug in the treatment group, as they should, and the upper right Q-Q plot reveals that the drawn values of placebo compliances in the treatment group (the values on the vertical axis) have essentially the same distribution as the actual values of placebo compliance in the control group, as they should too. The lower left plot indicates that according to the JR model, most of the men would take doses of the active drug that are within 90% of their placebo compliances. And the lower right plot suggests that the EF assumption of a deterministic relationship between dose of active drug and placebo compliance is not well supported by the data, at least under JR's more flexible model.

4.7 JR's Dose-Response Results

Figure 4.7 from JR displays the estimated dose-response curves at four selected values of placebo compliance: a perfect placebo complier, a 75th percentile placebo complier ($d_i = 0.97$), a median placebo complier ($d_i = 0.89$) and a 25th percentile placebo complier ($d_i = 0.60$). Compare the four plots at $Z_{Di} = 0.60$, which is the largest dose that would be assigned to a 60% placebo complier in our hypothetical experiment. The solid line is the expected dose-response, and the dotted lines give 95% posterior intervals. The poor placebo complier can expect nearly a 50 point

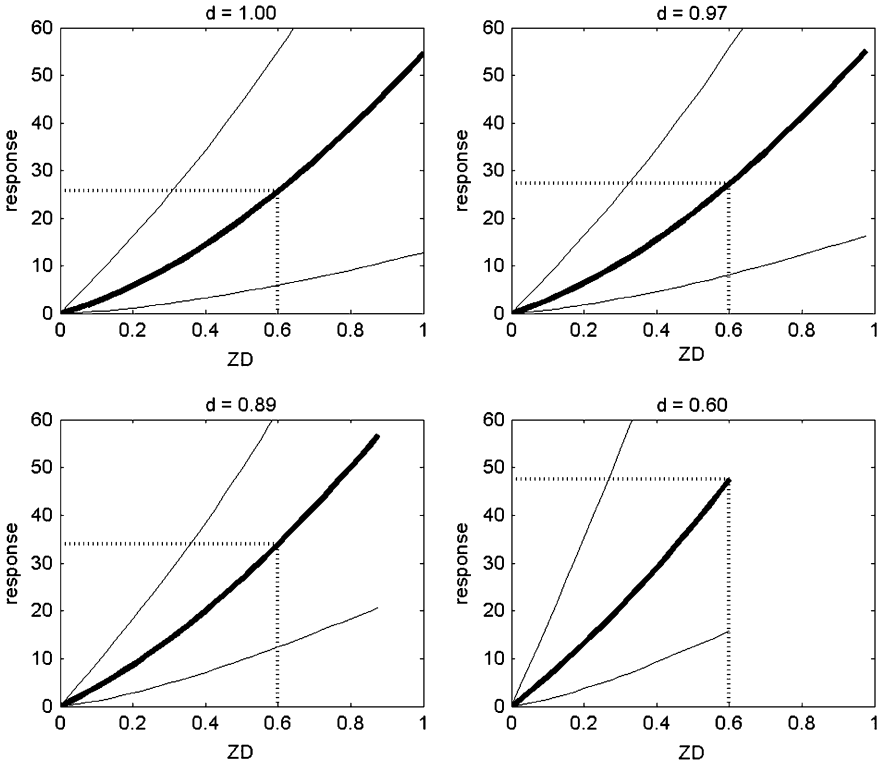

Fig. 4.7 Dose-response results for principal strata, maximum d, 75th d, median d, 25th d. From "Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data," by H. Jin and D. B. Rubin, 2008, *Journal of the American Statistical Association, 103*(481), p. 109. Copyright 2008 by the American Statistical Association. Reprinted with permission

reduction from a 60% dose, whereas a perfect placebo complier can expect about half that! Does this make sense? On reflection, I think so because the poor complier is probably doing very little to lower his cholesterol, other than taking the drug, which leaves more reduction available due to the drug alone. This result was at first surprising but eventually reinforced the utility of the model and approach being used by JR.

4.8 Discussion of the Dose-Response Conclusions

Under EF's assumptions, dose-response at each $d_i(C)$ is a point because $D_i(T)$ is a one-one function of $d_i(C)$ – very implausible. Instead, JR's dose-response results are causal under a debatable assumption. Is *Nature's randomization* of dose given

placebo compliance (i.e., the crucial latent ignorability assumption) plausible? Or do we need to condition further on background medical characteristics related to possible side effects of the drug? This issue is one that would be interesting to address by a sensitivity analysis in the context of a currently relevant treatment; cholestyramine is no longer of much interest because the class of drugs called statins appear to be much more effective, with typically fewer side effects. The model and analysis do suggest that in such encouragement designs, it is important to collect covariates that are predictive of outcomes and compliance behavior to reduce reliance on untestable assumptions. The framework presented here, however, appears to be superior to earlier attempts to address the same issue, and much of it was anticipated in Holland (1988).

Acknowledgement Thanks are due the reviewers of this piece, Sandip Sinharay, Jiahe Qian, and Derek Briggs for their helpful and thoughtful comments, and NSF Grant # SES-0550887 and NIH Grant # 10006441691 for their partial support in its production.

References

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association, 86*, 9–17.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment – noncompliance and subsequent missing outcomes. *Biometrika, 86*, 365–379.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58*, 20–29.

Holland, P. W. (1986). Statistics and causal inference (with discussion and rejoinder). *Journal of the American Statistical Association, 81*, 945–960.

Holland, P. W. (1988). Causal inference and path analysis. *Sociological Methodology, 18*, 449–484.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick Lord* (pp. 3–25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data. *Journal of the American Statistical Association, 103*, 101–111.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34–58.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association, 75*, 591–593.

Swinton, S. (1975). *An encouraging note*. Unpublished manuscript.

Chapter 5
The Role of Nonparametric Analysis in Assessment Modeling: Then and Now

Brian W. Junker

5.1 Item Response Models

Item response theory (IRT) is a family of statistical psychometric models for discretely scored responses of subjects (students, survey respondents, etc.) to items (questions) on exams, surveys, and so on, using a continuous latent variable to represent the general propensity of each subject to respond positively to each item or question. Although polytomous and partially ordered responses are considered in the IRT literature (e.g., van der Linden & Hambleton, 1997), this chapter will concentrate on ordered dichotomous response variables

$$\mathrm{X}_{ij} = \begin{cases} 1, & \text{if subject } i \text{ responds positively to item } j \\ 0, & \text{otherwise} \end{cases}, \tag{5.1}$$

$i = 1,\ldots,N$, $j = 1,\ldots J$, where a positive response might be a correct answer on a cognitive test, agreement on an attitudinal inventory, or other result. X_{ij} is a Bernoulli random variable, so that a model can be specified as

$$\begin{aligned} P[X_{ij} = 1|\theta_i, \beta_j] &= P(\theta_i, \beta_j) \\ P[X_{ij} = 0|\theta_i, \beta_j] &= 1 - P(\theta_i, \beta_j) \end{aligned},$$

where θ_i is the subject's latent variable, β_j are parameters describing the distribution of X_{ij} given θ_j, and $P(\theta_i, \beta_j)$ is some convenient probability function. The variable θ_i is sometimes referred to as a latent *proficiency* or *ability* (or, e.g., attitude, preference, etc., but this chapter will focus on cognitive assessment) variable and generally is interpreted as expressing the quantity of whatever is needed to respond positively to items; similarly β_j parameterizes features of item j.

B.W. Junker (✉)
Department of Statistics, Carnegie Mellon University, 132E Baker Hall,
Pittsburgh, PA 15213, USA
e-mail: brian@stat.cmu.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_5,

Here and throughout, θ_i and β_j (and other model parameters represented by Greek letters) may be scalar- or vector-valued in general. In cases where the distinction is important, it will be clear from context whether scalar or vector is meant.

In the simplest (but widely useful) IRT models, this chapter assumes θ_i is *unidimensional*, that is, θ_i is a scalar in $\Re$ (the real line), and that each $P(\theta_i, \beta_j)$ is *monotone*, that is, nondecreasing in θ_i. It is also usual to assume *local independence*, that is, a subject's responses are independent given the value of his/her latent variable θ_i, so that the IRT likelihood for one subject's responses is of the form

$$\begin{aligned} P[X_{i1} = x_{i1}, \ldots, X_{iJ} = x_{iJ} | \theta_i, \beta_1, \ldots, \beta_J] \\ = \prod_{j=1}^{J} P(\theta_i, \beta_j)^{x_{ij}} (1 - P(\theta_i, \beta_j))^{1-x_{ij}}. \end{aligned} \tag{5.2}$$

This model is generative in the sense that it embodies a bit of theory about how responses are generated: given θ_i (and the β_j's), item responses could be generated (simulated) as independent coin-flips with probability $P(\theta_i, \beta_j)$. The theory is not deep psychologically – that is, it is not grounded in a detailed account of how responses are generated by subjects – but it is useful in thinking about how item response data might impact inference about θ_i (or β_j).

In parametric IRT, $P(\theta_i, \beta_j)$ is a smooth function of θ_i and a low-dimensional β_j. For example, in the Rasch (1980) model, β_j is scalar (one-dimensional) and the model can be expressed as $P(\theta_i, \beta_j) = g(\theta_i, \beta_j)$where $g(x) = \exp(x)/[1 + \exp(x)]$. When θ is modeled as a random effect, the Rasch model is clearly a generalized linear mixed model (GLMM; e.g., McCulloch & Searl, 2001) for example. This model leads to one fruitful way in which IRT models have been incorporated into a larger modern statistical modeling framework (e.g., DeBoeck & Wilson, 2004; Johnson & Albert, 1999; and Skrondal & Rabe-Hesketh, 2004).

There are two basic inferential tasks for parametric IRT. The first task is to estimate the item parameters β_j in order to assess the quality of the items. Given a calibrated model (that is, well-estimated β_j's), the second task is to make inferences on the θ_i's for ranking, selection and, to the extent that items with differing cognitive content map to different parts of the θ scale, diagnostic purposes. Especially in large scale testing, it is standard practice to apply an expectation-maximization (E-M) algorithm (as reviewed by Tanner, 1996, for example) or a similar method to the marginal likelihood:

$$\begin{aligned} P[X_{i1} = x_{i1}, \ldots, X_{iJ} = x_{iJ} | \beta_1, \ldots, \beta_J, \lambda] \\ = \int \prod_{j=1}^{J} P[X_{i1} = x_{i1}, \ldots, X_{iJ} = x_{iJ} | \theta, \beta_j] f(\theta | \lambda) d\theta, \end{aligned}$$

to obtain maximum likelihood (ML) estimates of the β_j's [and the hyperparameters λ for the density $f(\theta|\lambda)$] and then use these estimates to obtain empirical Bayes

estimates[1] of the θ's. In smaller experimental or observational studies, estimating treatment or condition effects from the marginal model may be enough.

It is also possible to estimate β_j's and θ_i's simultaneously. This approach has a bad reputation among ML methodologists, since joint ML estimates based on replicating (5.2) for $i = 1 \ldots N$ subjects are usually inconsistent (asymptotically biased; e.g., Haberman, 1977) unless care is taken with the rates at which N and J tend to infinity in the asymptotic (e.g., Douglas, 1997). However, Holland's (1990) review of several ML approaches suggests that the finite-sample estimates of the β's can be quite similar across methods. Moreover, Bayesian joint estimation of the β_j's and θ_i's provides both a better idea of uncertainty involved in estimating θ in small samples, as well as a rationale for consistent Bayesian marginal estimates based on the joint estimation machinery (as sketched by Patz & Junker, 1999).

In nonparametric IRT, the assumptions of *unidimensionality*, *monotonicity,* and *local independence* are usually retained, but the function $P(\theta_i, \beta_j)$ is usually replaced with a general unspecified function $P_j(\theta_i)$. The resulting model is sometimes called the *monotone unidimensional IRT model* (e.g., Junker, 1993) in contrast to other models that relax one or more of these assumptions. One possible relaxation is to consider more general response variables than dichotomous (0/1) X's. For general (not necessarily dichotomous) observed variables $X = (X_1, \ldots, X_J)$, the monotonicity assumption is replaced with the assumption that each X_j is *stochastically ordered* by the unidimensional latent variable θ. If this form of monotonicity holds along with unidimensionality and local independence, the result is called the *monotone unidimensional latent variable model*.

5.2 Nonparametric Item Response Theory

Approaching IRT from a nonparametric point of view goes back at least to Meredith (1965) and arguably back as far as Loevinger (1947) or even farther. A broader view of nonparametric item response theory (NIRT) can be found in Junker and Sijtsma (2001a), but this chapter will concentrate on results from the 1980s and 1990s that helped to illuminate the observable structure of IRT models and lead to serviceable tests for monotone unidimensional IRT models for dichotomous response variables.

A key observation is that the assumptions unidimensionality, monotonicity, and local independence are not vacuous in the sense that they constrain the possible multivariate distributions of $(X_{i1}, \ldots, X_{iJ})$. Indeed, each pair of item response variables (X_{j_1}, X_{j_2}) must be positively correlated, analogous to the positive correlations among observed variables in a one-factor factor analysis model with all positive factor loadings.

[1] When the model admits of it, alternative approaches using conditional likelihood are also used (as reviewed by Sijtsma & Junker, 2006).

However, much more is true, and in much greater generality. Holland and Rosenbaum (1986) reviewed the relevant literature on associated random variables and probability inequalities and extended their earlier work in the area, beginning with Holland (1981), to prove that any vector of observable variables $X = (X_1, \ldots, X_J)$ satisfying a monotone unidimensional latent variable model must also satisfy *conditional association* (CA): For any partition of X into disjoint subsets of variables $Y = (Y_1, \ldots, Y_{J_1})$ and $Z = (Z_1, \ldots, Z_{J_2})$,

$$\text{Cov}\left(f(Y), g(Y) | h(Z) = c\right) \geq 0$$

for all coordinate-wise nondecreasing functions $f(\ )$ and $g(\ )$ and all functions $h(\ )$. CA imposes very strong conditions on the observed distribution of X: If $f(\ )$ and $g(\ )$ are thought of as subtest scores for the subtest Y, then any two subtest scores will be positively correlated given any information at all about the rest of the test Z.

The very strong coherence among item response variables implied by CA nearly characterizes monotone unidimensional latent variable models. It turns out that a complete characterization of these models requires a moderating assumption about the covariances among items, *vanishing conditional dependence* (VCD):

$$\lim_{J_2 \to \infty} \text{Cov}(X_a, X_b | X_{j_{+1}}, \ldots, X_{J_1+J_2}) = 0$$

for all $a, b \in \{1, \ldots, J_1\}$, for every J_1. Then one can show (Junker & Ellis, 1997) that as the test length $J = J_1 + J_2$ increases, CA is not only implied by, but also guarantees the existence of, a nontrivial monotone unidimensional latent variable model for the (now infinite) sequence of response variables $X = (X_1, X_2, \ldots)$.

The machinery of the Junker and Ellis (1997) proof also establishes that the latent variable θ must lie in the *tail σ-field* of the sequence X. That is, although θ can be perfectly estimated using infinitely many item responses, it can never be perfectly known from finitely many responses. This mathematical result is a useful way of thinking about what a latent variable in psychometric models "really is" (Ellis & Junker, 1997, p. 516), in contrast to a variable that is missing under a particular data collection design but could have been observed under a different design.

Insights like this are one reason that NIRT has been useful and pursued in the psychometric literature. Researchers learn things about the fundamental structure of IRT models, and latent variable measurement models generally, by considering models with a bare minimum of assumptions.

Results such as the Holland and Rosenbaum (1986) CA theorem also give us a hunting license for testing for a unidimensional IRT model without specifying a parametric form for the model. This may be useful if one suspects lack of fit may be due to the particular form of a *parametric* IRT model rather than the generic *generative* IRT assumptions. Although Bartolucci and Forcina (2005) developed order-constrained likelihood ratio tests of CA in log-linear models, this approach

becomes prohibitively slow computationally as the test length J increases; using a different testing approach, Yuan and Clarke (2001) discussed how the many conditions implied by CA and VCD may lead to prohibitively large sample size needs. More successful approaches have been the test of essential unidimensionality initiated by Stout (1987) and the Mokken scaling procedures elaborated and described by Sijtsma and Molenaar (2002) and their students and colleagues.

In addition, if one knows (or concludes, e.g., from the Stout or Mokken procedures) that the generic IRT assumptions apply but one does not specify or fit a parametric IRT model, it is still possible to make inferences about θ from the observed response data. For example Grayson (1988; see also Huynh, 1994) shows that $X_+ = \sum_{j=1}^{J} X_j$ stochastically orders θ in a monotone unidimensional IRT model. Basing inferences on X_+ rather than on a parametric estimate of θ can be especially useful for shorter item sets, for which parametric models can be more difficult to fit stably, and for very long item sets, for which parametric model fitting methods based on integrating or maximizing the likelihood may be too slow. These ideas generally seem to work for nonparametric polytomous IRT models as well (van der Ark, 2005), although the theoretical results are not as clean as in the dichotomous case.

5.3 Cognitive Diagnosis Models

In recent years, another set of models, so-called *cognitive diagnosis models* (CDMs) have attracted attention in the psychometric literature. The motivations for using these models are much the same as the motivations behind the report *Knowing What Students Know* (National Research Council, 2001): a desire to move summative testing from a norm-referenced to a criterion-referenced foundation for testing, a desire to provide formative feedback for teachers and other stakeholders at a finer grain size than total test score, and a desire to understand what really matters in student performance and to design tests around that. Models similar to CDMs have been embedded in online tutoring systems for many years, and indeed broad psychometric interest in these models was sparked in part by a conference (and the corresponding edited volume of Nichols, Chipman, & Brennan, 1995) on cognitively diagnostic assessment (linked largely to online systems) that made great use of examples from the automated tutoring literature.

CDMs provide a different account of subjects' cognitive status – and responses to test items – than traditional IRT, although as has been observed elsewhere (e.g., Junker & Sijtsma, 2001b; Rupp & Templin, 2008; and von Davier, 2008) and as evident in (5.3) below, CDMs are in fact a form of IRT model. As with traditional IRT, CDMs can be built for dichotomous or polytomous data (e.g., von Davier, 2008), but this chapter will again concentrate on dichotomous item scores as in Fig. 5.1. Instead of a continuous latent unidimensional variable θ_i expressing an undifferentiated *quantity* of proficiency or knowledge, CDMs typically contemplate

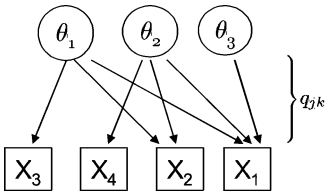

Fig. 5.1 A directed bipartite graph representation of the Q-matrix. A directed edge is drawn from skill k to item j if and only if $q_{jk} = 1$

a vector of dichotomous latent variables $\theta_l = (\theta_{i1}, \ldots, \theta_{iK})$ corresponding to K skills, knowledge components, or other cognitive features[2] needed to respond successfully to test items. Thus,

$$\theta_{ik} = \begin{cases} 1, & \text{if subject } i \text{ possesses skill } k \\ 0, & \text{otherwise} \end{cases}.$$

Thus, subject i is not placed on a continuum, but rather into one of 2^K latent classes labeled by $\theta_l = (\theta_{i1}, \ldots, \theta_{iK})$, indicating which skills the subject does or does not have.

In addition, because not all items require the same skills for successful response, CDMs employ a so-called Q-matrix (Barnes, 2005; Embretson, 1984; Tatsuoka, 1990), consisting of elements

$$q_{jk} = \begin{cases} 1, & \text{if subject } k \text{ possesses skill } j \\ 0, & \text{otherwise} \end{cases}.$$

One can think of Q as the adjacency matrix for a bipartite graph linking skills or cognitive attributes to items, as in Fig. 5.1. The likelihood for one subject's responses resembles the IRT likelihood in (5.2),

$$\begin{aligned} P[X_{i1} = x_{iJ}, \ldots, X_{iJ} = x_{iJ} | \theta_i, \beta_j, Q] \\ = \prod_{j=1}^{J} P(\theta_i, \beta_j, Q)^{x_{ij}} (1 - P(\theta_i, \beta_j, Q))^{1-x_{ij}}, \end{aligned} \tag{5.3}$$

except that, as indicated in the notation in (5.3), the response probability also depends on Q.

The variety of CDMs has grown considerably in recent years, as illustrated by the recent review of Rupp and Templin (2008), but for specificity this chapter will describe only one of the simpler models, the deterministic input, noisy *and* (DINA) model. Reviewed and named by Junker and Sijtsma (2001b), this model has antecedents going back to Embretson (1984) and Tatsuoka (1983) and is one of

[2] These features may encompass memorized facts, learned skills, higher-order concepts, and so on, but for brevity this chapter refers to them all as *skills*.

the first to be (re-)discovered whenever researchers want a simple conjunctive model (e.g., Pardos, Heffernan, Anderson, & Heffernan, 2006). The DINA model begins by combining the skill indicators for subject i and item j deterministically as

$$\xi_{ij} = \prod_{k=1}^{K} \theta_{ik}^{q_{jk}} = \begin{cases} 1 & \text{if } i \text{ has all skills for } j \\ 0 & \text{else} \end{cases}.$$

The 0/1 variable ξ_{ij} would be the *ideal response* to the item, if subjects' responses perfectly reflected the pattern of skills, relevant to the item, that they did and did not possess: If the subject possesses all the skills relevant to an item, the ideal response is 1 (success); otherwise, it is 0 (failure). Statistical error (inconsistent response, uncertainty, or even Q-matrix misspecification) is modeled in the probability of correct response as

$$P(\theta_i, \beta_j, Q) = (1 - s_j)^{\xi_{ij}} \, g_j^{1-\xi_{ij}} . \tag{5.4}$$

Here, $\beta_j = (s_j, g_j)$: the parameter s_j may be interpreted as the probability of *slipping*, given that the subject possesses all of the skills needed for the item; similarly g_j may be interpreted as the probability of getting the problem right by other means (e.g., guessing), given that some skills are missing. As long as $1 - s_j > g_j$, the model is *conjunctive*, in the sense that the subject needs all the skills associated with the item to have a high probability of positive response; otherwise, the probability of positive response is low.

The inferential tasks faced with CDMs are essentially the same as with IRT models. The first task is to estimate item parameters (e.g., the guess and slip parameters) in order to assess the quality of the items (the difference between $1-s$ and g indicates how well the item discriminates between subjects who do and do not possess the relevant skills, for example). Given a calibrated model (known Q-matrix, well-estimated s's and g's), the second task is to make inferences on the θ_{ik}'s – which skills or cognitive attributes do students possess (or not possess)? Current methods for estimating item parameters include marginal maximum likelihood (e.g., de la Torre, 2008; Templin, 2009; von Davier, 2008, and the references therein) and fully Bayesian methods (e.g., Hartz, 2002; Junker & Sijtsma, 2001b).

However, parametric estimation of CDMs becomes difficult as the size of the data, and/or the number of skills, grows. Published examples (e.g., de la Torre, 2008; Templin, Henson, Templin, & Roussos, 2008) tend to involve as few as two to four skills; when the number of skills grows, current estimation methods slow considerably (though von Davier, 2008, uses ML methods to estimate models with up to eight skills). The M-step of a straightforward E-M algorithm typically has to visit each of the 2^K latent classes labeled by the K dimensional binary vector θ_i, so that the E-M algorithm may be fast for few skills but slows exponentially as K grows. Markov chain Monte Carlo (MCMC; e.g., Gelman, Carlin, Stern, & Rubin, 2004,

Chap. 11) algorithms do not visit θ vectors with relatively low probability[3] but are still slow; Anozie and Junker's (2007) MCMC algorithm took 1 h per 100 steps for estimating a DINA model on approximately 300 items using approximately 100 skills[4] for approximately 600 students using data from the Assistments project (Junker, 2007; Razzaq et al., 2005) in which students typically answered 20–40 items each.

5.4 Nonparametric CDM

To my knowledge, a broad nonparametric theory for CDMs has yet to be developed, but nonparametric approaches to CDMs are beginning to take shape. This chapter considers two such approaches, both motivated by computational challenges in parametric inference with CDMs as the numbers of examinees, items, and/or skills grow.

5.4.1 Clustering to Make Inferences About Subjects' Skill Vectors

Henson, Templin, and Douglas (2007) observed that given a Q-matrix, the observed sum-scores

$$W_{ik} = \sum_{\{j:i \text{ answered } j\}} x_{ij} q_{jk}$$

are informative about subjects' skills, under a conjunctive model. (These sum-scores are like subscores in many operational tests, except that the W_{ik} can share items scores and most subscores are based on disjoint subsets of items. They are partly justified in this case because they are *complete conditional sufficient statistics for* the probability that student i knows skill k under the noisy input, deterministic, *and* gate (NIDA) model; see Junker & Sijtsma, 2001b.) Henson et al. (2007) investigated the use of cutoff scores using weighted and unweighted averages

[3] The E-M algorithm can be customized to prune out latent classes with low probability and thus exhibit similar behavior, once it is working near the *final* mode. Moreover E-M can be sped up to some extent with variational methods; see for example Minka (2009). These methods generally increase the size of the data space and/or latent space that can be dealt with, but they do not eliminate the computational explosion completely.

[4] The system designers had identified 126 skills of interest; this sample of items did not exercise all skills. A drawback of this analysis was that dependence between skills was not modeled. As de la Torre and Douglas (2004, Tables 9 and 10) suggested, however, ignoring dependence between skills may have minimal impact on estimates of item (slip and guess) parameters, and much greater impact on classifying students as masters or nonmasters of particular skills.

based on the W_{ik}'s in making inferences about subjects' corresponding latent skill indicators θ_{ik}.

Chiu (2008) investigated, theoretically and empirically, clustering the sum score vectors

$$W_i = (W_{i1}, \ldots, W_{iK})$$

using K-means and hierarchical clustering (Mardia, Kent, & Bibby, 1980), to try to reproduce the 2^K latent classes implied by the 0/1 skill vectors θ_i. In what is perhaps the first theoretical result in nonparametric CDM, Chiu (2008) showed, under suitable technical conditions, that as long as a nonvanishing proportion of single-skill items exists for each skill, then as J grows, all 2^K latent classes can be recovered.

Ayers, Nugent, and Dean (2008) worked instead with the normalized scores

$$B_{ik} = \frac{\sum_{\{j:\, i \text{ answered j}\}} x_{ij} q_{jk}}{\sum_{\{j:\, i \text{ answered j}\}} q_{jk}}, \tag{5.5}$$

which they called *capability scores*. They found, empirically, that clustering the capability vectors

$$B_i = (B_{i1}, \ldots, B_{iK}),$$

rather than clustering the W_i's, generally produces skill estimates closer to those of a fitted DINA, especially when data are missing. The normalization in (5.5) accommodates different numbers of questions for different subjects, and reduces the influence in the clustering of skills that appear in many items.

Clustering is, of course, also much faster. In a typical example, Ayers et al. (2008) found that for ten skills, estimating the skills by fitting the DINA model to the data using WinBugs (Lunn, Thomas, Best, & Spiegelhalter, 2000) takes approximately 1 day; estimating by fitting DINA using an E-M algorithm (de la Torre, 2008) takes approximately 15 min, and clustering capability vectors takes approximately 2 s.

To speed up clustering even more, in anticipation of much larger data sets such as might be encountered in data mining logs from online tutoring systems, Nugent, Ayers, and Dean (2009) developed a *bump-hunting* (Good & Gaskins, 1980) algorithm that Nugent et al. (2009) called *conditional subspace clustering*, which first seeks individual dimensions of B_i on which subjects can be separated into high-density clusters with low-density valleys between them. One-dimensional bump-hunting can be performed alone or as preprocessing to other clustering methods to speed up inferences about groups of subjects with similar skill profiles.

The three-dimensional scatter plot in Fig. 5.2 shows the capability vectors for three skills (for ease of visualization) for a sample of students in the Assistment project, plotted in the unit (hyper-)cube (the corners of this cube are the labels for the 2^K latent classes). The bump-hunting algorithm identifies three well separated high-density clusters in the evaluate functions dimension. Apparent clustering in

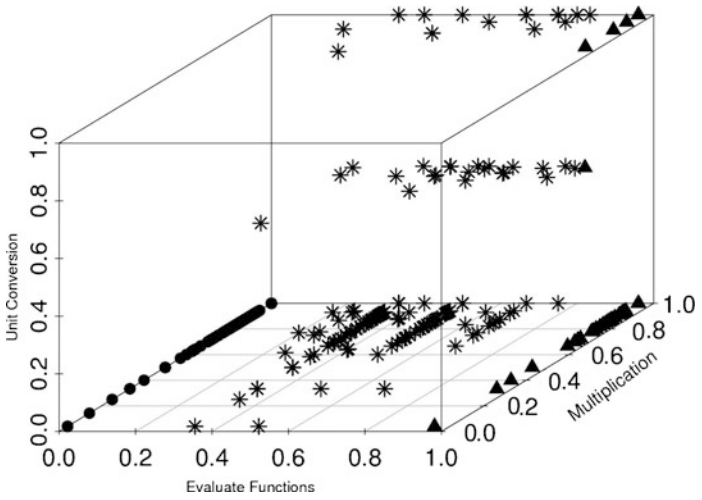

Fig. 5.2 Illustration of the Nugent et al. (2009) conditional subspace clustering algorithm. This scatter plot of the capability scores in the (hyper-)cube is defined by the skills being measured, with clusters discovered by "bump-hunting" identified by different shapes: disks, asterisks, and triangles

the unit conversion dimension is not identified by the algorithm because the secondary modes contain relatively few subjects; multiplication shows little bump-and-valley structure at all. Nugent et al. (2009) conjectured that clusters identified by the bump-hunting alone will often be instructionally relevant.[5] If additional clustering is desired, it can be carried out in the (complementary) subspaces conditional on the clusters identified by bump-hunting, as illustrated by the dendrogram in Fig. 5.3. This dendrogram was produced using a minimum-density linkage method that is particularly useful for visualizing high dimensional model-based (mixture of normals or other densities) clustering. The same symbols are used in Fig. 5.3 as in Fig. 5.2 to distinguish the bump-hunting clusters; one can see that an additional multiskill structure is present within each bump; see Nugent et al. (2009) for details.

5.4.2 *Using Observed Associations to Discover Skills and Q-Matrix Structure*

Of course, no method of estimating skills will be useful if we have the wrong set of skills for the items, or if the Q-matrix is not specified correctly. Some insight can be gained into the problem of *discovering* skills and/or Q-matrix structure by

[5] This of course depends on having a Q-matrix, among the many that may represent the data (e.g., (5.6) and subsequent discussion, as well as Maris & Bechger, 2009), that is itself instructionally relevant.

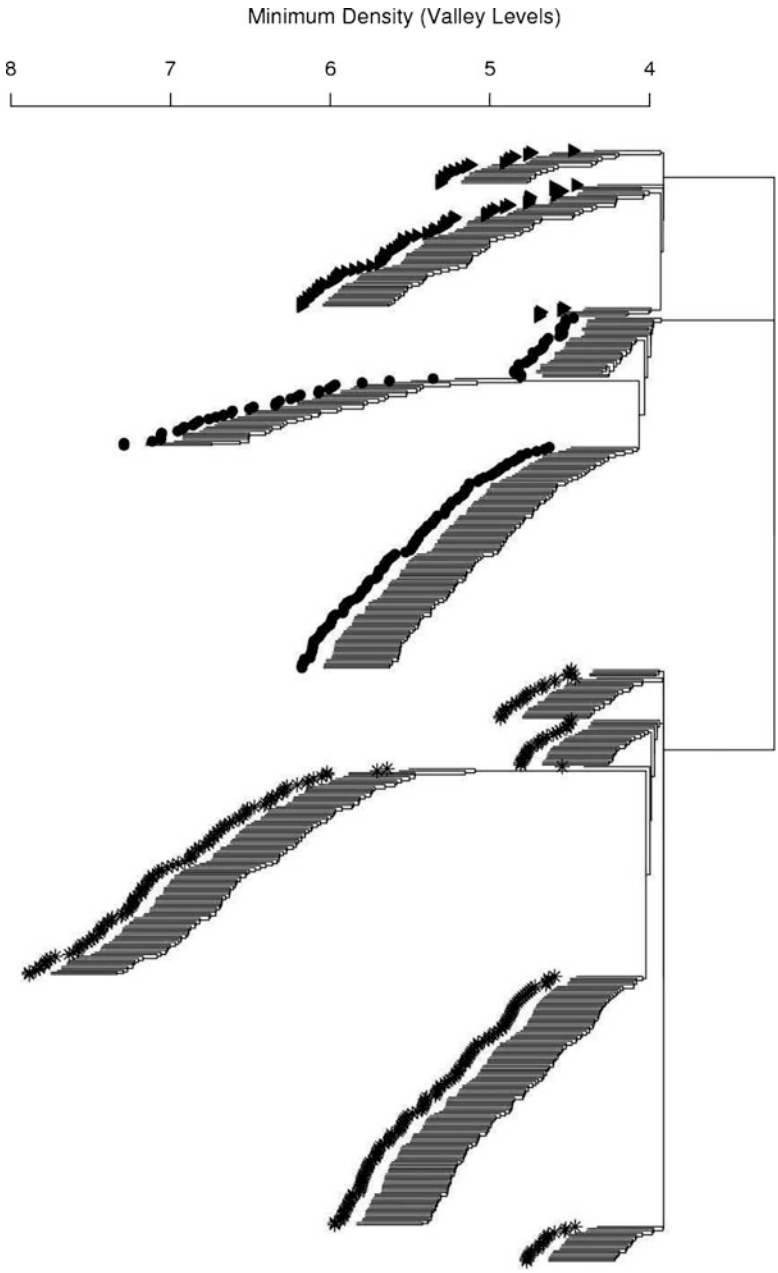

Fig. 5.3 Illustration of the Nugent et al. (2009) conditional subspace clustering algorithm. This dendrogram shows further model-based clustering within each of the subspaces defined by the one-dimensional clusters in Fig. 5.2

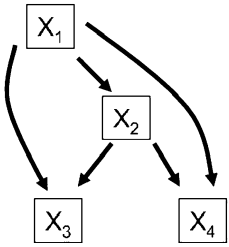

Fig. 5.4 Surmise relationships among items. We surmise positive performance on item b from positive performance on item a, if and only if a directed edge points from X_a to X_b ($a,b \in \{1,\ldots,J\}$)

considering the related problem of predicting performance on one item from performance on others, given a conjunctive Q-matrix structure.

For example, consider the Q-matrix

$$Q = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

illustrated in Fig. 5.1. We can surmise from the Q-matrix (or equivalently from Fig. 5.1) that if a subject responds positively to Item 1, the subject *knows* all three skills and should respond positively to all the other items (this is only a surmise, not a strict inference, to the extent that a probabilistic model like that of (5.4) may be in play). If the subject responds positively to Item 2, then we can surmise positive responses for Items 3 and 4, but we will be unsure of Item 1. Drawing an arrow to indicate each surmise here, we arrive at the directed graph in Fig. 5.4, which can also be represented by the incidence matrix

$$R = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $R_{ab} = 1$ if there is a directed edge from X_a to X_b and $R_{ab} = 1$ otherwise. Using ideas reviewed in van Mechelen, Lombardi, and Ceulemans (2007), we can also construct R from Q algebraically, as

$$R = (Q^c \circ Q^T)^c, \tag{5.6}$$

where Q^c is the element-wise complement of the Q-matrix (i.e., $q^c_{jk} = 1 - q_{jk}$), Q^T is the transpose of Q, and $\circ$ stands for Boolean matrix multiplication: In the dot product of each row with each column, "$\times$" is replaced with logical "and", and "+" is replaces with logical "or". If there were no noise in data, then R would be a kind of partial Guttman ordering: Given success on a subset of the items, we could use R to predict perfectly which other items a subject would succeed at. The graph in Fig. 5.4 or equivalently its incidence matrix R is an example of a *partially ordered knowledge structure* (POKS; Desmarais & Pu, 2005), a notion closely related to the *knowledge spaces* of Doignon and Falmagne (1999).

Tucker (2009) divided the task of discovering skills and Q-matrix structure into the following two subtasks:

Subtask I. Using raw performance data

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NJ} \end{bmatrix}$$

to estimate a surmise graph (incidence matrix) $\hat{R}$.

Subtask II. Factoring $\hat{R}$ as in (5.6).

For Subtask I, Desmarais and Pu (2005) observed that $X_a \rightarrow X_b$ implies

$$P[X_b = 1 | X_a = 1] \approx 1$$

and

$$P[X_a = 1 | X_b = 0] \approx 0,$$

and that these conditions can be tested from the table

	X_b	
X_a	n_{00}	n_{01}
	n_{10}	n_{11}

,

derived from X. In particular, we expect n_{10} to be small, relative to both n_{00} and n_{11}, if we can surmise success on X_b from success on X_a. Thus pairwise statistical tests can generate an estimated graph of surmise relationships, $\hat{R}$. This is Step 1 in Fig. 5.6.

An example of such an estimated surmise graph from Tucker's (2009) work is shown in Fig. 5.5, estimated from data for $N = 1{,}000$ simulated examinees, generated from a Q-matrix with $K = 10$ skills and $J = 20$ items; the true R matrix has 39 edges. Items are marked with their role in the original Q-matrix; for example, the item marked "m123" is a multiple skill item depending on (simulated) Skills 1, 2, and 3, while "s2.3" is the second single-skill item depending on Skill 3 only. The ratio of $J = 20$ item to $K = 10$ skills is likely too low to obtain stable inferences about all edges in R; further exploration of this approach using higher items-to-skills ratios is currently underway.

For Subtask II, it is always possible to set $Q = R^T$, but this produces a Q-matrix with a maximal number, J, of skills. On the other hand, finding the Q-matrix that factors R with the minimal number of skills is NP-hard (Leenen, van Mechelen, & DeBoeck, 1999). Tucker (2009) proposed a faster heuristic algorithm, which she conjectures will produce useful Q-matrixes for conjunctive CDMs with relatively sparse surmise graphs (numbered to match the corresponding steps in Fig. 5.6):

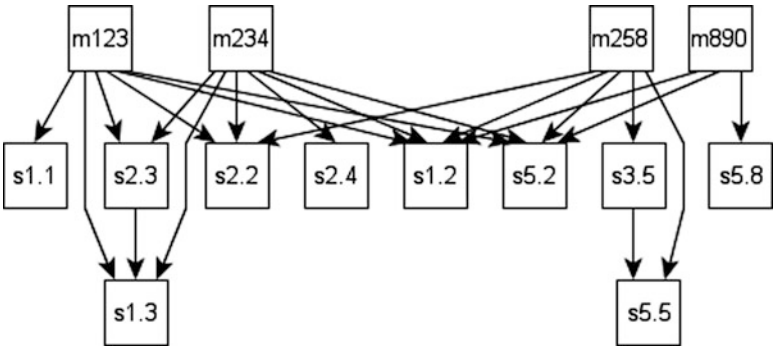

Fig. 5.5 Part of an estimated surmise graph among items created with the partially ordered knowledge structure (POKS) algorithm (Tucker, 2009). The item labels are motivated from an application in the Assistments (Junker, 2007; Razzaq et al., 2005) system; labels beginning with *m* indicate multiple-skill items, and labels beginning with *s* indicate single skill items

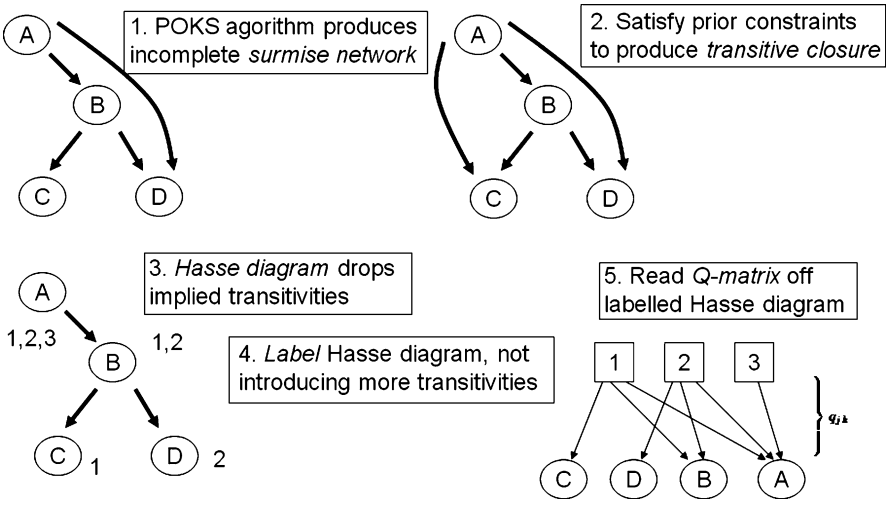

Fig. 5.6 Schematic of Tucker's (2009) skills discovery algorithm, starting from an estimated surmise graph among items. *POKS* partially ordered knowledge structure

(2) Produce the *transitive closure* of the estimated $\hat{R}$: For every pair of edges $X_a \rightarrow X_b$ and $X_b \rightarrow X_c$ in $\hat{R}$, ensure that $\hat{R}$ also has the edge $X_a \rightarrow X_c$. (3) Compute the *Hasse diagram* of $\hat{R}$, that is, the smallest directed graph $\hat{H}$ with $\hat{R}$ as its transitive closure. (4) Label the leaves of $\hat{H}$ with unique skills. For every parent node in $\hat{H}$, label it with the union of the skill labels of the child nodes, plus additional skills if needed to preserve non-transitivities. (5) Read the Q-matrix off the labeled Hasse diagram.

Tucker (2009) conducted a simulation study examining recovery of R from data, exploring both the effects of various levels of guess and slip parameters in the

DINA and related models, as well as the operating characteristics of binomial tests used to compare n_{10} with n_{00} and n_{11} in the table in (5.7). The next step in this work is to explore the recovery of the Hasse diagram H with $\hat{H}$, since H is the part of the true Q-matrix that is identifiable from the data (many Q-matrixes can be used to represent the distribution of the data; see (5.6), or more broadly Maris & Bechger, 2009) using the heuristic algorithm explained previously. Also important will be to modify the labeling algorithm to identify, among the many possible Q-matrixes for a particular example, one or more that are instructionally relevant.

5.5 Discussion

Nonparametric IRT has a relatively long history; some parts of it are closely related to modern nonparametric methods in statistics generally (e.g., Rossi, Wang, & Ramsay, 2002) but by and large nonparametric IRT has referred to methodology for (a) understanding the operating characteristics of IRT models generally and (b) developing formal and informal statistical tests for general IRT models without regard to parametric form. These methods, which this chapter reviewed briefly, have been most useful in situations where parametric model fitting is inconvenient, either because too little data or too much data exist.

Cognitive diagnosis models (CDMs) have recently attracted widespread attention among statistical psychometricians. Although parametric CDMs have been around for decades, and have been the object of relatively intense study in the past decade or so, parametric CDM methodology seems far from producing professional testing examples that would satisfy the same standards of reliability, validity, and distinctness (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) that high-stakes standardized tests must satisfy (Haberman & von Davier, 2007; Sinharay & Haberman, 2008). At the same time, CDM-like structures have been successfully used for years in online tutoring and related systems, going back at least to Nichols et al. (1995), where the social, educational and legal costs of lower reliability are not as keenly felt (e.g., as discussed by Junker, 1999). Part of the problem may be the unlikely hope that applying CDM methods to item response data generated by professional test developers to satisfy unidimensional IRT design constraints would produce rich fine-grained multidimensional latent structure (Luecht, Gierl, Tan, & Huff, 2006).

But it may also be that researchers do not yet understand the fundamental structure of CDMs; studies of the features of CDMs, apart from particular parameterizations, are not common, unlike nonparametric IRT. And, especially as data sets become large and latent skills models more complex, parametric CDMs also suffer computational efficiency problems. Thus a need for nonparametric CDM thinking exists,

both to understand the operating characteristics of CDMs and to assist with large-scale analyses.

Some specific nonparametric approaches to CDMs are beginning to take shape. This chapter considered two such new approaches to conjunctive CDMs. In one approach (Sect. 5.4.1 above; Ayers et al., 2008; Chiu, 2008; Nugent et al., 2009), given the assumption of conjunctive structure and a valid Q-matrix, cluster analysis is applied to identify groups of subjects (examinees, respondents) with similar patterns of cognitive attributes or skills. In another approach (Sect. 5.4.2; Desmarais & Pu, 2005; Tucker, 2009), observed association structure between items is first mined to discover surmise – or equivalently prerequisite – relations among items, and then these relations are factored to produce possible Q-matrix structure.

Both approaches are in their infancy, but they point to the possible significant advantages of nonparametric approaches. They exploit simple and interpretable data summaries that can be computed even when data sets become large, and they begin to suggest some of the operating characteristics of CDM models generally.

Acknowledgement Parts of this work were supported by US Department of Education Grants #R305K030140 and #R305B04063 and by National Science Foundation Award #DMS-0240019.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anozie, N., & Junker, B. W. (2007, April). *Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Ayers, E., Nugent, R., & Dean, N. (2008, June). Skill set profile clustering based on student capability vectors computed from online tutoring data. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: 1st International Conference on Educational Data Mining, Proceedings* (pp. 210–217). Retrieved from http://www.educationaldatamining.org/EDM2008/.

Barnes, T. (2005). Q-matrix method: Mining student response data for knowledge. In J. E. Beck (Program Chair), *Proceedings of the AAAI-05 Workshop on Educational Data Mining* (AAAI Technical Report #WS-05-02, pp. 39–46). Pittsburgh, PA: AAAI Press.

Bartolucci, F., & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70*, 31–43.

Chiu, C. (2008). *Cluster analysis for cognitive diagnosis: Theory and applications.* Unpublished doctoral dissertation, Department of Educational Psychology, University of Illinois at Urbana Champaign.

De la Torre, J. (2008). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

De la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

DeBoeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

Desmarais, M. C., & Pu, X. (2005). A Bayesian inference adaptive testing framework and its comparison with item response theory. *International Journal of Artificial Intelligence in Education, 15*, 291–323.

Doignon, J. P., & Falmagne, J. Cl. (1999). *Knowledge spaces*. New York, NY: Springer-Verlag.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62*, 7–28.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62*, 495–523.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, England: Chapman and Hall.

Good, I. J., & Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association, 75*, 42–56.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5*, 815–841.

Haberman, S. J., & von Davier, M. (2007). A note on models for cognitive diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Psychometrics, Vol. 26, pp. 1031–1038). Amsterdam, The Netherlands: Elsevier.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana Champaign.

Henson, J., Templin, R., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skill diagnoses. *Journal of Education Measurement, 44*, 361–376.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79–92.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577–601.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523–1543.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*, 77–79.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer-Verlag.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics, 21*, 1359–1378.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively relevant assessment*. Retrieved from http://www.stat.cmu.edu/~brian/nrc/cfa/.

Junker, B. W. (2007). Using on-line tutoring records to predict end-of-year exam scores: Experience with the ASSISTments project and MCAS 8th grade mathematics. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard settings*. Maple Grove, MN: JAM Press.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics, 25*, 1327–1343.

Junker, B. W., & Sijtsma, K. (2001a). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220.

Junker, B. W., & Sijtsma, K. (2001b). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Leenen, I., van Mechelen, I., & DeBoeck, P. (1999). A generic disjunctive/conjunctive decomposition model for n-ary relations. *Journal of Mathematical Psychology, 43*, 102–122.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 51*(Serial No. 285).

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1980). *Multivariate analysis*. New York, NY: Academic Press.

Maris, G., & Bechger, T. (2009). Equivalent diagnostic classification models. *Measurement: Interdisciplinary Research & Perspective, 7*, 41–46.

McCulloch, G. C., & Searl, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30*, 419–440.

Minka, T. (2009, July). *Automating variational inference for statistics and data mining*. Invited presentation at the 74th annual meeting of the Psychometric Society. Abstract retrieved from http://www.thepsychometricscentre.com/.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nugent, R., Ayers, E., & Dean, N. (2009, July). Conditional subspace clustering of skill mastery: Identifying skills that separate students. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining: Proceedings* (pp. 101–110). Retrieved from http://www.educationaldatamining.org/EDM2009/.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2006). *Using fine grained skill models to fit student performance with Bayesian networks*. Paper presented at the workshop in educational data mining, 8th international conference on intelligent tutoring systems, Jhongli, Taiwan.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Pelligrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expandedth ed.). Chicago, IL: University of Chicago Press.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). The Assistment project: Blending assessment and assisting. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th Artificial Intelligence in Education* (pp. 555–562). Amsterdam, The Netherlands: ISO Press.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences, 27*, 291–317.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic models: A review of the current state-of-the-art. *Measurement, 6*, 219–262.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*, 75–102.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response modeling*. Thousand Oaks, CA: Sage Publications.

Sinharay, S., & Haberman, S. J. (2008). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives, 7*, 46–49.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York, NY: Springer-Verlag.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 29*, 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Templin, J. (2009, April). *Estimation of diagnostic models with Mplus*. Presentation as part of a Diagnostic Models training session, National Council on Measurement in Education, San Diego, CA.

Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of unidimensional hierarchical modeling of discrete attribute association in cognitive diagnosis models. *Applied Psychological Measurement, 32*, 559–574.

Tucker, E. (2009). *Discovering partially ordered knowledge structures from student response data*. Unpublished senior honors thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283–304.

Van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Van Mechelen, I., Lombardi, L., & Ceulemans, E. (2007). Hierarchical classes modeling of rating data. *Psychometrika, 72*, 475–488.

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

Yuan, A., & Clarke, B. (2001). Manifest characterization and testing for certain latent properties. *Annals of Statistics, 29*, 876–898.

Chapter 6
What Aspects of the Design of an Observational Study Affect Its Sensitivity to Bias from Covariates That Were Not Observed?

Paul R. Rosenbaum

6.1 Introduction and Example

6.1.1 What Is Design Sensitivity?

In discussing observational or nonrandomized studies of treatment effects in his president's address to the Royal Society of Medicine, Austin Bradford Hill (1965) asked:

> Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation? (p. 295)

If Hill were correct, if there were actually aspects of association pertinent to judging evidence about causation, then a natural goal for the design of observational studies would be to ensure that these aspects are present in a decisive form. Hill proposed certain specific aspects to consider, as have others (e.g., Campbell, 1957, 1988; Meyer, 1995; Reynolds & West, 1987; Rutter, 2007; Shadish, Cook, & Campbell, 2002; Trochim, 1985; Vandenbroucke, 2004; Weed, 1997; Weiss, 1981, 2002). These proposals are useful and widely used, but they have been developed in an informal manner, with the consequence that it is often difficult to appraise the precise nature of the evidence provided, its usefulness and limitations, the quantitative magnitude of the evidence, the relative importance of evidence of different types, and the attending circumstances needed to ensure its validity.

A formal tool for thinking about issues of this sort is the design sensitivity (Rosenbaum, 2004). Where a *sensitivity analysis* is a statistical analysis of certain type performed on data from a particular study, the *design sensitivity* is a number

P.R. Rosenbaum (✉)
Department of Statistics, The Wharton School, University of Pennsylvania,
473 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104–6340, USA
e-mail: rosenbaum@wharton.upenn.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_6,

that evaluates the design of an observational study, that is, a particular data generating process and planned protocol for analysis. A sensitivity analysis asks: How far would a particular observational study have to depart from an analogous randomized experiment to materially alter the conclusions about treatment effects? In an observational study, if the treatment had been effective, and there was, in fact, no bias distorting the study's conclusions, then one could not be certain of this from the observed data; rather, at best, one might be able to report that the study's conclusions are insensitive to small and moderate biases. The design sensitivity is a number, $\widetilde{\Gamma}$, that anticipates the outcome of a sensitivity analysis. In this sense, the design sensitivity resembles the power of a test of a statistical hypothesis: It anticipates the results of analysis that will be performed when the data become available. A stronger design, one with a larger design sensitivity, $\widetilde{\Gamma}$, is expected to be less sensitive to unobserved biases if the treatment is effective and biases are absent. In this sense, the design sensitivity is a basis for appraising competing designs for observational studies.

Focusing on the simple special case of matched pairs, the concept of design sensitivity is reviewed and extended. Much of the material in the review is drawn from Rosenbaum (1997, 2003a, 2004, 2005, 2007), Small and Rosenbaum (2008), and Heller, Rosenbaum, and Small (2009), but some results are new.

After reviewing in Sects. 6.2.1 and 6.2.2 notation for treatment effects and randomization inference in randomized experiments, Sect. 6.2.3 discusses sensitivity analysis in observational studies, and Sect. 6.2.5 anticipates the results of a sensitivity analysis using the power of a sensitivity analysis and the related concept of design sensitivity. Later sections discuss factors that affect design sensitivity, including unit heterogeneity in Sect. 6.3, dose–response in Sect. 6.4, coherence among several outcomes in Sect. 6.5, and situations in which only a small part of the population is affected by treatment in Sect. 6.6. The example in Sect. 6.1.2 is used to illustrate several ideas: sensitivity analysis in Sect. 6.2.4, graduated dose–response in Sect. 6.4.2, extreme doses in Sect. 6.4.4, and coherence in Sect. 6.5.

6.1.2 Example: Genetic Damage Among Professional Painters

At several points, the following example will be used as an illustration: Paint and paint thinners contain several hazardous components, including lead and organic solvents, which may cause genetic damage. Pinto et al. (2000) compared male professional public building painters, working without masks or gloves, in Merida, Yucatan, Mexico, to male clerks, matched for age.[1] Pinto et al. examined several standard measures in genetic toxicology, including the frequency of micronuclei per 1,000 cells found in 3,000 oral epithelial cells gently scraped from the cheek of

[1] In Table 2 of Pinto et al. (2000), the identification numbers for the 22 pairs (painter, control) are (25, 48), (22, 50), (23, 47), (12, 44), (13, 45), (11, 42), (20, 43), (19, 41), (18, 39), (17, 38), (16, 37), (6, 36), (9, 33), (15, 34), (5, 35), (8, 32), (7, 31), (4, 30), (14, 29), (2, 28), (3, 26), (1, 27).

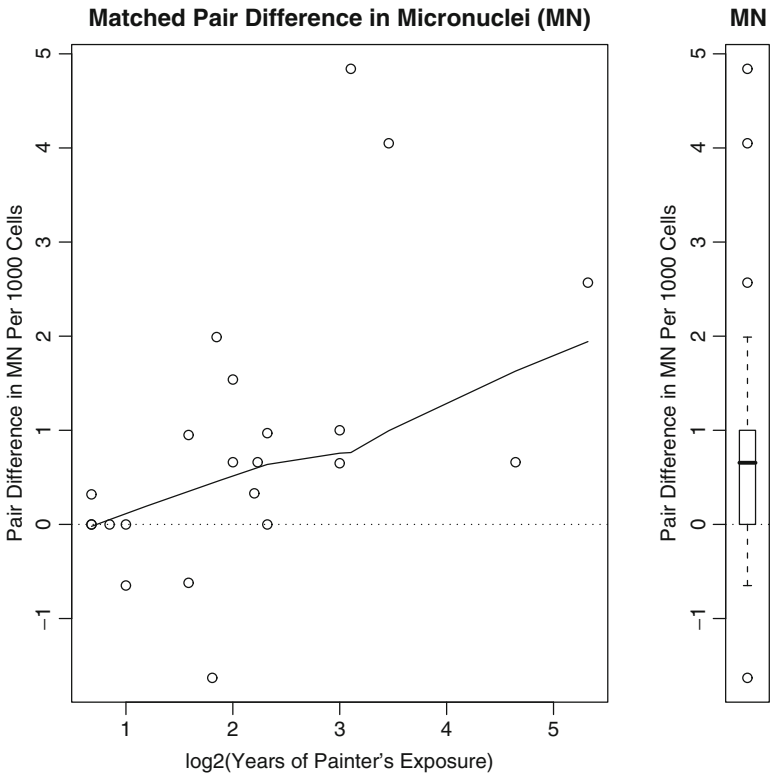

Fig. 6.1 Plot of 22 matched pair differences, painter-minus-control, in micronuclei frequency per 1,000 cells (MN), plotted against the $\log_2$ of years of exposure for the painter. There are also a marginal boxplot of the 22 differences in MN and a lowess smooth in the scatterplot

each individual. For 22 matched pairs, Fig. 6.1 plots the painter-minus-control difference in micronuclei against duration of exposure for the painter, measured as $log_2(years)$, so 2, 8, or 32 years of work as a painter corresponds with $log_2(2) = 1$, $log_2(8) = 3$, and $log_2(32) = 5$, as $log_2\left(2^k\right) = k$. Matching on age is important here, because a painter cannot have worked for 30 years if he is only 20 years old. The figure includes a marginal boxplot of the differences and a lowess smooth; see Cleveland (1994) for discussion of both the boxplot and the lowess smooth.

In Fig. 6.1, the matched pair differences in micronuclei tend to be positive, and they appear to be larger in pairs in which the painter has worked as a painter for a longer time. If Wilcoxon's signed rank test is applied to the matched pair differences, the one-sided significance level is 0.0032 and the associated Hodges–Lehmann (HL) point estimate of an additive effect is 0.645; see Lehmann (1998) for discussion of these standard statistical methods. These inferences would be appropriate in a randomized experiment, but the data in Fig. 6.1 are not from such an experiment.

6.2 Design Sensitivity for Matched Observational Studies

6.2.1 Notation for Treatment Effects and Treatment Assignments

Design sensitivity is a general concept, but it is most straightforwardly illustrated in the case of I matched pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, denoted $Z_{ij} = 1$, the other control, denoted $Z_{ij} = 0$, matched exactly for observed covariates $\mathbf{x}_{ij}$, so that $1 = Z_{i1} + Z_{i2}$ and $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i. In addition, in pair i, the treatment is applied to the treated subject at dose $d_i > 0$; however, unequal doses appear only in Sect. 6.4, so that in other sections $d_i = 1$ for $i = 1, \ldots, I$. Inevitably in an observational study, concern arises that matching may have failed to control a relevant covariate u_{ij} that was not observed, so that $u_{i1} \neq u_{i2}$. Each subject ij has two potential responses, $\left(r_{Tij}, r_{Cij}\right)$, where response r_{Cij} is observed if subject ij is assigned to control, $Z_{ij} = 0$, and r_{Tij} is observed if subject ij is assigned to treatment, $Z_{ij} = 1$, so subject ij exhibits response $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$, and the effect of the treatment, namely $r_{Tij} - r_{Cij}$, is not observed for any subject ij; see Neyman (1923) and Rubin (1974). Write $\mathbf{Z} = (Z_{11}, \ldots, Z_{I2})^T$, $\mathbf{R} = (R_{11}, \ldots, R_{I2})^T$, $\mathbf{r}_C = (r_{C11}, \ldots, r_{CI2})^T$, $\mathbf{r}_T = (r_{T11}, \ldots, r_{TI2})^T$ and $\mathbf{u} = (u_{11}, \ldots, u_{I2})^T$ for the $2I$-dimensional vectors. In a randomized experiment, Fisher's (1935) randomization test concerned the sharp null hypothesis of no treatment effect, which says that each subject is unaffected by treatment, $H_0 : r_{Tij} = r_{Cij},\ \forall ij$ or $H_0 : \mathbf{r}_C = \mathbf{r}_T$. Generally, the observed response, $\mathbf{R}$, changes with the treatment assignment, $\mathbf{Z}$, but if the null hypothesis H_0 of no effect is true, then $\mathbf{R} = \mathbf{r}_C$ does not change when $\mathbf{Z}$ changes.

Define $F = \left\{\left(r_{Tij}, r_{Cij}, d_i, \mathbf{x}_{ij}, u_{ij}\right), i = 1, \ldots, I, j = 1, 2\right\}$. Also define $\mathcal{Z}$ to be the set containing the 2^I possible values $\mathbf{z}$ of the treatment assignment $\mathbf{Z}$, so that $\mathbf{z} \in \mathcal{Z}$ if each z_{ij} is 0 or 1 and $z_{i1} + z_{i2} = 1$ for each i. For a finite set S, the number of elements of S is denoted $|S|$, so $|\mathcal{Z}| = 2^I$.

6.2.2 Randomization Distributions in Randomized Experiments

In a matched pair experiment, randomization ensures that $Pr(\mathbf{Z} = \mathbf{z} \,|\, F) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$. In a randomized experiment, under the null hypothesis of no effect, H_0, Fisher (1935) showed that any test statistic, $t(\mathbf{Z}, \mathbf{R})$, has a null distribution created by the randomization, specifically:

$$Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \,|\, F\} = Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \,|\, F\} = \frac{\left|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}\right|}{2^I}, \qquad (6.1)$$

because under H_0, $\mathbf{R} = \mathbf{r}_C$ is fixed by conditioning on F and $Pr(\mathbf{Z} = \mathbf{z} \,|\, F) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$.

In the discussion here, $t(\mathbf{Z}, \mathbf{R})$ will be Wilcoxon's signed rank statistic or some variant of this statistic, and ties of all kinds are assumed not to occur among the responses R_{ij}. Write Y_i for the treated-minus-control difference in observed responses in pair i, so $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, and write $sgn(w) = 0$ or 1 as $w \leq 0$ or $w > 0$, so that $sgn(Y_i) = 1$ if the treated subject in pair i had the higher response. Define two I-dimensional vectors, $\mathbf{A} = (|Y_1|, \ldots, |Y_I|)^T$ for the absolute differences, and $\mathbf{d} = (d_1, \ldots, d_I)^T$ for the doses. Let $q_i = q_i(\mathbf{A}, \mathbf{d}) \geq 0$ be a score for the i^{th} pair determined by $\mathbf{A}$ and $\mathbf{d}$; for instance, for Wilcoxon's signed rank statistic, $q_i = q_i(\mathbf{A}, \mathbf{d})$ is the rank of $|Y_i|$ among the $|Y_1|, \ldots, |Y_I|$. The statistics considered here are of the form $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$, which of course includes Wilcoxon's signed rank, among many others. Under the null hypothesis H_0 of no treatment effect, $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$ where $Z_{i1} - Z_{i2}$ is 1 or -1 and $|Y_i| = |r_{Ci1} - r_{Ci2}|$, so $\mathbf{A}$ and $\mathbf{d}$ are fixed in (6.1) by conditioning on F. Under H_0 in a randomized experiment, moreover, $sgn(Y_i) = 1$ or 0 each with probability $\frac{1}{2}$ independently for distinct i, and the null distribution of (6.1) of $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$ is the distribution of the sum of I independent random variables, $i = 1, \ldots, I$, taking values 0 and q_i each with probability $\frac{1}{2}$.

Associated with a Wilcoxon's statistic and its variants is a point estimate due to Hodges and Lehmann (1963). In its most familiar form, it is an estimate of an additive treatment effect, τ, so that $r_{Tij} - r_{Cij} = \tau$ for all i, j, with the consequence that $\mathbf{R} = \mathbf{r}_C + \tau \mathbf{Z}$ and $Y_i = \tau + (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \tau + (Z_{i1} - Z_{i2})\varepsilon_i$, say. If the q_i are some permutation of a fixed set of ranks, as is true for Wilcoxon's signed rank statistic, then with an additive effect, τ, in a randomized experiment, the expectation $E\{t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\} = E\left\{\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i\right\} = (1/2)\sum_{i=1}^{I} q_i$ is known, and the Hodges–Lehmann (HL) estimate is, in effect, the solution $\widehat{\tau}$ to the estimating equation $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i = (1/2)\sum_{i=1}^{I} q_i$ where q_i is the rank of $|Y_i - \tau|$. (Actually, Wilcoxon's $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i$ is a decreasing step function as τ increases, with many small steps for large I, so the solution $\widehat{\tau}$ is defined to be either the unique point at which $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i$ passes $(1/2)\sum_{i=1}^{I} q_i$ or midpoint of the interval on which $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i$ equals $(1/2)\sum_{i=1}^{I} q_i$).

6.2.3 *Sensitivity Analysis in Observational Studies*

In an observational study, treatments are not assigned at random, so there may be little basis for believing that $Pr(\mathbf{Z} = \mathbf{z} \,|\, F) = 2^{-I}$. A sensitivity analysis considers departures from random assignment of various magnitudes along with their impact on inferences about treatment effects. How much would significance levels, point estimates, or confidence intervals change if departures of a specified magnitude were made from $Pr(\mathbf{Z} = \mathbf{z} \,|\, F) = 2^{-I}$? A simple model for sensitivity analysis assumes that, in the population prior to matching, subjects are assigned to treatment independently with unknown probabilities $\pi_{ij} = Pr\left(Z_{ij} = 1 \,\middle|\, F\right)$ such that two

subjects, ij and ij', with the same observed covariate, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\left(1-\pi_{ij'}\right)}{\pi_{ij'}\left(1-\pi_{ij}\right)} \leq \Gamma, \tag{6.2}$$

and then return to the distribution of $\mathbf{Z}$ to $\mathcal{Z}$ by conditioning on $Z_{i1} + Z_{i2} = 1$. It is straightforward to show that this is equivalent to the model

$$Pr(\mathbf{Z} = \mathbf{z} \,|\, F) = \frac{exp(\gamma\, \mathbf{z}^T\mathbf{u})}{\sum_{\mathbf{b}\in\mathcal{Z}}\, exp\left(\gamma\, \mathbf{b}^T\mathbf{u}\right)} = \prod_{i=1}^{I} \frac{exp\{\gamma\,(z_{i1}u_{i1} + z_{i2}u_{i2})\}}{exp(\gamma u_{i1}) + exp(\gamma u_{i2})}, \quad \mathbf{u} \in U, \tag{6.3}$$

for $\mathbf{z} \in \mathcal{Z}$ where $\gamma = log(\Gamma)$ and $U = [0,1]^{2I}$ is the $2I$-dimensional unit cube. For discussion of this model, see Rosenbaum (1987) for the case of matched pairs and Rosenbaum (2002, Sect. 4) for extensions to other cases, and for the equivalence of (6.2) and (6.3), see Rosenbaum (2002, Sect. 6.4.2) where the unobserved covariate u_{ij} is constructed from π_{ij} as $u_{ij} = \left\{log\left(\pi_{ij}\right) - min_k log(\pi_{ik})\right\}/\gamma$. Expressed as (6.2), the sensitivity analysis is similar in spirit to method of Cornfield et al. (1959); see also Gastwirth (1992) and Wang and Krieger (2006). The parameter Γ may be reinterpreted in terms of two parameters: one describing the relationship between u_{ij} and the outcome, r_{Cij}, and the other describing the relationship between u_{ij} and the treatment, Z_{ij} – it is an alternative interpretation of the same analysis; see Rosenbaum and Silber (2009). Under (6.3), the distribution of $t(\mathbf{Z}, \mathbf{R}) = t(\mathbf{Z}, \mathbf{r}_C)$ under the null hypothesis H_0 of no treatment effect is

$$Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \,|\, F\} = \sum_{\mathbf{z}\in\mathcal{Z}} \chi\{t(\mathbf{z}, \mathbf{r}_C) \geq k\} \prod_{i=1}^{I} \frac{exp\{\gamma\,(z_{i1}u_{i1} + z_{i2}u_{i2})\}}{exp(\gamma u_{i1}) + exp(\gamma u_{i2})} \tag{6.4}$$

where $\chi(E) = 1$ if the event E occurs and $\chi(E) = 0$ otherwise. Here, (6.4) reduces to (6.1) when $\Gamma = 1$ and $\gamma = log(\Gamma) = 0$.

For a statistic of the form $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$, define $\overline{\overline{T}}_\Gamma$ as the sum of I independent random variables taking value q_i with probability $\Gamma/(1+\Gamma)$ and value 0 with probability $(1+\Gamma)^{-1}$. In parallel, define $\overline{T}_\Gamma$ as the sum of I independent random variables taking values q_i with probability $(1+\Gamma)^{-1}$ and value 0 with probability $\Gamma/(1+\Gamma)$. It is straightforward to show that under (6.3) and the null hypothesis H_0 of no treatment effect,

$$Pr\left(\overline{T}_\Gamma \geq k\right) \leq Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \,|\, F\} \leq Pr\left(\overline{\overline{T}}_\Gamma \geq k\right) \text{ for all } \mathbf{u} \in U, \tag{6.5}$$

where $\mathbf{R} = \mathbf{r}_C$ when H_0 is true; see Rosenbaum (1987, 2002, Sect. 4.3). For $\Gamma = 1$, the bounds in (6.5) are equal $Pr\left(\overline{T}_1 \geq k\right) = Pr\left(\overline{\overline{T}}_1 \geq k\right)$ and equal (6.1). For fixed $\Gamma > 0$, the null distribution of $t(\mathbf{Z}, \mathbf{R})$ is unknown but bounded by (6.5). For each

fixed $\Gamma \geq 1$, (6.5) yields an interval of possible significance levels, point estimates, and endpoints for confidence intervals. In (6.5), the upper and lower bounds are sharp: They are each attained for particular $\mathbf{u} \in U$, so to narrow the bounds one would need some additional information about $\mathbf{u}$.

The bounds in (6.5) may be computed exactly (see Rosenbaum, 2003b, appendix, for software), but as $I \rightarrow \infty$, the central limit theorem supplies approximate bounds. Because $E\left(\overline{\overline{T}}_\Gamma \middle| F\right) = \theta \sum q_i$ and $var\left(\overline{\overline{T}}_\Gamma \middle| F\right) = \theta(1-\theta)\sum q_i^2$ with $\theta = \Gamma/(1+\Gamma)$, the probability $Pr\left(\overline{\overline{T}}_\Gamma \geq k\right)$ in (6.5) is approximately

$$Pr\left(\overline{\overline{T}}_\Gamma \geq k\right) \approx 1 - \Phi\left\{\frac{k - \theta\sum q_i}{\sqrt{\theta(1-\theta)\sum q_i^2}}\right\}, \tag{6.6}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution. When $\Gamma = 1$, the randomization distribution of (6.1) equals $Pr\left(\overline{\overline{T}}_1 \geq k\right)$, and (6.6) yields the usual large sample approximation to the distribution of Wilcoxon's signed rank statistic.

Either (6.5) or (6.6) yields bounds on significance levels and, by inverting the hypothesis test, bounds on the endpoints of confidence intervals. For Wilcoxon's statistic for an additive treatment effect, τ, one obtains the interval that bounds the possible HL point estimates for all $\mathbf{u} \in U$ as the interval $[\widehat{\tau}_{min}, \widehat{\tau}_{max}]$ where $\widehat{\tau}_{min}$ solves $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i = \theta \sum_{i=1}^{I} q_i$ and $\widehat{\tau}_{max}$ solves $\sum_{i=1}^{I} sgn(Y_i - \tau)\, q_i = (1-\theta)\sum_{i=1}^{I} q_i$ where q_i is the rank of $|Y_i - \tau|$; see Rosenbaum (1993).

Various methods of sensitivity analysis in observational studies are discussed by Copas and Eguchi (2001), Cornfield et al. (1959), Gastwirth (1992), Gastwirth, Krieger, and Rosenbaum (1998), Imbens (2003), Manski (1990), Robins, Rotnitzky, and Scharfstein (1999), and Rosenbaum and Rubin (1983). For a few applications, see Aakvik (2001), Ahmed et al. (2008), Diprete and Gangl (2004), Foster, Wiley-Exley, and Bickman (2009), Origo (2009), Silber et al. (2005), and Slade et al. (2008).

6.2.4 *Example of Sensitivity Analysis*

Returning to the example in Sect. 6.1.2 and Fig. 6.1, Table 6.1 displays a sensitivity analysis for the one-sided significance level from Wilcoxon's signed rank test and for the associated HL point estimate $\widehat{\tau}$ of an additive effect τ. The case $\Gamma = 1$ reproduces the randomization inferences reported in Sect. 6.1.2; these would be appropriate in paired randomized experiments. When $\Gamma = 1$, there is only one p-value and only one point estimate $\widehat{\tau}$. If the analysis had failed to control an unobserved covariate u associated with a 50% increase in the odds of a career as a painter, $\Gamma = 1.5$, and perhaps a very strong association with micronuclei, then the interval of possible significance levels is [0.00017, 0.023] and the interval of possible point estimates is [0.33, 0.99]. If u were associated with a doubling of the odds of a career as a painter, $\Gamma = 2$, then the null hypothesis of no treatment effect

Table 6.1 Sensitivity analysis for the 22 matched pair differences, painter-minus-control, in micronulcei using Wilcoxon's signed rank statistic and the associated Hodges–Lehmann (HL) point estimate

Γ	Minimum p-value	Maximum p-value	Minimum $\widehat{\tau}$	Maximum $\widehat{\tau}$
1	0.0032	0.0032	0.64	0.64
1.5	0.00017	0.023	0.47	0.81
2	0.0000096	0.064	0.33	0.99

Note. For three values of Γ, the table gives the range of possible one-sided p-values for testing the null hypothesis of no treatment effect, and the range of possible point estimates of an additive effect τ. The null hypothesis of no treatment effect is barely plausible for $\Gamma = 2$ as the maximum p-value is 0.064, although the minimum point estimate, 0.33, is still positive

would be just barely plausible, as the largest possible significance level is 0.064, which exceeds the conventional 0.05 level.

Observational studies vary markedly in their sensitivity to unobserved biases. Hammond's (1964) study of heavy smoking as a possible cause of lung cancer becomes sensitive at about $\Gamma = 6$, while Jick et al.'s (1973) study of coffee as a cause of myocardial infarction becomes sensitive at about $\Gamma = 1.3$. Pinto et al.'s (2000) study of painters falls in between: The bias that would explain Table 6.1 is smaller than that for smoking and lung cancer but larger than that for coffee and myocardial infarction.

Are there features of the design on an observational study that would make it less sensitive to biases from unobserved covariates? The remainder of this paper will investigate this question.

6.2.5 *Power of a Sensitivity Analysis; Design Sensitivity*

Fix α, $0 < \alpha < 1$, where conventionally $\alpha = 0.05$. For each $\Gamma \geq 1$, there is a critical value, c_Γ, such that $t(\mathbf{Z}, \mathbf{R}) \geq c_\Gamma$ if and only if the maximum significance level is less than α for this Γ and for all $\mathbf{u} \in U$. From (6.5), c_Γ is the smallest number such that $Pr\left(\overline{\overline{T}}_\Gamma \geq c_\Gamma\right) \leq \alpha$, and from (6.6), for large I this is approximately

$$c_\Gamma \approx \theta \sum q_i + \Phi^{-1}(1-\alpha)\sqrt{\theta(1-\theta)\sum q_i^2}$$

with $\theta = \Gamma/(1+\Gamma)$, which for Wilcoxon's signed rank statistic without ties is

$$c_\Gamma \approx \frac{\theta I(I+1)}{2} + \Phi^{-1}(1-\alpha)\sqrt{\theta(1-\theta)I(I+1)(2I+1)/6}. \tag{6.7}$$

For fixed $\Gamma \geq 1$, at level α, the conditional power given F of the sensitivity analysis is the chance that the upper bound on the significance level is less than or equal to α, that is, $Pr\{t(\mathbf{Z}, \mathbf{R}) \geq c_\Gamma \,|\, F\}$. To compute this conditional power given F, one would need to know F, so for most purposes it is more practical to consider

the unconditional power, $Pr\{t(\mathbf{Z},\mathbf{R}) \geq c_\Gamma\} = E[Pr\{t(\mathbf{Z},\mathbf{R}) \geq c_\Gamma \,|\, F\}]$ where the expectation is with respect to a model that generates F. When $\Gamma = 1$, this reproduces the usual definition of the power of a randomization test in a randomized experiment when applied to matched pair differences sampled from a specific distribution, such as the normal.

Suppose the treatment is effective, increasing the responses of treated subjects, $r_{Tij} > r_{Cij}$, so that the null hypothesis of no effect, H_0, is false, and suppose the matching has been successful in removing bias, so there is no bias from an unobserved covariate u. Call this the *favorable situation*. If the favorable situation did occur, one would not know that it had occurred from the observable data. One would see that treated subjects had typically higher responses than matched controls, so the matched pair differences were typically positive, as in the boxplot in Fig. 6.1, but one would not know that this pattern was produced by a treatment effect without bias, as opposed to being produced by bias alone or a combination of effect and bias. The best one could hope to say is that the observed results were insensitive to moderately large unobserved biases. The power of the sensitivity analysis, computed in the favorable situation, is the chance that this will happen.

Consider the following simple case of the favorable situation: There is no bias from u, so in fact $\Gamma = 1$, and the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau > 0$, so the treated-minus-control matched pair differences are $Y_i = \tau + (Z_{i1} - Z_{i2}) \times (r_{Ci1} - r_{Ci2}) = \tau + (Z_{i1} - Z_{i2})\varepsilon_i$ where $\varepsilon_i = r_{Ci1} - r_{Ci2}$. In this case, given F, Y_i is $\tau \pm \varepsilon_i$ with equal probabilities. If the ε_i were a sample of size I drawn independently from a continuous distribution $F(\cdot)$, then the unconditional power $Pr\{t(\mathbf{Z},\mathbf{R}) \geq c_\Gamma\} = E[Pr\{t(\mathbf{Z},\mathbf{R}) \geq c_\Gamma \,|\, F\}]$ could be computed in a conventional manner using the unconventional critical value c_Γ. For instance, Lehmann (1998, Sect. 6.4.2) shows that the nonnull expectation $\mu_y$and variance σ_y^2 of the Wilcoxon's signed rank statistic $t(\mathbf{Z},\mathbf{R})$ are $\mu_y = I(I-1)p_1'/2 + Ip$ and

$$\sigma_y^2 = I(I-1)(I-2)\left(p_2' - p_1'^2\right) + \frac{I(I-1)}{2}\left\{2\left(p - p_1'\right)^2 + 3p_1'\left(1 - p_1'\right)\right\} + Ip(1-p), \tag{6.8}$$

where $p = Pr(Y_i > 0)$, $p_1' = Pr\left(Y_i + Y_j > 0\right)$ and, using the logical symbol $\wedge$ for "and," $p_2' = Pr\left(Y_i + Y_j > 0 \wedge Y_i + Y_k > 0\right)$ with $i < j < k$, so that the central limit theorem yields the approximate power of a one-sided sensitivity analysis as $Pr\{t(\mathbf{Z},\mathbf{Y}) \geq c_\Gamma\} \approx 1 - \Phi\left\{\left(c_\Gamma - \mu_y\right)/\sigma_y\right\}$.

In the favorable situation, under very mild conditions, there is a value, $\widetilde{\Gamma}$, called the *design sensitivity*, such that, as the sample size increases, the power of the sensitivity analysis tends to 1 for all $\Gamma < \widetilde{\Gamma}$ and to 0 for all $\Gamma > \widetilde{\Gamma}$; see Rosenbaum (2004). This says that, in sufficiently large samples, a particular design – for example, a particular data generating process and protocol for analysis – can distinguish a treatment effect from all biases $\Gamma < \widetilde{\Gamma}$, but not from biases $\Gamma > \widetilde{\Gamma}$. In the case of Wilcoxon's signed rank statistic, as $I \to \infty$, a little algebra applied to (6.7) and (6.8) shows

$\widetilde{\Gamma} = p_1^{'}/\left(1 - p_1^{'}\right)$; essentially, this algebra eliminates terms of order smaller than I^2 in μ_y and σ_y as irrelevant to the limit. More generally, the design sensitivity $\widetilde{\Gamma}$ is the limit as $I \to \infty$ of the solutions Γ to the equation

$$E\left(\overline{\overline{T}}_{\Gamma}\right) = E\{t(\mathbf{Z}, \mathbf{Y})\} \tag{6.9}$$

where the expectations are computed in the favorable situation. For Wilcoxon's signed rank statistic, with Y_i sampled independently from a distribution $F(\cdot)$, (6.9) becomes

$$\frac{\Gamma}{(1+\Gamma)} \frac{I(I+1)}{2} = \frac{I(I-1)p_1^{'}}{2} + Ip$$

which again yields the limiting solution $\widetilde{\Gamma} = p_1^{'}/\left(1 - p_1^{'}\right)$ as $I \to \infty$.

What features of the design of an observational study affect its design sensitivity $\widetilde{\Gamma}$? The remainder of this paper will examine several such features.

6.3 Unit Heterogeneity

6.3.1 An Old Controversy: John Stuart Mill and R. A. Fisher

In 1864, John Stuart Mill, in his *System of Logic: Principles of Evidence and Methods of Scientific Investigation* (1867), proposed four methods of experimental inquiry, including the method of difference: "If an instance in which the phenomenon . . . occurs and an instance in which it does not . . . have every circumstance save one in common . . . [then] the circumstance [in] which alone the two instances differ is the . . . cause or a necessary part of the cause" (III, Sect. 8).

Mill is saying that to establish cause and effect, one should drive out every source of variation except the cause under study. The use of nearly identical, genetically engineered mice in the biology laboratory is a modern expression of the method of difference; see Holland (1986) for discussion of Mill's method of difference. In stark contrast, in 1935, Sir Ronald Fisher, in his *Design of Experiments*, objected to Mill's method of difference. In discussing his famous experiment of "the lady tasting tea," Fisher (1935, p. 18) wrote:

> It is not sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation . . . These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences . . . because [they] . . . are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely

> eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment ...

Fisher is certainly correct that randomization permits causal inference with heterogeneous experimental units or subjects, but that does not settle the matter for observational studies when random assignment is infeasible. Is reducing unit heterogeneity relevant to causal claims?

Suppose that, in the design of an observational study, one could choose between a study with fewer subjects who were less heterogeneous or more subjects who are more heterogeneous. Here, heterogeneity refers to the variability of the treated-minus-control differences Y_i in outcomes in matched pairs. For illustration, imagine that the expected or typical size of the treatment effect is the same in both cases; only the sample size and the dispersion of the Y_is are different in the two studies. Mill's method of difference suggests the smaller, less heterogeneous study would be better. Much of the intuition developed from randomized experiments suggests that sample size and unit heterogeneity trade off against one another in the standard error of the estimated treatment effect, and that neither sample size nor dispersion of matched pair differences has much to do with bias from unmeasured covariates. Which view is correct when randomization is infeasible?

A common illustration of this choice is the use of identical twins (e.g., Ashenfelter & Rouse, 1998; Isacsson, 2007): Twins pairs are less heterogeneous, but few in number. In particular, Ashenfelter and Rouse compared the earnings of identical twins with different levels of education. Another illustration is from a study by Norvell and Cummings (2002) of the effects of helmets in motorcycle crashes, focusing on crashes in which two people rode one motorcycle, and one wore a helmet. Again, few such crashes have happened, but the two paired individuals were on the same motorcycle in the same crash. See Rosenbaum (2005) for additional illustrations, including the very clever study by Wright and Robertson (1976).

6.3.2 *Heterogeneity and Power of a Sensitivity Analysis*

As motivation, consider two simulated observational studies, both with constant treatment effect $\tau = \frac{1}{2}$ and, unknown to us, with no unobserved bias. In the larger and more heterogeneous study (LM), there are $I = 400$ pairs and $Y_i \sim_{iid} N\left(\frac{1}{2}, 1\right)$, while in the smaller and less heterogeneous study (SL), there are $I = 100$ pairs and $Y_i \sim_{iid} N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$. The mean of the I differences is $N\left\{\frac{1}{2}, \frac{1}{400}\right\}$ in both studies. If LM and SL were analyzed as if they were randomized experiments, then very similar 95% confidence intervals for τ are obtained from Wilcoxon's signed rank test, specifically [0.40, 0.60] for LM and [0.43, 0.62] for SL, and also very similar HL point estimates $\widehat{\tau}$ of 0.50 for LM and 0.52 for SL. However, if LM and SL were observational studies, then SL would be much less sensitive to unobserved bias than LM. Specifically, the upper bound on the one-sided significance level in (6.5) is above 0.05 for $\Gamma \geq 2.5$ for LM, but the upper bound is less than 0.05 for $\Gamma \leq 6$ for SL.

Table 6.2 Power of the sensitivity analysis under various assumptions

Errors	I Matched Pairs	τ	σ	$\frac{\sigma^2}{I}$	Power $\Gamma = 1$	Power $\Gamma = 1.5$	Power $\Gamma = 2$
Normal	120	$\frac{1}{2}$	1	1/120	1.00	0.96	0.60
Normal	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	1.00	1.00	0.96
Logistic	120	$\frac{1}{2}$	1	1/120	0.93	0.31	0.04
Logistic	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	0.93	0.61	0.32
Cauchy	200	$\frac{1}{2}$	1	1/200	0.98	0.32	0.02
Cauchy	50	$\frac{1}{2}$	$\frac{1}{2}$	1/200	0.95	0.60	0.28

Table 6.3 Design sensitivity $\widetilde{\Gamma}$ Wilcoxon's signed rank test when $Y_i \sim_{iid} N(\tau, \sigma^2)$

	$\sigma = 1$	$\sigma = \frac{1}{2}$	$\sigma = \frac{1}{4}$
$\tau = 0$	1.00	1.00	1.00
$\tau = 1/4$	1.76	3.17	11.71
$\tau = 1/2$	3.17	11.71	426.56

For $\Gamma = 2$, the range of HL point estimates $[\widehat{\tau}_{min}, \widehat{\tau}_{max}]$ is $[0.19, 0.81]$ for LM and $[0.37, 0.67]$ for SL. The SL study is less sensitive to unobserved bias than Hammond's (1964) study of smoking and lung cancer, but LM is much more sensitive. Is this an oddity of two simulated samples, or is it true in general? To answer this question, attention must turn away from data sets, say the simulated LM and SL, and toward the data generating processes (or designs) that produce the data.

Table 6.2 presents the power of the sensitivity analysis using Wilcoxon's signed rank test in the favorable situation in which treatments are actually randomized and the matched pair differences Y_i are independent and identically distributed so that $(Y_i - \tau)/\sigma$ has a normal, logistic, or Cauchy distribution. For each distribution, two situations are considered – one with a smaller σ and a smaller I, the other with a larger σ and a larger I – so that the "standard error" $\sigma/\sqrt{I}$ is the same. (Here, standard error is in quotes, because for the Cauchy distribution, there is a change in scale but there is no "standard error.") In a randomization test, $\Gamma = 1$, the power is about the same with small (σ, I) and with large (σ, I). In the sensitivity analysis, with $\Gamma > 1$, the power is much higher with small (σ, I). In this sense, Mill (1867) was correct: In the absence of randomization, causal inferences are less sensitive to unobserved biases when the unit heterogeneity is reduced, even at the expense of reduced sample size.

As $I \rightarrow \infty$, the power of the sensitivity analysis tends to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$ where $\widetilde{\Gamma}$ is the design sensitivity. Recall from Sect. 6.2.5 that the design sensitivity for Wilcoxon's signed rank statistic in the favorable situation is $\widetilde{\Gamma} = p_1'/(1 - p_1')$, where $p_1' = Pr(Y_i + Y_j > 0)$ when the Y_i are an independent and identically distributed (iid) sample. If $Y_i \sim_{iid} N(\tau, \sigma^2)$, then $p_1' = \Phi(\sqrt{2}\tau/\sigma)$, so with a treatment effect of fixed size, τ, reducing heterogeneity, σ, increases the design sensitivity, $\widetilde{\Gamma}$. Table 6.3 calculates $\widetilde{\Gamma}$ for several τ and σ. When it is feasible to reduce heterogeneity σ without altering the treatment effect τ, Mill's method of difference has the potential to greatly strengthen causal inferences.

6.3.3 *Heterogeneity and the Limiting Uncertainty in Point Estimates*

For the example in Fig. 6.1, for several values of Γ, Table 6.1 reported the interval $[\widehat{\tau}_{min}, \widehat{\tau}_{max}]$ of possible HL point estimates of a constant treatment effect τ. For fixed Γ, as $I \rightarrow \infty$, the interval of estimates $[\widehat{\tau}_{min}, \widehat{\tau}_{max}]$ converges in probability to a real interval $[\tau_{min}, \tau_{max}]$. The following proposition, which is proved in Rosenbaum (2005), says that the length of $[\tau_{min}, \tau_{max}]$ is strongly affected by the heterogeneity of the experimental units. Let $\Phi(\cdot)$ and $\Psi(\cdot)$ be, respectively, the standard normal and standard Cauchy cumulative distributions. Proposition 6.1 indicates what a sensitivity analysis yields, as $I \rightarrow \infty$ in the favorable situation when, unknown to us, there actually is no unobserved bias.

Proposition 1. *If* $(Y_i - \tau)/\sigma \sim_{iid} \Phi(\cdot)$ *then* $[\tau_{min}, \tau_{max}]$ *is* $\tau \pm \sigma\, \Phi^{-1}(\theta)/\sqrt{2}$, *where* $\theta = \Gamma/(1+\Gamma)$. *If* $(Y_i - \tau)/\sigma \sim_{iid} \Psi(\cdot)$ *then* $[\tau_{min}, \tau_{max}]$ *is* $\tau \pm \sigma\, \Psi^{-1}(\theta)$.

Proposition 1 confirms Mill's (1867) method. Proposition 1 describes the situation in which we would like to report as little sensitivity to bias as possible, because in fact the treatment worked with true effect τ and, unknown to us, there was no unobserved bias. In this situation, even after sampling variability has been driven out by letting $I \rightarrow \infty$, the uncertainty about unobserved bias, as reflected in $[\tau_{min}, \tau_{max}]$, is directly proportion to σ, the heterogeneity in the matched pair differences. In light of this, Mill's fanatical effort to reduce σ is, indeed, directly relevant to the evidence about *causality*, and does not merely reduce the *standard error*.

6.4 Dose and Response

6.4.1 *Does a Dose–Response Relationship Strengthen Causal Inference?*

Much has been written about dose–response relationships as evidence of cause and effect. Hill (1965) wrote:

> ... if the association is one which can reveal a biological gradient, or dose–response curve, then we should look most carefully for such evidence. For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers. (p. 298)

Not everyone agrees. Rothman (1986) wrote:

> Some causal associations, however, show no apparent trend of effect with dose; an example is the association between DES and adenocarcinoma of the vagina ... Associations that do show a dose–response trend are not necessarily causal; confounding can result in such a trend between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease. (p. 18)

Before offering a positive reinterpretation of Hill's (1965) argument, Weiss (1981) makes a similar observation: "... one or more confounding factors can be related closely enough to both exposure and disease to give rise to [a dose response relationship] in the absence of cause and effect" (p. 488).

An additional complication arises when observational studies are compared to experiments. Cochran (1965) had argued that observational studies should be patterned after simple experiments. In clinical trials, it is typically said (Peto et al., 1976, p. 590) that, within practical and ethical constraints, one should compare just two treatments that are as different as possible. How does this advice square with Hill's idea that a graduated dose–response relationship strengthens causal claims?

6.4.2 Example: Sensitivity Analysis with and Without Doses

The *signed rank statistic with doses* is $t(\mathbf{Z},\mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$ with q_i equal to the product of the dose d_i and the rank of $|Y_i|$; see van Eeden (1972) and Rosenbaum (1997, 2003a). For the example in Sect. 6.1.2 and Fig. 6.1, using $log_2(years)$ as the dose of exposure to professional painting, Table 6.4 contrasts the sensitivity analysis using Wilcoxon's signed rank statistic ignoring doses in Sect. 6.2.4 to the analogous sensitivity analysis using the signed rank statistic with doses. The statistic without doses is about as sensitive to a bias of $\Gamma = 2$ as the statistic with doses is to a bias of $\Gamma = 2.5$, because the upper bound on the one-sided significance level is about 0.065 in both instances. In this one example, to a very moderate degree, the presence of a dose–response relationship reduced sensitivity to unobserved biases.

Table 6.4 Sensitivity analysis with or without doses for the 22 matched pair differences, painter-minus-control, in micronulcei using Wilcoxon's signed rank statistic or the signed rank statistic with doses

Γ	Ignoring doses	Using doses
1	Max 0.0032	Max 0.0025
	Min 0.0032	Min 0.0025
2	Max 0.064	Max 0.038
	Min 0.0000096	Min 0.000014
2.2	Max 0.085	Max 0.049
	Min 0.0000031	Min 0.00000052
2.5	Max 0.12	Max 0.067
	Min 0.000000056	Min 0.0000016

Note. For four values of Γ, the table gives the range of possible one-sided p-values for testing the null hypothesis of no treatment effect, ignoring or using the log_2 of years of work as a painter as the dose. Without doses, the null hypothesis is barely plausible at $\Gamma = 2$ as the maximum p-value is 0.064. With doses, the null hypothesis is not quite plausible at $\Gamma = 2.2$ as the maximum p-value is 0.049. In this one case, a dose–response relationship has made the comparison slightly less sensitive to unobserved biases

6.4.3 Derivation of the Design Sensitivity with Doses of Treatment

The current section obtains a formula for the design sensitivity for the signed rank statistic with doses. The result is new, but is analogous to a result in Rosenbaum (2004) for a different statistic. In Sect. 6.4.4, the formula is used to evaluate the contribution of a dose–response relationship to evidence of cause and effect.

The signed rank statistic with doses is $t(\mathbf{Z},\mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$ with q_i equal to the product of the dose d_i and the rank of $|Y_i|$. Write

$$\begin{aligned} W_{ik} &= 1 \text{ if } |Y_i| \geq |Y_k| \text{ and } Y_i > 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

Then the signed rank statistic with doses may be written as

$$t(\mathbf{Z},\mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i = \sum_{i=1}^{I}\sum_{k=1}^{I} d_i W_{ik} \tag{6.10}$$

because $\sum_{k=1}^{I} d_i W_{ik} = 0$ if $sgn(Y_i) = 0$ and $\sum_{k=1}^{I} d_i W_{ik} = d_i rank(|Y_i|)$ if $sgn \times (Y_i) = 1$. The representation (6.10) is closely related to the representation of Wilcoxon's signed rank statistic as the number of positive Walsh averages, namely $(Y_i + Y_k)/2$. Also, write $V_{ik} = 1$ if $|Y_i| \geq |Y_k|$ and $V_{ik} = 0$ otherwise, so that for the signed rank statistic with doses, $q_i = \sum_{k=1}^{I} d_i V_{ik}$.

Proposition 2. *Suppose that (i) doses d_i in the I pairs are independently sampled from a discrete distribution with $L \geq 1$ possible doses, $\delta_1, \ldots, \delta_L$ with $Pr(d_i = \delta_\ell) = \eta_\ell$, $\ell = 1, \ldots, L$; (ii) treatments are randomly assigned within pairs, $Pr(\mathbf{Z} = \mathbf{z}\,|\,F) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$, (iii) treated-minus-control differences in responses Y_i are $Y_i = \beta d_i + (Z_{i1} - Z_{i2})\varepsilon_i$ where ε_i are independent of treatment Z_{ij} and doses d_i and independently sampled from a distribution $F(\cdot)$. Then as $I \to \infty$, the limiting sensitivity to unobserved bias using the signed rank statistic with doses is*

$$\widetilde{\Gamma} = \frac{\Upsilon_2}{\Lambda_2 - \Upsilon_2} \tag{6.11}$$

where $\Upsilon_2 = E(d_i W_{ik})$ and $\Lambda_2 = E(d_i V_{ik})$ with $i \neq k$.

Remark 1. Explicit forms for Υ_2 and Λ_2 are

$$\begin{aligned} \Upsilon_2 &= \sum_{\ell=1}^{L}\sum_{m=1}^{L} \eta_\ell \eta_m \delta_\ell \; Pr\{(|\beta\delta_\ell + \varepsilon| \geq |\beta\delta_m + \varepsilon'|) \wedge (\beta\delta_\ell + \varepsilon > 0)\}, \\ \Lambda_2 &= \sum_{\ell=1}^{L}\sum_{m=1}^{L} \eta_\ell \eta_m \delta_\ell \; Pr(|\beta\delta_\ell + \varepsilon| \geq |\beta\delta_m + \varepsilon'|), \end{aligned}$$

where ε and ε' are two independent observations from $F(\cdot)$. From these expressions, it is clear that $\Lambda_2 \geq \Upsilon_2$ with equality if and only if $1 = Pr(Y_i > 0) = Pr(\beta d_i + \varepsilon > 0)$.

In some instances, Υ_2 and Λ_2 may be determined explicitly, and it is always straightforward to evaluate these expressions by Monte Carlo. For instance, if $\beta = \frac{1}{2}$ and $F(\cdot)$ is the Cauchy distribution, if the doses are 1, 2, or 3 with equal probabilities, $L = 3$, $\delta_\ell = \ell$, $\eta_\ell = \frac{1}{3}$, for $\ell = 1, 2, 3$, then $\Upsilon_2 = .803$, $\Lambda_2 = 1.040$, and $\widetilde{\Gamma} = 3.4 = 0.80/(1.04 - 0.80)$. These values were obtained by sampling one million $(d_i, d_k, \varepsilon_i, \varepsilon_k)$ and taking Υ_2 as the mean of $d_i\,\chi\{(|\beta d_i + \varepsilon_i| \geq |\beta d_k + \varepsilon_k|) \wedge (\beta d_i + \varepsilon > 0)\}$ and Λ_2 as the mean of $d_i\,\chi(|\beta d_i + \varepsilon_i| \geq |\beta d_k + \varepsilon_k|)$ where $\chi(E) = 1$ if event E occurs and $\chi(E) = 0$ otherwise. Condition (iii) is actually used only to obtain these explicit forms and numerical values for $\Upsilon_2 = E(d_i W_{ik})$ and $\Lambda_2 = E(d_i V_{ik})$; however, condition (iii) could be replaced by other models for $Pr(Y_i|d_i)$, yielding different forms for $\Upsilon_2 = E(d_i W_{ik})$ and $\Lambda_2 = E(d_i V_{ik})$, but the same (6.11) for the design sensitivity $\widetilde{\Gamma}$ in terms of Υ_2 and Λ_2.

Remark 2. In particular, if the I paired control responses, (r_{Ci1}, r_{Ci2}), were *iid* from some bivariate distribution independent of d_i, condition (iii) would arise if the treatment effect were proportional to the dose, $r_{Tij} - r_{Cij} = \beta d_i$, because in this case,

$$Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2}) = \beta d_i + (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \beta d_i + (Z_{i1} - Z_{i2})\varepsilon_i\,,$$

where $\varepsilon_i = (r_{Ci1} - r_{Ci2})$.

Proof. Assume (i) and (ii) and let $\Upsilon_2 = E(d_i W_{ik})$, $\Lambda_2 = E(d_i V_{ik})$ with $i \neq k$. Also, define $\Upsilon_1 = E(d_i W_{ii}) = E\{d_i \chi(Y_i > 0)\}$, $\Lambda_1 = E(d_i V_{ii}) = E(d_i)$. From (6.10), the expectation of the signed rank statistic with doses is

$$\begin{aligned} E\{t(\mathbf{Z}, \mathbf{R})\} &= E\left(\sum_{i=1}^{I}\sum_{k=1}^{I} d_i W_{ik}\right) = \sum_{i=1}^{I} E(d_i W_{ii}) + \sum_{i=1}^{I}\sum_{k\neq i} E(d_i W_{ik}) \\ &= I\Upsilon_1 + I(I-1)\Upsilon_2. \end{aligned}$$

Because $q_i = \sum_{k=1}^{I} d_i V_{ik}$,

$$E\left(\overline{\overline{T}}_\Gamma \,\middle|\, F\right) = \theta \sum_{i=1}^{I} q_i = \theta \sum_{i=1}^{I}\sum_{k=1}^{I} d_i V_{ik},$$

where $\theta = \Gamma/(1+\Gamma)$, so that

$$\begin{aligned} E\left(\overline{\overline{T}}_\Gamma\right) &= E\left\{E\left(\overline{\overline{T}}_\Gamma \,\middle|\, F\right)\right\} = \theta \sum_{i=1}^{I}\sum_{k=1}^{I} E(d_i V_{ik}) \\ &= \theta\{I\Lambda_1 + I(I-1)\Lambda_2\}. \end{aligned}$$

The design sensitivity, $\widetilde{\Gamma}$, is the limit, as $I \to \infty$, of the solutions to $E\{t(\mathbf{Z}, \mathbf{R})\} = E\left(\overline{\overline{T}}_\Gamma\right)$, which is easily seen to be (6.11).

6.4.4 Evaluation of Design Sensitivity with Doses of Treatment

Table 6.5 evaluates the design sensitivity $\widetilde{\Gamma}$ for the signed rank statistic with doses in the case of L equally probable integer valued doses, $\ell = 1, \ldots, L$. (In the notation of Sect. 6.4.3, $\delta_\ell = \ell$, $\eta_\ell = \frac{1}{L}$, for $\ell = 1, \ldots, L$.) Table 6.5 describes the favorable situation: Specifically, $Y_i = \beta d_i + (Z_{i1} - Z_{i2})\varepsilon_i$ where the ε_i have either a normal or a logistic distribution and treatment assignment Z_{ij} is randomized. In this case, the marginal distribution of treated-minus-control difference Y_i is symmetric about $\beta(L+1)/2$, and this quantity is more relevant than β itself when comparing different numbers of doses, L. In the first three rows of Table 6.5, the typical value of Y_i is $\frac{1}{2} = \beta(L+1)/2$, but there are $L = 1$ or 3 or 5 doses. Table 6.5 addresses two questions raised in Sect. 6.4.1. Does a dose response relationship reduce sensitivity to unobserved biases when the typical effect remains unchanged? Would it be better to use just the largest doses, so the typical effect is larger, with no dose–response relationship?

In Table 6.5, holding $\beta(L+1)/2$ fixed while varying the number of doses provides some support for Hill's (1965) claim: There is somewhat less sensitivity to unobserved biases with additional dose levels. However, if $\beta = \frac{1}{4}$ and $L = 3$, as in row 2 of Table 6.5, the typical Y_i for individuals with dose $d_i = 3$ is $\beta d_i = 3/4$, rather than $1/2$. Would it be better to use only the subset of pairs with the extreme dose? In this situation with $\beta = \frac{1}{4}$ and $L = 3$, one could just use those pairs i with the largest dose $d_i = 3$, in which case there is just one dose level, and this is equivalent to row 5 of Table 6.5 with $\beta = 3/4$ and $L = 1$. Notice that the design sensitivity $\widetilde{\Gamma}$ is much higher for $\beta = 3/4$ and $L = 1$ in row 5 than for other rows of the table. The latter observation is consistent with the suggestion of Cochran (1965) and Peto et al. (1976, p. 590), mentioned in Sect. 6.4.1, that an observational study should resemble a simple experiment in which the treatment and control conditions are as different as possible. Even $\beta(L+1)/2 = 0.6$ with one dose level in row 4 of Table 6.5 is competitive with $L = 5$ dose levels and $\beta(L+1)/2 = 0.5$.

Three points should be kept in mind when thinking about Table 6.5. First, the design sensitivities $\widetilde{\Gamma}$ in Table 6.5 refer to the limiting case, as $I \to \infty$, so the loss of

Table 6.5 Design sensitivity for L equally probable doses, 1, 2, ..., L, with slope β with normal or logistic errors

L	β	$\frac{\beta(L+1)}{2}$	Normal	Logistic
1	$0.50 = 1/2$	0.5	3.18	1.95
3	$0.25 = 1/4$	0.5	3.77	2.17
5	$0.17 = 1/6$	0.5	3.99	2.25
1	$0.60 = 3/5$	0.6	4.05	2.24
1	$0.75 = 3/4$	0.75	5.91	2.74

Note. Here, $\beta(L+1)/2$ is the median treatment effect. With the same median effect, $\beta(L+1)/2$, a larger number of dose levels, L is slightly better, yielding a higher design sensitivity. However, one dose, $L = 1$ but with a larger median effect is often better. For instance $\beta = 0.25$ with $L = 3$ is inferior to $\beta = .75$ with $L = 1$; however, the latter is equivalent to insisting on the maximum dose of 3 in former situation

sample size due to focusing on extreme doses is not considered. For instance, if the investigator used only the largest doses when there are $L = 3$ equally probable doses, moving from row 2 of Table 6.5 to row 5, the investigator would have discarded two-thirds of the sample, and in practice this must be weighed against the improved design sensitivity. Power calculations, analogous to those in Table 6.2, can clarify the trade-off between sample size and design sensitivity. Second, the ordering of studies by design sensitivity agrees with the ordering by the power of a sensitivity analysis *for sufficiently large* I, so in sufficiently large studies, it is correct to discard pairs with small doses. If one followed the advice of Peto et al. (1976, p. 590) in designing a clinical trial, pairs with small doses would not have been collected in the first place. Third, the calculations in Table 6.5 are based on certain simple models, and other models are possible. For instance, in Sect. 6.1.2, if one believed that a man who worked as a professional painter for 2 years was very similar to a clerk in terms of u, but a man who worked as a professional painter for 10 years was very different from a clerk in terms of u, then low dose pairs would be less biased than high dose pairs, and this would provide additional information about $\mathbf{u}$ not used by (6.3).

For the example in Sect. 6.1.2, Fig. 6.2 divides the 22 pairs at the median years of work as a painter, namely 4 years, yielding ten pairs with less than 4 years of work

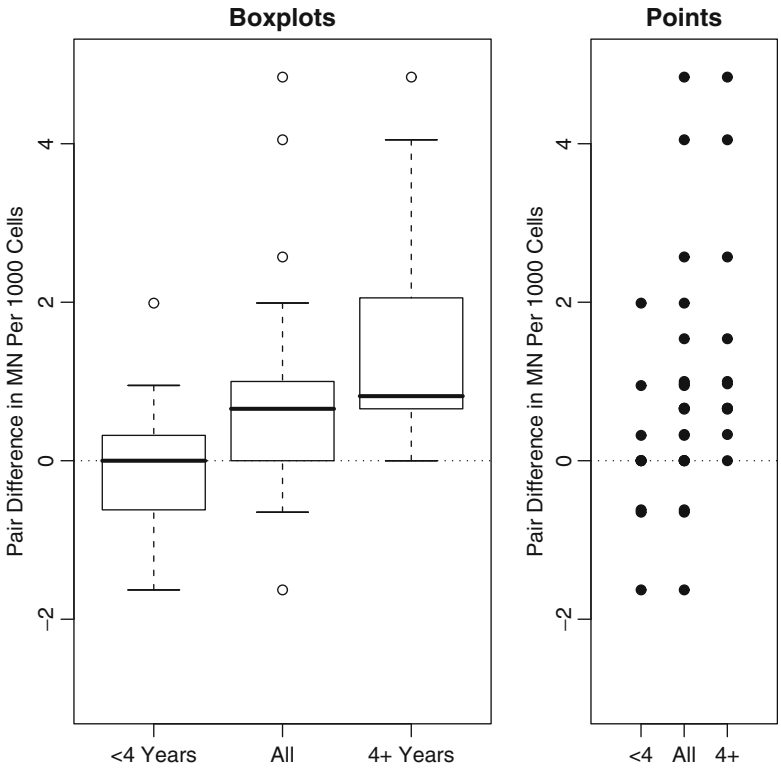

Fig. 6.2 Boxplots and point plots of the 22 matched pair differences, painter-minus-control, in micronuclei, split into pairs in which the painter had less than 4 year's work as a painter (ten pairs) or 4 or more year's work as a painter (12 pairs)

Table 6.6 Sensitivity analysis for micronuclei (mn) in painters and matched controls in three analyses: (a) Using all 22 pairs but ignoring doses with Wilcoxon's signed rank test, (b) Using all 22 pairs with the signed rank test with doses, and (c) Using the 12 high dose pairs with a painter who has worked for 4 or more years

	All pairs ($n = 22$)	All pairs ($n = 22$)	High dose pairs ($n = 12$)
Γ	Ignoring doses	Using doses	Ignoring doses
1	0.0032	0.0025	0.0012
2	0.064	0.038	0.016
2.5	0.12	0.067	0.028
3.3	0.22	0.12	0.048

as a painter and 12 pairs with 4 or more years of work as a painter. Performing the sensitivity analysis in Sect. 6.2.4 using just the 12 *high dose* pairs, yields the last column of Table 6.6. In Table 6.6, the analysis that focuses on the 12 high dose pairs is the analysis that is least sensitive to unobserved biases, becoming sensitive at about $\Gamma = 3.3$. Obviously, Table 6.6 concerns just one small example, but the pattern is consistent with the theoretical calculations in Table 6.5.

6.5 Coherence Among Several Outcomes

The association between a treatment and several outcomes is coherent if it is compatible with a mechanism through which the treatment is thought to produce effects. Campbell (1988) wrote that "inferential strength is added when each theoretical parameter is exemplified in two or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned" (p. 33). See also Reynolds and West (1987) and Trochim (1985).

For instance, in Sect. 6.1.2, in Pinto et al.'s (2000) study of genetic damage among professional painters, three standard measures of genetic damage were used, namely micronuclei frequency (MN), average chromosomal aberrations per cell (CA), and the sister chromatid exchanges (SCE). These three measures are very different measures of genetic damage. If all three measures were elevated among painters, the association might be judged as stronger, more coherent than if one was elevated among painters, another was elevated among controls, and the third exhibited no difference. The three measures and the relationships are depicted in Fig. 6.3 as 22 painter-minus-control matched pair differences. All three measures are elevated in painters compared to matched controls. In Fig. 6.3, for the matched pair differences, Kendall's rank correlation between MN and CA is 0.55, between CA and SCE is 0.14, and between SCE and MN is 0.13.

In its simplest form, the coherent signed rank statistic for three oriented outcomes is simply the sum of the three separate signed rank statistics, and with

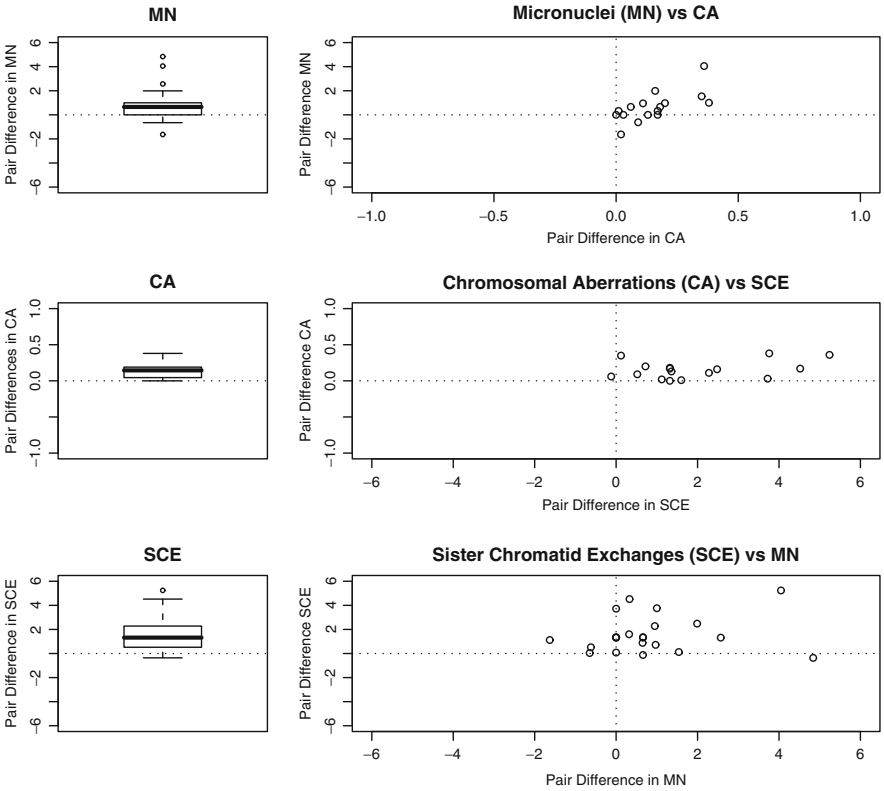

Fig. 6.3 Plot of 22 matched pair differences, painter-minus-control, for three measures of genetic damage: micronuclei frequency (MN), chromosomal aberrations (CA), and sister chromatid exchanges (SCE)

some attention to detail, the sensitivity analysis in Sect. 6.2.3 may be applied to this statistic (Rosenbaum, 1997). Table 6.7 is the sensitivity analysis for the three outcomes in Fig. 6.3. Comparing Tables 6.1 and 6.7, one sees that the coherent association in Table 6.7 is substantially less sensitive to unobserved biases than the association for micronuclei alone.

Various calculations of design sensitivity for coherence are given in Rosenbaum (2004) and Heller et al. (2009). Consider, for instance, four outcomes, with multivariate normal matched pair differences, each having expectation or treatment effect $\frac{1}{2}$, standard deviation 1, and the same intercorrelation ρ. If only the first outcome is used in Wilcoxon's signed rank test, the design sensitivity is $\widetilde{\Gamma} = 3.17$. If the signed rank test is applied to the average of the first two outcomes, the design sensitivity is $\widetilde{\Gamma} = 5.30$ for $\rho = 0$ and $\widetilde{\Gamma} = 4.39$ for $\rho = \frac{1}{4}$. If the signed rank test is applied to the average of all four outcomes, the design sensitivity is $\widetilde{\Gamma} = 11.71$ for $\rho = 0$ and $\widetilde{\Gamma} = 6.02$ for $\rho = \frac{1}{4}$. Of course, for $\rho = 1$, the design sensitivity is again $\widetilde{\Gamma} = 3.17$ for any subset of the four outcomes.

Table 6.7 Sensitivity analysis using the coherent signed rank statistic

Γ	*min* p-value	*max* p-value
1	0.000061	0.000061
2	1.8×10^{-8}	0.0041
3	6.1×10^{-12}	0.018
4	2.2×10^{-15}	0.039
4.5	$<10^{-15}$	0.050

Note. The coherent signed rank statistic is the sum of the signed rank statistics for micronuclei (MN), sister chromatid exchanges (SCE), and chromosomal aberrations (CA). For several values of Γ, the table gives the maximum and minimum significance level

6.6 Uncommon but Dramatic Responses to Treatment

6.6.1 Large but Rare Treatment Effects

Salsburg (1986) considered the possibility that a treatment might have a dramatic effect on some people and no effect at all on most people, where it is not possible to identify in advance of treatment the subgroup of people who will be affected; that is, $r_{Cij} = r_{Tij}$ for many ij, but $r_{Cij} \ll r_{Tij}$ for some ij. Building upon earlier work by Lehmann (1953), *Salsburg's unpaired model* assumes r_{Cij} are sampled independently from a continuous distribution $F(\cdot)$, while r_{Tij} are sampled from $(1-p)F(\cdot) + pF^m(\cdot)$ for $0 \le p \le 1$ and for integer $m \ge 2$, so only a fraction p of the population is affected by the treatment. Here $F^m(\cdot) = F(\cdot) \times \cdots \times F(\cdot)$ is the distribution of the maximum of m independent observations from $F(\cdot)$, so it is stochastically larger than $F(\cdot)$. If m is large, there is one sense in which the treatment effect is large and dramatic – perhaps insensitive to unobserved biases – and there is another sense in which, if p is small, the same effect is, in aggregate, quite small, because only a fraction of the population is affected. In the discussion here, for matched pairs, Salsburg's model is modified ever so slightly to permit dependence within pairs: *Salsburg's paired model* assumes $r_{Cij} - \xi_i \sim_{iid} F(\cdot)$ and $r_{Tij} - \xi_i \sim_{iid} (1-p)F(\cdot) + pF^m(\cdot)$, so that the pair parameter ξ_i creates dependence within pair i, but ξ_i is removed by taking matched pair differences.

Figure 6.4 depicts 1,000 matched pair differences from the Salsburg's paired model for several values of p and m. In the first boxplot, the treatment has no effect, because $p = 1$, so the differences have a normal distribution with expectation 0 and variance $1 + 1 = 2$. In the second boxplot, $p = 1/4$ of individuals respond to treatment with a response that equals the maximum of $m = 5$ responses to the control. Looking just at the boxplots, it would not be easy to discern that only a small fraction of the population is affected by the treatment.

Figure 6.5 depicts the corresponding density functions for the matched pair differences, together with the density function of a normal distribution having the same expectation and variance. The paired model and the normal differ discernibly for $p = 1/4$ and $m = 100$ and quite noticeably for $p = 1/4$ and $m = 500$.

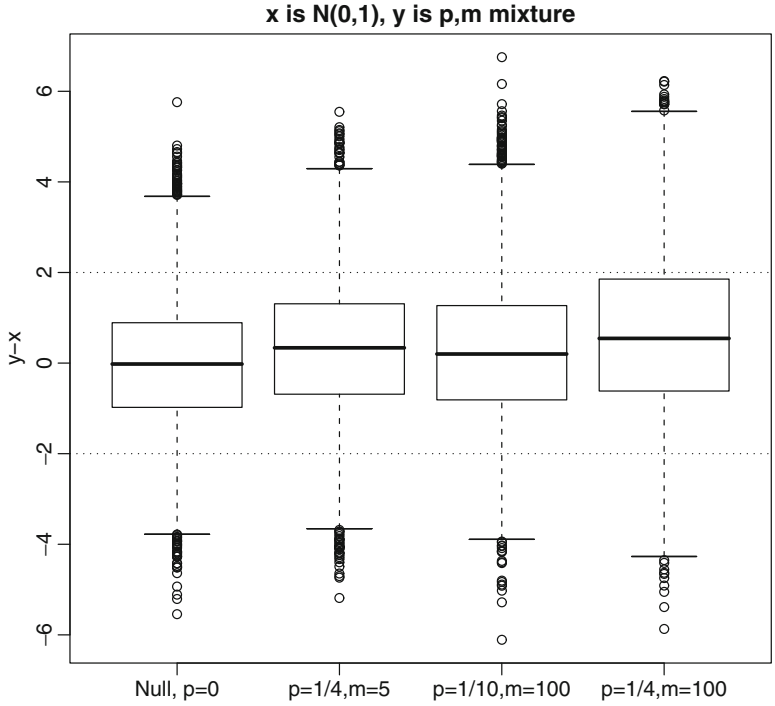

Fig. 6.4 Samples of 1,000 matched pair differences under Salsburg's paired model with $r_{Cij} - \Phi\,\xi_i \sim \Phi(.)$ and $r_{Tij} - \xi_i \sim (1-p)\Phi(.) + p\Phi^m(.)$, so that only a fraction p of subjects are affected by treatment receiving the maximum of m independent observations from $\Phi(.)$, where $\Phi(.)$ is the standard Normal cumulative distribution. In the first boxplot, there is no treatment effect. It is difficult to discern in the last three boxplots that most individuals are unaffected by treatment

6.6.2 Detecting Large but Rare Treatment Effects

Described informally, a locally most powerful rank test is one that has the highest possible power against small effects in large samples. Lehmann (1953) had shown that Wilcoxon's ranks, 1, 2, ..., I yield the locally most powerful rank test in the unpaired model with $m = 2$ as $p \to 0$ and $I \to \infty$. Conover and Salsburg (1988) derived the locally most powerful ranks for general m as $p \to 0$ and $I \to \infty$; these turn out to be a fairly unintuitive polynomial in the ordinary ranks whose highest power is $m - 1$, in agreement with Lehmann's result for $m = 2$. Conover and Salsburg's ranks closely resemble an intuitive second set of ranks due to Stephenson (1981), who was motivated by different considerations. Stephenson's ranks are also a polynomial in the ordinary ranks whose highest power is $m - 1$, and the two types of ranks exhibit nearly identical properties for large I. Because Stephenson's ranks are interpretable, they can be inverted to obtain confidence statements for the magnitude of effect. For detailed discussion of these issues, see Rosenbaum (2007).

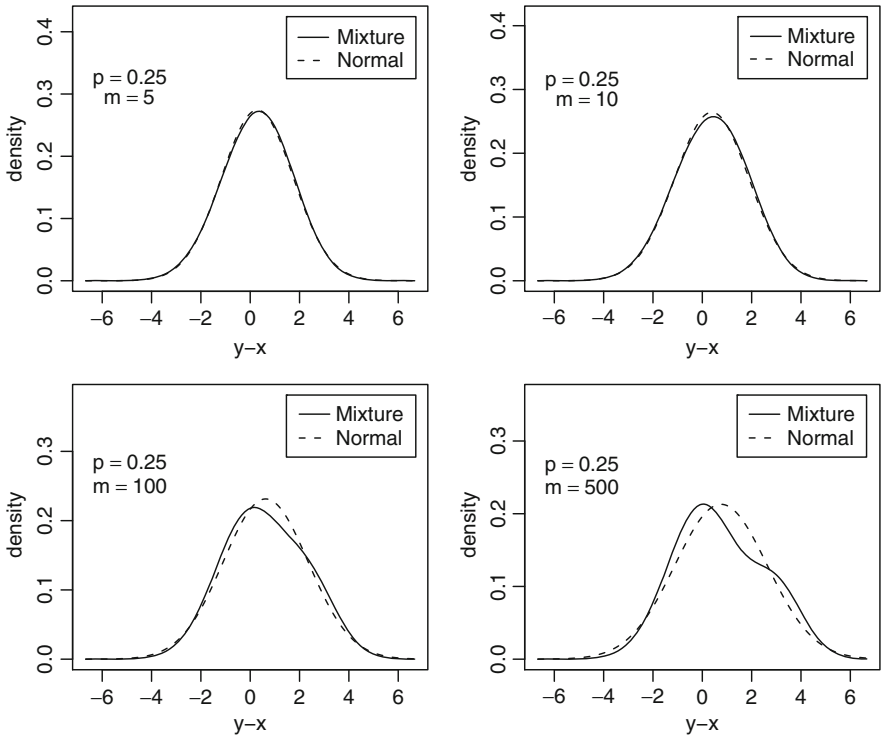

Fig. 6.5 The probability density function of matched pair differences Salsburg's paired model when $r_{Cij} - \xi_i \sim \Phi(.)$ and $r_{Tij} - \xi_i \sim (1-p)\Phi(.) + p\Phi^m(.)$ where $\Phi(.)$ is the standard normal cumulative distribution. The second curve is a normal density with the same expectation and variance as the paired model. The effect is clearly visible when $p = 1/4$ of individuals are affected, receiving the maximum of $m = 500$ observations from $\Phi(.)$

Ordinary ranks compare units two ($m = 2$) at a time. Among I untied units, a unit has rank k if it is the larger unit in $k-1$ of the $I-1$ possible comparisons with another unit. Stephenson's ranks compare units not two at a time but m at a time. Consider all $\binom{I}{m}$ comparisons of m units. Stephenson (1981) asked: In how many subsets of m units is unit i the largest? It is easy to check that among I untied units, a unit with conventional rank k is largest in $\binom{k-1}{m-1}$ subsets of size m, where $\binom{a}{b}$ is defined to equal zero for $a < b$, so Stephenson's rank is $\binom{k-1}{m-1}$.

In Stephenson's (1981) generalization, $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$, of Wilcoxon's signed rank statistic, if the i pair has k^{th} largest of the I untied absolute differences, $\mathbf{A} = (|Y_1|, \ldots, |Y_I|)^T$, then it is assigned rank $q_i = q_i(\mathbf{A}, \mathbf{d}) = \binom{k-1}{m-1}$. These ranks are relatively flat for small $|Y_i|$, but then rise steeply for large $|Y_i|$.

In Rosenbaum (2007), in data from two observational studies, an inference based on Stephenson's ranks was noticeably less sensitive to unobserved biases than an inference based on Wilcoxon's ranks. In those two examples, it seemed plausible that some treated subjects were strongly affected by the treatment, while others experienced little or no effect. How is the design sensitivity affected by uncommon but dramatic treatment effects?

6.6.3 Design Sensitivity for Large but Rare Effects

The design sensitivity will now be calculated under Salsburg's paired model with exponent $\overline{m}$ and mixing proportion p for the Wilcoxon and Stephenson tests with subsets of size m. Notice that $\overline{m}$ and m may differ, because the investigator picks m for use in the test not knowing $\overline{m}$ in the model.

Recall that design sensitivity is calculated in the favorable situation in which the treatment has an effect and there is no bias from the unobserved covariate u, so that $Y_i = Z_{i1}(r_{Ti1} - r_{Ci2}) + Z_{i2}(r_{Ti2} - r_{Ci1})$ where $Z_{i2} = 1 - Z_{i1}$ and $Pr(Z_{i1} = 1 \,|\, F) = \frac{1}{2}$.

Calculating the design sensitivity requires solving (6.9). Using Stephenson's signed rank statistic, $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} sgn(Y_i)\, q_i$, the ranks q_i are some permutation of $\binom{i-1}{m-1}$ with $\binom{i-1}{m-1} = 0$ for $i < m$, where obviously $\sum_{i=1}^{I} q_i = \sum_{i=m} \binom{i-1}{m-1} = \binom{I}{m}$. From this it follows that $E\left(\overline{\overline{T}}_\Gamma\right) = \{\Gamma/(1+\Gamma)\} \sum_{i=1}^{I} q_i = \Gamma\binom{I}{m}/(1+\Gamma)$. If the Y_i were sampled independently from some distribution $H(\cdot)$, define ζ_m as the probability that, in a sample of m independent observations from $H(\cdot)$, the observation with the largest absolute value has positive sign. For $m = 2$, in the notation for Wilcoxon's statistic in Sect. 6.2.5, $\zeta_2 = p_1'$. In the favorable situation, Stephenson's test statistic is the number of subsets of m observations in which the observation with the largest absolute value has positive sign, so $E\{t(\mathbf{Z}, \mathbf{R})\} = \zeta_m\binom{I}{m}$. Equation (6.9) is then

$$\frac{\Gamma}{1+\Gamma}\binom{I}{m} = \zeta_m\binom{I}{m}, \text{ with solution } \widetilde{\Gamma} = \frac{\zeta_m}{1-\zeta_m}, \tag{6.12}$$

in agreement with the result for $m = 2$ for Wilcoxon's statistic, where $\widetilde{\Gamma} = p_1'/(1-p_1')$.

Under Salsburg's paired model for r_{Tij} and r_{Cij}, the Y_i are I independent observations formed as the difference of independent observations from $(1-p)F \times (\cdot) + pF^{\overline{m}}(\cdot)$ and $F(\cdot)$, because differencing removes the pair parameters ξ_i. For specified p, $\overline{m}$, and $F(\cdot)$, the distribution of Y is straightforwardly, albeit perhaps numerically, obtained as convolution of two specified distributions. Figure 6.5 depicted the corresponding density in the normal case, $F(\cdot) = \Phi(\cdot)$. Calculating ζ_m for this distribution and using (6.12) produces Table 6.8.

Table 6.8 Design sensitivity for Stephenson's test with subsets of size m applied to Salsburg's paired model with $r_{Cij} \sim \Phi(\cdot)$ and $r_{Tij} \sim (1-p)\Phi(\cdot) + p\Phi^{\overline{m}}(\cdot)$ where $p = .25$ and $\Phi(\cdot)$ is the standard normal cumulative distribution

	Wilcoxon ($m = 2$)	Stephenson ($m = 5$)	Stephenson ($m = 10$)
$\overline{m} = 5$	1.6	1.8	2.0
$\overline{m} = 10$	1.8	2.2	2.5
$\overline{m} = 100$	2.2	3.6	5.5
$\overline{m} = 500$	2.4	4.7	8.9

Table 6.8 gives the design sensitivity, $\widetilde{\Gamma}$, for the sampling distributions depicted in Fig. 6.5. For $\overline{m} = 100$ or $\overline{m} = 500$, which differ visibly from the normal distribution in Fig. 6.5, the use of Wilcoxon ranks yields a much lower value of the design sensitivity, $\widetilde{\Gamma}$, than use of Stephenson's ranks with $m = 5$ or $m = 10$.

In short, if the treatment has no effect on many treated subjects but a dramatic effect on some subjects, then Wilcoxon's test may judge the results sensitive to smaller biases than will Stephenson's test with $m > 2$. When a treatment strongly affects a small fraction of treated subjects, it is important to use methods of analysis that can discern this pattern.

6.7 Sample Splitting As an Aid to Design

Sections 6.3–6.6 have shown that decisions about the design of an observational study strongly affect the study's sensitivity to biases from covariates that were not measured. Unfortunately, some of these decisions depend on issues that will be difficult to evaluate in the absence of data. Can unit heterogeneity be reduced by focusing on a subpopulation? Is the treatment effect much larger when the dose is larger? Would it be advantageous to look for coherence among several outcomes, and if so, which outcomes should be used? Does the treatment affect everyone about equally, or are the effects of the treatment confined to a subpopulation that cannot be identified in advance?

Heller et al. (2009) evaluated the splitting of a study into a small planning sample and a large analysis sample. Decisions about design are guided by the planning sample, which is then discarded, and a statistically independent analysis is based on the analysis sample. Heller et al. found that when it is possible to materially reduce sensitivity to unobserved biases through appropriate design decisions, it is often the case that a small planning sample correctly guides those decisions. The intuitive reason for this occurrence is that only dramatic issues can dramatically affect sensitivity to bias, and issues of that sort are often apparent even in a very small planning sample, perhaps 10% of the full sample. The technical reason is that the design sensitivity is unaffected by discarding a small analysis sample: In large samples, the power of the sensitivity analysis with $\Gamma > 1$ is determined by the design sensitivity, $\widetilde{\Gamma}$, not by the sample size, so increasing $\widetilde{\Gamma}$ is all important.

6.8 Summary

A research design is a decision to collect data in a particular way, which might be formalized as a decision to draw data from one data generating process rather than another. If several such processes were available, then which one is least sensitive to the claim that its results are not indicative of a treatment effect, but instead reflect a bias from failure to adjust for some unmeasured covariate? The design sensitivity

$\widetilde{\Gamma}$, is a single number that attaches to a data generating process and method of analysis, and it provides an answer when the sample size is large. Several design choices have been shown to influence the design sensitivity.

Acknowledgement This work was supported by grant SES-0849370 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation.

References

Aakvik, A. (2001). Bounding a matching estimator: the case of a Norwegian training program. *Oxford Bulletin of Economics and Statistics, 63*, 115–143.

Ahmed, A., Allman, R. M., Fonarow, G. C., Love, T. E., Zannad, F., Dell'Italia, L. J., et al. (2008). Incident heart failure hospitalization and subsequent mortality in chronic heart failure: A propensity-matched study. *Journal of Cardiac Failure, 14*, 211–218.

Ashenfelter, O., & Rouse, C. (1998). Income, schooling and ability: Evidence from a new sample of identical twins. *Quarterly Journal of Economics, 113*, 253–284.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Campbell, D. T. (1988). *Methodology and epistemology for social science*. Chicago, IL: University of Chicago Press.

Cleveland, W. W. (1994). *Elements of graphing data*. Summit, NJ: Hobart.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, A, 128*, 134–155.

Conover, W. J., & Salsburg, D. S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. *Biometrics, 44*, 189–196.

Copas, J., & Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, B, 63*, 871–896.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., & Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute, 22*, 173–203.

Diprete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects. *Sociological Methodology, 34*, 271–310.

Fisher, R. A. (1935). *Design of experiments*. Edinburgh, UK: Oliver and Boyd.

Foster, E. M., Wiley-Exley, E., & Bickman, L. (2009). Old wine in new skins: The sensitivity of established findings to new methods. *Evaluation Review, 33*, 281–306.

Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics, 33*, 19–34.

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika, 85*, 907–920.

Hammond, E. C. (1964). Smoking in relation to mortality and morbidity. *Journal of the National Cancer Institute, 32*, 1161–1188.

Heller, R., Rosenbaum, P. R., & Small, D. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association, 104*, 1090–1101.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*, 295–300.

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on ranks. *Annals of Mathematical Statistics, 34*, 598–611.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, 93*, 126–132.

Isacsson, G. (2007). Twin data vs longitudinal data to control for unobserved variables in earnings functions. *Oxford Bulletin of Economics and Statistics, 69*, 339–362.

Jick, H., Miettinen, O., Neff, R., Shapiro, S., Heinonen, O. P., & Sloan, D. (1973). Coffee and myocardial infarction. *New England Journal of Medicine, 289*, 63–77.

Lehmann, E. L. (1953). The power of rank tests. *Annals of Mathematical Statistics, 24*, 23–43.

Lehmann, E. L. (1998). *Nonparametrics*. Upper Saddle River, NJ: Prentice Hall.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review, 80*, 319–323.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13*, 151–161.

Mill, J. S. (1867). *A system of logic: The principles of evidence and the methods of scientific investigation*. Indianapolis, IN: Liberty Fund.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statistical Science, 5*, 463–480.

Norvell, D. C., & Cummings, P. (2002). Association of helmet use with death in motorcycle crashes: A matched-pair cohort study. *American Journal of Epidemiology, 156*, 483–487.

Origo, F. (2009). Flexible pay, firm performance and the role of unions. New evidence from Italy. *Labour Economics, 16*, 64–78.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., et al. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. *British Journal of Cancer, 34*, 585–612.

Pinto, D., Ceballos, J. M., Garcia, G., Guzman, P., Del Razo, L. M., Vera, E., et al. (2000). Increased cytogenetic damage in outdoor painter. *Mutation Research, 467*, 105–111.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11*, 691–714.

Robins, J. M., Rotnitzky, A., & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference. In E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology* (pp. 1–94). New York, NY: Springer.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika, 74*, 13–26.

Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association, 88*, 1250–1253.

Rosenbaum, P. R. (1997). Signed rank statistics for coherent predictions. *Biometrics, 53*, 556–566.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies (with discussion). *Statistical Science, 14*, 259–304.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R. (2003a). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics, 4*, 1–10.

Rosenbaum, P. R. (2003b). Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician, 57*, 132–138.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika, 91*, 153–164.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician, 59*, 147–152.

Rosenbaum, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics, 63*, 1164–1171.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, B, 45*, 212–218.

Rosenbaum, P. R., & Silber, J. H. (2009). Amplification of sensitivity analysis in observational studies. *Journal of the American Statistical Association, 104*, 1398–1405.

Rothman, K. J. (1986). *Modern epidemiology*. Boston, MA: Little, Brown.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Rutter, M. (Ed.). (2007). *Identifying the environmental causes of disease: How should we decide what to believe and when to take action?* London, UK: Academy of Medical Sciences.

Salsburg, D. S. (1986). Alternative hypotheses for the effects of drugs in small-scale clinical studies. *Biometrics, 42*, 671–674.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Chen, W., Zhang, X., Lorch, S., et al. (2005). Preoperative antibiotics and mortality in the elderly. *Annals of Surgery, 242*, 107–114.

Slade, E. P., Stuart, E. A., Alkever, D. S. S., Karakus, M., Green, K. M., & Ialongo, N. (2008). Impacts of age of onset of substance use disorders on risk of adult incarceration among disadvantaged urban youth. *Drug and Alcohol Dependence, 95*, 1–13.

Small, D., & Rosenbaum, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association, 103*, 924–933.

Stephenson, W. R. (1981). A general class of one-sample nonparametric test statistics based on subsamples. *Journal of the American Statistical Association, 76*, 960–966.

Trochim, W. M. K. (1985). Pattern matching, validity and conceptualization in program evaluation. *Evaluation Review, 9*, 575–604.

van Eeden, C. (1972). An analogue, for signed rank statistics, of Jureckova's asymptotic linearity theorem for rank statistics. *Annals of Mathematical Statistics, 43*, 791–802.

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomized experiments? *Lancet, 363*, 1728–1731.

Wang, L. S., & Krieger, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine, 25*, 2257–2271.

Weed, D. L. (1997). On the use of causal criteria. *International Journal of Epidemiology, 26*, 1137–1141.

Weiss, N. (1981). Inferring causal relationships: Elaboration of the criterion of "dose-response". *American Journal of Epidemiology, 113*, 487–490.

Weiss, N. (2002). Can the 'specificity' of an association be reahbilitated as a basis for supporting a causal hypothesis? *Epidemiology, 13*, 6–8.

Wright, P. H., & Robertson, L. S. (1976). Priorities for roadside hazard modification. *Traffic Engineering, 46*, 24–30.

Chapter 7
The Origins of Procedures for Using Differential Item Functioning Statistics at Educational Testing Service

Michael J. Zieky

7.1 Introduction

The chapter deals with the procedures that ETS developed to use the differential item functioning (DIF) statistic, not the statistic itself. To establish the context for the origins of the use of DIF at ETS, I first explain the Golden Rule Insurance Company settlement and its effects. Then I very briefly describe the precursors of DIF in use at ETS from the late 1960s to the mid 1980s. As those who know ETS might expect, decisions about the use of DIF were made by a committee. I describe the members and responsibilities of that committee.

In the remainder of the chapter, I describe some of the most divisive decisions that we made about how to use DIF at ETS. I do not discuss decisions about which there was early and complete agreement, even though some of those decisions were quite important, such as which groups to study. The focus is on the decisions that we argued about for extended periods, beginning with what to call the statistic. Other decisions I discuss in the chapter are which DIF statistics to use, whether to use impact data, what the criteria for flagging items should be, what sample sizes would require the calculation of DIF, the acceptability of balancing positive and negative DIF in a test, and how to treat items that favor groups commonly thought of as disadvantaged.

M.J. Zieky (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: mzieky@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_7,

7.2 Disclaimer

Because I was a participant in the events I describe in this chapter, I can make no claims to being an objective historian. I tried to be neutral in describing disagreements among committee members, even though I participated in those disagreements.[1]

7.3 Golden Rule

It is impossible to understand the urgency that impelled the development of DIF at ETS without knowledge of the events surrounding the lawsuit brought against the Illinois Department of Insurance and ETS by the unfortunately named Golden Rule Insurance Company.[2] The lawsuit began in 1976 alleging, based on differences in passing rates between African American and White test takers, that the test constructed by ETS to license insurance agents in Illinois was racially discriminatory.

Eight years later, in January of 1984, ETS and the Golden Rule Insurance Company were still engaged in pretrial legal maneuvering. The pretrial work was very expensive, and it consumed an inordinate amount of staff time that had to be diverted from more substantive work. Based on his interest in fair assessment and in an effort to stop a wasteful use of ETS resources, Gregory Anrig, then president of ETS, arranged to meet with J. Patrick Rooney, president of the Golden Rule Insurance Company. At the meeting, Anrig and Rooney agreed to instruct their attorneys to reach a settlement.

The parties agreed to a voluntary, out-of-court settlement in November of 1984. The most relevant aspects of the settlement for this chapter were that test items had to be divided into two groups: (a) items in which correct answer rates (percent correct) between Black and White test takers differed by no more than 15 percentage points, and (b) items in which correct answer rates between Black and White test takers differed by more than 15 percentage points. Test specifications were to be met, to the extent possible, by items in the first group. If two or more items were available to meet a test specification, the item with the smallest Black–White difference in percent correct must be used first. No item that was harder than 40% correct for either group could be used in a test.

The so-called Golden Rule Settlement was based on raw differences in percent correct on test items and paid no attention to the causes of those differences. Any difference in percent correct was considered bias even if it were caused by real and relevant differences between the groups in average knowledge of the tested subject.

[1] If readers cannot tell which side I was on, I have done my reporting job well.

[2] There is a built-in headwind when you try to argue against "the Golden Rule."

The Golden Rule procedure ignored even the possibility that the differences in percent correct could be valid.

The items that correlated highest with the total test score, that were the best at differentiating among test takers at different levels of knowledge and skill, and that were considered the best items by testing professionals and subject-matter experts were the items most likely to be eliminated by use of the Golden Rule procedures. Furthermore, all difficult items – not needed in a minimal-competency licensing test for insurance agents but crucial in tests designed for many other purposes such as selective admissions – were excluded.

For those reasons, ETS *never* believed that the Golden Rule procedure was a reasonable way to identify or to deal with test bias in general. However, ETS researchers (including Holland) demonstrated that acceptable tests could still be made using the Golden Rule procedures in the very limited context of a subset of the licensing tests that measured the minimal competency required to become an insurance agent in Illinois. Because the testing program could continue with reasonable quality, because the terms of the settlement were favorable, and because 8 years had already been spent in costly and time-consuming pretrial work, ETS agreed to the terms of the settlement.

ETS was (and, I believe, still is) far less skilled at politics than it is in psychometrics. ETS completely overlooked the possibility that Rooney would try to expand the Golden Rule settlement beyond its original, quite limited application. Rooney, however, funded a group called FairTest that vigorously lobbied legislators to mandate the Golden Rule procedure for *all* tests, regardless of their purposes. The gist of FairTest's argument was that ETS admitted that some of its tests were biased solely on the basis of Black–White score differences. ETS agreed to "de-bias" such tests using the Golden Rule procedure. According to FairTest, ETS had to use the same procedures for all of its tests that show Black–White differences in scores. FairTest alleged that if ETS failed to do so, it would be making biased tests on purpose.

A procedure that caused little harm in a minimal-competency licensing test for insurance agents would be devastating if applied to admissions tests for higher education such as the SAT® or GRE®, or even to licensing tests in more cognitively loaded areas such as teaching. Professional measurement organizations such as the National Council for Measurement in Education opposed the Golden Rule procedure. President Anrig acknowledged publicly that accepting the Golden Rule settlement had been an "error in judgment."

ETS, however, could not just renounce the Golden Rule procedure and do nothing in its place. For people without knowledge of measurement (almost all of the voting public and probably all legislators), the Golden Rule procedure had great allure. If two items measured the same specification, why not use the item with smaller Black–White differences? ETS needed an acceptable statistic to replace the raw differences in percent correct used in the Golden Rule procedure. If ETS had nothing to offer legislators in place of the Golden Rule procedure, test quality might be severely compromised by legislation that prohibited the use of difficult items and mandated the use of items with the smallest raw differences in

percent correct between Black and White test takers, regardless of the causes of those differences. The pressure to develop a psychometrically acceptable statistic to help ensure that tests were fair was intense and growing stronger by the day.

7.4 Precursors to Differential Item Functioning

ETS had a long tradition of searching for an empirical measure of fairness before DIF statistics were available. With the exception of item response theory (Lord, 1980), the statistical methods used to evaluate items for fairness in the 1960s, 1970s, and early 1980s are no longer used for that purpose. Therefore, I provide brief descriptions for readers who may not be familiar with the early techniques.

ETS researchers used analysis of variance techniques to search for interactions between performance on test items and group membership in the 1960s and 1970s (Angoff & Ford, 1973; Cardall & Coffman, 1964). This effort was done on a research basis, however, and was too cumbersome and difficult to interpret to be applied operationally to all items.

ETS also used delta plots to find items that might be unfair (Angoff, 1982). Delta is a measure of item difficulty in which percent correct for a question is expressed as a normal deviate. In plots of the deltas of a set of items for two groups, most items would fall very close to a straight line even if the items were much harder for one group than for the other. Some items, however, fell off the line because they had greater or lesser differences between groups than most items in the test had. The delta plots were easy to produce and subject-matter specialists found them easy to use. Falling off the line was not proof of bias, however. In fact, the best items in terms of high item-test correlations tended to be furthest away from the line formed by most items. Therefore, the delta plot technique was eventually discarded.

ETS investigated a form of the chi square statistic as an indicator of item fairness (Scheuneman, 1979). Methodological problems and arguments about whether the procedure was a true chi square statistic caused the method to fall out of favor, however.

At ETS, the way had been prepared for the use of a measure of DIF in the early to mid 1980s by work stretching back to the 1960s. The goal was an empirical means of distinguishing between real group differences in the knowledge and skill measured by the test and unfair differences inadvertently caused by biased aspects of items. Test developers wanted help in ensuring that items were fair, but each method tried so far either had methodological difficulties or was too unwieldy to use on an operational basis with a wide variety of tests and several groups of test takers. The threat of legislation that would mandate use of the Golden Rule procedure for all tests further motivated ETS staff members to adopt a practical measure of DIF. Clearly, ETS was primed to accept and implement a measure of DIF.

7.5 The Committee

In 1983, ETS Executive Vice President Robert Solomon, following the recommendation of Vice President of Research Samuel Messick, established an ad hoc committee to improve test development and statistical analysis (henceforth, the committee). The members of the committee at that time were Solomon, the directors of the three test development divisions at ETS (Al Carlson, Ernie Kimmel, and Cheryl Wild),[3] the directors of the three statistical analysis divisions (Al Carlson,[4] Gary Marco, and Nancy Petersen), the director of corporate development (Jerry Murphy), the director of systems (Barbara Foltin), and three researchers (Paul Holland, Samuel Messick, and Warren Willingham). Neil Dorans collaborated closely with the committee on DIF. I chaired the committee.

Among the tasks that Solomon assigned to the committee were increasing the professionalism of test developers and statisticians, improving training for test developers and statisticians, updating the ETS procedures for equating, improving item analysis and test analysis, writing guidelines for the development and use of constructed response testing, incorporating the use of item response theory in test development, and responding to proposed antitesting legislation.

The committee contained the lead test developers, statisticians, and systems analyst, and several prominent researchers including Holland – the people necessary to make decisions about the operational calculation and use of DIF. Therefore, when measures of DIF were to be incorporated into operational test development, Solomon gave the committee the tasks of (a) selecting the statistic, (b) developing statistical analysis procedures for calculating the statistic, (c) devising procedures for using the statistic in test development, and (d) disseminating the results of using the statistic.

The use of DIF has become a common and expected part of test development at ETS and at many other large-scale test publishers. It no longer generates excitement and controversy. At its inception, however, the use of DIF raised many difficult issues and engendered many intense arguments among the people given the responsibility of turning a measure of DIF into an operational reality. There are few left who recall the disagreements, debates, and disputes that preceded many of the decisions.[5] The remainder of the chapter will discuss some of the more controversial issues in developing procedures for the use of DIF at ETS.

[3] The three test development divisions and the three statistical analysis divisions included one division for all tests owned by the College Board, one for occupational tests, and one for K–12 and higher education tests not associated with the College Board. ETS is no longer organized in that way.

[4] Carlson served a dual role.

[5] Of the 12 people on the committee in 1983, all but one have died, retired, or left ETS.

7.6 Nomenclature

We could not even agree on what to call the statistic. The phrase *differential item functioning* and the acronym *DIF* have become so entrenched in the psychometric literature, and have become so common in the jargon of testing professionals, that it is difficult to conceive of a time when people argued about what name to use. The preferred phrase and acronym as late as 1986 were *differential item performance* and *DIP*. Some people (e.g., Dorans & Kulick, 1983, 1986) used the term *unexpected differential item performance* or *UDIP* to highlight the focus on differences in item performance that were greater than the differences that could be expected based on construct-related knowledge.

DIP fell out of favor because some of the extended uses of the acronym seemed too frivolous for the serious topic of item fairness. For example, "UDIP" spoken aloud sounded like an insult. Items that showed elevated values of differential item performance were referred to as "dippy." A tentative title for a chapter on DIP spoke of "doing the DIP" as a dance step. The executive vice president was not pleased.

An interim usage in place of DIP was *differential item difficulty*. It was simple, descriptive, and alliterative, but the acronym *DID* was difficult to use unambiguously in speech. ("We did DID.") Eventually Holland and Dorans proposed *differential item functioning* because it placed the emphasis on how the item functioned, rather than on how people performed. The focus on the item made sense to the members of the committee. Furthermore, the acronym *DIF* was distinctive, had the advantage of replicating the first syllable of *differential*, and was easy to use unambiguously in speech and writing. That name and the associated acronym were adopted by ETS and eventually by the entire testing community.

7.7 Selecting a Differential Item Functioning Statistic

Ironically, after so many years when a major problem at ETS was finding a usable DIF statistic, in the mid 1980s the committee was faced with the problem of deciding which of two good measures of DIF to adopt. The two choices were the standardization approach (Dorans & Kulick, 1983, 1986) that had become available in 1983 and an adaptation of the Mantel-Haenszel statistic (Holland & Thayer, 1986) that became available in 1985.

The technical differences between the two approaches are explained well in Dorans (1989) and will not be reiterated here. Few members of the committee were able to evaluate the technical adequacy of the two measures. The knowledgeable few assured the less technically sophisticated majority that the measures had different strengths and weaknesses, but that both were perfectly acceptable. The standardization measure was easier for test developers and subject matter

specialists to understand. It was also more useful for test developers because it could provide information on the relative attractiveness of the distracters in a multiple-choice item for test takers in matched groups. The adaptation of the Mantel-Haenszel statistic, however, had some better statistical properties.

Holland transformed the Mantel-Haenszel statistic, which was in the form of an odds ratio, into a difference on the delta scale of item difficulty commonly used at ETS. That transformation, called the Mantel-Haenszel Delta Difference (MH D DIF) was much more meaningful to test developers than was the original odds ratio, and it removed a major concern about the use of the Mantel-Haenszel statistic by test developers.

After extended but largely fruitless discussions, the committee as a whole failed to reach a decision about which measure of DIF to use. A more technically sophisticated subcommittee was formed (Holland, Marco, Petersen, and Messick, joined by Dorans). The subcommittee recommended that *both* DIF statistics be used operationally. Despite an early fear that test developers would suffer from information overload, both statistics remain in operational use to this day. The MH D DIF data are used to flag items that meet certain criteria, and the standardization data are used to help test developers decide what aspects of the items are possibly unfair, thus capitalizing on the strengths of each statistic.

7.8 Use of Impact Data

One of the early arguments about using DIF was whether or not to include impact data (raw differences in percent correct) in making decisions about items. No member of the committee believed that impact should be the sole criterion for identifying unfair items as in the Golden Rule procedure. The whole point of using DIF statistics was that impact conflated real, construct-related group differences with differences caused by potentially unfair aspects of items.

Some members of the committee, however, felt that among items with the same level of DIF, it made sense to use the items with smaller impact. They proposed, for example, dividing items into three levels of DIF and three levels of impact. The low DIF items should be used before the medium DIF items, and the medium DIF items should be used before the high DIF items. However, within each DIF group, the low impact items should be used before the medium impact items, and the medium impact items should be used before the high impact items. Thus, impact would clearly be secondary to DIF as a criterion for item selection, but impact would be allowed to influence item selection.

The committee, however, decided not to allow impact to have any influence on item selection because impact was simply not an appropriate measure of fairness.

7.9 Flagging Criteria

Now, many statisticians and test developers at ETS make daily use of the criteria by which items become flagged by MH D DIF. They take the criteria as a given and do not question their origins. Some researchers outside of ETS have adopted the ETS criteria as though they carried some intrinsic meaning. In reality, the criteria were admittedly arbitrary and were set by the committee following many arguments about the best way to use the data. The criteria were finally based on the size of the difference in Delta that test developers and statisticians judged to be meaningful based on their experience in working with items.

The criteria were set on a tentative basis to be revised as experience with the consequences of their use was gained over time. Almost a quarter of a century later, however, the criteria are unchanged and still in operational use![6] The flagging criteria that were finally adopted were based solely on the magnitude and significance of MH D DIF. (See Zieky, 1993, for a full description of the flagging criteria.)

Several models were proposed for using DIF data. The *strict cutscore* model set a single cutscore on MH D DIF. Below that point, items could be used freely and no attention would be paid to the amount of DIF. Above that point, items could not be used. The advantage would be quick and cheap decision making. Items would automatically be placed in go or no-go categories. There are, however, major disadvantages to the strict cutscore model. Any cutscore is arbitrary. Items adjacent to the cutscore on either side are very similar but would be treated very differently even if professional judgment found the item just above the cutscore to be superior to the item just below the cutscore.

The *flag and review* model set a single cutscore on MH D DIF. Below that point, items could be used freely and no attention would be paid to the amount of DIF. Above that point, items would be reviewed. Items that passed the review would be treated as though they belonged in the first category. The advantage of the flag and review model over the strict cutscore model is that acceptable items above the cutscore could be used after they passed review. As in the strict cutscore model, however, any cutscore is arbitrary. The review process adds time and costs to test development, and reviewers may not always make appropriate decisions about which items are acceptable.

The *graded* model set multiple cutscores to put items in several categories such as low, medium, and high DIF. Items in categories with the least DIF would be used to meet specifications. Items in higher DIF categories would be used only when necessary to meet specifications. The advantage is that the items in the lowest DIF categories would be used to assemble tests to the extent possible given the pool of

[6] Either the committee was remarkably prescient in selecting the optimal flagging criteria, or corporate decisions that do not lead to visible disasters tend not to be revisited in the press of all of the new decisions that must be made.

available items. The disadvantage is that multiple arbitrary cutscores must be set. Furthermore, the task of test developers becomes more complicated, particularly if several group comparisons are made on the same set of items. (For example, comparisons with White test takers could be made with African American, Asian American, Hispanic American, and Native American test takers in addition to comparisons between male and female test takers.)

In the *biserial* model, test developers would treat the DIF index as they treat the biserial correlation of the item with the test. Rules of thumb rather than strict cutscores are used. The rules may vary by subject matter and by test program. The index is weighed along with other item qualities in selecting items. The biserial model avoids arbitrary cutscores. It maximizes the role of professional judgment and minimizes mechanical constraints. The advantages lead directly to the disadvantages of increased time and expense in test assembly and increased time in resolution of disputes among test assemblers and reviewers who made different judgments about which items were most appropriate.

The model that the committee finally chose for test assembly when pretested items had DIF data is a combination of the graded model and the flag and review model. Items are divided into three categories on the basis of MH D DIF. The categories are essentially small, medium, and large DIF, labeled A, B, and C respectively. Items from group A are used before items from group B. If specifications cannot be met with group A items, then group B items may be used with preference given to the items with the smallest absolute values of DIF. Group C items may not be used unless necessary to meet specifications and the items have been reviewed and judged to be fair.

The model that the committee chose for use when DIF data are available only after an operational administration of the test is the flag and review model. A panel of diverse and disinterested people reviews any category C items flagged after a test administration. The flagged items are removed from scoring unless the panel judges them to be fair. Automatic deletion of the category C items was rejected because test takers had spent time responding to the items during an operational administration. That time should not be retroactively wasted if the items were judged to be perfectly acceptable.

Despite several intensive reviews of the DIF procedures in the following years, those models have remained in constant use for close to a quarter of a century with no revision.

7.10 Sample Sizes

The committee had to develop rules for when testing programs would have to use DIF. Now DIF is as routine as is item analysis, and no ETS program questions its necessity. When DIF was first introduced, however, testing programs were concerned about the additional expenses they would incur and about the additional

strain on their schedules. In addition, in the absence of any experience with DIF, programs feared that they would lose many good items.

Because of those concerns, programs hesitated to begin operational use of the DIF statistic. The committee, therefore, decided to set minimum sample sizes of test takers in the comparison groups that would trigger mandatory use of DIF. Programs would be free to do more DIF studies than were mandated, but they could not avoid the mandated DIF studies.

The committee members split into two groups. The statisticians and the statistically minded members wanted relatively large sample sizes to be set as triggers for DIF studies to reduce the amount of random error in the DIF statistic. An unstable statistic would be misleading, costly, and time consuming.

The people arguing for smaller samples were concerned that requiring large sample sizes would limit the use of DIF to only the largest testing programs for any comparisons except male–female. They believed that the increased random error of small sample sizes was acceptable because they were willing to accept the cost of a *good* (truly low DIF) item being mistakenly labeled as *bad* (high DIF) to obtain the benefit of DIF data on all items. The other type of error, a *bad* (truly high DIF) item being mistakenly labeled as *good* (low DIF), was seen as less of a problem than if DIF were not calculated at all, and all items were just assumed to be good.

The small sample size was adopted. Interestingly, the committee's decision about implementing DIF was one of the few that was later changed. The minimum sample sizes required to trigger the mandatory use of DIF were raised about 10 years later to increase the stability of the DIF statistics.

7.11 Balancing DIF

Committee members argued about whether to allow positive DIF items to offset negative DIF items in a test.[7] The clear advantage of allowing positive and negative items to balance one another was that more items would be available for use in a test. Some committee members believed that a test would be fair if the algebraic sum of DIF were close to zero.

One practical problem with allowing positive DIF to balance negative DIF is that DIF was routinely calculated for as many as five groups (African American, Asian American, Hispanic American, Native American, and female test takers) whenever sample sizes were sufficient. DIF results are not identical across groups. Balancing a negative DIF item for females with a positive DIF item for females might upset the balance between positive and negative DIF for Black test takers, for example.

[7] Positive DIF items favor the so-called *focal groups* – African American, Asian American, Hispanic American, Native American, and female test takers – over matched members of the so-called *reference groups* – White and Male test takers. Negative DIF items favor the reference groups over matched members of the focal groups.

Allowing balance would greatly complicate test assembly by forcing test developers to manipulate and to keep track of the algebraic sum of DIF for as many as five groups at once.[8]

A more substantive argument against allowing negative DIF to balance positive DIF was that DIF in either direction meant that the item was measuring something other than, or in addition to, the intended construct. The way to make a fair test was to keep the DIF for all items close to zero, not to use one departure from the construct to offset another departure from the construct. This view was most strongly defended by Messick, and many of the committee members were convinced by his arguments.

7.12 Positive and Negative DIF

The issue that caused more disagreement than any other among members of the committee was how to treat items with positive DIF. This issue is one of the few related to DIF that can still engender arguments among ETS staff.

What could be called the *psychometric* point of view was that DIF in either direction (positive or negative) was caused by departure from measurement of the intended construct. Adherents of that point of view argued for treating positive and negative DIF symmetrically. If an item with negative DIF was potentially unfair for the members of the focal group, then an item with positive DIF was potentially unfair for members of the reference group.

The *affirmative action* point of view was that positive DIF should be treated differently than negative DIF. The adherents of asymmetry in the use of DIF believed it inappropriate to find an item that favors focal group members and then not use the item.

ETS adopted the psychometrically sound symmetrical DIF policy. In practice, however, the members of the panels charged with reviewing items found to have high DIF after an operational administration very rarely find a positive DIF item to be unfair. The effect is to retain most items with positive DIF when the DIF data are obtained after an operational administration.

7.13 Conclusion

In 1986, DIF was used operationally for the first time. DIF was first used with the National Teacher Examination, a licensing test for aspiring teachers.[9] Since that time, DIF has been applied to almost every ETS test with sufficient numbers

[8] The widespread use of computers in test assembly that would make the balancing task manageable across multiple groups was still in the future.

[9] The NTE has since been replaced by Praxis™.

of test takers.[10] The use of Holland's MH D DIF and Dorans and Kulick's standardized P difference have become routine and accepted as a normal aspect of the test development process.

Few of the test developers who were at ETS when the DIF procedures were first instituted are still at work. For most test developers, the DIF procedures have been in place and have remained more or less constant ever since they were hired. (About half of ETS employees have been with the company for less than 5 years.) The decisions that were highly controversial in the mid 1980s are now accepted as a matter of course, as though no alternatives had ever been considered.

In short, the controversial decisions about the implementation of DIF made by the members of the committee with Doran's help close to a quarter of a century ago may not have been optimal, but they have endured and have become widely accepted and traditional.

This chapter focused on the development of procedures for the use of DIF at ETS, but we should keep in mind that DIF was just one of Holland's many contributions to the science of measurement.[11] It is entirely fitting that we honor the work of Paul Holland, who changed the way that tests are developed at ETS and at many other testing organizations and who contributed greatly to the fairness of tests taken by millions of people.

Acknowledgement Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–106.

Cardall, C., & Coffman, W. (1964). *A method for comparing the performance of different groups on the same items of a test*. Princeton, NJ: ETS (ETS Research Bulletin. No. RB-64-61).

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217–233.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach*. Princeton, NJ: ETS (ETS Research Rep. No. RR-83-9).

[10] TOEFL® was exempted from routine use of DIF because linguistic and cultural differences within the normally studied groups would make the DIF results meaningless.

[11] Holland freely shared his knowledge with his colleagues. One of the greatest postdoctoral educational opportunities available at ETS was to listen to Messick and Holland discuss the wine to order at dinner.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. Princeton, NJ: ETS (ETS Research Rep. No. RR-86-89).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143–152.

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Part IV
Holland the Berkeley Professor

Why I Left ETS and Returned

Paul W. Holland

It might seem strange after 17 years of successful work at Educational Testing Service (ETS), that in 1993 Roberta and I pulled up stakes and moved to Berkeley, California. There I joined the University of California faculty, only to move back to Princeton, New Jersey, after 7 years and rejoin the ETS research staff. My explanation is that in 1990 I was 50 years old and had been at ETS since 1975. I did not want to wake up when I was 65 and realize that I had been at ETS for 30 years without consciously intending to do that. So after looking around at the possibilities, the UC Berkeley Graduate School of Education seemed like a good choice because Berkeley had a distinguished statistics department as well. The truth is that when I arrived at Berkeley, I knew most of the members of the statistics department and only one or two of the education school faculty. Of course, this changed quickly, and after a semester I began to teach the required quantitative courses for the school. This was a great experience for me because I got to meet almost all of the graduate students in their first years of study. They were a remarkable group of young scholars who seemed to have more energy that I ever had. Two of them, Derek Briggs and Ben Hansen, have papers in this volume. Their contributions here are connected in a distant way to the course on causal inference that I taught when they were at Berkeley. Other students whom I remember fondly from that time are Eva Ponte, Pamela Paek, Laura Goe, and Insu Paek. Pam, Laura, and Insu also worked or still work for ETS, so I have been able to see them in action as professionals as well as students.

I must also mention one faculty member with whom I worked while at Berkeley, the late Nadine Lambert. She was the first member of the graduate school to introduce herself to me, and her massive longitudinal data-set on children who had been diagnosed with attention deficit/hyperactivity disorder (ADHD) years earlier became part of the material that I used regularly in my statistics courses.

In late 1999, the opportunity arose for me to retire from Berkeley and simultaneously rejoin the ETS staff. The new position was named in honor of ETS's most distinguished former psychometrician, Fred Lord, whom I greatly admired, and was too great an opportunity for me to resist. In 2000 Roberta and I packed up and moved back to Princeton. This volume is a fitting ending to that move.

Chapter 8
Cause or Effect? Validating the Use of Tests for High-Stakes Inferences in Education

Derek C. Briggs

Casual comparisons inevitably initiate careless causal conclusions.

—Paul Holland, 2004

8.1 Introduction

A good aphorism can, in a few words, capture an essential truth. Of the many good aphorisms Paul Holland has coined over the years, I have found myself invoking the one above frequently enough to worry that I should be paying out royalty fees, so it is only fitting that I use it as the starting point for some ideas I wish to explore in this paper.

It is fairly common for people to use the graphical shorthand $Z \rightarrow X$ to represent the inference that a change in some variable Z causes a change in another variable X. Yet without further explication, this sort of presentation is causally ambiguous. In his seminal presentation of what he termed *Rubin's causal model* (also known as the potential outcomes model or the Neyman–Rubin model), Holland (1986) clarified the elements necessary to define and estimate a quantity interpretable as an average causal effect. These elements include the units of analysis, the specific treatments to which units may or may not be exposed, the potential outcomes as a function of treatment exposure, the mechanism by which units are exposed to treatment conditions, and the approach taken to estimate an unbiased average causal effect. In theory, the application of Rubin's causal model for the design and analysis of an experiment or quasi-experiment should serve as a safeguard against drawing

D.C. Briggs (✉)
School of Education, University of Colorado at Boulder,
249 UCB, Boulder, CO 80309, USA
e-mail: derek.briggs@colorado.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_8,

careless causal conclusions. However, this safeguard has an Achilles Heel in the context of its application in educational research: the often-equivocal nature of test scores as measures of cognitive outcomes.

Rubin's causal model is agnostic about the measurement properties of the test used to define these potential outcomes: The role of a test score is to provide the units through which the estimate of an average causal effect can be quantified. My contention is that in many circumstances a failure to think carefully about test validity will serve to undermine inferences about an estimated average causal effect, whether or not this effect is unbiased in the statistical sense laid out by Holland (1986).

As measures, not all outcomes are created equally. For example, death and income are common outcome measures in epidemiology and economics, and are relatively straightforward to validate. In educational research, cognitive outcomes are typically of interest, but such outcomes are unobservable. Cognitive outcomes are measured with standardized tests, and the match between test scores and their intended interpretation and use has spawned a dense literature in psychometrics under the umbrella term of validity theory. In this paper, I will be making an argument that at first glance appears either circular or paradoxical: Causal inference in educational research depends upon establishing test validity, but test validity depends upon establishing a causal inference. The reason I developed this argument is because I think it can serve the purpose of helping to bridge the gap between validity theory and practice in the context of high-stakes test use in education. That is, once we see that causal inference and test validity have a symbiotic relationship, it becomes possible to kill two birds with one stone: In estimating the effect or effects of educational interventions, we may also gain valuable insights about what it is that tests are (and are not) really measuring.

Four sections follow. In Sect. 8.2, I provide a policy context for the kinds of causal inferences being made about education in the wake of the No Child Left Behind Act of 2001 (NCLB). I suggest that a focus on making causal inferences that are internally valid has overshadowed the important role played by the choice of test outcome in causal generalization. In Sect. 8.3, I provide a brief overview of current conceptions in test validation theory and contrast this with current state-level practices. I reintroduce an idea dating back to at least Cronbach (1971) that test validity might be fruitfully evaluated through the lens of causal inference and experimental design. In Sect. 8.4, I elaborate upon a validation design that uses the real-world context of NCLB-mandated tutoring as the basis for an evaluation of the item level instructional sensitivity of large-scale assessments. In the last section, I offer concluding comments.

8.2 The Context of Causal Inference in Educational Research

It would be difficult to overstate the impact NCLB has had upon state systems of educational accountability since its implementation in 2002. The stipulations of NCLB required all schools receiving Title 1 funds to test their students annually

in the subjects of math, English/language arts, and science in grades 3 through 8 and at least once during high school. The performance of students within a given school (disaggregated by demographic subgroups) is then evaluated relative to criterion-referenced thresholds for each subject-specific test. Students are subsequently classified into performance levels (e.g., unsatisfactory, proficient, advanced). By the year 2012, a target was set that 100% of students should demonstrate test performance that would place them in the proficient category or higher. To this end, states were asked to specify target school-level percentages of students classified as proficient or higher each year leading to 2012. Each year, if a school's aggregate percentage is below the target percentage for any student subgroup or test subject, they will have failed to demonstrate *adequate yearly progress* (AYP). High-stakes sanctions are attached to the NCLB law. If a school fails to make AYP in 2 consecutive years, it must offer parents the opportunity to choose a different public school for their child to attend. After 3 years of failing to make AYP, supplemental educational services (i.e., tutoring) must be provided for all students eligible for free or reduced lunches. After 5 years of failing to make AYP, schools become candidates for restructuring by an external agency.

The extent to which NCLB has had a positive or negative impact on the American educational system is unclear. However, the law has achieved one important ancillary outcome: It has established a tremendous infrastructure for evaluating the causal effects of educational interventions. When NCLB was authorized in 2002, relatively few states tested grade 3 through 8 students annually in multiple subjects, and only 18 had a statewide identification system in place that could link students, their test scores, and their schools over time. Five years later by 2007, virtually all states were testing students in grades 3 through 8 in math, English/language arts, and science, and had a statewide student identification system. Combined with the use of the Internet as a means of transferring large quantities of data electronically in a timely and secure manner, the upshot is the availability of longitudinal data for research and evaluation purposes on a scale previously only possible through federally funded surveys conducted by organizations such as the Department of Education's National Center for Education Statistics.

The scores from the various standardized tests being administered from state to state are now being used to facilitate a host of evaluative studies. I want to distinguish two types of prevalent studies. The first type of study is an evaluation in which a specific educational intervention has been implemented; the second type is one in which pre-existing teachers and/or schools are themselves under evaluation as an educational intervention. In both cases, causal inference hinges upon the following question: What is the effect of a given intervention on one or more cognitive outcomes? The answer to such a question can have high-stakes ramifications: Curricula may be adopted or abandoned; teachers may receive salary increases or get fired. Given that the causal inferences are high-stakes, it is clearly important to get the magnitude and direction of effect estimates right. But it is just as important to make sure that appropriate test scores are being used as outcome measures. I next describe two empirical examples from published studies, one for each of the study types defined above, in which the choice of outcome measure led

to very different causal inferences about the effect of an educational intervention. In both these examples, I focus on the domain of mathematics proficiency in the middle schools grades, and I put to the side the issue of whether any given causal effect estimate is in fact unbiased.

8.2.1 *Evaluating the Effects of the Connected Mathematics Curriculum*

Beginning in the late 1980s, the National Council for Teachers of Mathematics (NCTM) published a series of documents describing new standards for how math should be taught at different grades. The standards called for a greater emphasis on knowing when and how to use mathematical skills and concepts to solve real world problems. The Connected Mathematics Project (CMP) was funded by the National Science Foundation (NSF) to develop a reform-based math curriculum for grades 6 through 8, as described by Ridgway, Zawojewski, and Hoover (2000):

> The CMP curriculum is organised around problem settings. Activities are designed to involve groups of students with mathematical concepts and applications, and in discourse and reflective writing about these same ideas. Students are expected to observe patterns and relationships, make conjectures, discuss solutions and generalise from their findings. The goal is to immerse students in the mathematics and the styles of mathematical thinking needed for success in high school and eventually college. (p. 182)

As a means of evaluating changes in student understanding during exposure to reform-based mathematics curricula, the Balanced Assessment (BA) was developed in a concurrent project also funded by the NSF (Ridgway et al., 2000, citing Ridgway & Schoenfeld, 1994). According to Ridgway et al. (2000), the BA test was not designed such that its tasks ran in parallel with those on the CMP curriculum; rather, the aim was to assess transfer of learning according to the educational goals set out by the NCTM Standards. The BA tests consist entirely of open-ended items designed to assess reasoning, mathematical communication, connections, and problem solving. Because the open-ended items are time-consuming to complete, only a subset is administered to any given test-taker in one of five forms. Each form contains 10–15 individual items that are scored both holistically and analytically by trained raters.

Ridgway et al. (2000) reported on the results of a quasi-experimental evaluation of the CMP curriculum. The study employed a pre-post design with two different tests: One test was the BA described above; the other was the Iowa Test of Basic Skills (ITBS). The ITBS consists solely of multiple-choice items that focus on the mastery of technical skills in mathematics. A total of 500 grade 6 students, 861 grade 7 students, and 1,095 grade 8 students took grade-specific versions of these tests at the beginning of a fall semester and then again at the end of a spring semester. In each grade, some students were taught math using the CMP curriculum (reform-based treatment condition), while others used commercially available textbooks (nonreform-based control condition). The authors subsequently compared

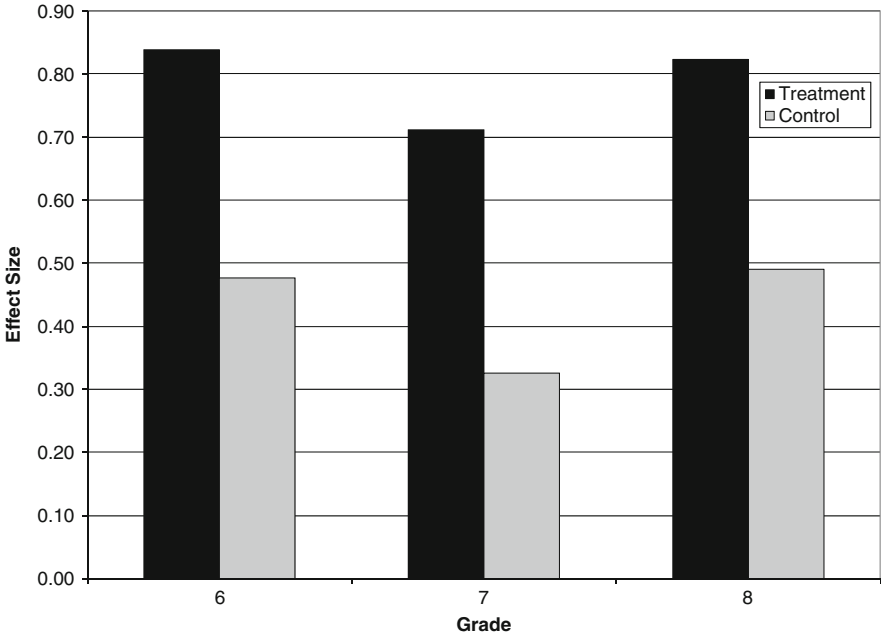

Fig. 8.1 Standardized gains on Balanced Assessment (BA) tests by grade and condition

the standardized gains for each group as a function of the outcome measure being used. These results are presented graphically in Figs. 8.1 and 8.2.

When evaluated using the BA tests, the results seem unequivocal. As shown in Fig. 8.1, students exposed to the CMP curriculum have considerably larger average gains than students exposed to traditional curricula. In contrast, when evaluated using the ITBS, far less compelling evidence exists to support the effectiveness of the CMP curriculum. There appears to be a negative effect in grade 6, no effect in grade 7, and a positive effect in grade 8.

A couple of comments are in order. First, the items on the ITBS are likely to be very similar to the types of multiple-choice items on the state-level tests administered to fulfill the requirements of NCLB. They are not necessarily bad items, nor is the ITBS necessarily an invalid test. However, the ITBS was not designed to evaluate the same cognitive outcomes for which the BA test was designed. If the ITBS were used as the sole outcome measure to estimate the effect of the CMP curriculum in grade 6, one would be likely to draw the conclusion that the curriculum should be abandoned. By contrast, were the BA test to be used, we would conclude that the CMP curriculum should be celebrated. Second, the different patterns of findings by test are the kinds of results that can lead to a greater understanding of the curriculum under investigation and how children are learning. A typical argument by those developing curricula that supposedly focus on depth of conceptual understanding is that this will not sacrifice surface understandings that are more procedural. The results from the Ridgway et al. (2000) study suggested

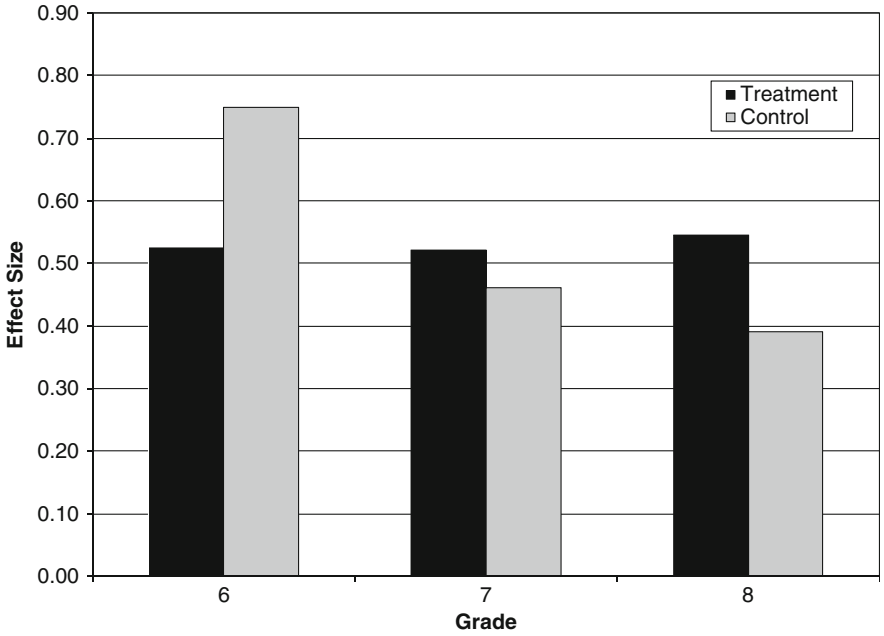

Fig. 8.2 Standardized gains on Iowa Test of Basic Skills (ITBS) tests by grade and condition

that procedural understanding (as measured by the ITBS) might suffer when students have only been exposed to 1 year of the program, but for students exposed to 3 years of the CMP curriculum, this gap reverses.

8.2.2 Evaluating the Effectiveness of Teachers with Value-Added Models

Value-added modeling (VAM) has become increasingly popular in the context of educational accountability systems because it offers the potential to estimate the effect of a specific teacher or school on student achievement independent of the influences of race, socioeconomic status, and other contextual factors. Currently, the most widely used program is the Educational Value-Added Assessment System (EVAAS; The SAS Corporation, n.d.). Some form of the EVAAS has been implemented (or is being considered for implementation) in over 300 school districts in 21 states. The statistical models that underlie VAM approaches such as the EVAAS are complex and incorporate techniques that, in theory, adjust for such factors as preexisting differences in the demographic and academic characteristics of students and the influence of previous schooling on test score

growth (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Sanders, Saxton, & Horn, 1997).

It is very unclear whether a VAM can be used to estimate quantities that can be reasonably interpreted as causal effects (Rubin, Stuart, & Zannato, 2004; Briggs & Wiley, 2008). A necessary condition for the use of a VAM to estimate teacher *effects* is the availability of longitudinal data on a collection of teachers with student test scores that have been linked over time. A statistical model can then be used to estimate the average score increment each teacher has contributed to the achievement of his or her students in a current year over and above the achievement that had been observed for students in prior years. These *increments* are not interpretable as causal effects in and of themselves. For this we must establish – for each teacher – a control group of students to represent the average test score increment that would have been observed had students not attended a class with the teacher being viewed as the educational treatment. In the EVAAS, this outcome is represented by the full sample of students across the collection of teachers being analyzed. As a result, value-added effects are estimated and interpreted relative to the average score gain contributed by all schools under analysis. The data employed for a value-added analysis are essentially an extreme version of an observational study in which students self-select the teacher (and by extension, schools) to which they are exposed. A key question of interest is whether different value-added models are better able to adjust for these sorts of selection biases than others.

The results from such a sensitivity analysis were presented by Lockwood et al. (2007). The authors examined 4 years of longitudinal data for a cohort of 3,387 students in grades 5 through 8 attending public schools in the state of Pennsylvania from 1999 to 2002. Of interest was the sensitivity of teacher effect estimates to the complexity of the VAM being specified. The authors chose four different VAMs in order of the complexity of their modeling assumptions: gain score, covariate adjustment, complete persistence, variable persistence. They also chose five different sets of control variables to include in the VAMs: none, demographics, base year test score, demographics plus base year test score, and teacher-level variables. Finally, they considered one novel factor seldom explored in prior VAM sensitivity analyses: the outcome measure. Students in the available sample had been tested with the Stanford 9 assessment across grades 5 through 8. Upon examining the items contained in the Stanford 9, Lockwood et al. disaggregated the test into two different subscores as a function of items that emphasized problem solving (40% of the test) and items that emphasized procedures (60% of the test). Having established three factors for their sensitivity analysis (type of VAM, choice of covariates, choice of test outcome), the authors estimated teacher effects for each three-way factor combination and asked the question: Which factor has the greatest impact on inferences about a given teacher's effect on student achievement?

What they found was that, by far, the choice of test outcome had the biggest impact on teacher effect estimates. Regardless of the choice of VAM or covariates, estimates of teacher effects tended to be strongly correlated (0.8 or higher). On the other hand, the correlations of teacher effects estimates by outcome were never greater than 0.4, regardless of the underlying VAM or choice of covariates.

8.2.3 Can Readily Available Standardized Tests Support Causal Conclusions?

I chose the two examples above because they illustrate the kinds of evaluative studies that are now being conducted thanks to the testing infrastructure spurred by NCLB. Administrators, parents, and policymakers are naturally going to want to use existing tests to address causal questions about the effectiveness of educational interventions. At this point, I think the question of whether the tests are up to the task – regardless of the quality of the underlying study design – is rather open. Imagine that each of the studies described above involved a randomized controlled experiment – the gold standard for estimating unbiased causal effects. This change would mean that in the Ridgway et al. (2000) study, schools were randomly assigned to the CMP or non-CMP curriculum, while in the Lockwood et al. (2007) study, students were randomly assigned both to schools and teachers. Assume further than the effects estimated in each study were unbiased estimates. Now if each study were conducted only using the test scores readily available to researchers through state testing programs – ITBS and Stanford 9 math test scores – we would miss a good chunk of the story about the effectiveness of the CMP curriculum and Pennsylvania teachers.

Most schools are eager to implement educational interventions that have been proven to work. To facilitate such decisions, the U.S. Department of Education has established the What Works Clearinghouse (WWC) as source where decision makers can turn to for evidence about a prospective intervention's effectiveness. The WWC is responsible for reviewing the quality of existing studies conducted to evaluate the effects of a wide range of educational interventions. However, such reviews focus almost exclusively on the internal validity of estimated causal effects (Briggs, 2008). Evidence that tests are valid for the causal inferences they are being used to support has been essentially delegated to state departments of education and their test contractors.

8.3 Building a Case for Test Validity: Theory and Practice

8.3.1 Test Validation in Theory

Perhaps the most famous and widely cited definition of what is meant by test validity comes from Messick's chapter on validity in the third edition of the book *Educational Measurement* (Messick, 1989). Messick wrote, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). Messick's contributions to validity theory, building upon the work of Cronbach (1971) and Cronbach and Meehl (1955), were

both influential and somewhat controversial because he rejected the formerly held trinitarian view of different types of validity (i.e., content, criterion, and construct) and emphasized the view that it is test scores, not the test itself, that are validated. In the process, he redefined the term *construct validity* as a single unitarian concept that encompassed content and criterion-related validity and made the consequences of testing a fundamental aspect of what is required to establish construct validity.

In the latest edition of *Educational Measurement*, Kane advances what he has described as an argument-based approach to validity (Kane, 1992, 2006). Kane's thesis, consistent in spirit with the perspectives of Cronbach (1971), Messick (1989), and Shepard (1993) before him, is that test validity is a matter of degree and depends upon the clarity, coherence, and plausibility of any interpretive argument that links test scores to the decisions and inferences for which they are to be used. The essence of the argument-based approach to validation is appealing: Be clear about how you plan to interpret and use test scores, build a case for why the test in question meets these needs, and defend yourself against alternative cases for why the test is inadequate. On the other hand, as a theory, the approach is incredibly broad and intentionally nonproscriptive.

8.3.2 Test Validation in Practice

This view of establishing test validity as the process of integrating different sources of evidence into a comprehensive argument has been formalized in Chap. 1 of the *Standards for Educational and Psychological Testing* [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME), 1999; hereafter referred to as "Test Standards for Validity"]. The "Test Standards for Validity" provided five categories of evidence from which an argument for or against the validity of any specific test score inference or consequence could be advanced: (a) test content, (b) the response processes of test-takers, (c) the internal structure of the test, (d) the relationship of test scores to other variables, and (e) the consequences of test use. If the "Test Standards for Validity" is to be taken seriously as a reflection of the consensus position on validity theory, then a critical question is to what extent it informs the practices of states, especially since NCLB was enacted. Two recent reviews have examined the gap between theory and state practices. Linn (2006) examined the validity evidence used to support test score inferences in the assessment programs of six states: California, Colorado, Florida, Ohio, South Carolina, and Washington. Using information submitted to the U.S. Department of Education as part of the NCLB peer review process (U.S. Department of Education, 2004), Linn compared the validity practices of each state against five categories of validity evidence described in the "Test Standards for Validity." Linn found that while the states generally provided a great deal of evidence about the content and internal structure of their standardized tests, and about the relationship of scores on these tests with other variables, little evidence existed to show that the states were

actively investigating the response processes of test-takers and consequences of test use (Linn, 2006). Ferrara (2006) conducted a similar review and concluded that "the types of evidence provided fall far short of current thinking and recent methodological developments relevant to developing validity evidence. Technical reports tend to describe evidence without integrating it into statements about the validity of various interpretations and uses" (p. 616).

My own analysis of the information and evidence that states make publicly available to support their testing programs have produced results that are consistent with the findings described Linn and Ferrara. However, the fact that a gap exists between validation theory and practice does not necessarily imply that tests are being invalidly used for high-stakes purposes. What can be safely concluded is that large-scale standardized tests administered from state to state

- Have items that were approved by committees of subject matter experts as being representative of a state's content standards,
- Have scores that are suggestive of high reliability, and
- Are developed to avoid obvious cultural biases.

Such information is valuable to be sure. However, these (and other) readily available pieces of information are only links from what should be a larger argumentative chain of reasoning. One important link that is missing is evidence showing the extent to which test scores are sensitive to formal instruction. Such an assumption seems implicit in the studies by Ridgway et al. (2000) and Lockwood et al. (2007) and would seem to be central to virtually all state tests used to support systems of educational accountability. Yet this assumption does not seem to be regularly validated.

8.3.3 Test Validation as Causal Inference

It seems to me that one principal reason it is so hard to validate the use of tests for high-stakes inferences is because the approach outlined in the "Test Standards for Validity" (AERA, APA, & NCME, 1999) essentially requires us to build an inferential argument by observing effects and then attributing them to a cause (which is daunting) rather than estimating the effects from a hypothesized cause (which is doable).

Figure 8.3 illustrates a typical psychometric conceptualization of the relationship between test items and a single latent construct underlying these test items. This conceptualization has an implied causal inference, where the idea seems to be that having more or less of the latent construct causes a test-taker to answer a given item correctly or incorrectly. This idea is formalized in item response theory with the conditional expectation $P(X = x_i|\theta)$. From this perspective, a necessary condition for establishing test validity is to establish that θ has a causal effect on item responses. The impediment, of course, is that θ is unobserved (and hence not manipulable). As a result, we can only observe differences in the item responses among test-takers and use these to make a causal attribution about θ. So θ is

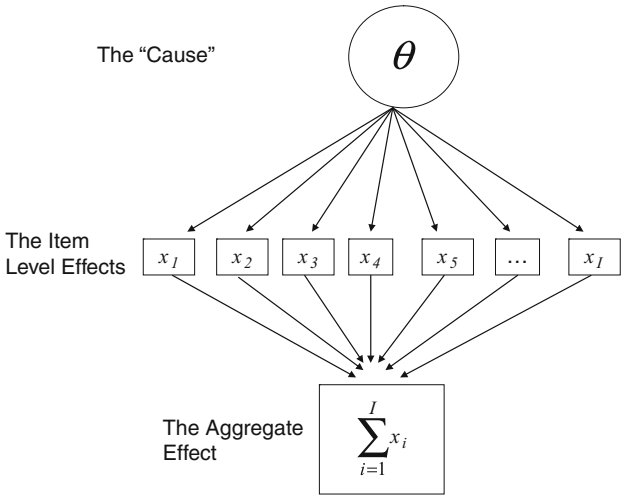

Fig. 8.3 Test validation as causal inference

operationally defined only by patterns of item responses. This result explains why the validity evidence typically provided by psychometricians in the technical reports of state testing programs rely so heavily upon evaluations of test item characteristics: their quality, their intercorrelations, and so on. A problem with such approaches is that it becomes possible to do analysis that is largely divorced from design. Because no hypotheses are being advanced for what we should expect to observe, almost any finding can be rationalized as acceptable within some bounds for acceptable (and perhaps arbitrary) ranges of item difficulty, point biserials, and reliability.

The notion that causal inference is implicit in test design and validation is not new. This idea can be found in recent manuscripts in the psychometrics literature (c.f., Borsboom, Mellenbergh, & Heerden, 2004; Wilson, 2005) and for decades in books and articles on structural equation modeling. However, in my view it is neither feasible nor necessary to model all the causes, latent or otherwise, that influence the item responses of test-takers on large-scale assessments. In his chapter "Test Validation" from the second edition of *Educational Measurement*, Cronbach (1971) pointed toward a more direct approach when he wrote:

> Experimental interventions in which something is deliberately done to change student scores, as a means of identifying influences to which test performance is sensitive, have been mentioned several times. The treatment may be a change in time limit, a special instruction, etc. The investigator, knowing of what his treatment consists, can predict its effect on the tests; the results confirm or challenge some part of his interpretation of the measuring instrument. (p. 474)

Cronbach was essentially proposing the substitution of an observed and well-understood educational intervention, Z, for the hypothesized latent construct θ in Fig. 8.3. By well-understood, I mean that Z should have been designed such that not

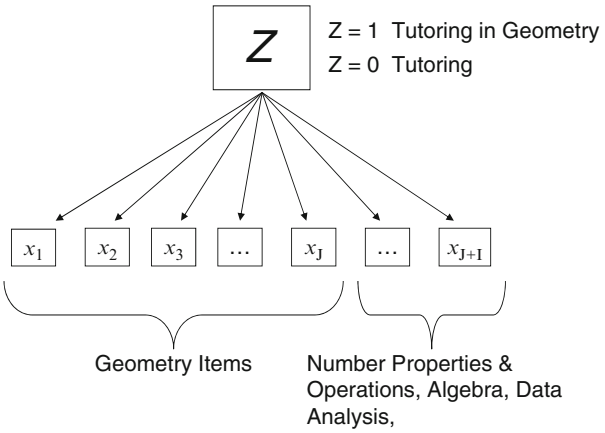

Fig. 8.4 Tutoring program as an indirect manipulation of the construct of measurement

only would exposure to it be expected to have an effect on overall test performance, but also that this effect could be properly hypothesized for specific items or item subsets. That is, if test-developers really understand what is being measured, it should be possible to imagine interventions that would (or at least should) increase the probability of students answering some test items correctly, but *not* increase the probability of answering other items correctly. I illustrate this notion in Fig. 8.4.

Here we imagine a scenario in which the middle-school students in a state are tested annually on large-scale assessments of math. The items on the test have been designed to measure different *content strands* according to the state's published standards framework, and these strands distinguish between the mastery of number properties and operations, algebra, data analysis, and geometry. Now, if we were to take a sample of students and randomly assign them to either a tutoring program that focused on instruction and practice in understanding geometric concepts (Group 1, $Z = 1$) or a tutoring program that focused on algebra (Group 2, $Z = 0$), we should expect that when the test performance of the two groups is compared, Group 1 students will have a significantly higher probability of answering geometry items correctly relative to algebra items, and vice-versa for Group 2 students. If we find this to be so, it would seem to bolster an argument that a manipulation of the underlying construct has had an effect on item response probabilities. A competing explanation that would need to be ruled out is that at least some portion of what the test measures is trivial (construct irrelevant) and can be manipulated through savvy coaching techniques (which results in what Koretz and Hamilton (2006) have called *score inflation*). If no significant differences in the average response probabilities exist between the groups, it would seem to suggest that whatever the test is measuring is not readily manipulable. Again, a competing argument would need to be ruled out: Perhaps the tutoring that was implemented differs from what was intended.

Note that in this brief example the central component of a validity argument becomes a matter of estimating effects rather than attributing cause. Of course, much hinges upon the defensibility of substituting Z in place of θ. But in my view,

being forced to make and defend this argument focuses important attention on the intended alignment between what is being taught and what is being assessed. If the substitution of Z in place of θ can be defended, then much of the theory and practice of causal effect estimation can be implemented at the item level. The resulting patterns would provide evidence for what a test is and is not measuring. And making item-level inferences would be possible (though challenging) even when students have not been randomly assigned into tutoring conditions.

8.4 Evaluating a Test's Instructional Sensitivity in Practice

The provision of supplemental educational services (which I hereafter refer to as tutoring) to low-income students in schools failing to make AYP under NCLB is just now beginning to attract the attention of educational researchers. In my view, it should really be attracting the attention of psychometricians. The tutoring that students are receiving is likely to be the purest form imaginable of teaching to the test. The theory of action behind NCLB and all systems of educational accountability is that a student who has a poor understanding of, say, algebra would have a better understanding if he or she had instead been exposed to some intervention (i.e., better teaching, more motivation, better diet, etc.). It follows from this that for educational accountability to achieve the consequences that are envisioned, there are two necessary conditions: the presence of good interventions and standardized tests that are instructionally sensitive.

In Colorado during the 2006–07 school year, approximately 1,500 students in grades 4 through 8 received tutoring beyond their normal school instruction in the subjects of math and reading. All of these students were receiving free lunch assistance, and 94% were Black or Hispanic students. There were many more students in the state in the same grades with the same demographic background and prior test performance who were similarly eligible to receive tutoring but either chose or were unable to take advantage of the tutoring services. Because both groups of students had taken the Colorado Student Assessment Program (CSAP) tests in 2006 and again in 2007, it is possible to estimate an effect of the tutoring. In an evaluation conducted by the OMNI Institute (2008), the tutoring appeared to have no aggregate effect on reading performance and a small effect on math performance. The effect found for math performance was not large enough to move any of the students from a performance level classification of unsatisfactory to proficient. These results are consistent with the few other evaluations of NCLB-mandated tutoring that have been conducted to date (Burch, Steinberg, & Donovan, 2007; Vergari, 2007). However, while the natural conclusion from such studies is that tutoring programs are largely ineffective, another conclusion must be entertained: Perhaps the programs are doing exactly what we would expect, and it is simply the case that the tests are not instructionally sensitive.

How could the principles described in the previous section be applied to this empirical context? To make this example as concrete as possible, imagine we have

access to the full population of grade 5 students in a single Colorado school district during the 2008–2009 school year. A subset of these students was eligible to receive tutoring services because they were low income and their schools failed to make AYP. To keep things simple in this illustration, we will focus just on math outcomes. Roughly 100 items are administered on the grade 5 CSAP math exam, and these have been mapped evenly into five designated content standards according to the state's department of education (number sense, algebra, statistics, geometry, and problem-solving). A first order of business would be to determine, through inspection of curricula or other analysis, the alignment between the tutoring programs and the CSAP math test. Does the program spend equal amounts of time on instruction that would map to each of the five item sets found on the CSAP? (If the tutoring company is being strategic, one might expect them to devote greater energy to the content with difficulty closest to the performance threshold that demarcates proficiency.) From this analysis a program-specific hypothesis can be generated about the types of items that should be most sensitive to tutoring. Now assume we have at least two tutoring programs to compare that have been determined to differ significantly in their relative alignment with the CSAP test.[1] In Program 1, a student has been exposed to a program with the greatest relative alignment to the 40 items emphasizing an understanding of number sense and algebra. In Program 2, a student has been exposed to a program with the greatest relative alignment to the 40 items emphasizing an understanding of statistics and geometry. Given such information, we can proceed to empirically compare the probability of correct item responses as a function of tutoring exposure after conditioning on math performance in prior grade(s).

One straightforward way this could be done would be to use the Mantel-Haenszel procedure described by Holland and Thayer (1988) for use in the context of diagnosing potential symptoms of item bias. Or, we could use logistic regression and an approximation technique (c.f., Swaminathan & Rogers, 1990) to estimate the area between curves as a function of tutoring exposure. Conditional on prior ability, students receiving more tutoring in number sense and algebra should outperform their counterparts receiving more tutoring in statistics and geometry on these test items, and vice-versa. Provisional conclusions about the instructional sensitivity of the test would hinge upon the results from these analyses. If the test appears to be instructionally sensitive, it bolsters the validity of its high-stakes use within an accountability system.

To be sure, many details of this approach would need to be ironed out:

- How big must an item-level difference between groups be before it is considered practically significant?
- Should the results be aggregated (for example, summed across all number sense and algebra items) or evaluated item by item for salient trends?

[1] It would also be possible to compare a single tutoring program to a control condition of no tutoring, but this comparison would introduce a clear source of bias in the sense that students enrolled in tutoring are likely to be more motivated than those who are not.

- Should an estimate of the current test score be used as a conditioning variable or only prior test scores? Should all available test score information be included? (Note: this increase in dimensionality could be reduced through propensity score estimation.)
- When students have not been randomly assigned to tutoring groups, what other variables are available for inclusion in the conditioning set?

Many of these questions have already been raised (and addressed) in the psychometric research literature on differential item functioning (DIF) techniques. An evaluation of DIF is standard practice for testing companies, but its interpretation is often highly equivocal because the categorical grouping variables employed are usually demographic. In contrast, the results are more readily interpretable for the present test validation context because the grouping variable is a manipulable treatment that serves as a proxy for the construct of measurement. While it is true that differences in average response probabilities might be due to selection bias (depending upon the reasons that some students choose to enroll in tutoring programs), a mitigating factor is the availability of longitudinal data and the fact that the students eligible for tutoring are, by definition, from low-income households. Furthermore, when the item-level performance of students in different tutoring programs is being compared, one might also be willing to assume that, on average, both sets of students are similarly motivated relative to students who were eligible for tutoring but did not enroll.

8.5 Some Final Comments

An important impetus for the test validation design proposed above is that a closer connection needs to be developed between the ways tests are designed and scores are interpreted. By looking for what are essentially causal effect estimates at the item level, we commit ourselves to an understanding of what we think is being taught in schools and what specific item sets we think will capture this learning. States such as Colorado should be able to say, for example, "The principle obstacle to being classified as proficient in mathematics as of grade 5 is an understanding of basic concepts in geometry and their application to solve measurement problems. So this should be the focus of our tutoring programs." If tutoring programs were to then respond by teaching geometric concepts and applications, we should expect to see causal effects on the associated geometry items, but not on items that focus, for example, on number sense. If we do, this is strong evidence in favor of test validity. If we do not, then I think we need to carefully consider that beyond the possible explanation that the tutoring is ineffective there is a possibility that the existing test is not valid for the high-stakes inferences inherent in accountability systems.

In conclusion, I think we can gain much more traction in validating the use of test scores for high-stakes inferences if we make our causal hypotheses complex but keep our analyses relatively simple. The evaluation of tutoring programs under

NCLB provides a unique opportunity for implementing this idea. In my view, these kinds of validation studies would be easy to convince states to do because they are at once theory driven and pragmatic – theory-driven because you have to know what it is your tutoring purports to teach and your tests purport to measure, but pragmatic because they may save states millions of dollars being spent on tutoring that does not help or on tests that are invalid for their proposed uses. When tests must be validated for use in supporting high-stakes causal inferences, the traditional sources of validity evidence are necessary but not sufficient. If we wish to avoid causal inferences that are careless, we proceed with business as usual at our own peril.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Validity. In *Standards for educational and psychological testing* (pp. 9–24). Washington, DC: American Educational Research Association.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Borsboom, D., Mellenbergh, G., & Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Briggs, D. C. (2008). Synthesizing causal inferences. *Educational Researcher, 37*(1), 15–22.

Briggs, D. C., & Wiley, E. (2008). Causes and effects. In L. Shepard & K. Ryan (Eds.), *The future of test-based educational accountability*. New York, NY: Routledge.

Burch, P., Steinberg, M., & Donovan, J. (2007). Supplemental educational services and NCLB: Policy assumptions, market practices, emerging issues. *Educational Evaluation and Policy Analysis, 29*(2), 115–133.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Ferrara, S. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). Westport, CT: American Council on Education/Praeger.

Holland, P. W. (1986). Statistics and causal inference (with discussion and rejoinder). *Journal of the American Statistical Association, 81*, 945–970.

Holland, P. W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. (2004). *Evidence for causal inference in education research*. Invited session on inference, Evidence and Scientific Research at the annual conference of the American Educational Research Association, San Diego, CA.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Linn, R. (2006). *Validity and reliability of student assessment results.* Unpublished manuscript.
Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–68.
McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/MacMillan.
No Child Left Behind Act of 2001, Pub. L. No. 107–110 § 115 Stat. 1425.
OMNI Institute. (2008). *Evaluation of supplemental educational services: 2006–07 academic year data.* Unpublished manuscript.
Ridgway, J., & Schoenfeld, A. H. (1994). *Balanced assessment: Designing assessment schemes to promote desirable change in mathematics education*. Keynote paper for the EARLI Email Conference on Assessment.
Ridgway, J., Zawojewski, J., & Hoover, M. (2000). Problematising evidence-based policy and practice. *Evaluation and Research in Education, 14*(3, 4), 181–192.
Rubin, D., Stuart, A., & Zannato, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.
Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.
Shepard, L. (1993). Evaluating test validity. *Review of Educational Research, 19*, 405–450.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.
The SAS Corporation. (n.d.). *SAS® EVAAS® for K–12*. Retrieved from http://www.sas.com/govedu/edu/k12/evaas/index.html.
U.S. Department of Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.
Vergari, S. (2007). Federalism and market-based education policy: The supplemental educational services mandate. *American Journal of Education, 113*, 311–339.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chapter 9
Propensity Score Matching to Extract Latent Experiments from Nonexperimental Data: A Case Study

Ben B. Hansen

9.1 Introduction

9.1.1 Purpose and Context of Paper

During the 1995–1996 academic year, investigators from the College Board surveyed a random sample of high school junior and senior SAT® takers to probe how they had prepared for the SAT. Among other questions, students were asked whether they had taken extracurricular test-preparation classes. Some 12% of respondents said that they had; the comparison of these students' SAT scores to those of the remaining 88% comprised the observational study reported by Powers and Rock (1999).

Attempts to estimate intervention effects without the benefit of randomization demand adjustment for covariates, potentially confounding variables. Powers and Rock's was no exception. Coached and uncoached students differed in educational preparation, race, class, and PSAT scores, among other relevant factors determined in advance of their decisions about coaching; each of these differences would have to be addressed. The most commonly used methods of adjusting for potential confounders involve using regression methods to model outcomes, here SAT scores, as functions of covariates and intervention variables.

Another method that may help is propensity score matching: estimate conditional probabilities of falling in the intervention group given the covariates, propensity scores (Rosenbaum & Rubin, 1983); match treatment group subjects to untreated controls whose estimated propensity scores are similar; then carry out the outcome analysis with adjustment for the propensity score matches.

B.B. Hansen (✉)
Statistics Department, 439 West Hall, University of Michigan,
Ann Arbor, MI 48109–1107, USA
e-mail: bbh@umich.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_9,

When treatment and control groups differ substantially on baseline measures, propensity matching reduces the differences, creating matched sets within which baseline treatment and control differences average to something closer to zero (Rosenbaum, 2001): it improves *covariate balance*, even when there are relatively many covariates. Well-balanced covariates lend credibility to an observational analysis. Cochran (1965, Sec. 3.1) suggested that good covariate balance be treated as a necessary precondition for analysis of an observational data set.

Propensity score matching secures this and other advantages at stages of the analysis that do not require, and often precede, use of measurements of the outcome. Thus propensity scoring and matching can be seen as parts of the design, as distinct from the outcome analysis, of an observational study (Rosenbaum, 2010; Rubin, 2008). Accompanying diagnostic procedures lead to tables and plots that, in contrast with regression diagnostics, directly bolster the credibility of the adjustment and are often of interest to scientific audiences in themselves.

The current paper revisits Powers and Rock's (1999) data set, using it in a demonstration of how specifically to combine propensity scores with modern matching and ordinary covariance adjustment techniques to obtain inferences about the coaching effect that are supported by the large-sample theory of a companion paper (Hansen, 2009). The method improves that of Hansen's (2004) analysis of the same data in several ways. It gives narrower confidence bounds; it is easier to implement. According to the supporting theory, it also better removes bias.

9.1.2 Workflow

Propensity score matching is an attempt to isolate pieces of a nonrandomized sample that resemble randomized experiments, at least in terms of the observed variables. The procedure is divisible into roughly six steps:

Step 1. *Select a small number of covariates for exact matching (stratification).* If among the variables that demand adjustment one or two seem to most influence selection into the treatment group, then it or they are the natural candidates for exact matching.

Step 2. *Evaluate plausibility of the proposition that treatment is effectively randomized within the strata created at Step 1.* This requires inspecting balance on those covariates not directly incorporated into the stratification and determining whether it is comparable with the balance that randomization would have produced. If so, then there is little reason to estimate or match on propensity scores: This part of the adjustment is complete.

Step 3. In the more likely event that stratification on just a few covariates does not balance the others, the next step is to *estimate propensity scores and, if necessary, make other preparations for matching*. The optional preparations

would include the construction of a *Mahalanobis distance*, as recommended by Rosenbaum and Rubin (1985b), Rubin and Thomas (2000), and others; however, the method recommended here avoids this extra step. Indeed, with this method it will ordinarily suffice to estimate just one propensity score, the conditional probability of being in the treatment group given all of the covariates, and to estimate it just once, using standard logistic regression techniques. (On the other hand, modern matching techniques allow one to match with attention to more than one propensity score, and it can be advantageous to match on several scores. An illustration appears in Sect. 9.4.3.)

Step 4. With a matching criterion in hand, *match intervention to control subjects*. With flexible techniques like the ones to be demonstrated in this paper, the matches need not be made only in pairs, or only in triples or in any other fixed configuration; rather, it is possible to tailor the structure of the matched sets to the contours of the sample. Thus, even when the treatment group is much smaller than the control group, there is no need either to forgo matching those controls that do not fit into pairs or to settle for poor matches in the interests of finding $k > 1$ distinct controls for each treatment subject. On the other hand, in order to avoid inadvertent extrapolation it is necessary to set aside those subjects that are unusually separated from all eligible candidates for matching in terms of the propensity score, and this ordinarily will necessitate leaving aside a few subjects. (Deciding just which subjects to leave aside is discussed in Sect. 9.4.2, below.)

Step 5. *Conduct diagnostic assessments of the match*. The diagnostics include: assessment of covariate balance, assessment of the size of the larger matched discrepancies on propensity scores, and assessment of the consequences of the matching for effective sample size. The first of these diagnostic criteria is much more frequently discussed than the other two, but in practice they are equally important.

Step 6. *Repeat from Step 1, Step 3, or Step 4, after suitable modifications to that step, until a satisfactory match is achieved*. With the approach to be presented here, it will not ordinarily be necessary or advantageous to return all the way to Step 1. A return to Step 3 is indicated if the match was found at Step 5 to insufficiently balance covariates that for some reason were excluded from the propensity score or to diminish effective sample size more than seemed necessary. A return to Step 4 is indicated if the match was found at Step 5 to insufficiently balance covariates that were included in the propensity score or to permit overly large matched distances on the propensity score.

To estimate treatment effects after matching, use any method of adjustment suitable for matched or finely stratified experiments: hierarchical linear or generalized linear modeling; if the outcome is continuous, perhaps linear regression with fixed effects; or randomization-based inference.

Many of these methods involve regression modeling. Of course, regression modeling itself can serve as a basis for confounder control without any propensity adjustment. The procedure is familiar; in outline:

Step A. Select variables for adjustment.
Step B. Specify and fit regression model.
Step C. Perform regression diagnostics and extract effect estimates from fitted model, returning as needed to Steps 1 and 2.

Adjustment with regression involves fewer steps, and has no need for a separate procedure to estimate treatment effects. Clearly, this brevity is an advantage; less clearly, it is a source of complication. At Steps A and B, one attempts to attend simultaneously to the reduction of bias associated with confounding, which may call for flexible specifications and adjustment for more variables, and to keep standard errors down, which calls for rigid specifications and fewer variables. In matching, by contrast, the guiding concern is control of confounding bias. One adjusts inclusively for those pretreatment variables thought to influence the outcome (Rubin & Thomas, 1996, 2000). Propensity score diagnostics make both benefits and costs of this adjustment easier to perceive, as I demonstrate below.

Although propensity score matching and regression are sometimes seen as competitors, it is quite possible to combine the methods, folding a matching into a regression analysis by adding fixed or random effects for the matched sets. The matching largely addresses bias due to observed confounders, freeing Steps A and B of the regression adjustment to focus on reducing error variance – which is generally better accomplished with more selective covariance adjustment for the main prognostic variables than by attempting to adjust for all potential confounders. In turn, such parsimony simplifies Step C; with just a few covariates, regression diagnostics are often routine. The combined method may have more steps than regression alone, but the greater focus of each step allows it to be executed with greater confidence.

9.1.3 Outline

After Sect. 9.1 introduction come two sections reviewing literature this paper draws on. Section 9.2 reviews the rich set of data Powers and Rock collected in order to estimate coaching effects. Section 9.3 reviews matching structures and algorithms, introducing the tradeoff between matching closely on the matching variable and maximizing effective sample size. Then Sect. 9.4 demonstrates two complementary ways of estimating propensity scores and, with them, a new and simplified way of managing the tradeoff between close matching and sample size. Section 9.5 presents permutation-based effect estimates and confidence intervals. Section 9.6 offers a rationale for the use of matching as a prelude and complement to covariance adjustment, drawing connections with supporting large-sample theory presented in a companion paper.

9.2 Highlighting the Strengths of a Strong Observational Study

Section 9.2.1 reviews the structure of Powers and Rock's fine study. Among other strengths, it collected an impressive array of descriptive information about coached and uncoached test takers. Adjusting for all of the potential confounders ought to enhance the credibility of the results; in practice, however, when the adjustments are made using regression, increasing the number of potential confounders for which adjustments are made may decrease the credibility of the analysis. One issue is that each additional covariate tends to have missing observations, so that cases with complete data on a smaller set of covariates have incomplete data on the larger collection of them; common responses to this situation in regression analysis may raise more questions than they settle. In contrast, with stratification-based adjustments including matching, a simple device addresses missingness on the covariate under relatively straightforward assumptions, as Sect. 9.2.2 will illustrate. Regression diagnostics are typically arcane, but central diagnostics for stratification-based adjustments convey information about the study design that is of independent interest to researchers. Section 9.2.3 demonstrates this in the process of attending to Steps 1 and 2 of Sect. 9.1.2's propensity matching workflow.

9.2.1 Powers and Rock's Data

The data to be analyzed derive from a stratified random sample of registrants for 1995–1996 administrations of the SAT-I test, details of which are given by Powers and Rock (1999). About 6,700 high school juniors and seniors received surveys asking whether and how they had prepared for the test; the replies of some 4,200 respondents were linked to the College Board's records of their scores on the 1995 or 1996 exams, as well as scores on previous SAT-I or PSAT tests and their answers to the Student Descriptive Questionnaire (SDQ), which all SAT-I registrants are asked to complete. Since its subjects were selected by a probability sampling design from the pool of all U.S. SAT test takers from a given period, as opposed to a convenience sample, the study supports inferences with greater external validity than is typical in evaluation research.

By their responses to questions about extracurricular SAT preparation, respondents were split into a treated and a control group. Nineteen in twenty of the survey respondents actually took the spring 1996 or fall 1995 exam for which they had registered. The analysis given below restricts itself to these 3994 students, using the corresponding SAT scores as outcome measures. Thus the record gives coaching status and SAT outcomes for all students in the sample to be analyzed; among the additional measures, each available for some fraction of the students, are pretest scores, racial and socio-economic indicators, various data about their academic preparation, and responses to a survey item that, by eliciting students' first choices

in colleges, recovered an unusually discriminating measure of students' educational aspirations. In all, there are 27 relevant pretreatment variables. The coached and uncoached groups differ appreciably in many of these recorded measures – as do high and low scorers on the SAT.

9.2.2 Missing Data; Initial Exact Match

Complete case analysis refers to the practice of handling by setting aside any subject for whom measurements on some covariate are not available. Although easy to implement, it can drastically reduce the sample size, particularly when many covariates are present, in a manner that risks adding to the bias as well as the variance of estimation. It is nearly as convenient to merge "missing" with an appropriate level of the covariate or to treat it as a category unto itself, acknowledging the absence of certain measurements without reducing the sample size. In adjustment based on matching or stratification, this is the same as making missingness part of the profiles according to which study subjects are sorted. Analysis based on this annotated data file will require stronger assumptions than would a parallel analysis without missing covariates; typically such assumptions are more credible than those of the complete-case analysis.

Step 1 involves identifying the one or two covariates that most threaten to confound the comparison in order to match exactly on them. A covariate's potential for confounding is partly a function of the size of the difference between its means in the treatment and control groups, with larger differences portending greater bias. In the Powers-Rock coaching sample, the variables that are most threatening in this sense are the race and socio-economic status (SES) variables, as seen in the left panels of Tables 9.2–9.4. The one race variable separates Asian-Americans (9%) from underrepresented minorities (8% Black, 3% Mexican-American, 1% Native American, 1% Puerto Rican, 3% other Hispanic, 3% other), collapsing the 6% of respondents who did not give their race with Whites (66%). To account for SES, SDQ responses give three potential stratifiers to choose from, namely parents' income and education levels of mothers and fathers. All three variables are probably measured with some error, but it seems that high school students are more likely to know and less likely to misreport their parents' education than their parents' income; and splitting the data into thirds at the 33, 67, and 100% quantiles of mother's and of father's education levels, father's education better separates both PSAT-Math and PSAT-Verbal scores. I stratify the College Board coaching data by race and father's education level, grouping students into three categories of father's education, plus an additional category for students not reporting it. Call this the Race-by-SES (Race $\times$ SES) subclassification; Table 9.1 shows sizes and compositions of its subclasses. Subclassifying in this way, no observations are rejected.

The strategy of creating missingness levels of covariates can also be used to construct propensity scores, leading to propensity scores which, when matched or stratified upon, balance both covariate-missingness and observed-covariate profiles

Table 9.1 Race × SES subclasses: sizes and control-to-treated-subject ratios

Father's education (by race category)	Percent of sample	Percent of treated	# controls per treated subject
White, or no race reported			
High school or less	26	9	21
AA or BA	20	15	10
Post-college	20	29	4.5
Not reported	7	10	4.5
white (all)	72	63	8.2
Under-represented minority			
High school or less	11	7	11
AA or BA	3	4	6.6
Post-college	3	5	3.6
Not reported	1	2	4.4
Under-represented minority (all)	19	18	7.2
Asian-American			
High school or less	4	6	3.8
AA or BA	3	5	3.4
Post-college	3	8	1.5
Not reported	0.4	0.2	15
Asian-American (all)	9	19	2.9
All	100	100	7.0

between treated and control groups. In effect, this addresses the missingness problem by a strengthening of the nonconfounding assumption, from an assumption that the collection of covariates deconfounds the comparison to an assumption that available covariates (along with indicators of their availability) deconfound the comparison (Rosenbaum & Rubin, 1984, Appendix). The strategy is well suited to missingness patterns in which observations tend to lack only a few of a large number of covariates. Such is the case here: On the 23 covariates other than pretest scores, only 32% of the College Board sample has complete data, but two-thirds are missing no more than two items, and 90% lack data on no more than five items. Our propensity score accommodates missing data in this way, in so doing retaining all 3,994 observations. (It also recodes as "missing" the pretest scores of 126 coached students whose pretests did not or may not have preceded their coaching, as well as the pretests of uncoached students whose pretests preceded their posttests by relatively short intervals; see Hansen, 2004, Sec. 1.2).

9.2.3 *What Would Cochran Do? Comparability on Covariates, With and Without Poststratification*

Section 3.1 of Cochran's (1965) landmark paper on observational studies took up the question of whether and how group differences in the distributions of covariates ought to inform decisions as to whether to adjust for them. One recommendation

Table 9.2 Coached versus uncoached on demographic variables, with and without race × SES subclassification

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
Parents' income Q1	0.27	0.14	−6.4	***	0.24	0.14	−5.3	***
Parents' income Q2	0.28	0.2	−4.1	***	0.25	0.2	−2.3	*
Parents' income Q3	0.15	0.12	−1.7		0.16	0.12	−1.9	
Parents' income Q4	0.16	0.35	1.3	***	0.19	0.34	7.8	***
Parents' income N/A	0.14	0.2	3.5	***	0.17	0.2	2.1	*
Dad's education = high school	0.43	0.23	−8.4	***	0.25	0.25	0	
Dad's education = some college	0.26	0.23	−1.7		0.24	0.24	0	
Dad's education = grad school	0.23	0.42	9.2	***	0.39	0.39	0	
Dad's education *n/a*	0.081	0.12	3.1	**	0.12	0.12	0	
Mom's education = high school	0.49	0.29	−8.5	***	0.4	0.3	−4.8	***
Mom's education = some college	0.27	0.29	1.0		0.29	0.29	0.4	
Mom's education = grad school	0.16	0.3	7.3	***	0.22	0.29	4.0	***
Mom's education *n/a*	0.071	0.12	3.7	***	0.1	0.12	2.1	*
1st language = english	0.8	0.69	−5.5	***	0.71	0.71	−0.5	
1st language = eng.+another	0.079	0.12	2.9	**	0.1	0.11	0.5	
1st language not english	0.075	0.11	2.7	**	0.11	0.1	−0.6	
1st language *n/a*	0.049	0.084	3.3	**	0.074	0.083	1.3	
Gender B	0.41	0.4	−0.5		0.43	0.4	−1.3	
Gender G	0.59	0.6	0.5		0.57	0.6	1.3	
Ethnicity = Asian	0.078	0.19	8.2	***	0.16	0.16	0	
Ethnicity = White	0.73	0.63	−5.1	***	0.65	0.65	0.0	
Ethnicity = URM	0.19	0.18	−0.3		0.19	0.19	0	

Note. Without stratification, the groups differ starkly in demographic terms. Exact matching on two demographic variables leaves highly significant imbalances on others. (URM = under represented minority)

was to conduct preliminary checks, comparing the groups on covariates before considering (perhaps before collecting) data on outcomes. Another recommendation was to assess group differences in covariate means in terms of corresponding t-statistics. For covariates of high or moderate prognostic value, t-statistics below 1.5 or so in magnitude would be okay, but statistics larger than that were potentially problematic. A few such imbalances might well be handled with regression adjustments once the outcome data become available, but many of them seemed to present a more fundamental problem. "If several x-variables show t-values substantially above 1.5," Cochran wrote, "this raises the question of whether the groups are suitable for comparison" (p. 243). How do Powers and Rock's coached and uncoached samples fare in these terms?

Table 9.3 Coached versus uncoached on scholastic preparation and achievement variables, with and without race × SES subclassification

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
PSAT-V	51	51	−0.3		51	51	−1.3	
min(PSAT-V, 40)	40	40	1.0		39	40	1.6	
max(PSAT-V, 60)	61	60	−1.9		61	60	−3.1	**
PSAT-M	50	51	2.3	*	50	51	0.3	
min(PSAT-M, 40)	39	40	2.4	*	39	40	2.2	*
max(PSAT-M, 60)	61	61	1.4		61	61	−0.5	
PSAT N/A	0.33	0.38	2.0	*	0.32	0.38	2.4	*
Prior SAT-V	480	479	−1.7		481	479	−2.4	*
min(prior SAT-V, 400)	399	399	−0.7		400	399	−1.2	
max(prior SAT-V, 600)	600	600	−0.7		600	600	−1.1	
Prior SAT-M	480	481	1.0		481	481	0.6	
min(prior SAT-M, 400)	400	400	0.1		400	400	−0.1	
max(prior SAT-M, 600)	600	600	1.1		600	600	1.0	
Prior SAT N/A	0.96	0.95	−0.6		0.96	0.95	−0.6	
GPA self-report Q4	0.074	0.048	−2.1	*	0.081	0.045	−2.8	**
GPA self-report Q3	0.32	0.38	3.0	**	0.32	0.38	2.3	*
GPA self-report Q2	0.45	0.4	−1.9		0.42	0.41	−0.4	
GPA self-report Q1	0.1	0.082	−1.5		0.094	0.085	−0.6	
GPA self-report *n/a*	0.056	0.082	2.3	*	0.079	0.081	0.2	
Avg. english = excellent	0.38	0.42	1.6		0.4	0.41	0.8	
Avg. english = good–fail	0.56	0.49	−2.7	**	0.52	0.5	−0.9	
Avg. english *n/a*	0.057	0.084	2.4	*	0.079	0.083	0.4	
Avg. math = excellent	0.34	0.37	1.1		0.35	0.36	0.4	
Avg. math = good–fail	0.6	0.55	−2.4	*	0.57	0.55	−0.6	
Avg. math *n/a*	0.055	0.086	2.8	**	0.079	0.085	0.6	
Avg. natural science = excellent	0.36	0.4	1.5		0.37	0.39	0.8	
Avg. natural science = good–fail	0.58	0.52	−2.6	*	0.55	0.52	−0.9	
Avg. natural science *n/a*	0.061	0.086	2.1	*	0.082	0.085	0.3	
Avg. social science = excellent	0.45	0.5	2.1	*	0.46	0.49	1.2	
Avg. social science = good–fail	0.49	0.42	−3.0	**	0.46	0.43	−1.2	
Avg. social science *n/a*	0.06	0.082	1.9		0.081	0.08	0	
# Yrs. english 0–2	0.17	0.17	−0.1		0.16	0.17	0.5	
# Yrs. english = 3–4	0.76	0.74	−0.7		0.74	0.74	−0.0	
# Yrs. english *n/a*	0.074	0.09	1.3		0.097	0.089	−0.7	
# Yrs. foreign language = 0–2	0.66	0.5	−7.0	***	0.63	0.51	−5.3	***
# Yrs. foreign language = 3–4	0.25	0.4	6.8	***	0.26	0.39	5.8	***
# Yrs. foreign language *n/a*	0.089	0.11	1.2		0.11	0.1	−0.4	
# Yrs. math = 0–2	0.29	0.2	−4.2	***	0.26	0.21	−2.7	**
# Yrs. math = 3–4	0.64	0.7	2.8	**	0.64	0.7	2.7	**
# Yrs. math *n/a*	0.071	0.096	2.0	*	0.095	0.092	−0.2	
# Yrs. natural science = 0–2	0.46	0.39	−2.9	**	0.43	0.4	−1.3	
# Yrs. natural science = 3–4	0.46	0.5	1.8		0.46	0.49	1.1	
# Yrs. natural science *n/a*	0.086	0.11	2.1	*	0.11	0.11	0.2	

(continued)

Table 9.3 (continued)

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
# Yrs. social science = 0–2	0.49	0.4	−3.4	***	0.47	0.4	−2.8	**
# Yrs. social science = 3–4	0.44	0.5	2.9	**	0.43	0.51	3.3	***
# Yrs. social science *n/a*	0.078	0.092	1.0		0.1	0.09	−1.0	

Note. Group differences in these variables are pronounced, if less so than for demographic variables. In most cases subclassification reduces large imbalances, if not to insignificance

Table 9.4 Coached versus uncoached on attitudes to college and to the SAT, with and without race × SES subclassification

	No stratification				Race × SES			
	Not coached	Coached	z		Not coached	Coached	z	
Avg. SAT at 1st choice college	1,059	1,098	9.5	***	1,067	1,097	7.0	***
Avg. SAT at 1st choice college *n/a*	0.36	0.36	0.2		0.35	0.36	0.4	
No previous score, or *n/a*	0.32	0.28	−1.9		0.31	0.28	−1.4	
Previous score seemed fair	0.22	0.14	−3.8	***	0.22	0.15	−3.6	***
Previous score seemed unfair	0.46	0.58	4.9	***	0.47	0.57	4.2	***
Nervous about SAT? (*n/a*)	0.21	0.24	1.5		0.21	0.24	1.7	
Nervous about SAT? – very	0.18	0.27	4.5	***	0.19	0.27	4.1	***
Nervous about SAT? – a bit	0.44	0.39	−2.0	*	0.44	0.39	−2.1	*
Nervous about SAT? – no	0.17	0.098	−3.9	***	0.16	0.1	−3.7	***
Score important? (*n/a*)	0.21	0.24	1.4		0.21	0.24	1.6	
Score important? – very	0.63	0.67	1.7		0.64	0.67	1.1	
Score important? – somewhat	0.15	0.086	−3.9	***	0.15	0.088	−3.5	***
Prefer 2- or 4-yr. college? (*n/a*)	0.11	0.1	−0.6		0.13	0.1	−2.1	*
Prefer 2- or 4-yr. college? – 4-yr	0.89	0.9	0.6		0.87	0.9	2.1	*
Degree goal: (*n/a*)	0.27	0.25	−0.8		0.27	0.25	−1.2	
Degree goal: < = BA	0.024	0.01	−2.0	*	0.02	0.011	−1.4	
Degree goal: BA	0.2	0.12	−4.1	***	0.18	0.13	−2.8	**
Degree goal: > = BA	0.51	0.61	4.5	***	0.53	0.61	3.6	***
Prefers public college	0.61	0.73	5.4	***	0.64	0.73	3.8	***
Public or private OK	0.39	0.27	−5.4	***	0.36	0.27	−3.8	***

Note. Subclassification on demographic variables fails to address large differences on these variables

The quoted passage makes sense only for unstratified comparisons, because of its reference to *t*-statistics. However, if the *t*-statistic can be replaced with an analogue that also makes sense in the stratified case, then we can also ask of

the subclassified Powers-Rock sample whether Cochran might have thought regression suitable to remove remaining observed covariate bias. To compare unstratified treatment and control groups on a covariate x, we scale the difference of coached and uncoached x-means, $\bar{x}_t - \bar{x}_c$, by the reciprocal of its permutational SD (i.e. the SD of the quantity $\bar{x}_t - \bar{x}_c$ under random permutations of the labeling of observations as treatment [t] or control [c]). (This permutational SD has the advantages that it can be calculated exactly, rather than estimated, and that its motivation does not require subjects to constitute a simple random sample of a population (Hansen & Bowers, 2008). Ordinarily it will be similar to the pooled standard deviation considered by Cochran.) For treatment-control comparisons of x that account for the Race×SES poststratification, we take weighted averages of stratum-wise differences of means on a covariate, then scale by their corresponding permutational SDs. (For these SDs, the relevant hypothetical randomizations are those that shuffle assignments to treatment or control within strata.) One has choices among weighting schemes when combining $\bar{x}_{ts} - \bar{x}_{cs}$ across strata s; our comparisons will weight strata by the harmonic mean of the numbers of treatment and control subjects they contain, which is the weighting implicitly used to construct the coefficient on the treatment variable in the ordinary least squares regression of the covariate in question on treatment and stratum dummies. (See Sect. 9.3.3 for a bit more discussion of harmonic weighting, or Hansen & Bowers, 2008, for a more systematic development of the issue.) Another slight modification of Cochran's suggestion, applying to comparisons both with and without stratification, has to do with our handling of the possibility of noncomparability on continuous measurements due to differences in spread or skewness of the variable, rather than mean differences. Whereas Cochran suggested comparisons on higher-order moments, we instead compare means of derived variables constructed to focus attention on continuous covariates' tails. For instance, rather than comparing treatment and control means in, say, PSAT-V, $(\text{PSAT-V})^2$, and $(\text{PSAT-V})^3$, we compare them in their means on PSAT-V and on min(PSAT-V, 40), a variable equal to PSAT verbal score if the score is less than 40 and equal to 40 otherwise, and on max(PSAT-V, 60). These derived variables track the presence and magnitude of PSAT-V score deviations (from the national mean of verbal PSAT scores, roughly 50) exceeding about one population SD.

Table 9.2 shows that coached and uncoached students differ quite sharply in demographic terms – differences between the groups are quite statistically significant on all but a few of the variables, with many of the z-scores well above 2.0 in absolute value. The Race×SES stratification markedly reduces the differences, eliminating them entirely on 7 of the 22 demographic indicators. In one exceptional case, that of the uppermost quartile of (student-reported) parents' income, stratification has made things worse, but for the remaining demographic variables it generally helps. Previously significant differences on whether English was the student's first language have been made negligible. As regards other variables that the Race×SES stratification does not specifically address, one could ask for a bit more: Controlling for father's education has reduced differences in mother's education, for instance, but 3 of the 4 mother's education

variables exhibiting significant differences before stratification remain significant after it. (Tables 9.2–9.4 were prepared using the RItools add-on package for R (Bowers, Fredrickson, & Hansen, 2010), which also helps with propensity score diagnostics.)

Table 9.3 shows scholastic achievement variables. There are many variables describing subjects' scholastic achievements around the time of their tests and coaching decisions; this is to the advantage of the study. But a good many of these variables have large z-values, which would appear to be to the disadvantage of the study. Stars appear where the z-value exceeds 2.0 in magnitude, multiple stars where it exceeds 2.6 or 3.3. Before stratification, fully 23 of 46 variables received at least one star. Stratification reduced this number to 12 of 46 variables. It is encouraging that controling only for demographic variables, ethnicity, and father's education controls implicitly for some of these nondemographic variables. However, even with this control well more than several variables appear by Cochran's criterion to demand adjustment. Among those imbalances that remain after subclassification are imbalances in the tails of the pretest distributions: in the constructed variable max(PSAT-V, 60), for example, where the uncoached students' mean exceeds the coached students' by more than three standard errors. PSAT scores are likely to be among the strongest predictors of the posttest.

If that isn't discouraging enough, Table 9.4 bears worse news. Without stratification the coached and uncoached differed significantly on 12 of 20 measurements describing subjects' attitudes to college and to the SAT; with Race $\times$ SES stratification they differed significantly on 13 of 20 such measurements. If addressing demographic differences helped implicitly with differences in scholastic achievement, it did little to nothing to help with differences in attitudes toward the test. These variables are of clear importance both to coaching decisions and to test performance, and their presence is one of Powers and Rock's data's most notable strengths.

Clearly, the Race $\times$ SES subclassification does not do enough. Indeed, by Cochran's standards, the situation now seems quite poor – not just "several" but tens of unsigned z values exceed 2 or more. The answer to the question of Sect. 9.1.2 Step 2 – Is it plausible that treatment is good as randomized within subclasses of the exact matching variable? – is a resounding no.

9.3 Matching Structures and Algorithms

How much better can propensity score matching do? In order to answer the question unequivocally, it is necessary first to review some modern matching methods.

By "matching" many will understand pair matching, the joining of unique treatment subjects to unique controls. After this, outcome analysis would be

based on paired differences, as would assessments of covariate balance.[1] With pair matching, and with generalizations of it to be discussed presently, matching amounts to arranging some part of the sample into finely grained strata; after matching, any of a variety of off-the-shelf estimation methods accommodating sparsely stratified data are available for diagnosing the match and then for using it to estimate treatment effects.

Pair matching generalizes easily enough to matched triples, the creation of subgroups consisting of a single treatment and two controls, and on to 1:k matching, wherein treatment subjects are joined to k controls each. Analysis might then be based on the differences between treatments' measurements and averages of controls' measurements. Another generalization is to matching with a varying number of controls, discussed by Ming and Rosenbaum (2000). Analysis can again begin with differences between treatment subjects' measurements and averages of their matched controls' measurements, although summarizing these differences across matched sets is less routine, as contributions from larger matched sets now call for upweighting relative to contributions from matched pairs, or matched sets with fewer controls. Given a weighting protocol to accommodate matched sets of varying structures, one can allow matched sets with multiple treatment subjects, i:1 matches with $i > 1$, in addition to sets with multiple controls. This allowance becomes helpful when there are values of the matching variable (or variables) that are better represented among treatment subjects than among controls – as is almost guaranteed to occur when one is matching on propensity scores.

9.3.1 Nearest-Available Versus Optimal Matching

Figure 9.1 presents an artificial data set modeled on an unpublished gender-equity study. Men and women university scientists within various departments were to be compared in terms of their lab space assignments, but first it was necessary to match them on factors that might confound the comparison. The actual study matched on total grant funding and several other factors, but to simplify the illustration we consider grant funding alone. The actual study used full matching, which will be reviewed in Sect. 9.3.2; however, this section uses the gender equity data to contrast two approaches to pair matching.

[1] In order that the paired differences be legitimately treated as independent, it is important that distinct treatment subjects be matched to distinct controls: When both A and B are matched to C, the A-C difference and the B-C difference cannot ordinarily be treated as independent. A few estimation techniques have been proposed for nearest-neighbor matching, which pairs subjects without regard to whether or how often they are paired to subjects elsewhere in the sample, permitting the pairs to overlap in arbitrary ways (Abadie & Imbens, 2006), but in the main methods for paired data assume no replacement, as does the remainder of this paper.

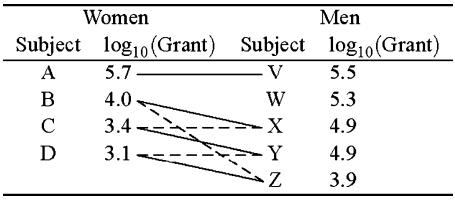

Fig. 9.1 Pair matching for a gender-equity study. Women and men scientists are to be matched on Grant Funding. Solid lines indicate the optimal pair match, for which the sum of matched differences on the matching variable is 3.4; dotted lines, a pair match determined using nearest-available matching, for which the corresponding sum is 3.6

Nearest-available, or *greedy*, matching algorithms move down the list of treated subjects from top to bottom, at each step matching a treated subject to the nearest available control, which is then removed from the list of controls available at the next step. Matchings are made at a given stage without attention to how they affect possibilities for later matchings. In the equity matching problem posed in Fig. 9.1, a nearest-available algorithm for pair matching would first match A to V, then B to Z, C to X, and finally D to Y, for a total *cost* (sum of absolute differences in log Grant Funding) of 3.6. Having matched A to V, Z is the nearest available potential match for B, but matching B to Z is in fact greedy, in that it forces C and/or D to be more poorly matched at the next stage. In contrast, optimal matching algorithms optimize global, rather than local, objectives. The optimal solution for the problem of pairing each of Fig. 9.1 women with one of its men joins A to V, B to X, C to Y, and D to Z, for a total cost of 3.4.

For pair matching with a large reservoir of controls, nearest-available algorithms often do nearly as well as optimal algorithms (Rosenbaum & Rubin, 1985b). But absent an excess of available controls, or with unfortunate orderings of the list of treated subjects, nearest-available algorithms can do much worse than optimal ones. Optimal pair matches are readily determined using the pairmatch function in R, a part of the optmatch add-on package (Hansen, 2007).

9.3.2 Full Matching and Full Matching with Restrictions

Full matching subdivides a sample into a collection of matched sets consisting either of a treated subject and any positive number of controls or a control subject and any positive number of treated persons. It generalizes pair matching and matching with multiple controls, and often leads to markedly closer matches. For example, one can readily verify that the optimum placement of the four women and five men in Fig. 9.1 into matched sets of one woman and one *or* two men matches A to V and W, B to X, C to Y, and D to Z, with total cost 3.8. The optimal full match, depicted in Fig. 9.2, reduces this sum to 3.6. Rosenbaum (1991) introduced full matching, Gu and Rosenbaum (1993) did a simulation study of it, and Marcus

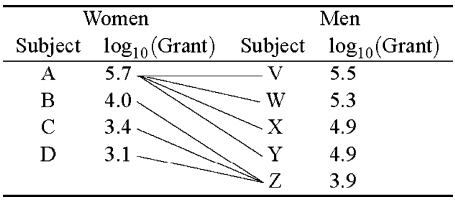

Fig. 9.2 Full-matching solution to the matching problem posed by Fig. 9.1

(2000) made use of it to assess the Head Start compensatory education program. Using R, optimal full matches can be found using the fullmatch function of the optmatch package.

Coincidentally, the optimal full match avoids matching any woman to a man whose grant funding differs from hers by more than a factor of 10 – a requirement that full matching enabled me to insist upon in the actual study on which the example is based. In terms of the matching variable, $\log_{10}$ of grant funding, the requirement was that matched subjects differ by no more than 1: I imposed a *caliper* of 1 on the log-grant variable.[2] In the example problem, matching within this caliper would have been compatible neither with pair matching nor with matching with one or two controls.

Had the caliper been narrower, say 1/2 rather than 1 unit of the matching variable, then it would become impossible to find matches for several subjects. Removing these subjects (D, X, and Y) from the matching problem, full matching becomes feasible, culminating in matched sets {A, V, W} and {B, C, Z}. On the other hand pair matching of the remaining subjects would not work, not because any one subject lacks a permissible match but because arranging permissible matches into nonoverlapping pairs is impossible. The distinction reflects a general and important feature of full matching for matching problems that involve calipers or other prohibitions of certain matches: Barring those units with no permissible matches, full matching is always able to arrange the remaining units into nonoverlapping matched sets, even when it is not possible to arrange those units into pairs, 1:k tuples or other specified matching structures. This generality makes full matching a useful starting point for matching within calipers.

9.3.3 *Matching Structures and Effective Sample Size*

A less desirable aspect of full matching is its tendency to collect many observations in a few rather lopsided matched sets. In Fig. 9.2, for example, full matching has created two matched sets, a 1:4 and a 3:1 structure, after which the data supports only two matched comparisons, whereas in principle it would have been possible to arrange for four matched comparisons (either four pairs, omitting a potential

[2] In this context, subject-matter intuition decided the width of the caliper. When matching on propensity scores, the data can be used to choose calipers; see Sect. 9.4.2.

control, or three pairs and a 1:2 triple). Four matched comparisons would have both increased the sum of matched discrepancies and violated the caliper of 1 on the matching variable, and these coarser matches could well translate into distortions in the matched comparisons that are the purpose of the exercise. Yet settling for fewer matched comparisons surely reduces the resolution of whatever picture eventually will emerge, even if it does this in the interest of promoting the faithfulness of the picture to what it aims to depict.

Because high resolution and low distortion are aims that are in competition with one another, it is useful to try to quantify them, in order to explicitly manage the trade-off. As a measure of the resolution supported by the match, translate the aggregate sizes of matched structures into matched pair equivalents, as follows: In each matched set – more generally, in each stratum, s – calculate the harmonic mean of the number of treatment groups subjects and the number of controls, $h(m_{st}, m_{sc}) = [(m_{st}^{-1} + m_{sc}^{-1})/2]^{-1}$; add these harmonic means across strata to determine the *effective sample size*. The units of this measure are matched-pair equivalents: A matched pair contributes $h(1,1) = 1$, so that in matched pair designs the effective sample size is simply the number of matched sets. A matched quadruple contributes somewhat more than a matched pair, $h(1,3) = 1.5$, but less than twice as much, fitting with the intuition that two matched pairs would enable two distinct treatment-control comparisons whereas the matched triple enables only one – pairs of pairs add more resolution than do single matched quadruples. A matched set, that is a stratum s with either $m_{st} = 1$ or $m_{sc} = 1$ or both, never contributes more than twice what a matched pair contributes, as $h(1,x) = h(x,1) > 2$ for all $x < \infty$, and no stratum s contributes unless both $m_{st} > 0$ and $m_{sc} > 0$. Competing candidates for the designation effective sample size, such as the number of matched sets, the number of matched treated subjects, or the number of subjects of both kinds, lack comparable graces.[3]

The full match shown in Fig. 9.2 has an effective sample size of $h(1,4) + h(3,1) = 3/2 + 8/5 = 3.1$ pairs. Each of the pair matches depicted in Fig. 9.1, on the other hand, consists of four pairs and accordingly has effective sample size 4. The reduced distortion (to the eventual comparison of matched men and women in measures of their working conditions) that is bought by better matches on grant funding comes at a price of decreased resolution, and comparing effective sample sizes quantifies that price. Looking aside from issues of bias,[4] to reduce the standard

[3] A more general motivation for the formula is that in the ordinary least squares regression of a variable v on the treatment variable, allowing separate intercepts for each stratum, the standard error of the treatment coefficient is inversely proportional to the square root of the sum of these harmonic means. This coefficient is in turn interpretable as an average of matched differences $\bar{v}_{st} - \bar{v}_{sc}$, weighted in proportion with $h(m_{st}, m_{sc})$, which is the minimum-variance estimate of the contrast in the homoskedastic linear model with constant effects of treatment on v across strata (see e.g., Kalton, 1968).

[4] Incidentally, in this example making sense of bias is particularly thorny, as the example involves contrasts on a trait, gender, which is not readily manipulable. (See the excellent discussions of Holland, 1986a, 1986b; Rubin, 1986.)

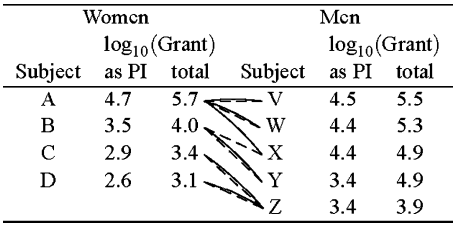

Fig. 9.3 Matching with restrictions (*dotted lines*) and on an alternate matching variable selected to reduce separation between the groups (*solid lines*) for the problem posed by Fig. 9.1. (The restrictions are `min.controls=1/2`, `max.controls=2`; matching on the alternate variable, here log grant as PI (principal investigator), is done within calipers of 1.0 on the original matching variable, log total grant)

errors of gender contrasts made on the basis of the full match to the levels of those that either pair match would support, one expects to have to find one additional pair of treatment and control subjects who are suitable to be matched.

Optimal full matching reduces matched discrepancies to the lowest possible levels (Rosenbaum, 1991). Insofar as controlling the matching variable reduces bias of matched comparisons, this prevents distortions that might otherwise be present in them; but it does so at the expense of effective sample size. Hansen (2004) used full matching with structural restrictions: explicit limits on the numbers of controls that could be matched to one treated subject and on the number of treated subjects permitted to be matched to a single control. The dotted lines in Fig. 9.3 demonstrate the result of full matching research scientists under the restrictions that no more than two controls share a match in the treatment group and no more than two treatment group members share a matching control: in the syntax of optmatch, `fullmatch(loggrant, max.controls=2, min.controls=1/2)`. The structural restrictions improve unrestricted full matching's effective sample size from 3.1 to 4 pair-equivalents.

Unfortunately, structural restrictions bring an undesirable complication to the workflow of full matching: The combination of structural restrictions with calipers may render a matching problem infeasible. In Figs. 9.1 or 9.3, the restrictions `max.controls=2` and `min.controls=1/2` are jointly compatible with a caliper of 1.0 on the matching variable, for example, but not with a caliper of 0.8, which would prevent B from being matched to X or Y. This is a complication, not a limitation, because the matching algorithm implemented by optmatch finds and quickly reports such infeasibility when it is present, and one adapts by simply reducing or lifting the structural restrictions that caused it. Narrowing the caliper on the log of grant funding to 0.8 in Fig. 9.3, for instance, `fullmatch()` reports infeasibility unless `max.controls` is at least 4 and `min.controls` is no more than 1/3. Finding each of these cutoffs requires a line search, however; although optmatch has dedicated functions to conduct the line searches, `minControlsCap` and `maxControlsCap`, the process is a bit more time consuming, requiring up to a few minutes in problems for which matching alone takes seconds.

An alternate strategy to limit full matching's profligacy with effective sample size is to find a primary matching variable on which the treatment and control groups are less separated. In Fig. 9.3, the as-PI grant funding variable plays this role: On it, men's and women's means differ by 80% of a pooled SD, whereas the two groups were separated by 95% of a pooled SD on the original matching variable, log total grant funding. The solid, curved lines in Fig. 9.3 represent an optimal full match on the new matching variable with calipers of 1.0 on the original matching variable. Because no additional structural restrictions are used (i.e. `min.controls` or `max.controls` arguments to `fullmatch`), this matching problem is always feasible, in the sense that full matching finds matches for each matching candidate with an opposite-group counterpart within caliper distance of it. This approach may be more or less sparing with effective sample size than matching with restrictions, depending upon the alternate matching variable; in this case it is a bit less sparing, yielding an effective sample size of 3.8 as opposed to 4.0 pair-equivalents.

To summarize: Optimal full matching on a variable is a very effective strategy for setting up comparisons between subjects with similar values of the variable. It places each subject into some matched set, except perhaps if specified potential matches have been forbidden in advance, in which case any subjects for whom all possible matches have been forbidden are, necessarily, excluded from matching. A drawback is that it may lead to relatively small effective sample sizes, even when it makes use of most or all of the available sample. One remedy is to match with structural restrictions, as demonstrated in an earlier analysis of Powers and Rock's data (Hansen, 2004); an operationally simpler remedy is to full match on another matching variable, a variable on which the groups are less separated, perhaps within calipers of the original matching variable.

9.4 Estimating and Matching on Propensity Scores

This section narrates the creation and refinement of several related propensity-score full matches of Powers and Rock's sample. A documented transcript of R code used to create the propensity scores, the matches, and the accompanying diagnostics is available from the author upon request.

9.4.1 Matching the Full Sample on an Ordinary Propensity Score

In an observational study, one seeks to measure and adjust for a collection of *pre-exposure* variables, $\mathbf{X} = (X_1, \ldots, X_k)$, with the property that conditional on $\mathbf{X}$ the assignment to treatment conditions (Z) is independent of potential responses, Y_c and Y_t (Holland, 1986b, Sec. 4.5). Our candidate for such a collection of x-variables is the union of those appearing in Tables 9.2–9.4. There are far too many to attempt to match on all of them at once. This is where propensity scores come in.

The propensity score is the conditional probability of assignment to treatment given the covariates, $\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x})$, or a monotonic transformation thereof. In particular, $\varphi(x) := logit(\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x}))$ is sometimes called the *linear propensity score*, as the most common model for the regression of Z on $\mathbf{X}$, that it is logistic-linear in $\mathbf{x}$, entails that $\varphi(\mathbf{x})$ is linear in $\mathbf{x}$. Rosenbaum and Rubin (1985b) suggested matching on estimates of $\varphi(\mathbf{x})$, rather than estimates of $P(Z = 1|\mathbf{X} = \mathbf{x})$, because estimates of $\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x})$ often cluster near 0 and 1 while linear propensity scores remain more dispersed; [15] offers additional considerations in support of this recommendation.

If the form of the regression of the treatment variable on covariates were known, we would certainly use that knowledge in our estimate of the propensity score. More commonly, as here, we know nothing about that regression. In these cases, Hansen (2009) suggested that the important thing is to avoid gross misspecification of the model. If using logistic regression, or something similar, one should aim to chose predictors in such a way as to bring the (true) linear propensity score, $\varphi(\mathbf{x})$, within their linear span. If this can be achieved, or nearly achieved, then moderate overfitting or underfitting of the score is unlikely to be harmful. To this end, we model Z as logistic in all of the variables appearing in Tables 9.2–9.4, expanding each of the measurement variables into natural cubic splines (Ruppert, Wand, & Carroll, 2003, Sec. 3.7.2) with four degrees of freedom (d.f.). It would also be possible to add interactions to the model. Hansen (2004) used stepwise regression to select from among the many possible interactions; with Bayesian methods, one could use more of them by incorporating penalties on the second-order terms.

Logistic regression is likely to over fit, making the treatment and control groups more separated on the estimated propensity score than they would be on the true propensity score, if it were available. Indeed, it can be shown that whenever a linear combination of a covariate exists such that the two groups have no overlap on that linear combination, then logistic regression will return a linear predictor on which the two groups are fully separated (Hastie, Tibshirani, & Friedman, 2001, p. 111), even if overlap on the true propensity score is substantial. The downside of logistic regression's tendency to exaggerate separation is that it can make it hard to match closely on estimated propensity scores. This turns out to be less of a problem than it might at first seem, however, because precise matching on the estimated propensity score will turn out not to be necessary. The upside to logistic regression's tendency to separate the groups is that a plot comparing the groups on $\hat{\varphi}(\mathbf{x})$, as seen in Fig. 9.4, reveals immediately whether the groups can be separated by a linear combination of x-variables. If they can, then propensity matching is made difficult; but by the same token any comparison between the groups is, at least in terms of $\mathbf{X}$, inherently extrapolative.

Matching is performed separately within each Race$\times$SES subclass. This matching within subclasses forces exact matching on race and SES, as was decided at Step 1 (see Sect. 9.1.2) of the matching procedure, as executed in Sect. 9.2.2; furthermore, trading one large matching problem for many smaller ones drastically improves computation time. The algorithm optmatch employs requires on the order of $n^3 \log(n)$ operations, where n is the number of subjects to

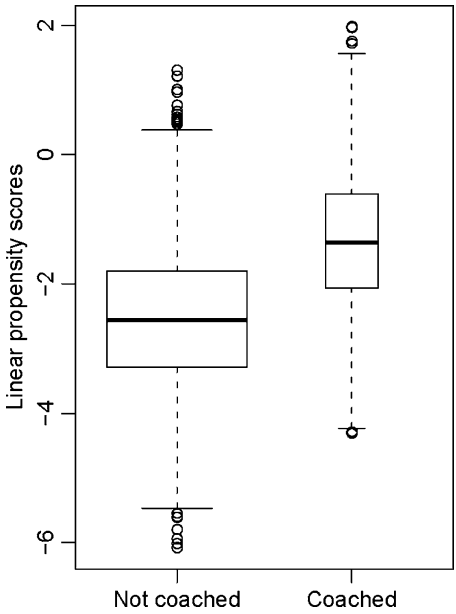

Fig. 9.4 Estimated linear propensity scores, $\hat{\varphi}(\mathbf{x})$, in the coached and the uncoached groups

be matched (Hansen & Klopfer, 2006); exchanging $n = 4{,}000$ for 12 n's summing to 4,000 reduces this time estimate by 97%. Subdivided in this way, the whole full matching problem takes a second or two to solve using a modern computer.

Full matching on the propensity score does wonders for the imbalances on covariates found at Step 2 of the matching workflow, demonstrated in Sect. 9.2.3. Examining how full matching changes these imbalances is part of Step 5. With no stratification whatsoever, the chi-square statistic combining the imbalances (Hansen & Bowers, 2008) is extremely significant, 486 on 66 d.f.; the root mean square (RMS) of the covariate-wise measures of imbalance is, accounting for correlations among them, $(486/66)^{1/2} = 2.7$ – well above even Cochran's more generous benchmark (2.0). Furthermore, although at Step 2 the Race × SES subclassification was found to cut χ^2 nearly in half, to 287 on 61 d.f., even so the RMS of the z-statistics was still 2.1: worse than Cochran would have thought salvageable, and markedly worse than what one would expect under randomization. At this first pass through Step 5, we find full matching on the propensity score to have reduced covariate imbalance measurements by an order of magnitude, to $\chi^2 = 17$, on 69 d.f., for an RMS z-measure of 0.5. Cochran's criterion is easily met, and moreover, the randomization p-value is indistinguishable from 1: Balance on *observed* covariates is now better than what randomization would be expected to produce. (Unobserved variables are another matter.)

Step 5 also requires that we assess matched propensity score discrepancies and the effect of the matching on effective sample size, and on both of these counts the full match within Race × SES subclasses leaves something to be desired. Outlyingly large matched discrepancies exist on $\hat{\varphi}(\mathbf{x})$: Although half are lower than 3%

of a pooled SD in $\hat{\varphi}(\mathbf{x})$, some matched subjects are separated by as much as 2.4 SDs of $\hat{\varphi}(\mathbf{x})$. The effective sample size is only 679 pair-equivalents. With 500 treatment subjects and 3,500 controls, this is only slightly better than what 1:2 matched triples would have given, the sample size equivalent of 667 matched pairs. On two counts, then, Step 6 directs us to try again from an earlier stage of the procedure. Nothing calls into question the choice of stratifying variables or the propensity model, so there is no need to back up as far as Steps 1 or 3. Rather, we revisit Step 4, attending first to the large matched discrepancies.

9.4.2 Matching Within Propensity Score Calipers

Figure 9.4 shows a few coached students whose propensity scores fall outside of the range of propensity scores for uncoached students and a good number of uncoached students with estimated scores below those of anyone who received coaching. Having an estimated propensity score outside of the range of propensity scores estimated for the comparison group is sometimes taken as a sign of a subject that must be excluded from the analysis (e.g. Dehejia & Wahba, 1999). However, the fact that propensity scores are ordinarily known to be overfitted suggests that falling outside of the comparison range in this way may often be an artifact of the fitting routine. The asymptotic theory of Hansen (2009) suggested that a weaker criterion is more appropriate: impose calipers on the propensity score, calipers strict enough to prevent outlying matched discrepancies on it. The calipers are imposed for all subjects, but for those subjects near the extremes of their groups' $\hat{\varphi}(\mathbf{x})$ distributions, they have the side effect of excluding the subject if it has no counterpart within caliper distance.

In full matching without calipers, the largest matched discrepancy exceeds the 95th percentile of matched discrepancies by a factor of six (2.38 as compared to 0.40 pooled SDs in the propensity score). Let us reduce this factor to something closer to, say, 2. Imposing a caliper of half of a pooled SD in $\hat{\varphi}(\mathbf{x})$, s_p, has several effects: It slightly reduces the effective sample size, from 679 to 676; it makes imbalance even less, moving χ^2 from 17 to 11; it brings the maximum matched discrepancy on $\hat{\varphi}(\mathbf{x})$ down to 0.499, just more than double the 95th percentile of such discrepancies (0.24); and it excludes 10 coached students, 2% of the treatment group, and 140 uncoached students from matching.

As the objective of the analysis is to estimate the benefit of coaching, leaving aside a few potential controls is not a problem. Rejecting treatment group members may be a problem, as the final matched analysis will be unable to speak to effects of the treatment on them (Rosenbaum & Rubin, 1985a). To avoid rejecting treatment group members, Rosenbaum and Rubin (1985b) first imposed a propensity score caliper and then lifted it for those treatment group subjects more separated from any member of the control group than the width of the caliper. This strategy makes sense, but only within reasonable limits. If the data contain no suitable comparisons for a member of the treatment group, then no basis exists for matched estimation of

the treatment's effect on it. Rather than pretending otherwise, it would be better to restrict the scope of the analysis, estimating treatment effects only for a proper subset of the intervention group.

Narrowing the scope of the analysis to only 98% of the intervention group, as full matching with an $s_p/2$ caliper would force us to do, changes the focus of the analysis only very slightly. However, as optmatch's `caliper` function permits the user to relax a caliper requirement selectively, we can easily take a moderate step in the direction of Rosenbaum and Rubin (1985b), permitting those coached students without counterparts within the $s_p/2$ caliper to be matched to uncoached students as far as s_p away. This brings 5 of the missing 10 back into the analysis, so that only 1% of treatment group subjects are rejected. The 95th percentile of matched discrepancies is about 0.24 s_p. As the five newly matched coached students are matched to controls much farther from them than $s_p/2 > 2*0.24\ s_p$, our outlier condition is violated; but the violation is contained, affecting only five cases, and the remaining matched discrepancies are all less than $s_p/2$. Balance is slightly diminished but remains excellent:$\chi^2 = 14.1$, on 69 d.f.; p is effectively 1.

9.4.3 *Focusing the Propensity Score*

The first match we considered, in Sect. 9.4.1, had another potential shortcoming in addition to its large matched discrepancies on the propensity score: Its effective sample size was disappointingly small. The modified match of Sect. 9.4.2 removed outlying matched discrepancies at the expense of slightly worsening effective sample size. This section returns to the sample size issue, addressing it by incorporating a second propensity score in the matching criterion.

The propensity score used thus far strives just as much to balance variables of possible prognostic relevance, for instance the number of semesters of foreign language taken, as it does to balance variables of clear prognostic relevance, such as pretest scores. Alternatively, one could estimate and match on a propensity score based only on a subset or lower-dimensional reduction of the x-variables, call it $\tilde{\mathbf{X}}$, selected or constructed so as to summarize prognostic information in $\mathbf{X}$. Hansen (2008b) discussed prognostic summaries of this type, describing conditions under which if it is sufficient to adjust for $\mathbf{X}$ then it is sufficient to adjust for $\tilde{\mathbf{X}}$. Methods for confidently isolating $\tilde{\mathbf{X}}$ from within $\mathbf{X}$ are a topic of current research, and are somewhat beyond the scope of this paper. What is within the scope of the paper is to demonstrate how even a crude prognostic reduction of $\mathbf{X}$ can be used to complement and focus a propensity score formed in the ordinary way.

As a prognostic reduction of the covariate, I take all of the pretest measures, PSAT math and verbal and, where available, prior SAT math and verbal scores, along with indicators of availability of both of these sets of scores. (I have imputed median values on these variables to those students who either did not take the relevant test or did take it, but may have done so after their coaching or after their decision not to obtain coaching; see Hansen, 2004.) In addition, I include two

functions of these and the remaining covariates, constructed as follows: First, fit two regression models to the control group, one predicting math scores on the posttest while the other predicts verbal posttest scores, both using all covariates as predictors. (For fitting, I simply use ordinary least squares.) Second, extrapolate the fitted regressions to the entire sample. These $\hat{y}_c(\mathbf{x})$'s, the estimated conditional mean potential responses to control, are the prognostic summaries which, taken together with the pretest variables themselves, form $\tilde{\mathbf{x}}$. One then fits a *focused* propensity score, $\hat{\varphi}(\tilde{\mathbf{x}})$, by logistic regression of z on $\tilde{\mathbf{x}}$, as opposed to $\mathbf{x}$.

Extrapolating conditional mean fits from controls to the treatment group is risky when the groups exhibit separation on $\mathbf{x}$, as indeed they do in this example. It is ironic that we are led to extrapolate this way in the interest of focusing the propensity score: A central purpose of propensity adjustment itself is to reduce and mitigate this sort of extrapolation, to which regression methods are notoriously vulnerable (Rubin, 1997). But we need not rely on the focused propensity scores alone. To retain advantages of adjustment for the ordinary propensity score, which does not inherit $\hat{y}_c(\cdot)$'s vulnerability to separation, we can match on our focused propensity score but within the propensity score calipers that were described in Sect. 9.4.2.

Figure 9.5 shows that the coached and uncoached groups are considerably less separated on $\hat{\varphi}(\tilde{\mathbf{x}})$ than on $\hat{\varphi}(\mathbf{x})$. The groups' means differ on the ordinary propensity score by 1.1 SDs in it, but by only 42% of an SD in the focused

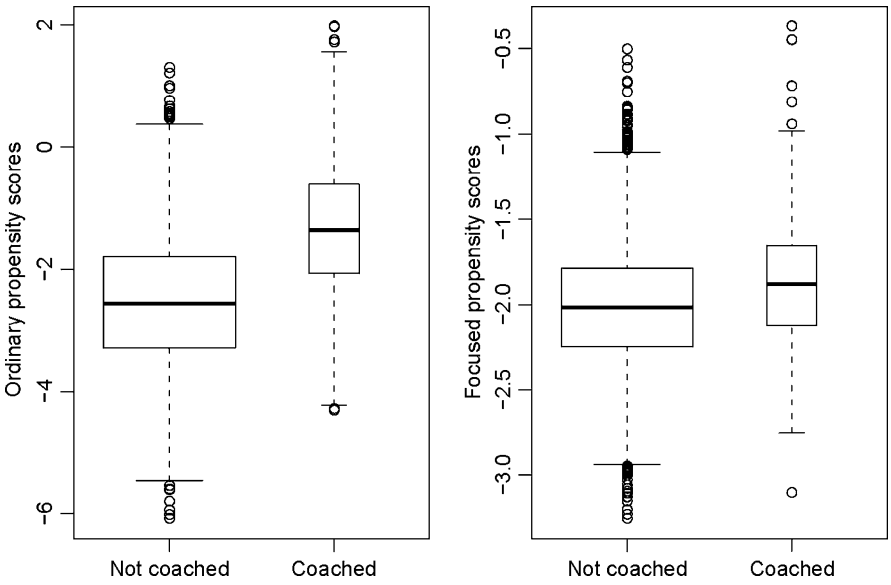

Fig. 9.5 Coached and uncoached students' (**a**) propensity scores, $\hat{\varphi}(\mathbf{x})$, at left; and (**b**) focused or "prognostic" propensity scores, $\hat{\varphi}(\tilde{\mathbf{x}})$, at right. Focusing the propensity score markedly reduces the apparent separation between the groups: At left, the group means differ by 1.1 pooled SDs of $\hat{\varphi}(\mathbf{x})$; at right, by only .42 pooled SDs of $\hat{\varphi}(\tilde{\mathbf{x}})$

propensity score. Recall that in the illustration in Sect. 9.3.3, effective sample size was increased by replacing full matching on one variable with full matching on a second variable, a variable on which the groups were less separated, with the added constraint of calipers on the original variable. In like fashion, full matching on $\hat{\varphi}(\tilde{\mathbf{x}})$ within calipers of $\hat{\varphi}(\mathbf{x})$, rather than on $\hat{\varphi}(\mathbf{x})$ itself within calipers on the same variable, increases effective sample size from 677 to 701 matched-pair equivalents. Imbalance overall is increased, from $\chi^2 = 14$ to $\chi^2 = 56$ on 69 d.f., but not past Cochran's limits (the RMS of z-measures is $(56/69)^{1/2} = 0.9$) nor to a level incommensurate with what randomization might have produced ($p = 0.9$). Imbalance on the eight pretest and derived variables, the targets of our second propensity score's focus, is quite effectively controlled: $\chi^2 = 1.2$ on 8 d.f., for an RMS imbalance of 0.4 on these central prognostic variables; looking at these variables alone, the randomization p-value is indistinguishable from 1.

9.5 Results

9.5.1 Matched, Permutation-Based Estimates of the Treatment Effect

Had coaching been allocated at random within matched sets, more or less assumption-free inferences about the treatment effect could be made using randomization-based permutation tests. Without randomization, permutation tests are not assumption free, but they dispense with various ancillary assumptions (Rosenbaum, 2002b). In particular, according to Hansen (2009), if potential responses in the absence of treatment (Holland, 1986b), Y_c, were known to be conditionally independent of assignment to treatment conditions, Z, given the covariates, $\mathbf{X}$, then propensity score matching could bring about a situation comparable to that of randomized studies. If the propensity match balances all of the covariates at least crudely and balances the main prognostic variables well, and if it avoids outlying matched differences on the propensity score, then asymptotic inferences based on a normal approximation are valid under similar data conditions as would be needed to justify the approximation after random assignment. These balance and outlier requirements are, of course, precisely what the matching and diagnostic procedures of Sects. 9.2 and 9.4 sought to ensure.

Under the conditional independence assumption, then, we can test hypotheses about coaching effects, at least for the 99% of coached students included in our match. For simplicity, we consider only hypotheses according to which coaching effects are the same for everyone. (Hypotheses stipulating varying treatment effects somewhat complicate notation and calculations.) Consider for instance the hypothesis $H : Y_t \equiv Y_c + 50$, that coaching would increase any subject's SAT math score by 50 points. For each coached student i in the sample, the observed posttest measure, y_i, reveals i's potential response to treatment, y_{ti}. According to H, his potential response to control is implicitly revealed to be $y_i - 50$. For the purpose of

testing this H, for each i compute $\tilde{y}_i$ as $y_i - 50$, if i was coached, or as y_i itself, if i was not coached. Now calculate the matched correlation of $\tilde{y}$ and z, the indicator of coaching, and compare it to the distribution of such correlations as z is independently permuted within matched sets. Then H is sustained at level α, and 50 goes inside our $(1 - \alpha)100\%$ confidence interval for coaching's putatively constant effect on SAT math score, if and only if $\rho_{\tilde{y},z|match}$ falls within the central $(1 - \alpha)$ 100% of this permutation distribution. By repeating the procedure with hypotheses to the effect that $Y_t - Y_c \equiv 60$, 70, and so forth, as well as 40, 30, and so on, one bounds the extent of the confidence interval; continuing to iterate over a finer grid of hypothesized treatment effects, one delimits the interval with arbitrary precision.

Applying this approach with our matched sample (and with grids 1 SAT point wide) gives 95% confidence intervals of [−10,9] and [12,30] for effects of coaching on verbal and on math scores, respectively. For point estimates, define α' to be the largest α such that the $(1 - \alpha)100\%$ confidence interval has positive extent. The centroids of the $(1 - \alpha')100\%$ confidence intervals, also known as Hodges-Lehmann point estimates, are 0 and 21, respectively. These results are quite similar to those Hansen (2004) reported using a somewhat different match and a model-based method of analysis.

9.5.2 *Matched Outcome Analysis with Permutation Tests and Covariate Adjustment*

Rubin (1979) and Rubin and Thomas (2000), among others, recommended that propensity score adjustments for all potentially relevant preexposure variables be combined with regression adjustments for a few of the most important ones. This idea can be combined with permutation-based inference for treatment effects. When regression, matching, and permutation-type inferences are suitably combined, the validity of the inferences need not depend on the correctness of a working model for Y, or for Y_c or Y_t. Instead, it can be warranted by the combination of the assumption of conditional independence of Y_c and Z given $\mathbf{X}$, the trueness of our implementation of the propensity scoring, matching and diagnostic methods, and by supporting theory asymptotic theory. This can happen in at least two ways.

In the first, for covariate adjustment we use the coefficients fitted when estimating $\hat{y}_c$'s (Sect. 9.4.3). Equivalently, for each i calculate $e_i = y_i - \hat{y}_c(\mathbf{x}_i)$; now calculate tests, confidence intervals and point estimates for the treatment effect in the manner of Sect. 9.5.1, substituting e's for y's throughout.[5] The method yields

[5] This is not quite a permutation test, because $\hat{y}_c(\cdot)$ is determined by the composition of the control group, so that even under the hypothesis of no effect whatsoever, $Y_c \equiv Y_t$, e_i's vary as treatment and control labels are permuted. However, let μ_{ci} be the expectation of $\hat{y}_c(\mathbf{x}_i)$ as treatment labels are permuted and let $\varepsilon_i = y_i - \mu_{ci}$, for each i. In a closely related context, Hansen and Bowers (2009) reviewed arguments to the effect that in large samples, differences between $\rho_{e,z|match}$ and $\rho_{\varepsilon,z|match}$ are negligible, and the permutation distribution of $\rho_{\varepsilon,z|match}$ is well approximated by treating e s as if they were ε s.

similar estimates of the coaching effect to those of Sect. 9.5.1. The verbal effect is again estimated at 0 points, but with a 95% confidence interval of [−7,6]. Covariate adjustment reduces the width of the interval by more than 30%. The covariate-adjusted estimate of the math effect is also sharper: The point estimate remains at 21 while the 95% confidence interval again shrinks by about a third, from [12,30] to [14,26].

The second method is discussed by Rosenbaum (2002a). Let $\mathbf{X}_1$, typically a proper subset of variables in $\mathbf{X}$, represent those preexposure variables for which covariate adjustment is desired. Rather than testing hypotheses about the treatment effect by referring matched correlations of $\tilde{y}$ and z to their distributions under permutations of z, as in the previous section, one computes matched, *partial* correlations of $\tilde{y}$ and z, adjusted for covariance of $\tilde{y}$ and $\mathbf{x}_1$, and refers them to their distributions under permutations of z. In the present example, taking $\mathbf{X}_1$ to consist of the pretest scores and their missingness indicators leads to the same point estimates as did the first method, but with modestly wider confidence intervals: [−7,7] rather than [−7,6]; [14,27] as opposed to [14,26]. The confidence intervals remain much narrower than those calculated without covariate adjustment.

9.5.3 The Coaching Debate

Powers and Rock's (1999) study was published in the midst of ebullient claims on behalf of coaching's benefits to SAT scores. The Princeton Review (2004) has long said its students' average benefit is 140 points in combined SAT score, and during the 1990s, Kaplan Educational Centers claimed average benefits of 120 points (Zehr, 2001). The coaching companies' figures appear to be based on studies conducted for them by outside accounting or consulting firms (Princeton Review, 2004); but since neither these studies nor methodological descriptions of them are published or publicly available, the integrity of their conclusions was difficult to assess. In contrast, Powers and Rock found much weaker coaching effects: about 20 points on the math section and 10 on the verbal.

Applying methods similar to those of this paper to Powers and Rock's data, Hansen (2004) estimated somewhat higher math effects and somewhat lower verbal effects, for a net coaching benefit similar to what Powers and Rock had found. The present analysis finds math effects more similar to Powers and Rock's original estimates, alongside verbal effect estimates that remain lower than theirs. Briggs (2001) and Domingue and Briggs (2009) have used NELS:88 and ELS:02 data to study SAT coaching, arriving at similar conclusions about the magnitude of its effects.

9.6 Discussion

9.6.1 Matching as a Basis for Confounder Control

Regression adjustment for confounder control is ordinarily motivated by the idea that a general rule relating the covariates and intervention variable to outcomes, a "response schedule" (Freedman, 2004), can be specified in outline a priori and then estimated in detail from the data. Holland (1979) presented an alternative interpretation in which the role of model fitting is to construct smoothed representations of within-sample patterns of multivariate association; similar conceptualizations of the role of regression are a staple of recent texts on causal inference in the social science (Angrist & Pischke, 2009; Morgan & Winship, 2007). Still, confounder control from regression can be expected to succeed only if the model beneath the regression manages to do one of these things, accurately represent a general rule relating its independent and dependent variables or accurately represent potentially subtle patterns of multivariate association between the variables.

Matching prior to regression adjustment seems to make it easier for regression to do its work. In a paper contemporaneous with Holland's (1979), Rubin (1979) compared regression adjustments after matching to regression alone in a situation of moderate misspecification, finding that the regression after matching more reliably discovered and corrected for multivariate associations than did regression unassisted by matching. The finding was explained in large part by matching's tendency to reduce the possibility of extrapolation between groups being compared, even extrapolation in terms of combinations of covariates that wouldn't be seen in a comparison of the groups on any one covariate. Such extrapolation can be difficult to identify, but estimating and matching on propensity scores quite dependably reveals and mitigates it (Rubin, 1997).

Yet propensity matching has been presented, here and elsewhere (Rubin, 1991; Rosenbaum, 2001), as a primary countermeasure to bias due to measured confounding, not just as an assistant to regression adjustment. At first blush, propensity matching seems to have requirements similar to those of regression adjustment since it involves specifying and fitting a regression of its own. Indeed, its requirements would appear to be tougher: The regression it involves has a binary dependent variable while the outcome regression's dependent variable may be continuous, in which case more informative diagnostic procedures may be available for it; and the matching procedure, an extra step which confounder control from regression does not need, is inherently inexact, whereas theory supporting the method seems to require matching exactly on the propensity score.

It may well be that propensity matching requires stringent, scarcely attainable conditions if it alone is to remove confounding bias in precisely the manner suggested by the original theory of propensity scores (Rosenbaum & Rubin, 1983). But the theory separately predicts, and in practice it has frequently been reaffirmed, that propensity matching can alleviate an intervention and a control group's incompatibilities in terms of observed confounders. Newer theory (Hansen, 2009)

says that in large samples, this alleviation of differences suffices to remove bias due to observed confounders. If knowing the precise form of the outcome's dependence on observed confounders would suffice to estimate treatment effects, that is, then the combination of a well-implemented propensity matching with a simple covariate adjustment – requiring prior knowledge neither of the outcome's nor of the treatment's regression on covariates – also suffices to validly estimate treatment effects.

9.6.2 A Comparison of Matching Strategies: Focus Versus Restrictions

The recommendations of this paper are to full match, in order to make use of as much of the sample as possible; to match within calipers of an linear propensity score estimated in the usual way, so as to balance many variables at once and to avoid large matched discrepancies; but within those calipers to match on a second, more focused propensity score, in order to avoid inordinately reducing the effective sample size. Confronting the same matching problem as discussed here, Hansen (2004) addressed similar concerns by full matching on just one estimated propensity score but using structural restrictions (reviewed briefly in Sect. 9.3.3), which the current recommendation does not involve. How do the two approaches compare?

In this instance, results from the two approaches do not differ greatly, but the obtaining differences are instructive. Balance overall is similar but a bit better for focused matching, $\chi^2 = 56$ versus $\chi^2 = 79$ for matching with restrictions, on 69 d.f. in both cases. Effective sample size is similar but very slightly better for matching with restrictions, 701 versus 705. Figure 9.6 shows that both did well at balancing the main prognostic variables, although balance on predictive uncoached verbal scores, the $\hat{y}_c(\mathbf{x})$'s for SAT-V, is markedly better under the focused approach. Propensity matching with focus on selected prognostic variables appears to more reliably balance those variables.

Both of these matches incorporate propensity-score calipers, which Hansen's, 2004 paper's matches did not. It seems best to use calipers, which help to ensure favorable large-sample properties by heading off outlying differences on the (true) propensity score (Hansen, 2009) and have also been recommended by other authors (Haviland, Nagin, & Rosenbaum, 2007; Rosenbaum & Rubin, 1985b; Rubin & Thomas, 2000). However, certain combinations of calipers with structural restrictions make matching impossible, as discussed in Sect. 9.3.3. This problem does not arise when full matching without restrictions. Indeed, when the calipers used in Sects. 9.4.2 and 9.4.3 above are combined with the specific restrictions used by Hansen (2004), matching is impossible in 10 of the 12 subclasses, and only 119 subjects can be matched to one another. It becomes necessary first to determine what restrictions are compatible with the caliper, a task that is more computationally intensive than anything required by the method of this paper (see Sect. 9.3.3),

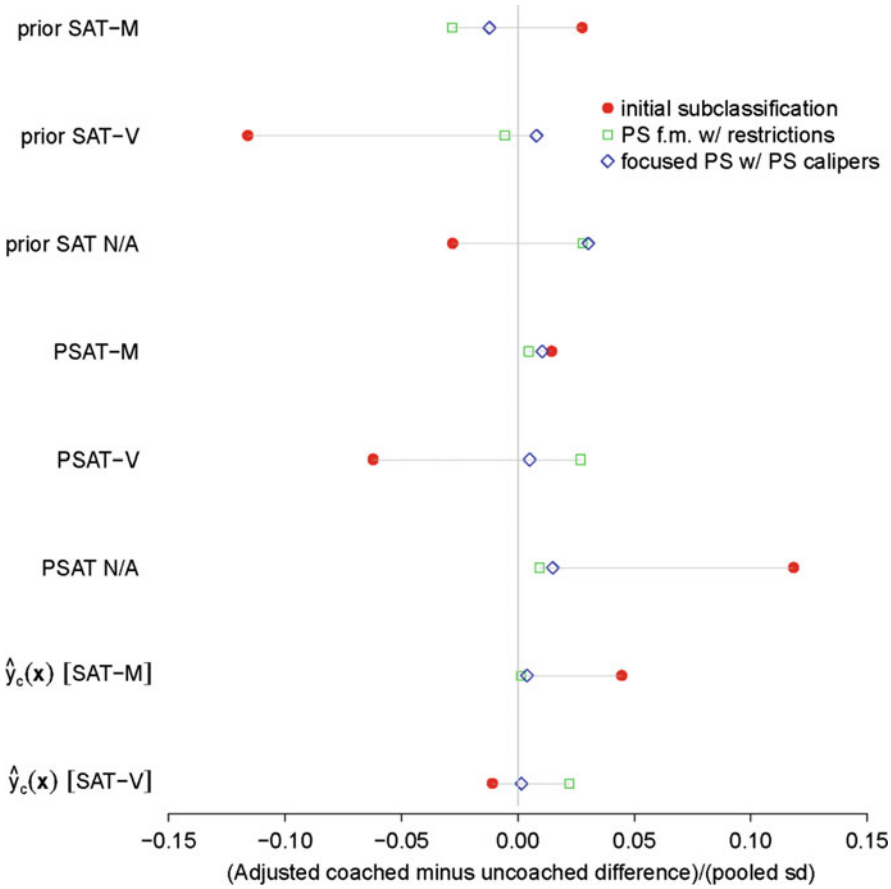

Fig. 9.6 Treatment-control differences on key prognostic covariates, adjusted for the initial Race × SES stratification and for two propensity score matches that subdivide the Race × SES strata: the match produced by full matching with calipers and structural restrictions, and the match produced by full matching on the focused propensity score within propensity score calipers

and then to adjust these restrictions with attention to balance, a task that is more labor intensive than anything required so far. The matching strategy recommended here requires substantially less effort on the part of the statistician.

9.6.3 Why Match: Particularly If We're Going to Use Regression after Matching Anyway?

Recall Cochran's criterion for when an imbalanced comparison might and might not plausibly be rectified by statistical adjustments: The adjustment would have a fighting chance, in Cochran's assessment, if there were a few imbalances large

enough to give *t*-statistics greater than 2, but not necessarily otherwise. A sample like Powers and Rock's (1999), with its tens of highly significant imbalances, appears hopeless from his perspective; nonetheless, propensity score full matching reduces most all of the imbalances to insignificance. (See Hansen, 2008a, for discussion of the meaning and uses of *statistical significance* in this setting.)

Part of what informed Cochran's pessimism about these situations was the difficulty of adjusting for tens of confounding variables using the methods of his day. Matching seemed equipped to handle no more than a few confounders. This paper has reviewed a combination of propensity scoring, matching and diagnostic techniques that enables analysts to address confounding on quite large numbers of covariates, sharply and simultaneously reducing covariate imbalances by poststratifying more finely and with greater focus than do poststratifications of the kind that would have been familiar to Cochran. (Our initial exact matching on race and SES is such a poststratification.) One can only speculate about what Cochran would have thought of the method, but it has handily addressed a problem, substantial covariate imbalances on a large number of variables, that his 1965 paper regarded as both intractable and damning.

Multiple regression, on the other hand, was well known to Cochran – he contributed importantly to the development of the technique as it is known today – and unlike exact matching, it was feasible with more than two or three covariates. However, Cochran did not recommend it as a remedy for covariate imbalances like the Powers-Rock study's. Of course, few of today's diagnostic techniques for assessing and adjusting a multiple regression specification were available in Cochran's day. Why not adjust for the many confounders using some form of regression-based covariance adjustment, using modern diagnostics (Cook & Weisberg, 1982; Fox, 2005) to ensure that the model fits reasonably well? One reason to hesitate runs as follows. Thoroughly applying diagnostics to a multiple regression adjusting for many or all of the covariates would be a daunting chore, rivaled in tediousness only by the task of reviewing that it had been properly done. Because no one undertakes that task willingly, few journals would be willing to publish the many plots and other materials needed to be ensure that it had been done correctly; and because the journal wouldn't be publishing the material anyway, it will rarely be checked by peer reviewers. The reader of an article reporting results from a large multiple regression is thus forced to trust that all of the necessary diagnostics have been adequately done. Because the diagnostics take time, and little credit to be had for actually doing them, savvy readers expect that they will have been done hastily or perhaps not at all. Researchers can try to counter such perceptions by laying out their analytic procedure in detail, as Powers and Rock's research reports admirably do; but in other ways claiming to have diagnosed one's regression model with great care tells against the credibility of one's observational study. It raises the possibility of data dredging, when a researcher fiddles excessively with his regression specification under a conscious or unconscious stopping rule that favors statistically significant or otherwise desirable estimates of the treatment effect. Taken together, these conflicting threats and pressures engender a quite rational cynicism about multiple regression-adjusted treatment effect estimates.

On the other side of the ledger, with stratification-based adjustments, the relevant diagnostics are of the form we've just seen, assessments of whether the stratification renders treatment and control groups indistinguishable in terms of the covariates. Such comparisons are themselves of interest to scientific audiences of studies, particularly when they are set alongside unstratified comparisons of the groups, as in Tables 9.2–9.4, because they convey relevant information about data characteristics as well as information about statistical adjustments. When they suggest adjustments to the stratification or matching, those adjustments are made prior to any outcome analysis, greatly mitigating the threat of unconscious or semiconscious data dredging. Diagnostics for propensity score matching are better suited to the scholarly record than regression diagnostics, and more likely to enhance the credibility of the research.

Author's Note Portions of this chapter are reprinted with permission from the *Journal of the American Statistical Association*. Copyright 2004 by the American Statistical Association. All rights reserved. This research was supported by the NSF (DMS-0102056 & SES-0753164), whose support does not entail endorsement of conclusions of the research.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*, 235–267.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bowers, J., Fredrickson, M., & Hansen, B. (2010). RItools: Randomization inference tools. R package version 0.1-9 [Computer software]. Available from http://www.jakebowers.org/RItools.html.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance, 14*, 10–21.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, A, 128*, 234–266.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman and Hall.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*, 1053–1062.

Domingue, B., & Briggs, D.C. (2009, April). *Using linear regression and propensity score matching to estimate the effect of coaching on the SAT*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Fox, J. (2005). *Regression diagnostics*. Newbury Park, CA: Sage Publishers.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review, 28*, 267.

Gu, X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*, 609–618.

Hansen, B. B. (2007). OPTMATCH: Flexible, optimal matching for observational studies. *R News, 7*, 18–24.

Hansen, B. B. (2008a). The essential role of balance tests in propensity-matched observational studies: Comments on a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 by Peter Austin. *Statistics in Medicine, 27*, 2050–2054.
Hansen, B. B. (2008b). The prognostic analogue of the propensity score. *Biometrika, 95*, 481–488.
Hansen, B. (2009). *Propensity score matching to recover latent experiments: Diagnostics and asymptotics* (Tech. Rep. No. 486). Ann Arbor: University of Michigan, Statistics Department.
Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*, 219–236.
Hansen, B. B., & Bowers, J. (2009). Attributing effects to a cluster randomized Get-Out-The-Vote Campaign. *Journal of the American Statistical Association, 104*, 873–885.
Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*, 609–627.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction: With 200 full-color illustrations*. New York, NY: Springer.
Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*, 247.
Holland, P. W. (1979). The tyranny of continuous models in a world of discrete data. *IHS-Journal, 3*, 29–42.
Holland, P. W. (1986a). Statistics and causal inference: Rejoinder. *Journal of the American Statistical Association, 81*, 968–970.
Holland, P. W. (1986b). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.
Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Applied Statistics, 17*, 118–136.
Marcus, S. M. (2000). Estimating the long-term effects of head start. In S. Oden, L. J. Schweinhart, & D. P. Weikart (Eds.), *Into adulthood: A study of the effects of Head Start* (ch. F, pp. 179–200). Ypsilanti, MI: High/Scope Press.
Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*, 118–124.
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, England: Cambridge University Press.
Powers, D., & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93–118.
Princeton Review. (2004). *SAT classroom courses for class of 2005*. Available from http://www.princetonreview.com.
Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, 53*, 597–610.
Rosenbaum, P. R. (2001). Observational studies: Overview. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 10808–10815). Amsterdam, The Netherlands: Elsevier/North-Holland.
Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science, 17*, 286–327.
Rosenbaum, P. R. (2002b). *Observational studies* (2nd ed.). New York, NY: Springer.
Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516–524.
Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*, 33–38.
Rosenbaum, P. R., & Rubin, D. B. (1985b). The bias due to incomplete matching. *Biometrics, 41*, 103–116.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318–328.
Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962.
Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics, 47*, 1213–1234.
Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757–763.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2*, 808–840.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–64.
Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*, 573–585.
Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge, England: Cambridge University Press.
Zehr, M. (2001, April 4). Study: Test-preparation courses raise scores only slightly. *Education Week*.

Part V
Holland Rebuilding ETS

Returning to ETS from Berkeley

Paul W. Holland

One of the joys of returning to Educational Testing Service (ETS) was not only to reconnect with old friends, but also to meet the new and much younger staff that had arrived at roughly the same time as my return. Sandip Sinharay had just arrived the year before and added a burst of energy, which he continues to display. Working with him on several papers was a very gratifying collaboration.

When I returned, I served as an interim group head for the members of statistics and psychometrics research. I was very pleased that Shelby Haberman eventually agreed to come to ETS from Northwestern University to continue his many lines of research as well provide a more permanent leadership role for this group. Shelby was an excellent addition to ETS, and his research continues to add to serious contributions to both theory and practice, something that I value highly.

Shortly after I arrived back at ETS I decided to take the work that I had done on kernel equating years earlier and turn it into a book. At about the same time, I met Alina von Davier, who was interested in working at ETS, and I decided to hire her to help me write the book. This was a very fortunate choice because not only did the book materialize due to her skills and enthusiasm (as well as those of Dotty Thayer), but ETS also got a great employee and the field of educational measurement got a skilled scientist who has continued to work on many problems of importance to the testing enterprise.

Chapter 10
Log-Linear Models as Smooth Operators: Holland's Statistical Applications and Their Practical Uses

Tim P. Moses

10.1 Overview

Paul Holland's statistical applications have produced important answers to several problems encountered in equating practice. This paper focuses on one of Holland's far-reaching applications: his application of log-linear models as a smoothing method for equipercentile equating. Section 10.2 describes the context for Holland's initial equating work and the practical problems inevitably faced when researching and doing equipercentile equating. Section 10.3 describes the application of log-linear models as a smoothing technique for addressing equipercentile equating problems. The developments that were introduced to adapt this application to test equating are also described. Section 10.4 summarizes some of the collaborative investigations by Holland and Moses of the use of log-linear models for equating.

10.2 Initial Work with Equipercentile Equating

Paul Holland's equating work began in 1978 when ETSers Donald Rubin and Robert Solomon started the Program Statistics Research Project (von Davier, Holland, & Thayer, 2004, p. vii). The purpose of this initiative was to focus the statistical research interests of the newly formed Research Statistics Group on problems relevant to the work of the testing programs at ETS. Holland was given responsibility for the test equating research part of this initiative.

Holland's research responsibilities amounted to a fairly large task because equating research and practice had been ongoing pursuits at ETS for several years (e.g., Lord, 1950, 1955; Wilks, 1961). His initial work in test equating produced

T.P. Moses (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: tmoses@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_10,

a conference and book (Holland & Rubin, 1982) and mathematical analyses of equating methods commonly used at ETS (Braun & Holland, 1982). New equating methods were also developed in response to arising needs (e.g., section pre-equating, see Holland & Thayer, 1981). This work appears to have laid the groundwork for future work that would address some of the more long-standing problems in test equating.

One long-standing problem encountered in equipercentile equating is test scores that are possible to attain but unobserved in sample data. Table 10.1 illustrates the issue, showing the relative and cumulative relative frequencies for some test scores of a hypothetical test. In Table 10.1, test score 8 is not observed and therefore has a relative frequency of 0.00. One result of this unobserved test score is that more than one score will have the same cumulative relative frequency (i.e., scores 7 and 8). This result creates a problem for equipercentile conversions based on finding scores with matching percentiles because these conversions rely on each score's percentile being unique.

The difficulties of unobserved test scores can create problems that are more serious than Table 10.1's ambiguities, particularly for the so-called post stratification equipercentile equating method (also known as the frequency estimation or the direct equipercentile method). The post stratification equipercentile method is based on using administration groups' marginal anchor score distributions to estimate the test score distributions for a hypothetical, synthetic population. For example, if test X and anchor A were administered to population P, test Y and A were administered to population Q ($P \neq Q$), the joint (X,A) probability distribution in synthetic population T ($=wP + (1 - w)Q$, $0 \leq w \leq 1$) would be estimated as,

$$\text{Prob}_T(X, A) = w\text{Prob}_P(X, A) + (1 - w)\frac{\text{Prob}_P(X, A)}{\text{Prob}_P(A)}\text{Prob}_Q(A). \quad (10.1)$$

Similarly, the joint (Y,A) probability distribution in T would be estimated as,

$$\text{Prob}_T(Y, A) = (1 - w)\text{Prob}_Q(Y, A) + w\frac{\text{Prob}_Q(Y, A)}{\text{Prob}_Q(A)}\text{Prob}_P(A). \quad (10.2)$$

Equations (10.1) and (10.2) require that the A scores be observed in both P and Q, otherwise divisions by zero undermine the estimation of the synthetic population distributions.

Table 10.1 Part of the frequency distribution for a hypothetical test

Score	Relative frequency	Cumulative relative frequency
10	0.12	1.00
9	0.17	0.88
8	0.00	0.71
7	0.07	0.71

In his initial equating research, Holland realized that equipercentile methods have practical limitations (Braun & Holland, 1982, p. 22). He was also aware of several ad hoc practices being used to avoid the equipercentile method's difficulties, some of which include

- Adding small constants to all test scores (Hanson, 1990; Kolen & Brennan, 1995),
- Averaging unobserved test scores with observed test scores (Kolen & Brennan, 1995; Livingston, 2004),
- Substituting observed parts of the score distribution for the unobserved parts (Jarjoura & Kolen, 1985),
- Tukey-Cureton Smoothing (Angoff, 1984; Cureton & Tukey, 1951),
- Negative hypergeometric and Beta4 smoothing (Keats & Lord, 1962; Lord, 1965), and
- Smoothing by hand drawings and graph paper (Angoff, 1984).

Discussions of the ad hoc practices have detailed their potential for inaccuracy, their arbitrariness, and their lack of usefulness for particular types of test score distributions (e.g., Kolen, 1991). Braun and Holland's (1982) quote, "Modern methods of data smoothing should have important contributions to make" (p. 22), revealed Holland's interest in finding a different method for addressing unobserved test scores in equipercentile equating.

10.3 Log-Linear Models as a Smoothing Technique

Whereas most practices for using equipercentile equating methods when test scores are unobserved appear to be based on pragmatic motivations (i.e., adding small constants, pooling scores, etc.), Holland's perspective was that the estimation of test score distributions should be treated as a statistical problem. He drew on his background in categorical data analysis (Bishop, Fienberg, & Holland, 1975) and proposed a particular class of log-linear models for use with test score distributions (Holland & Thayer, 1987). This application could be expected to produce modeled test score distributions that satisfy statistical criteria, including the following:

- Consistency: With increasing sample size, models' estimates converge to the population values.
- Efficiency: Given the sample size, the deviations of the estimated distributions are as small as possible.
- Positivity: All test scores' probabilities are greater than zero.
- Integrity: The observed moments in the test data (i.e., means, variances, correlations) are preserved in the modeled distribution.

Because log-linear models are based on meeting statistical criteria, the modeled test score distributions could be used to avoid the equipercentile equating method's difficulties with unobserved test scores (positivity), and they would likely be more accurate estimates than those produced by other, previously proposed approaches.

In addition, log-linear models are extremely flexible, supporting the fitting of a wide variety of test score distributions. Finally, the modeled distributions produced by a log-linear model can be smooth. The desirable statistical properties, flexibility and smoothness outcome, resulted in *log-linear smoothing* becoming a widely used technique at ETS and elsewhere.

10.3.1 Log-Linear Models and Univariate Test Score Distributions

This section illustrates how log-linear modeling may be applied to estimate and smooth the distribution of a single test, X, with possible scores $x_1,\ldots,x_J$, or x_j, with $j = 1,\ldots, J$. The transposed row vector of observed score frequencies, $\boldsymbol{n} = (n_1,\ldots,n_J)^t$, sums to the total sample size, N. There are N independent observations of the discrete random variable X, and $\boldsymbol{n}$ is assumed to follow the multinomial distribution M(N, $\boldsymbol{p}$), where $\boldsymbol{p}$ denotes the population probabilities corresponding to the $\boldsymbol{n} = (n_1,\ldots,n_J)^t$ values. The log-linear model expresses the log of the expected (not actual) relative frequencies in terms of a polynomial function of the test scores,

$$\log_e(p_j) = \beta_0 + \sum_{i=1}^{I} \beta_i x_j^i, \tag{10.3}$$

where the x_j^i are functions of the possible test scores (e.g., x_j^1, x_j^2, x_j^3, …, x_j^I), β_0 is a normalizing constant that forces the sum of the expected relative frequencies (p_j) to equal 1, and the β_i are parameters to be estimated in the model-fitting process. The value of I determines the extent of smoothing and, when maximum likelihood estimation is used, the number of moments of the actual test score distribution that are preserved in the smoothed distribution. If $I = 1$ then the smoothed distribution preserves only the first moment (the mean) of the observed distribution. If $I = 4$ then the smoothed distribution preserves the first, second, third, and fourth moments (mean, variance, skewness, and kurtosis) of the observed distribution. The value of I also determines the extent to which the smoothed frequencies, $m_j = Np_j$, approximate the observed frequencies, n_j.

Figure 10.1 presents the frequency distribution for a hypothetical 20-item test with a possible score range from 0 to 20. This distribution exhibits previously described issues that would be problematic in equipercentile equating, in that scores 0, 13, 15, 17, and 20 are unobserved (i.e., have frequencies of 0). In addition, fluctuations in the frequencies can be attributable to each frequency being based on nine or fewer examinees.

The log-linear model in (10.3) may be used to model the major characteristics observed in Fig. 10.1's frequency distribution to smooth out fluctuations attributable to sampling variability and to provide reasonable estimates of the frequencies at unobserved scores 0, 13, 15, 17, and 20. Figures 10.2–10.6 plot the observed

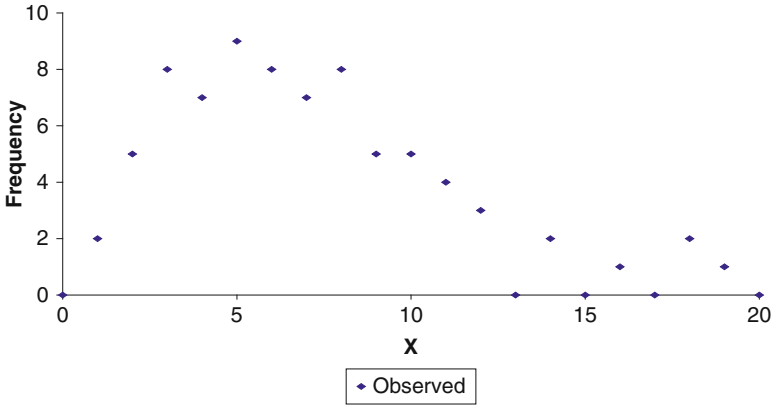

Fig. 10.1 Observed frequency distribution for a hypothetical 20-item test

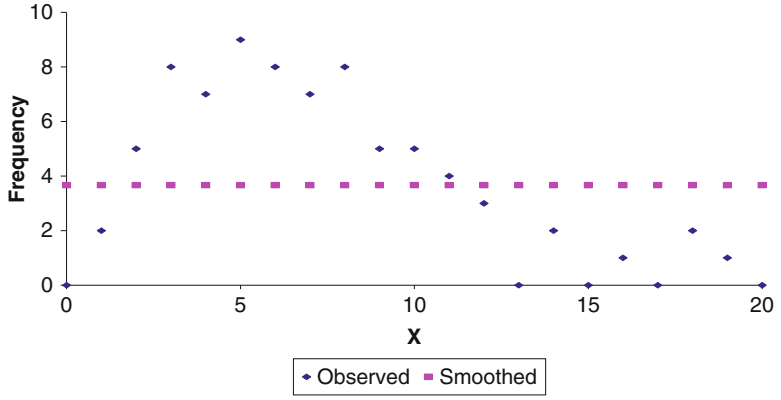

Fig. 10.2 Observed and modeled frequency distributions ($I = 0$)

distribution shown in Fig. 10.1 along with estimated distributions produced by fitting five log-linear models to the observed distribution. These five models differ in their number of parameters, ranging from $I = 0$ to 4.

Figures 10.2–10.6 illustrate the major features of univariate log-linear models. Figure 10.2 gives the *null* model upon which most log-linear models are based, showing that models based on $I = 0$ produce a uniform distribution where the overall sample size is the only observed characteristic preserved in the modeled distribution. Figures 10.3–10.6 are based on fitting more features in the observed distribution, beginning with the mean ($I = 1$, Fig. 10.3), the mean and variance ($I = 2$, Fig. 10.4), the mean, variance, and skewness ($I = 3$, Fig. 10.5), and the mean, variance, skewness, and kurtosis ($I = 4$, Fig. 10.6). The figures show that modeled distributions based on more parameters (larger I values) fit the observed distribution more closely. The figures can inform decisions of the most appropriate

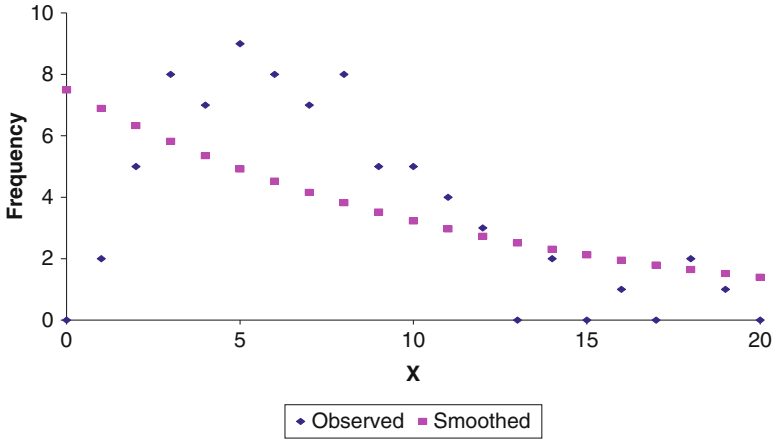

Fig. 10.3 Observed and modeled frequency distributions ($I = 1$)

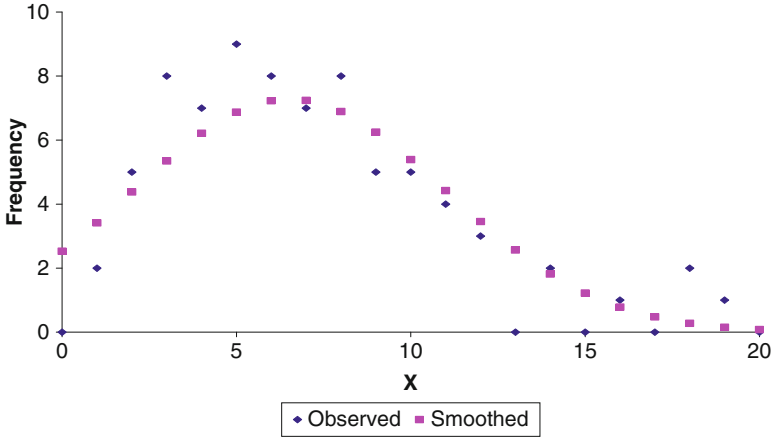

Fig. 10.4 Observed and modeled frequency distributions ($I = 2$)

model, suggesting that the model based on fitting the mean, variance, and skewness ($I = 3$ in Fig. 10.5) fits the observed distribution considerably better than the simpler models (Figs. 10.2–10.4) and is perhaps almost as good as the model based on fitting the mean, variance, skewness, and kurtosis ($I = 4$ in Fig. 10.6). Comparisons of the models' fit statistics could also inform model selection (see Sect. 10.4).

The log-linear model in (10.3) can be expanded in several ways. One important class of models can fit score-specific features and overall features of test score distributions. For example, structures such as lumps (abnormally large frequencies) and bimodality can occur in distributions due to different types of scoring and

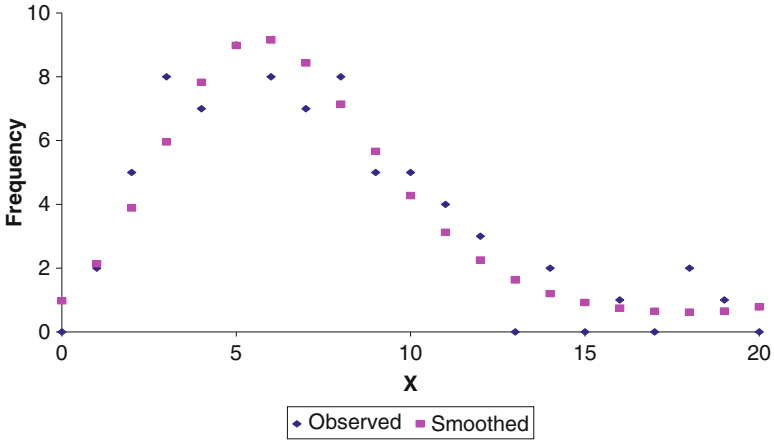

Fig. 10.5 Observed and modeled frequency distributions ($I = 3$)

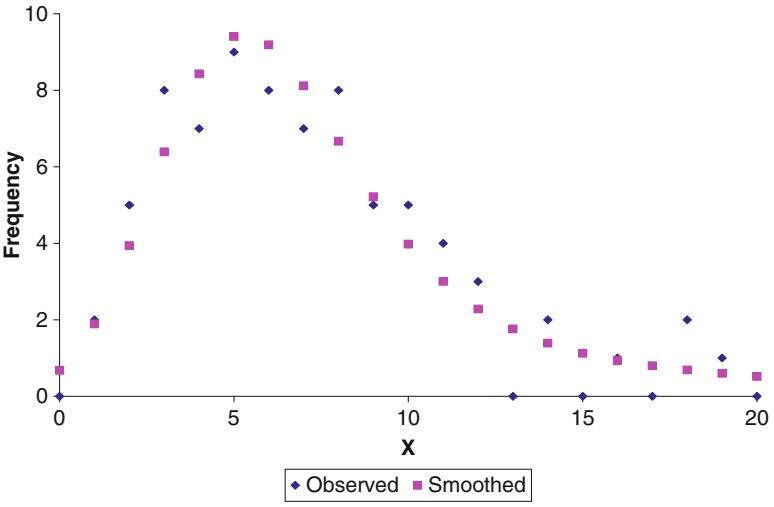

Fig. 10.6 Observed and modeled frequency distributions ($I = 4$)

rounding practices or to heterogeneous examinee groups. To account for these structures, (10.3) can be expanded by adding an indicator function, $S(j)$,

$$\log_e(p_j) = \beta_0 + \sum_{i=1}^{I} \beta_i x_j^i + \beta_{I+1} S(j). \tag{10.4}$$

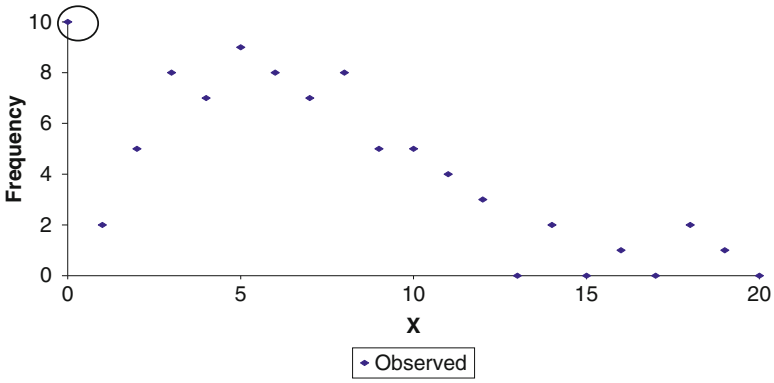

Fig. 10.7 Figure 10.1's observed frequency distribution for a hypothetical 20-item test with an abnormally large frequency at score 0

For the test score(s) in a subset, $S(j)$ is set equal to 1, and for the test scores not in the subset, $S(j)$ is set equal to 0. The result is that (10.4) will fit I overall moments in the observed distribution and the frequencies of the scores defined in the subset. Equation (10.4) can be expanded to model the moments of the subset distribution through products of $S(j)$ and x_j^i.

To illustrate the use of (10.4), the observed test score distribution shown in Fig. 10.7 is modeled. This is the same distribution as shown in Fig. 10.1 except the frequency at test score 0 is abnormally large, amounting to a lump at 0. Most of this distribution appears to have an overall shape that would be appropriately modeled by fitting the mean, variance, and skewness. The frequency at score 0 disrupts the overall shape of the distribution. Equation (10.4) might be used by defining an indicator function $S(j)$ such that it is equal to 1 at score $x_j = 0$ and is equal to 0 otherwise. Figure 10.8 shows the observed and modeled distributions based on such a model with $I = 3$. This model appears to fit the observed distribution adequately, and the abnormally large frequency at score 0 does not interfere with the model's fit with the overall distribution.

10.3.2 Log-Linear Models and Bivariate Test Score Distributions

An important extension of univariate log-linear models such as (10.3) and (10.4) is the modeling of bivariate distributions, as would be encountered when a group of examinees takes two tests. A bivariate log-linear model for a distribution where examinees take test X and anchor A can be expressed as,

$$\log_e(p_{jl}) = \beta_0 + \sum_{i=1}^{I} \beta_{x,i} x_j^i + \sum_{h=1}^{H} \beta_{a,h} a_l^h + \sum_{d=1}^{D} \sum_{e=1}^{E} \beta_{xa,de} x_j^d a_l^e, \tag{10.5}$$

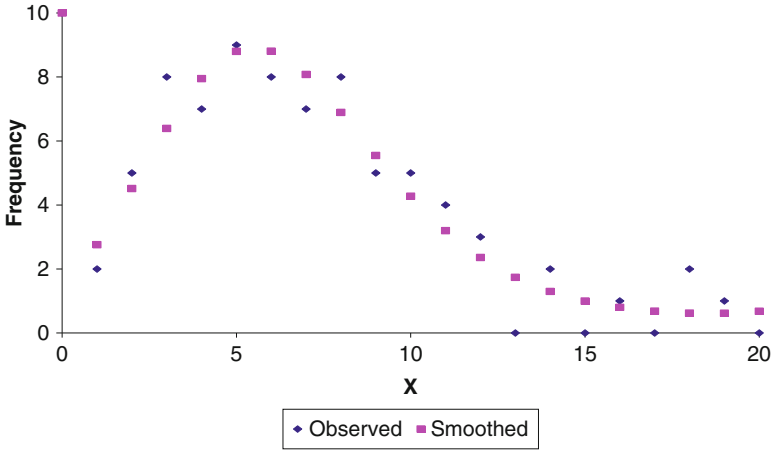

Fig. 10.8 Figure 10.7's observed frequency distribution modeled with $I = 3$ and an indicator function for score 0's abnormally large frequency

where p_{jl} is the expected probability of examinees obtaining test score x_j and anchor score a_l, I is the number of moments observed in X's univariate distribution fit, H is the number of moments observed in A's univariate distribution fit, and the D and E values determine the number of cross-moments observed in the joint (X,A) distribution fit. When $D = E = 1$, (10.5) will fit the observed XA covariance, or the conditional means of X given the A scores and of A given the X scores. Higher values of D and E can be used to model the conditional standard deviations and skewness of X given A and of A given X, conditional moments that can vary in complex ways in joint distributions of bounded test scores.

The fitting of bivariate log-linear models such as (10.5) is typically more challenging than the fitting of univariate models such as (10.3) and (10.4). The tests may be fairly long, amounting to many possible score combinations in the joint distribution. For example, with a test of 76 possible scores and an anchor with 36 possible scores, 2,736 score combinations are possible. Fitting algorithms for log-linear models typically requires the forming of square matrices where all possible score combinations are paired with each other (Holland & Thayer, 2000; SAS Institute, 2002). This means, for example, that several 2,736 × 2,736 matrices might be needed, each of which would contain more than seven million cells. Other difficulties are that the majority of the possible score combinations in bivariate distributions are likely to be unobserved in test data due to the volumes of some examinees at testing programs' typical test administrations and also due to moderate or high correlations between the tests being modeled. The size of the model fitting task and the size of the available examinee data can result in nonconverging bivariate models.

10.3.3 Practical Issues in Fitting Log-Linear Models

In his proposals of the use of log-linear models for smoothing test data (Holland & Thayer, 1987, 2000), Holland described several developments directly suited to the difficulties of test data. Holland proposed specific developments for addressing difficulties such as the size of large bivariate modeling problems, including

1. Scaling strategies other than the power functions of the test scores shown in (10.3)–(10.5).
2. Convergence criteria that focused directly on the likelihood function and the models' moment-matching results rather than on changes in the β's.
3. Algorithms for performing computations in ways that avoid the forming of extremely large matrices.

My personal experience with available algorithms such as SAS Proc GENMOD (Moses, von Davier, & Casabianca, 2004; SAS Institute, 2002) has found that Holland's algorithm can often produce converged solutions in situations where other algorithms fail.

Not only are Holland's algorithms directly suited to the difficulties of test data, they are also suited to the interests of those working with test data (Holland & Thayer, 1987, 2000). Holland recognized that the log-linear modeling results of most practical value when used to smooth test data involve the smoothed probabilities and frequencies rather than the β's. The use of a model convergence criterion that focused directly on moment-matching in the smoothed distribution helped ensure that smoothed distributions would reflect user-specified features of observed data. A factorization of the variance-covariance matrix of the smoothed results (i.e., the C-matrix) resulted in an efficiently storable matrix that made it possible to estimate asymptotic standard errors of smoothed equating functions (Holland, King, & Thayer, 1989; Moses & Holland, 2008).

10.4 Research on Using Log-Linear Models for Equating: A Summary

Since the introduction and use of log-linear smoothing methods, concerns have arisen for how to best utilize their flexibility. Although it is certainly true that log-linear models' wide ranges of parameterizations support the modeling of many types of test score distributions, the short timelines typical of equating practice have not necessarily supported the elaborate search processes typically demonstrated for comparing and selecting appropriate models (e.g., Holland & Thayer, 2000; von Davier et al., 2004). Useful applications of log-linear models to equating practice need to simultaneously satisfy concerns of flexibility, accuracy, and efficiency. This section describes some of the research studies by Holland and Moses that have

considered the selection of log-linear models in sample data and the implications of log-linear models for equating function accuracy.

10.4.1 Implications of Selecting Univariate Log-Linear Models for Equating Function Accuracy

Initial studies by Holland and Moses considered how to select the parameterizations for univariate log-linear models (i.e., the I value in (10.3)) in sample test data (Moses & Holland, 2007a, 2009, 2010a). Relevant prior work had focused on repeated use of particular models not selected based on sample data (Livingston, 1993; Skaggs, 2004) or on using one of many possible statistical strategies for selecting models' parameterizations in sample data (i.e., likelihood ratio chi-square tests, see Hanson, 1990). Simulations were designed to evaluate the repeated use of several statistical strategies for selecting univariate models' parameterizations in sample data and for smoothing test data prior to computing equipercentile equating functions.

For realistic simulations, Holland and Moses obtained a range of univariate test score distributions from different testing programs, found well-fitting log-linear models for these distributions and used them as population distributions, drew random samples of different sizes from the populations, used the statistical strategies to select the appropriate parameterizations of the log-linear models in the sample data, and computed equipercentile equating functions for two tests where the log-linear models were selected by some statistical selection strategy.

Some of the considered statistical strategies are based on a search process described in Haberman (1974) for comparing one of four chi-square statistics (likelihood ratio, Pearson, Freeman-Tukey, and Cressie-Read). Others are based on minimizing a combination of model fit and model parameterization, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Consistent Akaike Information Criterion (CAIC). An index of model fit sometimes attributed to Goodman (Agresti, 2002) that is routinely produced in statistical software packages (SAS Institute, 2002) was also considered. See Bozdogan (1987) for formulas for these statistics.

Figure 10.9 shows one of the population distributions considered in the studies by Holland and Moses. This distribution is produced from a univariate log-linear model where $I = 6$, a model that was determined to fit a particular test score distribution fairly well. In the studies, several hundred samples of various sizes were drawn from the $I = 6$ and other population distributions, and the selection strategies were used to select the log-linear models from a range of models (e.g., $I = 2, 3, \ldots, 10$) fit to the sample datasets.

Table 10.2 presents the average I value selected in 500 datasets of sample sizes of 100, 500 and 2,500 for the eight statistical strategies mentioned above. The results in Table 10.2 show that the statistical strategies can be differentiated based on their tendencies to select log-linear models with more or fewer parameters.

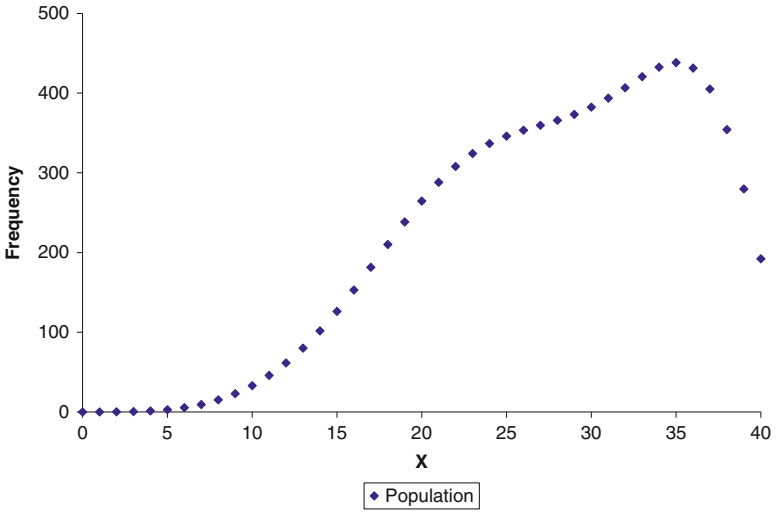

Fig. 10.9 An $I = 6$ population distribution

Table 10.2 $I = 6$ in the population

	Average parameters selected (500 replications)		
Selection strategy	$N = 100$	$N = 500$	$N = 2{,}500$
Likelihood ratio chi-square	2.62	3.14	5.08
Pearson chi-square	3.04	3.31	5.27
Freeman-Tukey chi-square	2.41	2.96	4.99
Cressie-Read chi-square	2.73	3.13	5.15
AIC	3.72	4.95	6.28
BIC	2.27	2.75	4.78
CAIC	2.15	2.52	4.65
Goodman	4.85	4.39	6.31

Note. *AIC* Akaike information criterion, *BIC* Bayesian information criterion, *CAIC* Consistent Akaike information criterion

The AIC and Goodman strategies tended to select models with more parameters, and the chi-square, BIC, and CAIC strategies selected fewer parameters. All strategies are more accurate when sample sizes are large (i.e., 500 or 2,500 rather than 100). These results were consistent across several distributions. In general, the AIC strategy appeared to be the most accurate strategy across all the conditions considered in the study, though it had a tendency to select models that overfit very simple distributions.

In other simulations (Moses & Holland, 2007a, 2009), the repeated use of the statistical strategies for selecting log-linear models was evaluated with respect to equipercentile equating accuracy. These evaluations involved forming pairs of the population test score distributions, smoothing test data sampled from the populations using the statistical strategies to select the log-linear models, computing

equipercentile equating functions to link the pairs of test scores with the smoothed distributions, and comparing the sample estimated equating functions to the population equating functions in terms of bias and variability. The results reflected the bias vs. variability tradeoff often described in smoothing discussions (Kolen, 1991; Kolen & Brennan, 2004): The use of statistical strategies that tended to select relatively few parameters in sample data produced sample equating functions with more bias and less variability, whereas the use of other strategies that tended to select more parameters produced sample equating functions with less bias and more variability. Across all of the equating and sample size conditions considered in the Moses and Holland (2007a, 2009) studies, the AIC strategy produced the most accurate equating functions, with the least bias, and only slightly more variability than the equating functions produced using other statistical strategies.

Although the results of the equating function evaluations in the Moses and Holland studies (2007a, 2009) were understandable in terms of the statistical strategies' selection tendencies and tradeoffs in bias and variability, the actual influence of selection strategies on equating function accuracy was smaller than expected. Specifically, the results suggested that most of the practically important equipercentile equating accuracy could be realized when the selection strategy tended to select models with I values of at least 3. The accuracy gains diminished greatly among selection strategies that varied in I selections greater than 3, even when the population model considered had an I value of 6.

Some analyses were done to assess why the selection of log-linear models such as those with I values of 3 might produce equating functions acceptably accurate for practice even when the population I value might be 6. The follow-up analyses in the Moses and Holland studies (2007a, 2009) showed that equipercentile equating functions can be somewhat robust to an overly simple log-linear smoothing model because they are based on cumulative frequency distributions rather than the frequency distributions modeled with the log-linear models. For example, Fig. 10.10 shows a poorly chosen smoothing model fit to Fig. 10.9's population distribution model 9 ($I = 3$ rather than $I = 6$). This model has clear deficiencies in fitting the population model, and with sample sizes of 1,000 or greater, many statistical strategies would consider this model too simple (Table 10.2 in this paper). However, the smoothing model shown in Fig. 10.10 is used in equipercentile equating only after it is used to calculate a cumulative frequency distribution. The cumulative frequency distributions are calculated from Fig. 10.10's frequency distributions and shown in Fig. 10.11. They are then rescaled so they can be evaluated with respect to the frequency distributions (Fig. 10.12). The figures show that the cumulative distributions used in equipercentile equating can approximate the population cumulative distributions even when they are based on poorly fitting frequency distributions.

To summarize, the results of the Moses and Holland studies suggested the following,

- Statistical selection strategies vary in their tendencies to select more or fewer parameters, with the chi-square, BIC, and CAIC strategies selecting models with fewer parameters, and the AIC and Goodman strategies selecting models with more parameters.

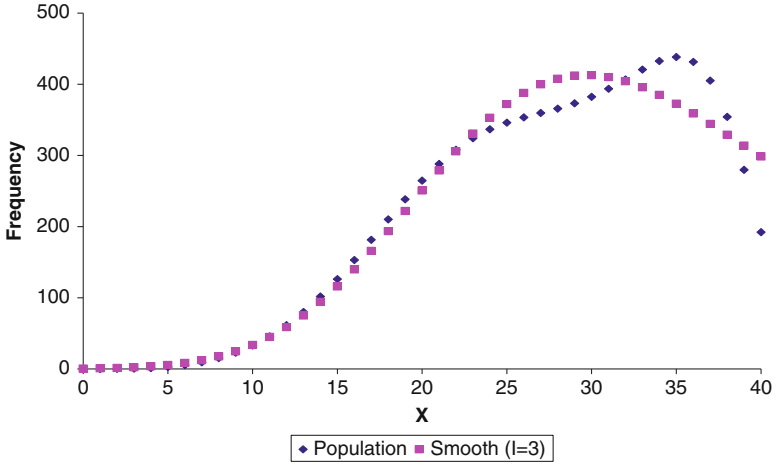

Fig. 10.10 Figure 10.9's $I = 6$ population distribution and an incorrect $I = 3$ model of the $I = 6$ distribution

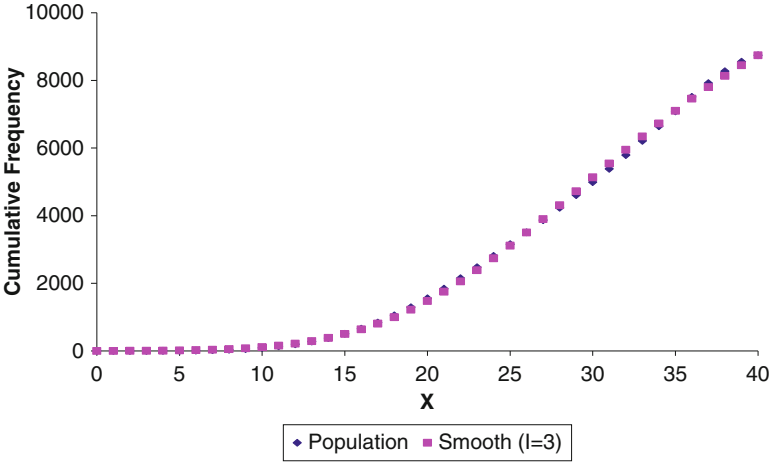

Fig. 10.11 The cumulative versions of Fig. 10.10's $I = 6$ and $I = 3$ distributions

- The best selection strategy for equating function accuracy is the AIC strategy, which favored more parameters, producing equating functions with the least bias and negligibly higher variability.
- The selection strategies do not have to make perfectly accurate model selections in order to produce equating functions that are sufficiently accurate for practice. The accuracy implications are greater for the consideration of models with $I = 2$ vs. $I = 3$ than for considerations such as models with $I = 5$ vs. $I = 6$. These accuracy implications are partially due to the possibility that accurate cumulative distributions can be calculated from inaccurate frequency distributions.

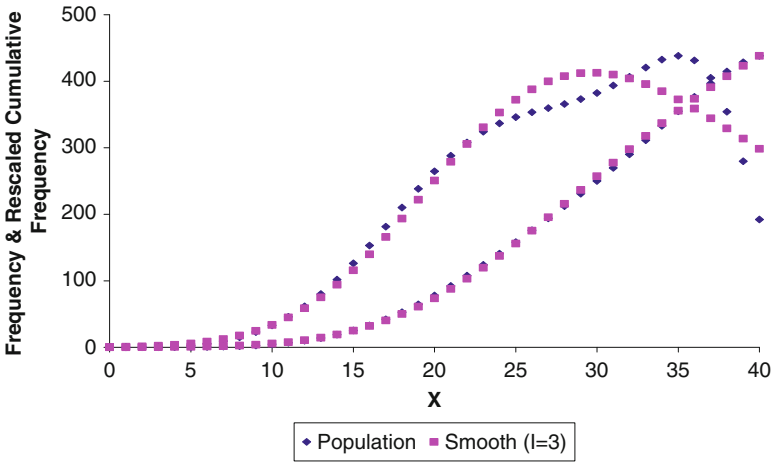

Fig. 10.12 The $I = 6$ and $I = 3$ distributions from Figs. 10.9 and 10.10

10.4.2 Implications of Selecting Bivariate Parameterizations for Equating Function Accuracy

Other studies by Holland and Moses extended initial focus on univariate distributions and equating functions to bivariate distributions and the equating functions based on bivariate distributions (Moses & Holland, 2010a, 2010b). These studies were carried out by finding well-fitting log-linear models such as (10.5) for actual test data, treating these models as populations from which to draw samples of particular sizes, and then making selections of the bivariate parameterizations with different statistical strategies and computing equipercentile equating functions based on the selected models. The model selection process focused only on the bivariate terms, where models were considered with $D = E$ values of 0, 1, 2, or 3. The statistical strategies evaluated were the same ones as evaluated in the studies of univariate distributions. Two equipercentile equating functions were evaluated: the post stratification equating function described in Sect. 10.2 and the *chained* equipercentile equating function, an equating function that uses marginal distributions and not bivariate distributions in its equating.

The model selection results for selections of bivariate parameterizations were somewhat different from those obtained for selections of univariate parameterizations described in Sect. 10.4.1. Whereas in the Moses and Holland (2010a) univariate results the chi-square strategies were fairly similar and the AIC and Goodman strategies selected relatively more parameters (Table 10.2), in the Moses and Holland (2010a, 2010b) bivariate results the Pearson and Cressie-Read chi-square strategies tended to select more parameters than the other strategies ($D = E = 2$ or 3) and the Goodman strategy tended to select the fewest parameters

($D = E = 0$ or 1). Further examination of the simulation results showed that the Pearson chi-square, Cressie-Read chi-square, and Goodman strategies were affected by extreme sparseness that tends to arise in bivariate distributions of test data. The Pearson and Cressie-Read chi-square strategies used chi-square statistics that required divisions by smoothed frequencies. When these chi-square statistics were calculated based on log-linear smoothing models that mis-fit the observed bivariate data, they could be greatly inflated in such a way that models with the largest numbers of bivariate parameters would be selected at high rates. The Goodman strategy was based on evaluating the likelihood ratio chi-square statistic with respect to its degrees of freedom, an evaluation that would result in the selection of very few bivariate parameters because chi-square statistics are often considerably smaller than their degrees of freedom in sparse bivariate test data.

The influence of the selection strategies for bivariate log-linear smoothing models on equating function accuracy showed a bias and variability tradeoff: Equating functions based on selection strategies that tended to select more bivariate parameters (i.e., Pearson and Cressie-Read chi-squares) were less biased and more variable than equating functions based on strategies that selected fewer bivariate parameters (i.e., Goodman). In terms of practical implications, the post stratification equating function was more strongly influenced than the chained equating function by the use of different bivariate selection strategies. This result was not surprising given that the post stratification equating function makes direct use of bivariate test distributions whereas the chained equating function does not.

10.4.3 Implications of Log-Linear Models for Standard Error Estimation

Another evaluation of the use of log-linear models was on the estimation accuracy of asymptotic standard errors of smoothed equating functions (Moses & Holland, 2007b). In this simulation, different log-linear smoothing models were fit to hundreds of sample datasets and the standard deviations of the equated scores were compared to the averages of the standard error estimates. Similar to the studies described in Sects. 10.4.1 and 10.4.2, the results showed that equating functions based on log-linear models with more parameters were more variable (i.e., had larger standard errors) than equating functions based on log-linear models with fewer parameters. Interestingly, the accuracy of the standard error estimation was less influenced by the choice of log-linear model than by the accuracy of the equating function. Log-linear models that were incorrect with respect to the population test score distributions produced quite accurate standard error estimates, a result that has been shown by others (Liou & Cheng, 1995; Liou, Cheng, & Johnson, 1997).

10.5 Discussion

As described in this paper, Paul Holland has been an active contributor to the theory and practice of test equating for several years. His application of log-linear models is one of the most recommended smoothing techniques for equipercentile equating (Kolen & Brennan, 2004; Livingston, 2004). The widespread use of his application is no doubt related to his efforts to develop a practical algorithm suited to the difficulties of test data and to integrate smoothing as an explicit step within the entire equating process (e.g., von Davier et al., 2004). I was obviously very fortunate to have worked with Paul in evaluating different aspects of using log-linear models to equate tests. Our collaborative work has increased the understanding of how this statistical application performs in practical situations and has helped me become a better equating practitioner.

Acknowledgement Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: ETS.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC). The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic.

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. *American Psychologist, 6*, 404.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30*, 589–600.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions*. Iowa City, IA: American College Testing (ACT Research Rep. No. 90-4).

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions*. Princeton, NJ: ETS (ETS Research Rep. No. RR-89-06).

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

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating: The Graduate Record Examination*. Princeton, NJ: ETS (Program Statistics Research Technical Rep. No. 81-13).

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions*. Princeton, NJ: ETS (ETS Technical Rep. No. TR-87-79).

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133–183.

Institute, S. A. S. (2002). *The GENMOD procedure (Version 9) [Computer software manual]*. Cary, NC: Author.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10*, 143–160.

Keats, J. S., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59–72.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*, 257–282.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*, 259–286.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*, 349–369.

Livingston, S. (1993). Small-sample equatings with log-linear smoothing. *Journal of Educational Measurement, 30*, 23–39.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Lord, F. M. (1950). *Notes on comparable scales for test scores*. Princeton, NJ: ETS (ETS Research Bulletin No. RB-50-48).

Lord, F. M. (1955). Equating test scores: A maximum likelihood solution. *Psychometrika, 20*, 193–200.

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 34*, 239–270.

Moses, T., & Holland, P. W. (2007a). *Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Moses, T., & Holland, P. (2007b). *Kernel and traditional equipercentile equating with degrees of presmoothing*. Princeton, NJ: ETS (ETS Research Rep. No. RR-07-15).

Moses, T., & Holland, P. W. (2008). *Notes on a general framework for observed score equating*. Princeton, NJ: ETS (ETS Research Rep. No. RR-08-59).

Moses, T., & Holland, P. W. (2009). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement, 46*, 159–176.

Moses, T., & Holland, P. W. (2010a). A comparison of statistical selection strategies for univariate and bivariate loglinear models. *British Journal of Mathematical and Statistical Psychology, 63*(3), 557–574.

Moses, T., & Holland, P. W. (2010b). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47*(1), 76–91.

Moses, T., von Davier, A. A., & Casabianca, J. (2004). *PROC GENMOD: A numerical approach using SAS*. Princeton, NJ: ETS (ETS Research Rep. No. RR-04-27).

Skaggs, G. (2004). *Passing score stability when equating with very small samples*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.

Wilks, S. S. (1961). *Scaling and equating College Board tests*. Princeton, NJ: ETS.

Chapter 11
Chain Equipercentile Equating and Frequency Estimation Equipercentile Equating: Comparisons Based on Real and Simulated Data

Sandip Sinharay

11.1 Introduction

The nonequivalent groups with anchor test (NEAT) design, also known as the *common item, nonequivalent groups design* (Kolen & Brennan, 2004), is used in equating scores of several large-scale tests such as the SAT® and the certification examinations conducted by the American Society for Quality. The two observed-score equating (OSE) methods popular with the NEAT design are chain equating (CE) and poststratification equating (PSE). Here, we consider their nonlinear versions, that is, the frequency estimation equipercentile equating (FEEE) for PSE, and the chained equipercentile equating (CEE) method for CE (see Kolen & Brennan, 2004, for further details on these methods).

Von Davier, Holland, and Thayer (2004a, 2004b) showed that both the CEE and FEEE methods are examples of OSE methods under different assumptions about the missing data in the NEAT design. These assumptions cannot be directly evaluated using the data that are usually available under a NEAT design. In practical situations, the FEEE and CEE methods tend to produce different results when the two non-equivalent groups of examinees differ substantially in performance on the anchor test. The weaker the correlation between the test and anchor scores, the bigger the difference. Naturally, practitioners would like to know which of the two methods leads to more accurate equating so that they can employ that method. This paper attempts to answer that question by discussing a collection of results found by several researchers on the comparison of the FEEE and CEE methods.

The next section describes the NEAT design and discusses the two equating methods. The three sections that follow discuss the comparison of the FEEE and CEE methods based on (a) theoretical arguments, (b) simulated data, and (c) operational test data. Conclusions and discussion are provided in the last section.

S. Sinharay (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: ssinharay@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_11,

11.2 The Nonequivalent Groups with Anchor Test Design and the Two Equating Methods

In the NEAT design, two operational tests, the new test $\boldsymbol{X}$ and the old test[1] $\boldsymbol{Y}$, are given to two different samples of examinees from different test populations (denoted here by $\boldsymbol{P}$ and $\boldsymbol{Q}$). In addition, an anchor test $\boldsymbol{A}$ is given to both samples. Test $\boldsymbol{X}$ is observed on $\boldsymbol{P}$ but not $\boldsymbol{Q}$, and $\boldsymbol{Y}$ is observed on $\boldsymbol{Q}$ but not $\boldsymbol{P}$; data for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$ are always missing in a NEAT design. The data collection design is shown in Table 11.1.

The task is to equate the scores of $\boldsymbol{X}$ to those of $\boldsymbol{Y}$. Both external anchor tests (anchor tests whose scores do not contribute to the reported examinee score) and internal anchor tests (anchor tests whose scores contribute to the reported examinee score) are considered in this study.

The *target population* $\boldsymbol{T}$ for the NEAT design, which is the population on which the equating is supposed to be defined, is the *synthetic population* based on $\boldsymbol{P}$ and $\boldsymbol{Q}$ (Braun & Holland, 1982), in which $\boldsymbol{P}$ and $\boldsymbol{Q}$ are given weights w and $(1 - w)$ that indicate their degree of influence on $\boldsymbol{T}$. Following Braun and Holland (1982), $\boldsymbol{T}$ is denoted by

$$T = wP + (1 - w)Q, \quad 0 \leq w \leq 1. \tag{11.1}$$

The *total* or *combined population*, often denoted by $\boldsymbol{P} + \boldsymbol{Q}$, that is obtained by pooling the samples from $\boldsymbol{P}$ and $\boldsymbol{Q}$, is the synthetic population. The weight w in (11.1) is usually taken as proportional to the sample size from $\boldsymbol{P}$, i.e. $w = N_{\boldsymbol{P}}/(N_{\boldsymbol{P}} + N_{\boldsymbol{Q}})$, where $N_{\boldsymbol{P}}$ and $N_{\boldsymbol{Q}}$ denote the sample sizes from $\boldsymbol{P}$ and $\boldsymbol{Q}$.

In our discussion, we will let F, G, and H denote the *cumulative distribution functions* (cdfs) of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{A}$, respectively, and will use the subscripts P, Q, and T to indicate the populations that the cdfs refer to. For example, $F_{\boldsymbol{P}}(x)$ denotes the proportion of examines in $\boldsymbol{P}$ for which $\boldsymbol{X}$ is less than or equal to the value x, i.e., $F_{\boldsymbol{P}}(x) = \mathrm{P}\{\boldsymbol{X} \leq x|\boldsymbol{P}\}$.

We take the position, as in von Davier et al. (2004a), that in order to justify an equating method as an OSE method, it is necessary and sufficient to show that the

Table 11.1 The design table for the nonequivalent groups with anchor test (NEAT) design

	$\boldsymbol{X}$	$\boldsymbol{A}$	$\boldsymbol{Y}$
$\boldsymbol{P}$	✓	✓	
$\boldsymbol{Q}$		✓	✓

Note. X = old test, A = anchor test, Y = new test, P and Q = two different samples of examinees from different test populations

[1]Note that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are often different forms of the same test (for example, forms A and B of SAT) rather than being different tests. We call them *tests* rather than *test forms* for simplicity. The *new test* is often referred to as the test/form to be equated, and the *old test* is referred to as the test/form to be equated to.

method is equivalent to an *equipercentile equating function* defined on the target population, that is, for some $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$,

$$\text{Equi}_{XY;T}(x) = {G_T}^{-1}(F_T(x)). \tag{11.2}$$

A basic requirement for developing an OSE method for the NEAT design is to make sufficiently strong and not directly testable missing-data assumptions that allow $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ to be estimated in order to apply (11.2). The assumptions of CEE and FEEE, formalized in von Davier et al. (2004a), are described next.

In FEEE, it is assumed that the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{P}$ is the same as the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{T}$ for any choice of $\boldsymbol{T} = w\boldsymbol{P} + (1 - w)\boldsymbol{Q}$. An analogous assumption holds for $\boldsymbol{Y}$ given $\boldsymbol{A}$ in $\boldsymbol{Q}$ and in $\boldsymbol{T}$.

Using these assumptions, the marginal distribution of $\boldsymbol{X}$ in $\boldsymbol{T}$, $\text{P}\{\boldsymbol{X} = x_j|\boldsymbol{T}\}$, is estimated as

$$\text{P}\{\boldsymbol{X} = x_j|\boldsymbol{T}\} = \sum_j \text{P}\{\boldsymbol{X} = x_j|\boldsymbol{A} = a_l\boldsymbol{T}\}\text{P}\{\boldsymbol{A} = a_l|\boldsymbol{T}\}, \tag{11.3}$$

where $\text{P}\{\boldsymbol{A} = a_l|\boldsymbol{T}\} = \text{wP}\{\boldsymbol{A} = a_l|\boldsymbol{P}\} + (1 - w)\text{P}\{\boldsymbol{A} = a_l|\boldsymbol{Q}\}$ is the marginal distribution of $\boldsymbol{A}$ in $\boldsymbol{T}$. A similar expression for the marginal distribution of $\boldsymbol{Y}$ in $\boldsymbol{T}$, $\text{P}\{\boldsymbol{Y} = \text{y}_k|\boldsymbol{T}\}$, can be obtained. The next steps are to compute the corresponding cdfs $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ from $\text{P}\{\boldsymbol{X} = x_j|\boldsymbol{T}\}$ and $\text{P}\{\boldsymbol{Y} = \text{y}_k|\boldsymbol{T}\}$, continuize the cdfs, and employ (11.2) to obtain the FEEE equating function.

In CE, it is assumed that the equipercentile function computed in $\boldsymbol{P}$ for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ is the same as that for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{T}$ for any choice of $\boldsymbol{T} = w\boldsymbol{P} + (1 - w)\boldsymbol{Q}$. An analogous assumption holds for the links from $\boldsymbol{A}$ to $\boldsymbol{Y}$ in $\boldsymbol{Q}$ and in $\boldsymbol{T}$.

The assumptions lead to the relationship

$${H^{-1}}_T(F_T(x)) = {H^{-1}}_P(F_P(x)). \tag{11.4}$$

From the definition of inverse functions, (11.4) is equivalent to defining the cdf, $F_{\boldsymbol{T}}(x)$, by

$$F_T(x) = H_T({H^{-1}}_P(F_P(x))). \tag{11.5}$$

In a similar manner, the CEE assumptions lead to

$${G^{-1}}_T(u) = {G^{-1}}_Q(H_Q({H^{-1}}_T(u))). \tag{11.6}$$

Equations (11.5) and (11.6) may be combined to form the chain equipercentile equating function linking $\boldsymbol{X}$ to $\boldsymbol{Y}$ as

$$\begin{aligned} {G^{-1}}_T(F_T(x)) &= {G^{-1}}_Q(H_Q({H^{-1}}_T(H_T({H^{-1}}_P(F_P(x)))))) \\ &= {G^{-1}}_Q(H_Q({H^{-1}}_P(F_P(x))))). \end{aligned} \tag{11.7}$$

The above descriptions make it clear that both the FEEE and CEE methods can be expressed as OSE methods.

The FEEE assumptions imply missing data assumptions that are conditional on the anchor test. The CEE assumptions require some manipulation to see their implication for the missing data. No simple connection exists between these two sets of assumptions. Researchers von Davier et al. (2004a) gave an example where the means of $\boldsymbol{A}$ in $\boldsymbol{P}$ and $\boldsymbol{Q}$ differed by about a third of a standard deviation and the two methods produced results that were different enough to have practical consequences. In such an example, it is impossible for both sets of assumptions to be simultaneously satisfied – one or both sets must be violated.

The next three sections discuss the comparison of the FEEE and CEE methods based on (a) theoretical arguments, (b) simulated data, and (c) operational test data.

11.2.1 Theoretical Comparison

It was discussed above that both the FEEE and CEE methods can be expressed as OSE methods. Researchers von Davier et al. (2004b) showed that both of these methods produce essentially identical results under two extreme conditions: (a) the two populations are very similar or (b) the anchor test is perfectly correlated with both tests. The second of these conditions is never satisfied in practice, while the first is sometimes satisfied for a small number of carefully controlled tests like the SAT. Hence one can expect the two methods to produce different results in most practical situations.

The CEE method uses two equatings of unequally reliable tests and then chains them together for the final result. Several experts thought that such a procedure might inherit some problems because of the unequal reliability issues of each link. Harris and Kolen (1990) declared the FEEE method to be preferable to the CEE method on theoretical grounds. In their recent book on equating, Kolen and Brennan (2004, p. 146) referred to CEE as having "theoretical shortcomings." However, Holland, Sinharay, von Davier, and Han (2008) commented, "We have attempted to discover what these problems might be, but currently regard such efforts as pointless. The theoretical basis of CEE is exactly like that of FEEE and consists of sets of assumptions about the missing data in the NEAT design that, in turn, allow CEE to be interpreted as an OSE equipercentile function for the NEAT design" (p. 38).

On the other hand, Livingston (2004) explained that when the correlation coefficient between the tests to be equated and the anchor test is considerably lower than 1 and $\boldsymbol{P}$ and $\boldsymbol{Q}$ differ substantially in ability, the FEEE method does not adjust enough for the difference in difficulty of the tests to be equated and hence leads to biased equating. Livingston commented that this problem does not occur for the CEE method. Livingston, Dorans, and Wright (1990) explained that stratifying on observed scores in the FEEE method is a fallible approximation to stratifying on true scores, and this fallibility results in an under-correction of the group differences that increases as the anchor-score difference between the groups increases.

11.2.2 Comparisons Based on Simulated Data

Wang, Lee, Brennan, and Kolen (2008) and Sinharay and Holland (2007) compared the bias, variability, and root mean squared error (RMSE) of the CEE and FEEE equating functions under several conditions by generating data from item response theory (IRT) models. Wang et al. generated data from the three-parameter logistic (3PL) model while Sinharay and Holland generated data from the two-parameter logistic (2PL) model and also from its multivariate version with four dimensions. Both of these studies found that the CEE method tends to show less bias and less RMSE than the FEEE method when large differences are present between the two groups, while the FEEE method has slightly less variability than the CEE method for all simulation conditions. Sinharay and Holland also found that for data produced under the multivariate IRT model, the FEEE method leads to slightly less bias, variability, and RMSE than the CEE method when the new form population is more able than the old form population in some content areas and less able in some other content areas (a situation observed by, for example, Klein & Jarjoura, 1985). Some results for data simulated from the 2PL model by Sinharay and Holland are discussed next.

Sinharay and Holland (2007) varied the following factors in their simulation study that simulated data from the 2PL model:

1. *Test length*. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are always of equal length and the length takes one of the values (45, 60, or 78) to emulate three operational tests: (a) a 45-item basic skills test, (b) the 60-item mathematics section of an admissions test, and (c) the 78-item verbal section of the same admissions test. The factor that is denoted by test length refers to more than simply the length of the tests to be equated. Each test length has its own set of item parameters that were estimated from an operational test data set and were used to simulate the data for the comparison of the two equating methods. Moreover, the length of the anchor test for each test length is different as indicated below.
2. *Sample size*. The sample sizes for $\boldsymbol{P}$ and $\boldsymbol{Q}$ are equal and are equal to one of three values: 100 (small), 500 (medium), and 5,000 (large).
3. *Difference in the mean ability (denoted as Δ_a) of the two examinee populations $\boldsymbol{P}$ and $\boldsymbol{Q}$*. Four values were used: -0.2, 0, 0.2, and 0.4. Units are in standard deviation (SD) of θ.
4. *Difference in the mean difficulty (denoted as Δ_d) of the two tests $\boldsymbol{X}$ and $\boldsymbol{Y}$*. Three values were used: 0, 0.2, and 0.5. Units are in SD of θ.

The values for the above four factors were chosen after examining data from several operational tests. The anchor test is of length 20 for the 45-item basic skills test and is the same length as the operational administrations of the two admissions tests – 35 for the 78-item test and 25 for the 60-item test. Sinharay and Holland (2007) considered three types of anchor tests, but here we will only discuss the results for the *minitest* (an anchor that is a representative of the total test in content coverage and difficulty). The average difficulty of an anchor test was always

centered at the average difficulty level of $\boldsymbol{Y}$, the old test form. The item parameters for test $\boldsymbol{Y}$ were computed from real data sets. The item parameters for the tests $\boldsymbol{X}$ and $\boldsymbol{A}$ were simulated from a multivariate normal distribution fitted on the item parameters used for test $\boldsymbol{Y}$. For any simulation condition determined by a level of each of the four factors, the bias, variability, and RMSE of the FEEE and CEE methods were computed by comparing the equating functions produced by these two methods to the CEF, which was easy to compute for these simulated data.

Figures 11.1 and 11.2 compare the bias and variability of the two methods for nine simulation conditions (all combinations of three test lengths [45, 60, and 78] and three Δ_as [0, 0.2, and 0.4]) when the sample size is 500 and $\Delta_d = 0$. The results were very similar for other values of Δ_d. In any panel of these two figures, the values

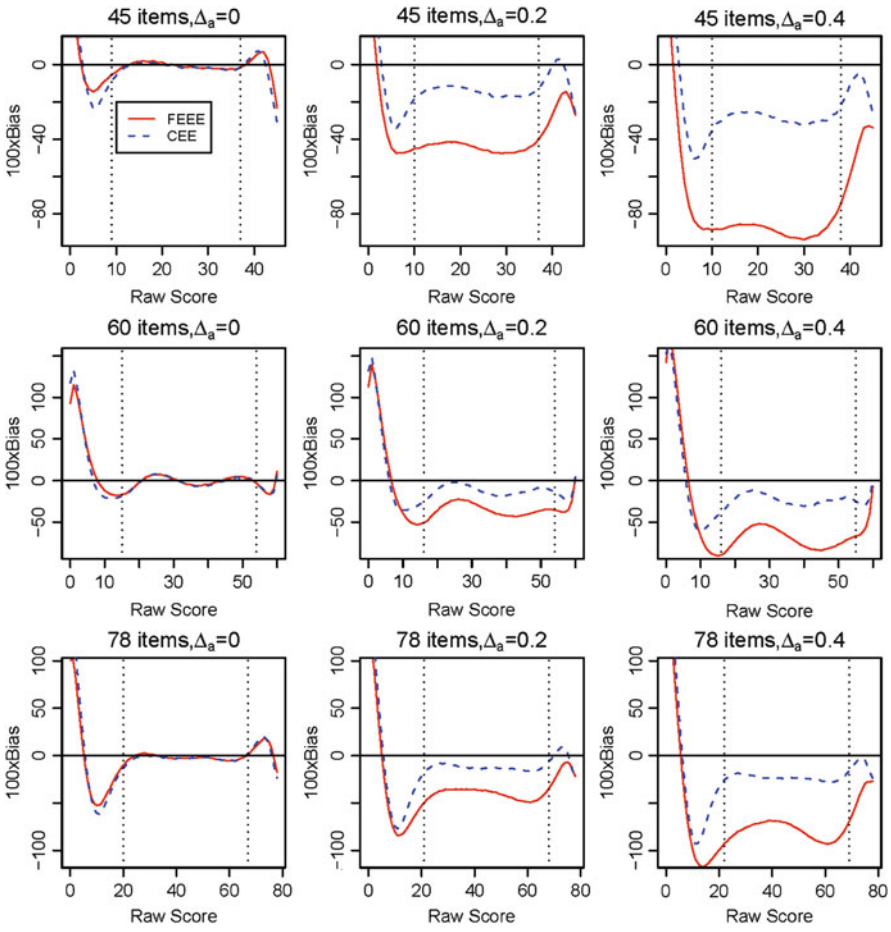

Fig. 11.1 Comparison of the bias of the frequency estimation equipercentile equating (FEEE) and chained equipercentile equating (CEE) methods for sample size of 5,000

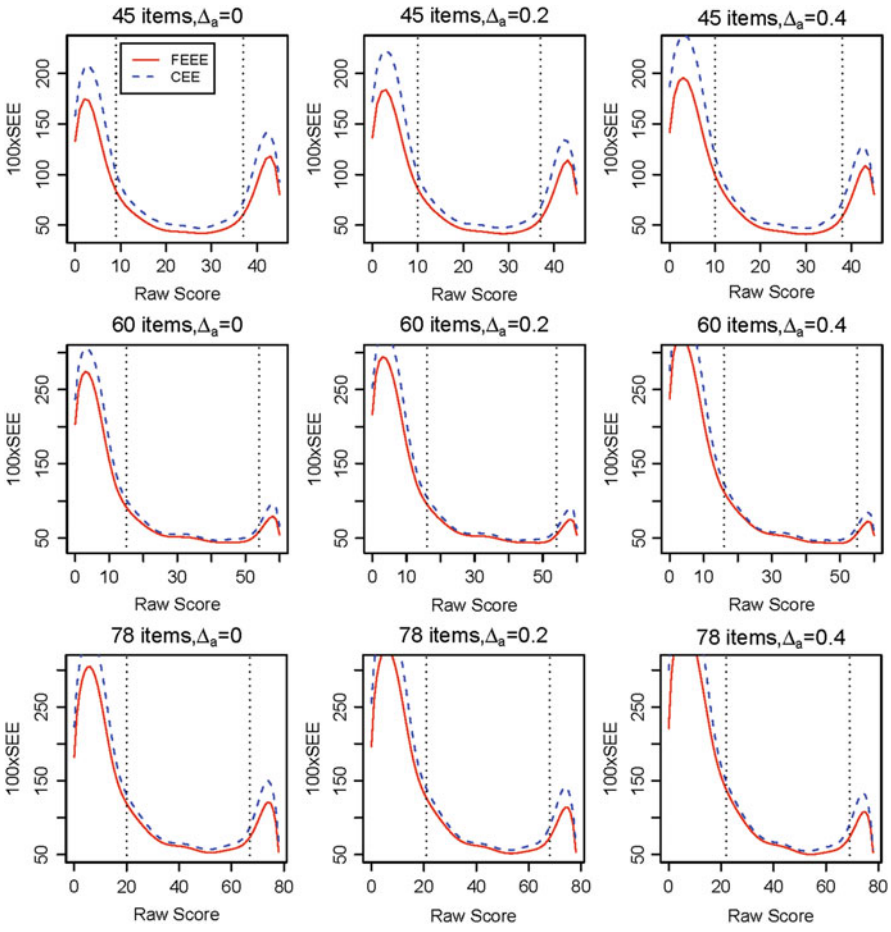

Fig. 11.2 Comparison of the standard error of equating (SEE) of the frequency estimation equipercentile equating (FEEE) and chained equipercentile equating (CEE) methods for sample size of 500

of bias or standard error of equating (SEE), multiplied by 100, for FEEE and CEE are plotted along the Y-axis for all possible scores that are plotted along the X-axis. Each row corresponds to a test length. Figure 11.1 shows that when Δ_a (the ability difference between the two groups) increases, the bias of the CEE method increases, but the bias of the FEEE method increases much more rapidly so that the difference in bias of these two methods is substantial when $\Delta_a = 0.4$. Figure 11.2 shows that the SD of the FEEE method is always smaller than that of the CEE method, but only very slightly. Sinharay and Holland (2007) found similar results for other simulation conditions when data were generated from a 2PL model.

Two problems with comparing the methods using data generated from an IRT model are that the simulated data have an uncertain relationship to operational data and that the simulated data may favor one of these methods. For example, Wang

et al. (2008) acknowledged that their simulation procedure may have disadvantaged the FEEE method to an unknown extent. If the data were simulated in another manner, the FEEE method might have exhibited less bias than that in the studies of Wang et al. and Sinharay and Holland (2007). Two ways to overcome this limitation are (a) to use operational test data or (b) to simulate data to reflect reality as closely as possible. The following section discusses the first of these ways.

11.2.3 Comparisons Based on Operational Test Data

Several researchers compared the FEEE and CEE methods based on operational test data. Marco, Petersen, and Stewart (1983) compared the FEEE and CEE methods using data from SAT-V (the verbal section of the SAT). Livingston et al. (1990) compared the FEEE and CEE methods using data from the SAT-V and SAT-M (the math section of the SAT). Harris and Kolen (1990) compared the FEEE and CEE methods using data from a certification test. von Davier et al. (2004a) compared the two methods using data from a high volume testing program. Ricker and von Davier (2007), Sinharay and Holland (2007), and Holland et al. (2008) compared the FEEE and CEE methods using a specially designed data set from a licensure test.

Harris and Kolen (1990) compared the results of the FEEE and CEE methods and concluded that the results were different enough to have practical implications. In von Davier et al. (2004a), the equating functions of CEE and FEEE were compared and found to differ significantly. Holland et al. (2008) compared the marginal distributions of the missing data (that is, the marginal distribution of $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$) predicted by the two methods to the corresponding observed values (both were known because of the unique design of the study) and found that the predictions from the two methods were close, but the predictions from the CEE method were closer to the corresponding observed values. Marco et al. (1983), Livingston et al. (1990), Ricker and von Davier (2007), and Sinharay and Holland (2007) compared the equating functions obtained from the FEEE and CEE methods to suitably chosen *criterion equating functions* (CEF).

Obtaining the CEF for real data applications is not straightforward – these researchers employed several methods to obtain one. For example, Livingston et al. (1990) sampled from a large group of examinees who were divided into two random groups by spiraling of two test forms $\boldsymbol{X}$ and $\boldsymbol{Y}$. The equipercentile equating of these two test forms (assuming the two groups of examinees receiving the two test forms were random groups) was used as the CEF. Livingston et al. sampled examinees from the large examinee group in different ways to form nonequivalent groups and then equated $\boldsymbol{X}$ to $\boldsymbol{Y}$ using anchor tests which were also taken by the examinees. Sinharay and Holland (2007) used a data set from von Davier et al. (2006) where the usually missing data in the NEAT design (that is, data for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$) were available so that it was possible to obtain the CEF as the single group equating function of $\boldsymbol{X}$ to $\boldsymbol{Y}$ in $\boldsymbol{P} + \boldsymbol{Q}$. Ricker and von Davier (2007) found

little difference between the results from the CEE and FEEE methods. The common finding from Marco et al. (1983), Livingston et al., and Sinharay and Holland is that the CEE method performs better than the FEEE method when the two examinee groups $\boldsymbol{P}$ and $\boldsymbol{Q}$ differ substantially, whereas the FEEE method performs better than the CEE method when the two examinee groups do not differ much.

Researchers von Davier, Holland, and Thayer (2003), von Davier et al. (2004a, 2004b), and von Davier (2003) compared the FEEE and CEE methods with respect to SEE and the degree of population invariance using several operational test data sets. These studies showed that both the FEEE and CEE methods appear to be similar in their SEE and in their degrees of population invariance.

Thus, the findings from operational data also slightly favor the CEE method. However, the results from these two methods do not differ too much. For example, Holland et al. (2008) compared these methods under a difficult equating situation where the two tests differed substantially in difficulty and the two population differed substantially in ability but found only a slight difference between the results of the two methods. Next, we discuss some results from operational data from Sinharay and Holland (2007).

The operational data used in Sinharay and Holland (2007) are those used by von Davier et al. (2006) and are from one form of a licensing test for prospective teachers. The form included 120 multiple-choice items, about equally divided among four content areas – language arts, mathematics, social studies, and science. Ordinarily, the total score from different forms of this test are equated through a NEAT design with an internal anchor test. The form of the test used here was administered twice, and the two examinee populations played the role of $\boldsymbol{P}$ and $\boldsymbol{Q}$.

The mean total scores (the number right) of the examinees taking the test at these two administrations differed by approximately one-fourth of a standard deviation, as can be seen from the second column of Table 11.2.

Construction of the pseudo-tests. These data were used to construct two pseudo-tests ($\boldsymbol{X}$ and $\boldsymbol{Y}$) as well as three different pseudo-anchor tests ($\boldsymbol{A1}$, $\boldsymbol{A2}$, and $\boldsymbol{A3}$) of different lengths. A pseudo-test consists of a subset of the test items from the original 120-item test, and the score on the pseudo-test for an examinee is found from the responses of that examinee to the items in the pseudo-test. The pseudo-tests $\boldsymbol{X}$ and $\boldsymbol{Y}$ each contain 44 items: 11 from each of the four content areas. Tests $\boldsymbol{X}$ and $\boldsymbol{Y}$, having no items in common, were made parallel in content, but test $\boldsymbol{X}$ was constructed to be much easier than test $\boldsymbol{Y}$.

The external anchor test cases. To create data sets with external anchor tests, a basic set of 24 items (6 items from each content area) was selected to be representative of the original test and to serve as the largest external anchor $\boldsymbol{A1}$. This anchor test has no items in common with either $\boldsymbol{X}$ or $\boldsymbol{Y}$. The two other anchor tests, $\boldsymbol{A2}$ and $\boldsymbol{A3}$, were formed by deleting 4 and 8 items, respectively, from $\boldsymbol{A1}$ in such a way that $\boldsymbol{A2}$ is a 20-item subset of $\boldsymbol{A1}$, and $\boldsymbol{A3}$ is a 16-item subset of $\boldsymbol{A2}$. Furthermore, to maintain parallelism in content, test $\boldsymbol{A2}$ had five items from each content area, while $\boldsymbol{A3}$ had four. The mean percent correct of the anchor tests approximately equaled that for the original test.

Table 11.2 N*s*, means, standard deviations, reliabilities, and average proportions correct for the scores on the total and Pseudo-tests on *P*, *Q*, and the Combined Group, $P + Q$

Test	Total (120 items)	*X* (44 items)	*Y* (44 items)	*A*1 (24 items)	*A*2 (20 items)	*A*3 (16 items)	$X1 = X + A1$	$Y1 = Y + A1$
P	82.3	35.1	26.6	16.0	13.7	10.8	51.2	42.6
$N = 6{,}168$	(16.0)	(5.7)	(6.7)	(4.2)	(3.6)	(3.0)	(9.3)	(10.3)
		[0.81]	[0.81]	[0.75]	[0.71]	[0.68]	[0.88]	[0.88]
		0.80	0.60	0.67	0.69	0.68	0.75	0.63
Q	86.2	36.4	28.0	17.0	14.5	11.5	53.4	45.0
$N = 4{,}237$	(14.2)	(4.8)	(6.3)	(3.9)	(3.3)	(2.8)	(8.0)	(9.6)
		[0.77]	[0.79]	[0.73]	[0.69]	[0.66]	[0.85]	[0.87]
		0.83	0.64	0.71	0.73	0.72	0.79	0.66
$\boldsymbol{P + Q}$		35.6	27.2	16.4	14.0	11.1	52.1	43.6
$N = 10{,}405$		(5.4)	(6.6)	(4.1)	(3.5)	(3.0)	(8.9)	(10.1)
		[0.80]	[0.80]	[0.75]	[0.71]	[0.68]	[0.87]	[0.87]
		0.81	0.62	0.68	0.70	0.69	0.77	0.64

Note. Numbers in parentheses ( ) are standard deviations; numbers in brackets [ ] are reliabilities

Table 11.2 also gives the *N*s, means, standard deviations, reliabilities (Cronbach's alpha), and average proportion correct for the scores on $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{A1}$, $\boldsymbol{A2}$, and $\boldsymbol{A3}$, and for the two sums $\boldsymbol{X1} = \boldsymbol{X} + \boldsymbol{A1}$ and $\boldsymbol{Y1} = \boldsymbol{Y} + \boldsymbol{A1}$ (that play a role for the internal anchor cases to be discussed shortly) for the examinees in $\boldsymbol{P}$, $\boldsymbol{Q}$, and the combined group. $\boldsymbol{X}$ is considerably easier than $\boldsymbol{Y}$ (the average percent correct on $\boldsymbol{X}$ ranges from 80 to 83% while on $\boldsymbol{Y}$ the range is 60 to 64%). The mean score on $\boldsymbol{X}$ for the combined group is 127% of a standard deviation larger than the mean score on $\boldsymbol{Y}$. In addition, all three anchor tests show differences of approximately a quarter of a standard deviation between $\boldsymbol{P}$ and $\boldsymbol{Q}$. The reliabilities of the three anchor tests behave as expected, with $\boldsymbol{A1}$ being the most reliable and $\boldsymbol{A3}$ the least reliable. However, the range of these reliabilities is modest – from 0.68 to 0.75 on the combined group.

The pseudo-test data were designed to lead to a *difficult* equating problem for which CEE and FEEE were expected to give *different* answers. The large difference in difficulty between $\boldsymbol{X}$ and $\boldsymbol{Y}$ made the equating problem non-linear. The difference in the test performance of $\boldsymbol{P}$ and $\boldsymbol{Q}$ was intentionally chosen to be as large as possible; this difference ensured that CEE and FEEE would give different results.

The internal anchor test cases. To create data sets that had internal anchor tests, we formed $\boldsymbol{X1} = \boldsymbol{X} + \boldsymbol{A1}$ and $\boldsymbol{Y1} = \boldsymbol{Y} + \boldsymbol{A1}$. Then we paired $\boldsymbol{X1}$ and $\boldsymbol{Y1}$ with $\boldsymbol{A1}$, $\boldsymbol{A2}$, or $\boldsymbol{A3}$ as the three internal anchor test cases. Because $\boldsymbol{A2}$ was a subset of $\boldsymbol{A1}$ and $\boldsymbol{A3}$ was a subset of $\boldsymbol{A2}$, each of the three anchor tests is internal to the tests $\boldsymbol{X1}$ and $\boldsymbol{Y1}$. This approach allows one to keep the total tests of the same size ($44 + 24 = 68$ items) as one varies the lengths (and therefore the reliabilities) of the anchor tests.

Mimicking the nonequivalent groups with anchor test (NEAT) design. Because all the examinees in $\boldsymbol{P}$ and $\boldsymbol{Q}$ took all 120 items on the original test, all of the examinees in $\boldsymbol{P}$ and $\boldsymbol{Q}$ have scores for $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{X1}$, and $\boldsymbol{Y1}$ as well as for each of the three anchor tests, $\boldsymbol{A1}$, $\boldsymbol{A2}$, and $\boldsymbol{A3}$. In order to mimic the structure of the NEAT design, it was pretended that scores for $\boldsymbol{X}$ or $\boldsymbol{X1}$ were not available for the examinees in $\boldsymbol{Q}$ and that scores for $\boldsymbol{Y}$ or $\boldsymbol{Y1}$ were not available for the examinees in $\boldsymbol{P}$. However, because all scores are, in fact, available for the pseudo-test data, they allow one to compare the frequencies predicted by the CEE and FEEE assumptions with the actual frequencies in the data.

Ricker and von Davier (2007) found little difference between the equating functions of the FEEE and CEE methods for these data. What follows is a comparison of the predictions made by CEE and FEEE with the observed data for $\boldsymbol{X}$ or $\boldsymbol{X1}$ in $\boldsymbol{Q}$ and for $\boldsymbol{Y}$ or $\boldsymbol{Y1}$ in $\boldsymbol{P}$ (see Holland et al. 2008 for a description of how the predictions are computed). The comparisons are divided into three parts, as described below.

Comparisons of the observed and predicted frequencies. Figure 11.3 shows the observed and predicted frequencies for CEE and FEEE for $\boldsymbol{X}$ and $\boldsymbol{X1}$ in $\boldsymbol{Q}$ and for $\boldsymbol{Y}$ and $\boldsymbol{Y1}$ in $\boldsymbol{P}$, for the case of the longest anchor test, $\boldsymbol{A1}$. (All of the graphs for the shorter anchor tests look very similar and are omitted.) The solid lines connect the observed frequencies.

It is evident that the predicted distributions for CEE and FEEE are very similar and that they depart from the observed frequencies by similar amounts and in similar directions. In general, the agreement between the observed and predicted

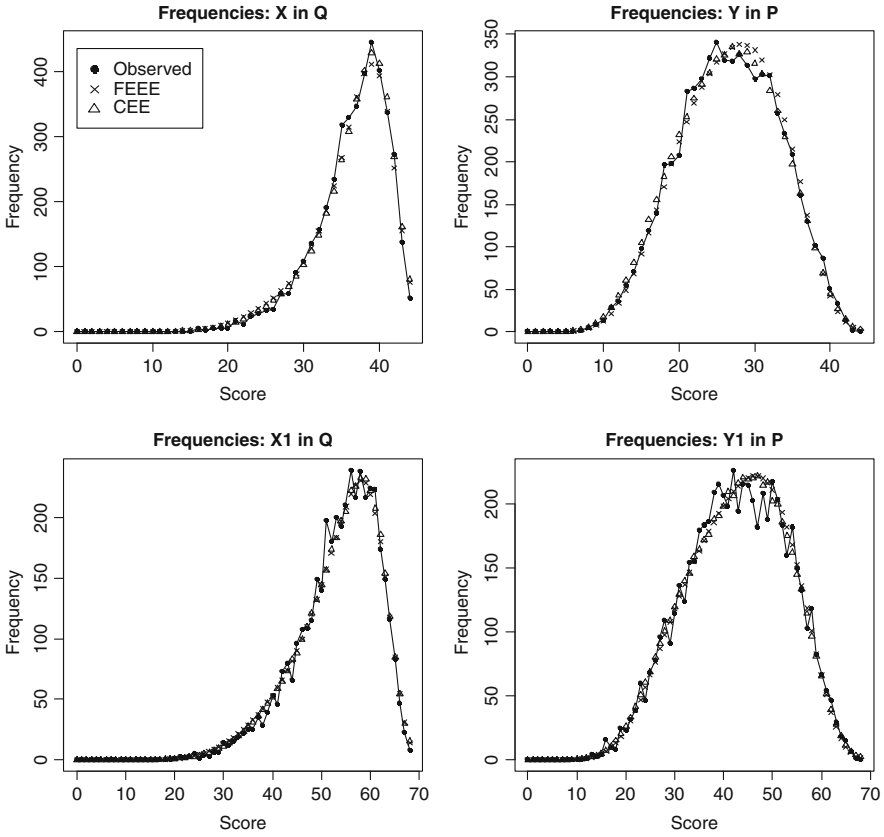

Fig. 11.3 Frequencies for X in Q and Y in P for external anchor test $A1$ (*top row*) and for X1 in Q and Y1 in P for internal anchor test A1 (*bottom row*). *Note*. FEEE = frequency estimation equipercentile equating, CEE = chained equipercentile equating

frequencies is quite good, indicating that both CEE and FEEE make predictions that are reasonably close to the data. To look at the differences in more detail, we used the Freeman-Tukey (FT) residuals, which are defined as

$$\sqrt{n_i} + \sqrt{n_i + 1} - \sqrt{4m_i + 1},$$

where n_i is the observed frequency for score i and m_i is the corresponding predicted frequency. The FT residuals are graphed in Fig. 11.4. The figure shows that the patterns of the FT residuals for CEE and FEEE are very similar and appear fairly random, well within the expected range for well-fitting predictions. However, the residuals for CEE often are smaller than those for FEEE. This finding is clearest in the middle range of scores in the top row of plots in Fig. 11.4. In summary, both CEE and FEEE track the data fairly well and both sets of predictions appear to be somewhat more similar to each other than they are to the observed data.

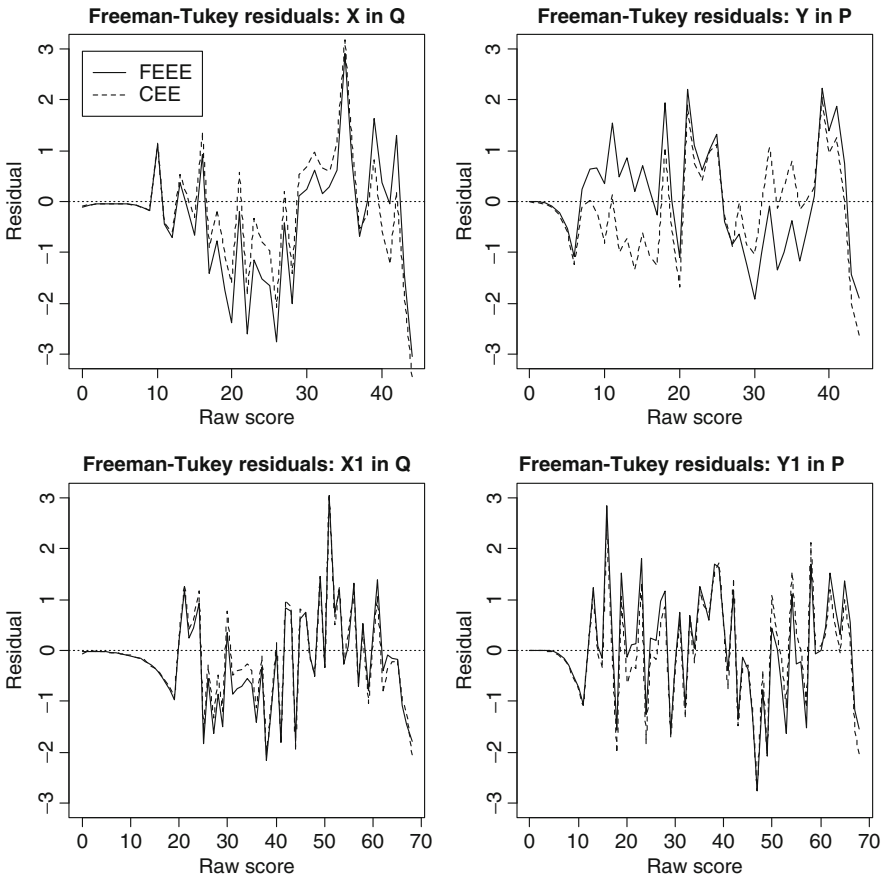

Fig. 11.4 Freeman-Tukey residuals for X in Q and Y in P for external anchor test $A1$ (*top row*) and for X1 in Q and Y1 in P for internal anchor test A1 (*bottom row*). *Note*. FEEE = frequency estimation equipercentile equating, CEE = chained equipercentile equating

Comparisons of the goodness-of-fit measures. Table 11.3 gives the values of χ^2_{FT}, which is given by

$$\chi^2_{FT} = \sum_i \left(\sqrt{n_i} + \sqrt{n_i + 1} - \sqrt{4m_i + 1}\right)^2,$$

for all the cases in the study. Table 11.3 shows, just like Fig. 11.4, that the predictions of CEE are somewhat closer to the observed frequencies than the FEEE predictions. In all cases, χ^2_{FT} is smaller for CEE than for FEEE. Thus, while the CEE and FEEE predictions are very similar, as seen in Fig. 11.3, those of CEE are, on average, slightly closer to the observed frequencies.

In addition, there is a consistent tendency for χ^2_{FT} for FEEE to get smaller as the length of the anchor test increases. Thus, it is evident that the length (and the

Table 11.3 The Freeman-Tukey goodness-of-fit measure

External anchor		Internal anchor	
Tests and anchors	χ^2_{FT}	Tests and anchors	χ^2_{FT}
X*, *A*1: *Q		***X*1, *A*1: *Q***	
FEEE	66.3	FEEE	56.0
CEE	52.1	CEE	49.1
***X*, *A*2**		***X*1, *A*2**	
FEEE	76.0	FEEE	69.4
CEE	62.9	CEE	66.9
***X*, *A*3**		***X*1, *A*3**	
FEEE	85.1	FEEE	77.7
CEE	60.4	CEE	69.4
Y*, *A*1: *P		***Y*1, *A*1: *P***	
FEEE	49.9	FEEE	72.3
CEE	39.2	CEE	69.2
***Y*, *A*2**		***Y*1, *A*2**	
FEEE	58.0	FEEE	83.5
CEE	47.4	CEE	79.4
***Y*, *A*3**		***Y*1, *A*3**	
FEEE	67.4	FEEE	93.7
CEE	45.0	CEE	85.0

Note. *A*1 is the longest anchor test, *A*3 is the shortest. *FEEE* frequency estimation equipercentile equating, *CEE* chained equipercentile equating

reliability) of the anchor test has a distinct and measurable effect on improving the predictions of FEEE. This finding is consistent with the argument of Livingston (2004) that the FEEE method provides biased results for a low value of the anchor-test-to-total-test-correlation. The predictions for CEE do not show this trend for the external anchor test cases, but they do show it for the internal anchor test cases.

Comparisons of the moments. Another comparison of the predictions for CEE and FEEE made in Holland et al. (2008) concerned those of the mean and SD of the observed frequency distributions. The values of these moments are given in Table 11.4. The table also shows the *percent relative differences* (% rel. dif.) between the observed and predicted moments. The % rel. dif. is the predicted moment minus the observed moment divided by the *absolute value* of the observed moment times 100. Thus, positive values indicate *overprediction*, while negative values indicate *underprediction*.

In almost every case in Table 11.4, in terms of the absolute value of the % rel. dif., the CEE predictions are closer to the observed data than are the FEEE predictions. The predictions of the means are quite accurate for both methods; the means have the consistently smallest percent relative differences in the table, but the differences for the CEE predictions are always smaller. For the SDs, the percent relative differences are generally a little larger than that for the means, but again, those for CEE are always smaller.

Table 11.4 Observed (obs) and predicted moments and percent relative difference (% rel. dif.) for external and internal anchor-test cases

External anchor cases					Internal anchor cases				
Observed and predicted	Mean	Mean % rel. dif.	SD	SD % rel. Zdif.	Observed and predicted	Mean	Mean % rel. dif.	SD	SD % rel. dif.
X,A1:Q					***X1,A1:Q***				
Obs	36.38		4.77		Obs	53.38		8.04	
FEEE	36.16	−0.6	5.15	7.9	FEEE	53.17	−0.4	8.47	5.3
CEE	36.43	0.1	4.98	4.4	CEE	53.31	−0.1	8.35	3.8
X, A2					X1, A2				
Obs	36.38		4.77		Obs	53.38		8.04	
FEEE	36.13	−0.7	5.19	8.8	FEEE	53.11	−0.5	8.57	6.5
CEE	36.41	0.1	5.03	5.5	CEE	53.29	−0.2	8.44	4.9
X, A3					X1, A3				
Obs	36.38		4.77		Obs	53.38		8.04	
FEEE	36.04	−0.9	5.26	10.2	FEEE	52.94	−0.8	8.67	7.8
CEE	36.33	−0.1	5.09	6.7	CEE	53.16	−0.4	8.52	5.9
Y,A1:P					***Y1,A1:P***				
Obs	26.59		6.68		Obs	42.62		10.31	
FEEE	26.79	0.8	6.56	−1.7	FEEE	42.82	0.5	10.17	−1.3
CEE	26.44	−0.6	6.72	0.6	CEE	42.62	0.0	10.26	−0.4
Y, A2					Y1, A2				
Obs	26.59		6.68		Obs	42.62		10.31	
FEEE	26.82	0.9	6.52	−2.3	FEEE	42.89	0.6	10.18	−2.2
CEE	26.45	−0.5	6.64	−0.5	CEE	42.64	0.0	10.16	−1.4
Y, A3					Y1, A3				
Obs	26.59		6.68		Obs	42.62		10.31	
FEEE	26.91	1.2	6.49	−2.8	FEEE	43.08	1.1	10.00	−3.0
CEE	26.55	−0.2	6.62	−0.9	CEE	42.80	0.4	10.11	−1.9

Note. *A*1 is the longest anchor test, *A*3 is the shortest. *FEEE* frequency estimation equipercentile equating, *CEE* chained equipercentile equating

As seen earlier for the goodness-of-fit measures, there is a consistent tendency for the accuracy of the FEEE-predictions of the means and SDs to increase as the length of the anchor test increases. The CEE predictions for the mean and SD for the internal anchor test show the same consistent improvement as the length of the anchor test increases.

11.2.4 Conclusions and Discussion

This study discusses a collection of results that compare the two most common OSE methods for the NEAT design – CEE and FEEE. The research works discussed here used a variety of ways to compare the two methods – some used theoretical arguments, some used simulated data, and some used operational data.

The existing research suggests unequivocally that the CEE method performs slightly better than the FEEE method under most circumstances. The only situations when the CEE method performs worse than the FEEE method are when the two populations are the same in ability and when they differ differentially in the various content areas of the test (that is, one population is better than the other in some content areas and worse in some other content areas) – both of these situations are rare in practice.

Linear equating methods were not considered in this paper. However, results discussed in this paper are in agreement with the existing results for linear equating. For example, Livingston et al. (1990) and Puhan (2010) showed using operational data that the chain linear equating leads to more accurate equating in general than the Tucker equating method (which can be viewed as the linear version of the FEEE method).

What should a practitioner do regarding the choice of CEE versus FEEE in an operational testing situation? As mentioned earlier, the psychometric basis for CEE was questioned for a long time. However, accumulating evidence suggests that the missing data assumptions of CEE are reasonable and likely to be useful in a variety of circumstances. In addition, as discussed in this paper, an ever increasing set of findings slightly favors CEE over FEEE methods for simulated data and in real test situations, and we do not expect that result to change with further research. While further research is surely needed to help distinguish situations where one of these methods is to be preferred, it is certainly the case that CEE is a clear competitor to FEEE and the other OSE methods for the NEAT design.

Several related issues could be explored in future. The existing studies compared the FEEE and CEE method with respect to the bias, SEE, and RMSE of the equating function and population invariance. It will be useful to consider other equating criteria such as the same distributions property and the first- and second-order equity properties (e.g. Tong & Kolen, 2005). Also, more research is needed regarding simulation of data that will not favor any of these two methods – a comparison of the two methods for this type of data will provide a clear picture of the relative performance of the two methods.

Acknowledgments This work was funded by Educational Testing Service. The author thanks Dan Eignor, Paul W. Holland, Rick Morgan, and Skip Livingston for helpful comments, and Ayleen Stelhorn and Kim Fryer for editorial help. Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic Press.

Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50*, 61–71.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*, 17–43.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197–206.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York, NY: Academic Press.

Puhan, G. (2010). A comparison of chained linear and post stratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*, 54–75.

Ricker, K., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a non-equivalent group design*. Princeton, NJ: ETS (ETS Research Rep. No. RR-07-44).

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*, 249–275.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*, 418–432.

von Davier, A. A. (2003). *Notes on linear equating methods for the non-equivalent groups design*. Princeton, NJ: ETS (ETS Research Rep. No. RR-03-24).

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data*. Princeton, NJ: ETS (ETS Research Rep. No. RR-06-02).

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus poststratification methods for equating and test linking. In N. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program® Examinations*. Princeton, NJ: ETS (ETS Research Rep. No. RR-03-27).

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The kernel method of test equating*. New York, NY: Springer.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The chain and poststratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15–32.

Wang, T., Lee, W., Brennan, R. J., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement, 32*, 632–651.

Chapter 12
An Observed-Score Equating Framework

Alina A. von Davier

12.1 Introduction

Paul Holland has made remarkable contributions to equating theory and practice and has influenced the work of many researchers and psychometricians. In this paper, it is argued that the methodology introduced by Holland and Thayer (1989) and von Davier, Holland, and Thayer (2004b), along with the kernel method of test equating, involves more than simply a continuization method for test score distributions: It has introduced a powerful equating framework[1] for all observed-score equating (OSE) methods. This framework has already proven to be useful for various research purposes outside of Gaussian kernel equating (KE). Referred to in this paper as the observed-score equating (OSE) framework, it is one example of the application of Holland's work to the practice of equating.

Identifying a framework that connects the methods used in OSE practice is part of the continuous search for a theory of equating (see also Holland & Hoskens, 2003; von Davier, 2011). This equating framework, the OSE framework, together with Dorans and Holland's five requirements of an equating procedure (Dorans & Holland, 2000), is the closest to a theory that is available for OSE.

This paper starts with a brief history of KE methods (in particular, Gaussian KE) and the development of the OSE framework. A few formal aspects of the equating process are included here, along with the introduction of the analytical standard errors from the KE procedure. The practical advantages of the modular structure of the OSE framework are discussed. It is shown how the framework extends beyond

[1]"Conceptual frameworks (theoretical frameworks) are a type of intermediate theory that has the potential to connect to all aspects of inquiry (e.g. problem definition, purpose, literature review, methodology, data collection and analysis). Conceptual frameworks act like maps that give coherence to empirical inquiry" (Conceptual framework, 2010, para 2).

A.A. von Davier (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: avondavier@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_12,

the Gaussian kernel to include all existing OSE methods. The paper also covers the rich body of equating research that has evolved around the OSE framework and points out directions for new research.

12.2 History of Kernel Equating Methods

To equate test forms, psychometricians often use the percentile rank method (or the equipercentile equating function in conjunction with linear interpolation for continuizing the otherwise discrete distribution functions). One of the consequences of this method is that the linearly interpolated cumulated distribution functions (cdfs) and the equating function have irregularities – that is, the functions are not smooth (see Kolen & Brennan, 2004, Figs. 2.4, 2.5, and 2.10). Another issue that arises in equipercentile equating with linear interpolation is that the equated scores are assigned arbitrarily when no examinees are recorded at a particular score. To address these issues, research over the past 30 years has focused on procedures for smoothing the data prior to equating (presmoothing) and after the equating (postsmoothing) on alternative procedures for continuizing the cdfs and on new equating functions.

In 1989, Holland and his colleagues published two ETS research reports that described a new equating method that they called the kernel method of test equating (Holland, King, & Thayer, 1989; Holland & Thayer, 1989). These reports made several important contributions, the main one being the application of statistical techniques for continuizing discrete distributions using Gaussian kernels to test score distributions in the field of psychometrics. This new continuization method resulted in several advantages over the traditional linear interpolation method of continuization, which had been and continues to be broadly used by practitioners and researchers. These advantages include the fact that the KE method can potentially result in the following: (a) a family of equating functions with linear equating as a special case, (b) smooth and differentiable equating functions with a parameter or bandwidth that controls the degree of smoothing, and (c) analytical asymptotic standard errors that are defined everywhere on the domain of the function. The equating steps described in Holland and Thayer (1989) were outlined as follows: (a) presmoothing (using log-linear models), (b) continuization (using the Gaussian kernels), (c) computation of the equating function, and (d) computation of the standard error of equating (SEE). The fact that Holland and Thayer discussed the steps (noted in next section) to be followed in a general OSE process went almost unnoticed. As is shown later in the paper, these steps provide the basic structure for the OSE framework (see Appendix 12.1 for the differences between OSE framework between 1989 and 2004).

In 1993, Livingston applied the theoretical work of Holland and Thayer and wrote two papers (Livingston, 1993a, 1993b) that led to equating practitioners adopting the use of log-linear models to presmooth data. One of the papers (Livingston, 1993a) showed the significant impact on the SEE of the use of a well-fit log-linear model to presmooth the data. The other paper (Livingston, 1993b) showed that the

two methods of continuization of discrete test score distributions – the traditional linear interpolation and the Gaussian kernel, gave very similar results – with the Gaussian kernel method achieving slightly better results (closer to the equating criterion and with lower SEE). At the time Livingston wrote those papers, there was no automatic way to select the degree of smoothing within the Gaussian kernel method, and therefore, a practitioner had to try various smoothing degrees manually and somehow fix on an ad hoc basis the values of the bandwidth that controlled smoothing. That is, at that time, the process was impractical and the choice of the degree of smoothing was arbitrary.

In 2004, von Davier et al. (2004b) published *The Kernel Method of Test Equating*, which to some degree changed the way researchers talk (and maybe think) about equating. This shift did not happen overnight, but over time the change became noticeable in the terms of the framework and type of vocabulary used in research papers on equating (see Appendix 12.1). While some of these concepts were available since 1989, practitioners and researchers adopted them only more recently, after the kernel equating book was published. This paradigm shift and the ongoing research are the focus of this paper (see Appendix 12.1). Appendix 12.1 describes the research agenda around the OSE framework in the past, present, and future.

In the beginning, not even the authors of the book realized the full set of implications of their work. For example, they called the book *The Kernel Method of Test Equating*, while a more appropriate title perhaps should have been *A New Equating Framework and the Application of Gaussian Kernel Continuization* to emphasize the level of generalizability of the proposed framework. This framework easily includes not only the Gaussian kernel, but also other kernels, as well as the traditional linear interpolation method. The framework described in von Davier et al. (2004b) also introduced several new concepts, such as (a) the design function (DF), (b) the standard error of equating differences (SEED), (c) explicit use of the assumptions that underlie common data collection designs, (d) new approaches for employing data collected from these designs, (e) ways to model the impact of the data collection strategy by using DF, (f) new names or acronyms for known designs, such as the equivalent groups (EG) design and the nonequivalent groups with anchor test (NEAT) design, (g) and the fact that the Gaussian kernel is only one possible choice for continuization and only one of many efficient applications within the framework. The most important feature in the framework, although it is implicit, is an approach to equating that makes use of a statistical parametrical model – an approach that relies on a set of assumptions, makes inferences about the data, includes an estimation method for the parameters, and has accuracy and diagnostic measures.

Since the book was published (von Davier et al. 2004b), the research on OSE has been revived, and to a significant extent, it has shifted from direct empirical applications to theoretical development of new equating models, new continuization models, and new work extending the accuracy measures to building tests of linear hypotheses of equating functions (see Appendix 12.1). Many of these recent or current studies rely on OSE framework.

The purpose of the paper is to (a) identify the components of the conceptual framework of OSE methods, (b) discuss how this framework encompasses all OSE methods, (c) emphasize the practical advantages of the modular structure of the framework, (d) indicate how this framework has already proven to be useful for various research purposes outside of Gaussian KE, and (e) discuss potential further developments in the equating field that could be derived from this OSE framework.

12.3 A Brief Description of the Observed-Score Equating Framework

The process of observed-score kernel equating was described in von Davier et al. (2004b) as consisting of the following: (a) presmoothing (using log-linear models), (b) estimation of the score probabilities (using DF), (c) continuization (using the Gaussian kernels), (d) computation of the equating function and new diagnostic measures, and (e) computation of accuracy measures, such as SEE and the newly developed SEED. The equating process described in von Davier et al. was enhanced in several ways since its first description in Holland and Thayer (1989) (see above the four equating steps from Holland & Thayer, 1989, and see Appendix 12.1 for a tabular comparison of the two descriptions of the equating process).[2]

In this paper, OSE framework follows the five steps in the OSE process as described in von Davier et al. (2004b) and also includes an explicit description of the relationship between the observed-score equipercentile and linear equating functions. Next the notation and the OSE framework are introduced.

Explicitly or implicitly, in most OSE methods (in particular for the nonlinear methods), the equating functions depend on the score probabilities for each of the two test distributions to be equated on a target population, called T here. The two tests are denoted here by X and Y, and their score values are denoted by x_j (with $j = 0, \ldots, J$) and y_k (with $k = 0, \ldots, K$), respectively. The vectors of the score probabilities are denoted by $\mathbf{r}$ and $\mathbf{s}$ on T:

$$\boldsymbol{r} = (r_1, ..., r_J), \quad \text{and} \quad \boldsymbol{s} = (s_1, ..., s_K) \tag{12.1}$$

and each r_j and s_k are defined by

$$r_j = P\{X = x_j|T\} \quad \text{and} \quad s_k = P\{Y = y_k|T\}. \tag{12.2}$$

[2] The appendix contains a summary of the three generations of the OSE framework. The first column shows the steps employed within the framework, while additions made in 2004 and in 2009 are included in the next two columns. Examination of the appendix reveals, for example, that Step 2 of the current approach was not added until 2004 (hence the odd numbering in column one of 1, 3, 4, and 5). The table also reveals that another major shift between the 1989 and 2004 was the addition of the SEED to the framework.

The score probabilities for X are associated with the X raw scores, $\{x_j\}$, and those for Y are associated with the Y raw scores, $\{y_k\}$. Based on the equating design employed, the score probabilities $\boldsymbol{r}$ and $\boldsymbol{s}$ are computed through the design function, which can range from the simple identity function to more complex functions for anchor test methods (von Davier et al. 2004b, Chap. 2).

The steps in the OSE framework are covered in more detail in the following subsections.

12.3.1 Presmoothing (Step 1)

The score probabilities are first either estimated through various procedures such as fitting log-linear models to the observed-score test probabilities or by estimating them using the sample frequencies; either way, they are subsequently collected as part of a row vector, $\hat{\mathbf{u}}$. The estimated marginal score probabilities $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ are actually computed (explicitly or not) using DF. A description of log-linear model presmoothing is not given here because (a) it is richly documented in the literature (Holland & Thayer, 1987, 1989, 2000; Moses & Holland, 2008), (b) it is an equating step that is already widely followed and understood by practitioners of equating, and (c) in theory (and consistent with the goals of this paper), it can be achieved using other methods and models that can easily be made to match the results obtained from OSE framework.

12.3.2 Estimating the Score Probabilities (Step 2)

The estimated equating function can be written to express the influence of the data collection design as

$$\hat{e}_y(x) = e_y(x;\ \mathrm{DF}(\hat{\mathbf{u}})). \tag{12.3}$$

Or it can equivalently be written as

$$\hat{e}_y(x) = e_y(x; \hat{\mathbf{r}}, \hat{\mathbf{s}}), \tag{12.4}$$

where $\mathbf{u}$ is a generic notation of the data-vector that reflects the way the data are collected, and $\hat{\mathbf{u}}$ denotes its estimate. For example, if the data are collected from an EG design, then the data are in the form of two univariate distributions; in this case design function is the identity function and $\mathbf{u} = (\mathbf{r}, \mathbf{s})$. If the data are collected following a single group (SG) design, where the same group of test takers takes both test forms X and Y, then $\mathbf{u}$ is the vector whose components are the joint probabilities from the single bivariate distribution. In this case, design function is a linear

function that computes the marginal probabilities **r** and **s** from this bivariate distribution. The design function becomes more complex for the various equating methods for the NEAT design, but the results of its application to vector **u** are always the score probabilities vectors, **r** and **s** on T.

12.3.3 Continuization (Step 3)

In OSE framework for KE (Gaussian or others), the kernel functions are continuous random variables added to the original discrete variable. Consider $X(h_X)$ as a continuous transformation of X such that

$$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_{XT}, \tag{12.5}$$

where

$$a_X^2 = \frac{\sigma_{XT}^2}{\sigma_{XT}^2 + \sigma_V^2 h_X^2} \tag{12.6}$$

and h_X is the bandwidth controlling the degree of smoothness. In (12.5), V is a continuous (kernel) distribution with variance σ_V^2 and mean 0. The mean and the variance of X on T are denoted by $\mu_{\boldsymbol{XT}}$ and σ_{XT}^2, respectively. The role of a_X in (12.5) is to ensure that the first two moments of the transformed random variable $X(h_X)$ are the same as the first two moments of the original discrete variable X. When h_X is large, the distribution of $X(h_X)$ approximates the distribution of V; when h_X is small, $X(h_X)$ approximates X, but as a continuous function. In von Davier et al. (2004b), V followed a standard normal distribution (that is, a Gaussian kernel, with mean 0 and variance 1) and this is why the terms Gaussian KE and KE are sometime used interchangeably. However, Lee and von Davier (2008) discussed the use of alternative kernels for equating, and in their approach, V is a generic continuous distribution. The Y distribution is continuized in a similar way.

One important property of OSE framework that was developed for KE functions (Gaussian or other kernels) is that by manipulating the bandwidths for the new distributions one can obtain a family of equating functions that includes linear equating (when the bandwidths are large) and equipercentile equating (when the bandwidths are small) as special cases. The choice of bandwidth balances the fit and the smoothness of the new continuous function. In von Davier et al. (2004b), the bandwidth is obtained by minimizing a penalty function that has two parts and the user can choose to select the first part only or both parts to minimize. By using the first part of the penalty function, one ensures that the continuous function stays close to the discrete density; by adding the second part of the penalty function, one also penalizes for roughness. If the user chooses to select only the first part of the

penalty function, then the resulting equating function will be close to the traditional equating function obtained through linear interpolation.

The continuized function $X(h_X)$ can be evaluated or diagnosed by comparing its moments to the moments of the discrete score distribution, in this case, of X. The residuals at the score points can be also investigated. However, more research is necessary on the degree of smoothing needed for the density functions.

12.3.4 Computing the Equating Function (Step 4)

Once the discrete distribution functions have been transformed into continuous cumulative distribution functions, then the observed-score equipercentile equating function that equates X to Y is computed as

$$\hat{e}_y(x) = e_y(x; \mathrm{DF}(\hat{\mathbf{u}})) = {G_{Tc}}^{-1}(F_{Tc}(x; \hat{\mathbf{r}}); \hat{\mathbf{s}}), \tag{12.7}$$

where G_{Tc} is the continuized cumulative distribution function of Y on the target population T and F_{Tc} is the continuized cumulative distribution function of X on T. The equating function e_Y in (12.7) can have different formulas (linear or nonlinear, for example). In a NEAT design, it can take the form of chained equating, post-stratification equating, Levine equating, and so on.

The equating function can be evaluated by comparing the moments of the equated scores distribution $\hat{e}_y(x)$ to the moments of the targeted discrete score distribution, in this case, of Y. Other commonly used diagnostic measures involve accuracy measures (see below) and historical information available about the equating results from previous administrations of the assessment.

12.3.5 Computing Accuracy Measures (Step 5)

The SEE and SEED are described next. von Davier, Holland, and Thayer (2004a) applied the theorem known as the *delta method* (Kendall & Stuart, 1977; Rao, 1973) to obtain both the SEE and the SEED. The delta method was applied to the function from (12.7) that depends on the parameter vectors $\mathbf{r}$ and $\mathbf{s}$ on T. According to the delta method, the analytical expression of the asymptotic variance of the equating function is given by

$$\mathrm{Var}(\hat{e}_y(x)) = \mathrm{Var}(e_y(x; \mathrm{DF}(\hat{\mathbf{u}})) \sim \mathbf{J}_{e_y}\mathbf{J}_{\mathrm{DF}}\hat{\Sigma}J^t_{\mathrm{DF}}\mathbf{J}^t_{e_y}, \tag{12.8}$$

where Σ is the estimated asymptotic covariance of the vectors $\mathbf{r}$ and $\mathbf{s}$ after the presmoothing, $\mathbf{J}_{e_y}$ is the Jacobian vector, that is, the vector of the first derivatives of $e_y(x; \mathbf{r}, \mathbf{s})$ with respect to each component of $\mathbf{r}$ and $\mathbf{s}$, and $\mathbf{J}_{\mathrm{DF}}$ is the Jacobian matrix,

that is, the matrix of the first derivatives of the design function with respect to each component of vector **u**.

The asymptotic SEE for $e_y(x)$ is the square root of the asymptotic variance in (12.8), and it depends on three factors that correspond to the data collection and manipulation steps carried out so far: (a) the data collection design through $\mathbf{J}_{\mathrm{DF}}$, (b) presmoothing (using a log-linear model, for example) through estimating the **r** and **s** and their estimated covariance matrix Σ, and (c) the combination of continuization and the mathematical form of the equating function from Step 4 (computing the equating function) in the OSE framework.

Moreover, (12.8) makes obvious the modular character of OSE framework (and therefore, of the software created for it): If one changes the data collection design, the only thing that will change (12.8) will be $\mathbf{J}_{\mathrm{DF}}$. If one changes the equating method (linear or nonlinear, chained versus frequency estimation, etc.), the only piece that will change in (12.8) is $\mathbf{J}_{e_y}$. Finally, if one chooses a different log-linear model, then what will change in (12.8) is Σ.

Hence, the formula of the estimated asymptotic variance of the equating function from (12.8), that is

$$\mathbf{J}_{e_y}\mathbf{J}_{\mathrm{DF}}\hat{\Sigma}J^{t}_{\mathrm{DF}}\mathbf{J}^{t}_{e_y}, \tag{12.9}$$

could be seen simplistically as the formal representation of OSE framework.

In addition to the five steps in the equating process described above that are synthesized in (12.9), the OSE framework also includes an explicit description of the relationship between the observed-score equipercentile and linear equating functions, which is described in the next section.

12.4 The Relation Between Linear and Equipercentile Equating Functions

Following von Davier et al. (2004a, 2004b), all OSE functions linking X to Y on T can be regarded as equipercentile equating functions that have the form shown in (12.7) and (12.10):

$$\mathrm{Equi}_{XYT}(x) = G^{-1}_{Tc}(F_{Tc}(x)), \tag{12.10}$$

where $F_{Tc}(x)$ and $G_{Tc}(y)$ are continuous forms of the cdfs of X and Y on T, and $y = G^{-1}_{T}(p)$ is the inverse function of $p = G_{\boldsymbol{T}}(y)$. Different assumptions about $F_{Tc}(x)$ and $G_{Tc}(y)$ lead to different versions of $\mathrm{Equi}_{XYT}(x)$ and, therefore, to different OSE functions (for example, chained equating, frequency estimation, etc.).

Let μ_{XT}, μ_{YT}, σ_{XT}, and σ_{YT} denote the means and standard deviations of X and Y on T that are computed from $F_{Tc}(x)$ and $G_{\boldsymbol{Tc}}(y)$, as in $\mu_{XT} = \int x dF_{Tc}(x)$, and so on.

In general, any linear equating function is formed from the first two moments of X and Y on T as

$$\text{Lin}_{XYT}(x) = \mu_{YT} + (\sigma_{YT}/\sigma_{XT})(x - \mu_{XT}). \tag{12.11}$$

The linear equating function in (12.11) that uses the first two moments computed from $F_{Tc}(x)$ and $G_{Tc}(y)$ will be said to be compatible with $\text{Equi}_{XYT}(x)$ in (12.10). It is the compatible version of $\text{Lin}_{XYT}(x)$ that appears in Theorem 1. The issue of compatible linear and equipercentile equating functions is covered in more detail in von Davier, Fournier-Zajac, and Holland (2007). Theorem 1 is proved in von Davier et al. (2004b), and it connects the equipercentile function, $\text{Equi}_{XYT}(x)$, in (12.11) to its compatible linear equating function, $\text{Lin}_{XYT}(x)$, in (13.11).

In addition to outlining the relationship between the compatible equipercentile and linear equating functions, Theorem 1 also provides a translation into formal language of the well-known fact that when $F_{Tc}(x)$ and $G_{Tc}(y)$ have the same shape, the equipercentile equating function is identical to the linear equating function.

Theorem 1. *For any population, T, if $F_{Tc}(x)$ and $G_{Tc}(y)$ are continuous cdfs, and F_0 and G_0 are the standardized cdfs that determine the shapes of $F_{Tc}(x)$ and $G_{Tc}(y)$; that is, both F_0 and G_0 have mean 0 and variance 1 and*

$$F_{Tc}(x) = F_0\left(\frac{x - \mu_{XT}}{\sigma_{XT}}\right) \text{ and } G_{Tc}(y) = G_0\left(\frac{y - \mu_{YT}}{\sigma_{YT}}\right), \tag{12.12}$$

then

$$\text{Equi}_{XYT}(x) = G_{Tc}^{-1}(F_{Tc}(x)) = \text{Lin}_{XYT}(x) + R(x), \tag{12.13}$$

where the remainder term, R(x), is equal to

$$\sigma_{YT}\, r\left(\frac{x - \mu_{XT}}{\sigma_{XT}}\right), \tag{12.14}$$

and r(z) is the function

$$r(z) = G_0^{-1}(F_0(z)) - z. \tag{12.15}$$

When $F_{Tc}(x)$ and $G_{Tc}(y)$ have the same shape, it follows that $r(z) = 0$ in (12.15) for all z, so that the remainder in (12.13) satisfies $R(x) = 0$, and thus, $\text{Equi}_{XYT}(x) = \text{Lin}_{XYT}(x)$.

However, it is not always true, for the various methods used in the NEAT design, that the means and standard deviations of X and Y used to compute $\text{Lin}_{XYT}(x)$ are the same as those from $F_{Tc}(x)$ and $G_{Tc}(y)$ that are used in (12.8) to form $\text{Equi}_{XYT}(x)$. The compatibility of a linear and equipercentile equating function depends on both the equating method employed and how the continuization process for obtaining $F_{Tc}(x)$ and $G_{Tc}(y)$ is carried out.

The continuization method for kernel equating post-stratification (KE-PSE) methods ensures that the means and standard deviations of $F_{Tc}(x)$ and $G_{Tc}(y)$ are the same as those of the underlying discrete distributions for any choice of bandwidth. KE-PSE includes the frequency estimation and Tucker methods in OSE framework and uses a kernel to continuize the discrete score distributions (see von Davier et al. 2004b, for details). As mentioned earlier, in KE, $\text{Lin}_{XYT}(x)$ corresponds to large bandwidths, whereas $\text{Equi}_{XYT}(x)$ corresponds to smaller bandwidths that optimize a penalty function (von Davier et al. 2004b). Thus, in KE-PSE, the four moments underlying $\text{Lin}_{XYT}(x)$ are the same as those of the $F_{Tc}(x)$ and $G_{Tc}(y)$ that underlie $\text{Equi}_{XYT}(x)$, and the linear and equipercentile functions are compatible. The compatibility of linear and nonlinear equating functions does not hold for all classes of equating methods. For example, the traditional method of continuization by linear interpolation (Kolen & Brennan, 2004) does not reproduce both the mean and variance of the underlying discrete distribution. The piecewise linear continuous cdf that the linear interpolation method produces is only guaranteed to reproduce the mean of the discrete distribution that underlies it. The variance of the continuized cdf is larger than that of the underlying discrete distribution by 1/12 (Holland & Thayer, 1989). Moreover, the four moments of X and Y on T that are implicitly used by the chained linear or the Tucker linear method are not necessarily the same, nor are they the same as those of the continuized cdfs of frequency estimation or the chained equipercentile methods. The issue of compatibility of the various linear and equipercentile methods used in practice for the NEAT design is a topic worthy of further research.

In conclusion, the OSE framework introduced here includes the five steps of the equating practice formally described in (12.9), and it also incorporates both the linear and nonlinear equating functions together with a description of their relationship.

12.5 Recent and Current Research on Equating Based on the Observed-Score Equating Framework

This section describes several new research directions and their links to OSE framework. Each of these research projects is discussed in this section in the context of a step or feature of OSE framework. Various new studies take advantage of the formal and coherent formulation and the modular characteristics of OSE framework and focus on application of OSE framework to address particular equating issues. Although the KE method is often mentioned, the research studies described here use OSE framework more than the Gaussian KE method itself, and the research that was carried out is innovative in ways that have gone outside the scope of the KE-Software (ETS, 2006, 2007, 2010). Some other studies have focused on replacing various procedures used in the five steps of the equating process. These studies are described next.

12.5.1 Studies that Address Step 1, Presmoothing

The studies mentioned in this section focus on investigating further the use of log-linear models in the presmoothing step in the OSE process. The presmoothing methodology is reflected in OSE framework through the covariance matrix in (12.9). The studies focus either on (a) expanding the methodology described in Holland and Thayer (2000) and von Davier et al. (2004b) or (b) investigating the impact of presmoothing on the equating results. The first group of research studies includes (a) an analysis by Moses and Holland (2008) on extending the KE framework to include the possibility of unsmoothed data; (b) a paper by Moses and von Davier (2006) on extending the procedure to presmoothing trivariate distributions, where the third variable could be, for example, a background variable such as gender; and (c) studies by Holland and Moses (2007) and Chen, Yan, Hemat, Han, and von Davier (2007) on selection algorithms for log-linear models. The studies in the second group focus on the impact of the misfit of the log-linear model on equating results and include papers by Puhan, von Davier, and Gupta (2008) and Mekhael and von Davier (2007).

The work by Cui and Kolen (2007) on investigating the cubic-B-spline method for presmoothing in equating could be considered in this second group, although the authors did not use OSE framework.

12.5.2 Studies that Address Step 2, Estimating the Score Probabilities

The three studies mentioned in this section focus on expanding the OSE framework to new data collection designs. Formally, the data collection design is reflected in the OSE framework through the Jacobian matrix of the design function in (12.9).

Shen and von Davier (2007) and Duong and von Davier (2008) presented a method for equating tests from bimodal data collected from an SG design using OSE framework that was developed for KE in a counterbalanced (CB) design. The practical benefit of this newly developed approach applied within OSE framework is that it explicitly accounts for the bimodal distribution in the equating results, which has not been done before in OSE.

Moses, Deng, and Zhang (2010) extended OSE framework with the post-stratification equating function to include a second anchor. In their study, they expanded upon the work on post-stratification equating in von Davier et al. (2004b, Chap. 2) to include a bivariate anchor distribution and to condition the test distributions on this new bivariate anchor. The practical benefit of this approach is that the differences in ability between the two populations of test takers can be better adjusted for by including the second anchor. The advantage of the OSE framework here is that it allows for the computation of the SEED between the equating functions based on two anchors and those based on one anchor. These accuracy measures in turn can help practitioners decide whether considering an additional anchor would significantly increase the precision of equating.

12.5.3 Studies that Address Step 3, Continuization

Several studies that propose alternative continuization methods are briefly reviewed. Formally, the continuization methodology is reflected in OSE framework through the Jacobian vector of the equating function in (12.9).

The motivation for these recent studies on continuization was to simplify the algorithm of continuization and to improve the resulting equating function in terms of bias and error. The studies discussed here fall into three categories: (a) obtaining the continuized distribution using other kernels, (b) using other methods of continuization, or (c) improving the KE by fine-tuning the Gaussian kernel continuization.

The paper by Lee and von Davier (2008) falls in the first category of studies. The paper uses implicitly the OSE framework and investigates alternative distributions to be used as kernels for continuizing the distribution function. As a potential alternative to the Gaussian kernel, Lee and von Davier (2008) discussed the possibility of using a logistic kernel. One of the advantages of the logistic kernel is that the analytical form of the derivatives required for computing the SEE and the SEED is very simple. This kernel also simplifies the analysis of the behavior of the equating functions (and the two cdfs) in the tails of the distributions. The results of this study, however, do not support the claim that the use of a Gaussian kernel distorts the higher moments of the distribution.

The works of Wang (2011) and Haberman (2008) fall in the second category of studies. They both use an exponential family of functions to approximate and continuize the discrete distribution. This approach does not naturally result in a family of equating functions as KE does, and therefore, it does not include linear equating as one of the options. Wang (2004) continuized the discrete probability distribution by using the polynomial log-linear function (from the presmoothing step) divided by the area under it, to ensure that the distribution is a probability distribution function. The method is called the *continuized log-linear (CLL) method*. The CLL method also assumes that the possible values of the discrete distribution are equally spaced or are consecutive integers.

Haberman (2008) introduces a new way to continuize discrete distribution functions using exponential families of functions. In his study, a distribution from the continuous (univariate or bivariate) exponential family is used for continuization of the discrete test score distributions. The continuous distribution is estimated by constraining several of its consecutive moments to match the equivalent moments of the original discrete test score distribution(s). Once the continuous distribution is obtained DF is used (explicitly or not) to obtain the marginal distribution, and then OSE framework continues from Step 4. In the Wang (2011) and Haberman (2008) studies, presmoothing and continuization are done in one step, and the continuization does not depend on a bandwidth. Wang's and Haberman's methods are conceptually very similar; the numerical approximations and operational implementations are different.

More recently, some studies have been looking into alternative numerical ways for determining the bandwidth in the kernel framework (the third category of studies).

Cid and von Davier (2009) investigated the use of adaptive kernels where the bandwidth is allowed to vary across the score range to account for the differences in sample sizes at different scores. Liang and von Davier (2009) applied the cross-validation technique to estimate the bandwidth that balances the closeness of the continuous distribution to the discrete distribution with the smoothness of the continuous function. Cross-validation is commonly used in density estimation procedures.

12.5.4 Studies that Address Step 4, Equating

Formally, the composition of the equating function with the design function is reflected in OSE framework through the Jacobian vector of the equating function and the Jacobian matrix of DF in (12.9). The research carried out recently focused on the evaluation of the equating results and it is mentioned below. The studies fall under these research directions: (a) equating evaluation and (b) equating criteria.

Equating evaluation. In von Davier et al. (2004b), the percent relative error diagnostic indexes were described for all equating methods but the chained equating methods (both linear and nonlinear). One recent study focused on expanding the index to chained equating (Jiang, von Davier, & Chen, 2011). Another study (Moses, 2008) is investigating the following questions: What are the characteristics of a good equating method? Should the equating function reflect the irregularities in the data or should it be smooth? Should a procedure attempt to balance the two?

Equating criteria. Other research directions under equating encompass the research on establishing an equating criterion in simulation or special studies. Shen and von Davier (2007) discussed the choice and development of an equating criterion for equating functions in the NEAT design. Inspired by the work of Holland, von Davier, Sinharay, and Han (2006) and Holland, Sinharay, von Davier, and Han (2008), Shen and von Davier recommend the creation of a synthetic single group design with the equipercentile method as the equating criterion for equating the two tests. The synthetic single group is constructed to be similar to the target population in the NEAT design, that is, to be a weighted average of the two nonequivalent samples from the NEAT design. In order to use such an equating design, one needs to have data for both tests to be equated in both nonequivalent samples. This is only possible in a special study or with simulated data. Routine operational administrations cannot provide the data structure needed for this purpose.

12.5.5 Studies that Address Step 5, Accuracy Measures

Formally, the accuracy methods (SEE and SEED) are directly reflected in OSE framework via the computation in (12.9) or a slight adaptation of it. Here, studies that focus on extending the use of SEE and SEED are briefly reviewed. These include

(a) studies that extend the application of the SEE and SEED to other equating functions and (b) studies that extend the use of the SEE and SEED to hypotheses testing.

The formulas in OSE framework for the computation of SEE and SEED for KE functions can easily be adapted to derive these indexes for all the OSE functions. Recent research already showed that OSE framework can be used to compute standard errors of the traditional equipercentile equating (Wang, 2004) or the SEED between a KE method and a traditional equating method (Moses, Deng, & Zhang, 2010).

The SEED has been shown to have practical uses: It aids in the decision to be made between equating functions that are (a) linear and nonlinear or (b) based on different assumptions, such as post-stratification and chained equating (see von Davier & Kong, 2005; von Davier et al. 2004b). In addition to these already established uses, SEED can be extended to construct omnibus statistical tests to decide between two equating functions (see Rijmen, Qu, & von Davier, 2008).

12.5.6 Studies that Address the Relation Between Linear and Equipercentile Equating Functions in the Observed-Score Equating Framework

Formally, the relationship between the linear and equipercentile equating functions is only captured in Theorem 1 and is not directly reflected in (12.9). This subsection mentions two studies that propose nonlinear versions of the Levine OSE function that have used the five steps from OSE framework and Theorem 1 in their development. The motivation for looking into other equating methods based on classical test theory (CTT) is due to the search for a theory of equating based on CTT concepts, such as true and observed-scores and reliability, and measurement errors (see also Holland & Hoskens, 2003).

The Levine method has been known to be a linear function without a curvilinear analogue and without a version in KE. Nevertheless, the Levine OSE method is often computed in practical applications for comparison purposes. Under certain circumstances it might be more accurate than the other linear equating methods and, hence, used operationally for score reporting (see Petersen, Marco, & Stewart, 1982). These circumstances refer to the quality of the tests and the anchor and to a situation where a linear equating is appropriate (i.e. the distributions of the two variables X and Y have similar shapes). However, situations do exist where the tests and the anchor are very carefully constructed but the two test distributions differ in shape (see von Davier et al. 2006). In such a case, a nonlinear version of the Levine function would be needed.

Chen and Holland (2008) and von Davier et al. (2007) presented different ways of constructing a nonlinear Levine observed-score and true-score equating methods in OSE framework. The newly developed nonlinear Levine equating functions from von Davier et al. and Chen and Holland (2008) rely on Theorem 1 and OSE framework.

12.6 Future Directions

This paper demonstrates that the method introduced by Holland and Thayer (1989) and von Davier et al. (2004b) represents more than simply a continuization method. This paper shows that von Davier et al. (2004b) introduced a powerful equating framework for all OSE methods that has already proven useful for various research purposes. Moreover, through its modular features, the OSE framework facilitates the manipulation of the software developed to compute the KE. KE Software (ETS, 2006, 2007, 2010) can be easily enhanced by adding routines to replace different parts of (12.9). Acknowledging that the OSE fits under a modular framework can make the operational infrastructure flexible and efficient.

Although the KE method has been around for almost 20 years, it has been slowly adopted in operational practice (for instance, at ETS, it was first adopted by several programs in 2008) despite the theoretical and practical advantages KE offers. The arguments for supporting a more extensive use of OSE framework in operational practice are (a) the accuracy and diagnostic measures available within the framework; (b) the framework's modular system, which translates readily into a modular software package; and (c) the easy-to-use software interface. Moreover, the OSE framework has the potential to introduce automatic procedures (with incorporated decision steps) and therefore can reduce some of the present routine equating workload for psychometricians and data analysts.

Research focused on decision aids and automatic equating procedures is needed to simplify the equating process and to make it more efficient. Developing or refining indexes and tests – such as (a) testing linear hypotheses about equating differences on specific intervals, (b) using SEED for aiding in the process of comparing equating functions, and (c) developing indexes for deciding among log-linear models in presmoothing, as well as (d) making attempts to develop procedures to improve the fit of log-linear models (and therefore improve the stability of equating results) for score ranges that matter to a particular program – are of vital importance in the practice of equating. In addition, researchers should focus on (e) expanding the research on OSE framework to include the scaling process, (f) monitoring scale drift, and (g) studying ways to equate tests that have bimodal distributions. For example, the scaling process (that is, the process by which the equated raw-scores are placed onto a reporting scale) can be formalized as part of the OSE framework. The scores on the reporting scale are, mathematically speaking, obtained from the composition of the scaling function and equating function. As such, one could eventually formalize this step by including the Jacobian of the scaling function with respect to the equated scores into (12.9). However, the discreteness of the scaling function and the routine of rounding scores for reporting purposes would definitely be a challenge in formalizing this step.

In this paper, the OSE framework was laid out, past research was linked to steps in the framework, and areas where additional research might be beneficial for practical use were mentioned.

Acknowledgments My thanks go to Dan Eignor and Skip Livingston for their valuable feedback and suggestions on previous versions of the manuscript. I also thank Kim Fryer for her help with the editorial work. Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Chen, H., & Holland, P. W. (2008, March). *True score equating under the KE framework, the associated log-linear model and its relation with Levine equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Chen, H., Yan, D., Hemat, L., Han, N., & von Davier, A. A. (2007). LOGLIN/KE user guide (Version 3.0) [Computer software manual]. Princeton, NJ: ETS.

Cid, J., & von Davier, A. A. (2009, April). *Examining potential boundary bias effects in kernel smoothing on equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Conceptual framework. (2010). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Conceptual_framework.

Cui, Z., & Kolen, M. (2007, April). *An introduction of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306.

Duong, M., & von Davier, A. A. (2008, March). *Kernel equating with observed mixture distributions in a single-group design*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

ETS. (2006). *KE-software (Version 1) [Computer software]*. Princeton, NJ: Author.

ETS. (2007). *KE-software (Version 2) [Computer software]*. Princeton, NJ: Author.

ETS. (2010). *KE-software (Version 3) [Computer software]*. Princeton, NJ: Author.

Haberman, S. J. (2008). *Continuous exponential families: An equating tool* (ETS Research Rep. No. RR-08-05). Princeton, NJ: ETS.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*, 123–149.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (ETS Research Rep. No. RR-89-06). Princeton NJ: ETS.

Holland, P. W., & Moses, T. P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (ETS Research Rep. No. RR-07-15). Princeton, NJ: ETS.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*, 17–43.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton NJ: ETS.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Research Rep. No. RR-89-07). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate log-linear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133–183.

Holland, P.W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design* (ETS Research Rep. No. RR-06-17). Princeton, NJ: ETS.

Jiang, Y., von Davier, A. A., & Chen, H. (2011). *Evaluating equating results: Percent relative error for chained kernel equating*. Manuscript submitted for publication.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). New York, NY: Macmillan.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling and linking* (2nd ed.). New York, NY: Springer.

Lee, Y.-H., & von Davier, A. A. (2008). *Comparing alternative kernels for the kernel method of test equating: Gaussian, logistic and uniform kernels* (ETS Research Rep. No. RR-08-12). Princeton NJ: ETS.

Liang, T., & von Davier, A. A. (2009, July). *Alternative methods to determine the optimal bandwidth for the kernel equating function*. Paper presented at the international meeting of the Psychometric Society, Cambridge, UK.

Livingston, S. (1993a). Small-sample equatings with log-linear smoothing. *Journal of Educational Measurement, 30*, 23–39.

Livingston, S. (1993b). *An empirical tryout of kernel equating* (ETS Research Rep. No. RR-93-33). Princeton NJ: ETS.

Mekhael, M., & von Davier, A. A. (2007, April). *The effects of log-linear models on kernel equating results*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Moses, T. (2008). *An evaluation of statistical strategies for making equating decisions* (ETS Research Rep. No. RR-08-60). Princeton NJ: ETS.

Moses, T., Deng, W., & Zhang, Y. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (ETS Research Report No. RR-10-23). Princeton, NJ: ETS.

Moses, T. P., & Holland, P. W. (2008). *Notes on the general framework for observed score equating* (ETS Research Rep. No. RR-08-59). Princeton NJ: ETS.

Moses, T. P., & von Davier, A. A. (2006). *A SAS macro for log-linear smoothing: Applications and implications* (ETS Research Rep. No. RR-06-05). Princeton, NJ: ETS.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York, NY: Academic Press.

Puhan, G., von Davier, A. A., & Gupta, S. (2008). *Impossible scores resulting in zero frequencies in the anchor test: Impact on smoothing and equating* (ETS Research Rep. No. RR-08-10). Princeton, NJ: ETS.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

Rijmen, F., Qu, Y., & von Davier, A. A. (2008, March). *Hypothesis testing of equating differences in the KE framework*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Shen, X., & von Davier, A. A. (2007, April). *An exploration of constructing criteria equating for simulation studies comparing kernel and IRT equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 1–17). New York, NY: Springer.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating* (ETS Research Rep. No. RR-07-14). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data* (ETS Research Rep. No. RR-06-02). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and poststratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*(1), 15–32.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. New York, NY: Springer.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the nonequivalent groups design. *Journal of Educational and Behavioral Statistics, 30*(3), 313–342.

Wang, T. (2011). *An alternative continuization method: the continuized log-linear method*. In A. A. von Davier (Ed.), Statistical models for Test Equating, Scaling, and Linking (pp. 141–158). New York, NY: Springer.

Part VI
Holland: From Mentor to Colleague

Great Colleagues Make a Great Institution

Paul W. Holland

When Don Rubin and I decided to hire Henry Braun into the research statistics group, we had no inkling of just how far Henry would go at Educational Testing Service (ETS). He became my boss as the director of the division and then my boss's boss when he rose to the vice president for research. I liked having a statistician up there in the administrative stratosphere of the organization. Moreover, he was somehow able to continue to do research while in these roles and has made many contributions to educational research with a current emphasis on educational policy at the national level.

In the late 1980s, Neil Dorans and I were both interested in developing good methods for detecting test questions that exhibited differential item functioning (DIF). Neil came at it as part of his work on the SAT®, while I had begun to think about providing DIF methods that had good statistical properties but were easy to compute. I was unhappy with what the field of psychometrics had come up with regarding measuring DIF in those days, and Neil wanted something that could be used with the huge data sets that arose with the SAT. We came up with different but closely related approaches that are now widely applied. Over the years Neil and I have had many opportunities to collaborate. One of the most satisfying for me was our chapter on linking and equating in the fourth edition of *Educational Measurement* edited by Robert Brennan. In every respect, Neil has been an outstanding example of the great ETS research combination of scholarship and practical work.

Chapter 13
An Exploratory Analysis of Charter Schools

Henry I. Braun, Christina Tang, and Kathleen M. Sheehan

13.1 Introduction and Overview

Charter schools are publicly supported schools to which parents can opt to send their children. The goal of the charter school movement is to offer parents a choice of schools within the public school sector. In general, charter schools are freed from many of the regulations under which traditional public schools operate – with the hope that the increased flexibility will result in superior student achievement.

The first charter schools opened their doors in 1991, and since then thousands more have been started. The rules governing charter schools vary from state to state, but in all states charter schools are established under the auspices of authorizers approved by the state. The authorizers can be public or private, including universities, school districts, for-profit entities, individuals, and so on.

Charter schools have attracted a great deal of attention, and their relative effectiveness in comparison to public noncharter schools has been the object of much study (see Carnoy, Jacobsen, Mishel, & Rothstein, 2005). Within-state comparisons have produced mixed findings. In part, these findings can be attributed to the difficulty in conducting research in this area. The technical difficulties include data insufficiency and the perennial problem of making credible causal inferences from observational or quasi-experimental studies. These and other difficulties are addressed in the report of the Charter School Achievement Consensus Panel (2006).

Responding to the interest in charter schools, the National Assessment Governing Board asked the National Center for Education Statistics (NCES) to oversample charter schools for the grade 4 assessment conducted in 2003. Students from approximately 150 charter schools were included in the assessment. Analysis of the resulting data for both reading and mathematics (National Center for Education

H.I. Braun (✉)
Lynch School of Education, Boston College, 140 Commonwealth Avenue,
Chestnut Hill, MA 02467, USA
e-mail: braunh@bc.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_13,

Statistics, 2005) revealed that, on average, students enrolled in charter schools did not perform as well as students enrolled in public noncharter schools. Furthermore, a comparison of grade 4 reading scores between charter schools associated with a local education authority (C/LEA) and charter schools not associated with an LEA (C/nLEA) showed that students enrolled in the former performed about 10 points better, on average, than those enrolled in the latter.[1] In mathematics, the difference was about 11 points, again in favor of charter schools associated with an LEA.

Both results were somewhat surprising to proponents of charter schools. A plausible explanation for the first result was that students enrolled in charter schools were, on the whole, more disadvantaged than those enrolled in public noncharter schools and that the NCES analyses had not properly accounted for these differences. However, a subsequent reanalysis of this data using hierarchical linear models (Braun, Jenkins, & Grigg, 2006) did not materially change the findings described above. That is, in both reading and mathematics, after adjusting simultaneously for differences in all measured student characteristics, the average of charter school means was about 4 points lower than the average public noncharter school means. Both differences were statistically significant.

Following the example of the NCES report, Braun et al. (2006) also compared the results for the two types of charter schools. When school means for grade 4 reading were adjusted for measured differences in their student populations, the average school mean for C/LEA schools was slightly more than 4 points greater than the average school mean for C/nLEA schools. In mathematics, the gap in (adjusted) means was slightly less than 4 points. The differences were not statistically significant.

Braun et al. (2006) also conducted a multilevel analysis of data drawn only from charter schools. The intent was to examine the relationships between various characteristics of charter schools and the achievement of the students enrolled in those schools. In reading, a combination of student and school characteristics accounted for about 82% of the between school variance. Only three charter school specific variables were in the model (including charter school type). In mathematics, student and school characteristics accounted for about 69% of the between school variance. In this case, several charter school specific variables were in the model (including charter school type).[2]

As pointed out by Braun et al. (2006), estimating school effectiveness from cross-sectional data is problematic, even with the possibility of incorporating individual student characteristics into the model. In point of fact, it is difficult to compensate adequately for the absence of prior measures of achievement.

[1]Local education authorities (LEAs) are usually school districts that are established by one or more political entities, such as towns or cities. In the presentation that follows, the former set of schools will be denoted by C/LEA and the latter set by C/nLEA. In Braun, Jenkins, and Grigg (2006), they are referred to as PSD-affiliated and non-PSD-affiliated, respectively.

[2]Charter school type was retained in both models because of policy maker interest. The corresponding regression coefficients were not significant in either case.

Nonetheless, taking these findings at face value, the straightforward interpretation is that the distinction between C/LEA and C/nLEA schools was not particularly useful in explaining the variation in average scores among charter schools. However, this distinction remained of interest, at least for some stakeholders. In particular, the expectation was that C/nLEA schools would typically have more flexibility and, consequently, yield better results. As reported above, this expectation was not realized.

Of course, one possible explanation was collinearity between the charter school type indicator and the various measured school characteristics. This speculation then gave rise to the question as to whether the collection of school characteristics could be employed to reliably distinguish between the two types of charter schools (C/LEA and C/nLEA); in particular, which charter school specific characteristics would prove to be useful predictors. A related question concerned the relationships between charter school characteristics and (adjusted) school means.

The purpose of this note is to describe the results of a set of exploratory analyses intended to address these questions. With respect to the these two questions, the analyses drew both on general school data and on the information derived from a special charter school questionnaire that was completed by school personnel at the time of the National Assessment of Educational Progress (NAEP) administration. The information comprised responses to six sets of characteristics of charter schools:

- Monitoring
- Exemptions (Waivers)
- Constitutes (Progress reporting)
- Strong law state
- Charter school size
- Teacher certification

Each set was constructed from the responses to a number of related queries. For example, with respect to *monitoring*, school personnel were queried about whether they were monitored by the school's authorizer in one or more of the following areas: instructional practices, student achievement, student behavior, student attendance, school governance, school finances, and compliance with state or federal relations. Responses were *yes*, *no*, or *don't know*. *Waivers* concerns the number and type of state or district policies for which the school was granted a waiver or exemption. *Progress reporting* is a count of the number of stakeholders to which the school had to make a report. *Strong law state* categorized the home state as having or not having a strong charter school law. *School size* classified school enrollment as low or high. Finally, *teacher certification* indicated the percentage of certified teachers in the school. For further details consult Appendix A of Braun et al. (2006, pp. 52–53).

The analysis was carried out in three phases:

1. Nonlinear models were fit to the full set of 148 charter schools. The best fitting model is only moderately successful in distinguishing between C/LEA and C/nLEA schools.

2. In view of the possibility that charter schools with higher adjusted score means may exhibit different patterns than charter schools with lower adjusted score means, the full set was divided into two groups: Those schools with adjusted reading means > 211 and those schools with adjusted reading means < 211. (Note: 211 is approximately the average of all charter school reading means.) Models were fit separately to each group. The outcomes were similar to those in Phase 1. A parallel analysis was conducted using a split based on adjusted mathematics means, with comparable results.
3. With the results of Phases 1 and 2 in hand, it was decided that it would be useful to determine if the average difference in means between school types was, in some sense, unusually large. Accordingly, the full charter school set was successively divided into two groups, based on all possible splits of four of the available school characteristics.[3] For both reading and mathematics, the absolute value of the difference in average school means resulting from each split was recorded and the empirical distribution of these absolute differences was compiled. When this empirical distribution is used as a reference distribution, it appears that in either case the average difference in school means between C/LEA and C/nLEA schools is not particularly unusual.

13.1.1 Analyses: Phase 1

In view of the nature of the data, traditional linear discriminant models would likely not be satisfactory. Accordingly, a nonlinear (or tree regression) approach, which is more flexible, was followed (Breiman, Friedman, Olshen, & Stone, 1988). The outcome variable is the indicator for whether the school is C/LEA or C/nLEA. The predictors are the various school characteristics. The fitting algorithm successively searches among the predictors for the best cut-point or split, where best is determined by a statistic related to the accuracy in discriminating between the groups. Once the best split has been found, the search is repeated for each of the two subsamples that have been created. The process continues until an appropriate stopping point has been reached.

To begin, a number of analyses were conducted to determine which of the measured charter school specific characteristics were potentially useful predictors and, for those characteristics, whether a single cut along the scale could be employed. For example, the variable monitor appeared to be a good predictor. Schools were assigned a score of 0–6, depending on the number of areas monitored.[4] Based on the

[3]The other two characteristics are dichotomous and so do not accommodate multiple splits.

[4]One area of monitoring was excluded due to missing data.

preliminary analyses, we decided to dichotomize the variable, with schools monitored in three or fewer areas in one category (labeled L) and schools monitored in four or more areas in another category (labeled H). Exemptions were dichotomized as none (labeled N) or one or more (labeled 1M). With respect to constitutes, schools were assigned a score of 0–6, depending on the number of groups to which they had to report.[5] The other two variables, strong law and school size, were already dichotomous. These reductions and transformations enabled us to produce simpler and more interpretable regression models.

Drawing on the results of Braun et al. (2006), a number of general school characteristics were added to the database. In particular, for the variable teacher certification, schools were dichotomized as either none (N) or some or all (AS) of the school's teachers were certified.

When the amount of data is not large compared to the number of predictors, a common problem in regression is overfitting. To guard against this, we divided the full sample into five mutually exclusive subsamples of about 30 schools each. These subsamples were denoted by the letters A through E, respectively. Using a fixed predictor pool, a regression model was successively fit to the data, leaving out each subsample in turn. This resulted in five fitted models.[6] Each model was cross-validated on the corresponding excluded subsample, then the results were combined over the five cross-validations to produce an overall cross-validated estimate of the predictive accuracy of the model.

In nonlinear regression with a dichotomous outcome, the splits were compared on the basis of their deviance. In this case, each split yielded two nodes, and the deviance was defined as the sum of the squared differences between the observed and the predicted values summed over the nodes. The observed value was 1 (C/LEA) or 0 (C/nLEA), and the predicted value was the proportion of schools in the node that were C/LEA. One measure of the utility of the fitted tree was R^2, which is defined as the squared correlation between the observed values and the predicted values.

The results for this phase are presented in Appendix 1. Figure 13.1 displays the regression tree for 119 schools (leaving out subsample E). The first split is on the monitor variable. The monitor = L branch (with 24 schools) is not further subdivided. The monitor = H branch (with 95 schools) is subdivided by exemptions. The exemptions = N branch (with 39 schools) is not further subdivided. The exemptions = 1M (with 56 schools) is further subdivided by the teacher certification variable. There are no further splits. Indeed, it is noteworthy that only three predictors are included in the final tree, which has only four nodes.

The vertical placement of a node indicates the deviance associated with that node, and the horizontal placement indicates the empirical probability of schools in that node being C/LEA. For purposes of prediction, schools in nodes with that probability greater than 0.5 are classified as C/LEA and schools in nodes with

[5]Two groups were excluded due to missing data.

[6]Fortunately, each of the five models incorporated the same subset of predictors.

that probability less than 0.5 are classified as C/nLEA. In this case, schools in the nodes with monitor = L or with monitor = H, exemptions = 1M, and teacher certified = AS are classified as C/LEA since these nodes have empirical probabilities greater than 0.5. The remaining schools are classified as C/nLEA. Note that all four nodes are located well away from the threshold of 0.5. The results of the analyses are found in Appendix 1.

For each of the five estimation samples, Table 13.1 displays the number of schools in each node, and Table 13.2 displays the proportions of schools in each node that are in C/LEA. The results are very consistent across the five samples. Table 13.3 then presents the results for the cross-validated predictions. Predictions of C/nLEA were correct nearly 74% of the time, while predictions of C/LEA were correct 67% of the time. (This difference most likely reflects the positions of the nodes relative to 0.5.) The overall prediction accuracy was approximately 0.7. Since chance agreement is 0.5, this yielded a kappa value of 0.40 indicating fair to moderate accuracy.

Although NAEP scores are not involved in the fitting algorithm, it is of interest to note the distribution of adjusted school means for the two subsets of schools (Fig. 13.2). It is evident that there is both considerable heterogeneity within each subset and substantial overlap between them. Table 13.4 shows the averages and standard deviations for each subset. These findings are then compared with the predicted classifications based on the final model. The distributions of adjusted means for the two classifications are presented in Fig. 13.3 and the summary statistics in Table 13.5. Thus, the average difference in school means between the school types is slightly larger than the average difference between the two pairs of nodes.

13.1.2 Analyses: Phase 2

In this phase, schools were first categorized by the magnitude of their adjusted reading mean. Because of the reduced sample size, only one model was fit to each half-sample. The results of these analyses are found in Appendix 2.

The regression model for higher performing schools ($N = 84$) is displayed in Fig. 13.4. The tree structure is similar to that found for the full data set. Table 13.6 shows the number of schools in each node, and Table 13.7 shows the empirical probabilities of a school being in C/LEA associated with each node. Table 13.8 presents the predictions of the model cross-classified with the actual designations. The kappa = 0.40 as before, though this is *not* based on cross-validation.

Figure 13.5 displays the distributions of adjusted school means for the C/LEA and C/nLEA subsets. Again, we see considerable heterogeneity within each subset and substantial overlap between subsets. Table 13.9 presents the averages and standard deviations for each subset. Note that the means are nearly identical, although the distributions have rather different shapes. These findings are then compared with the predicted classifications based on the final model. The

distributions of adjusted means for the two classifications are presented in Fig. 13.6, and the summary statistics are presented in Table 13.10.

The regression model for lower performing schools ($N = 64$) is displayed in Fig. 13.7. The tree structure is different from that found for the full data set in that there are only three nodes rather than four nodes. Table 13.11 shows the number of schools in each node, and Table 13.12 shows the empirical probabilities of a school being in C/LEA associated with each node. Table 13.13 presents the predictions of the model cross-classified with the actual designations. The kappa = 0.40 as before, though this is *not* based on cross-validation.

Figure 13.8 displays the distributions of adjusted school means for the C/LEA and C/nLEA subsets. Again, we see considerable heterogeneity within each subset and substantial overlap between subsets. Table 13.14 presents the averages and standard deviations for each subset. Note that in this case the average for C/nLEA is about 0.5 points greater than the average for C/LEA. These findings are then compared with the predicted classifications based on the final model. The distributions of adjusted means and the summary statistics for the two classifications are displayed in Fig. 13.9, and are presented in Table 13.15.

13.1.3 Analyses: Phase 3

In this phase, charter schools were successively divided into two groups based on a split at different levels on each of four school characteristics. The results of these analyses are found in Appendix 3.

Table 13.16 displays the splits and, for each split, the corresponding absolute difference in the averages of the school means for the two groups formed by the split. Figure 13.10 presents a smoothed version of the empirical distribution of these absolute differences. (Note that this distribution is not a standard sampling distribution because of the dependencies generated by repeatedly splitting the same data set.) The difference of 3.5 in adjusted means between C/LEA and C/nLEA (cf., Table 13.4), which is highlighted in the figure, does fall near the tail of the reference distribution.

Phases 2 and 3 were repeated using data from the mathematics assessment. The results were quite comparable to those based on the reading assessment. In particular, the difference between C/LEA and C/nLEA did fall near the tail of the reference distribution.

13.2 Conclusions

We conclude that measured school characteristics are of limited utility in distinguishing between C/LEA and C/nLEA schools, with the likely reason that the variation across authorizers within states, as well as differences between states, undermine any attempt to make general statements about charter school types.

Looking for clearer patterns when the schools are divided into those with higher and lower adjusted school means is no more successful. Finally, the observed difference based on the C/LEA–C/nLEA classification is not found to be extreme when compared to differences based on other characterizations. Although charter school affiliation may be of substantive interest, little statistical support exists for further investigation using the available data.

Acknowledgments The work reported here was supported in part by the Research and Development Division of the Educational Testing Service. Any opinions expressed here are those of the authors and not necessarily of Educational Testing Service.

Appendix 1

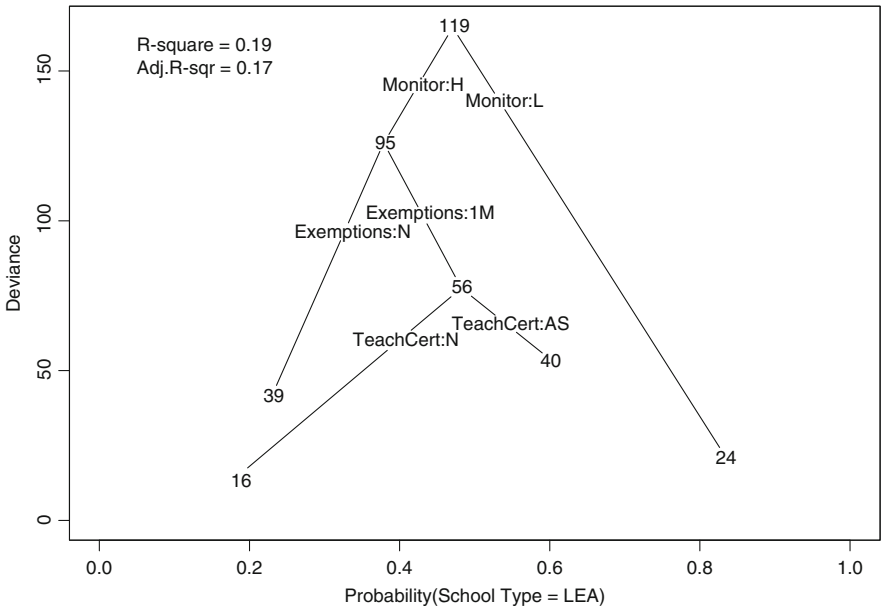

Fig. 13.1 Tree model estimated from all schools *except* the 29 schools assigned to sample E. *Note*. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate; LEA = local education authority

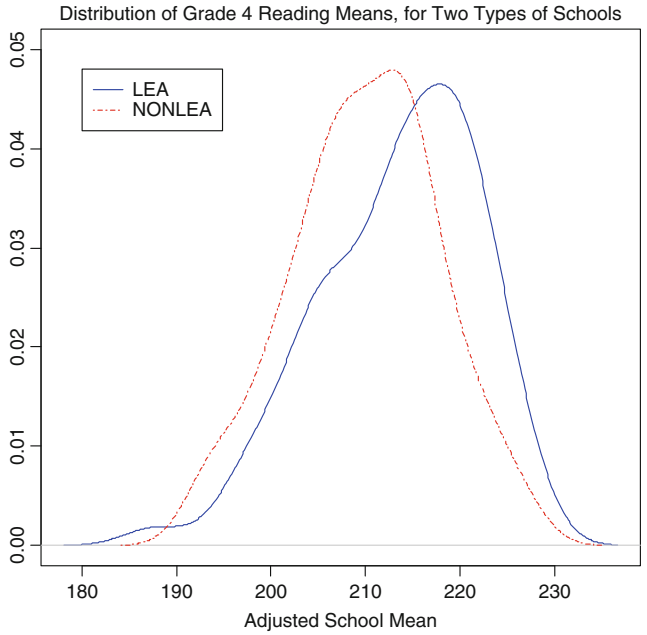

Fig. 13.2 Distribution of adjusted school means for the 70 local education authority (LEA) schools and the 78 non-LEA schools

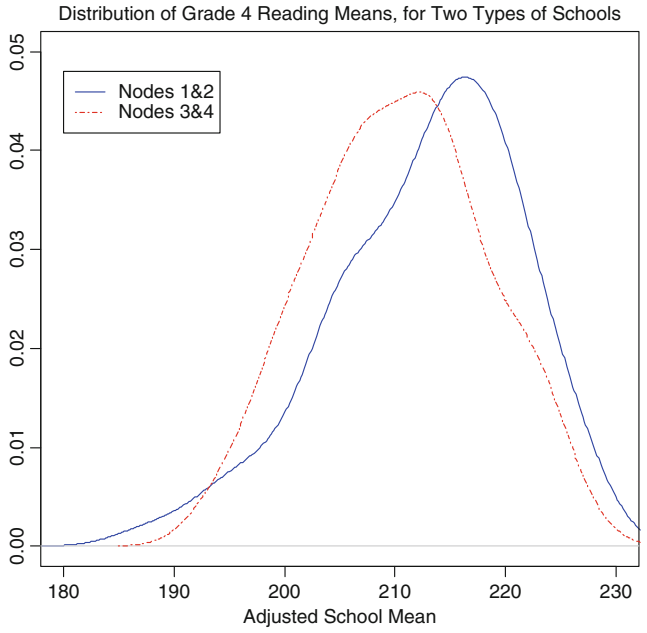

Fig. 13.3 Distribution of adjusted school means for the 76 schools classified at nodes 1 and 2 and the 72 schools classified at nodes 3 and 4

Table 13.1 Numbers of schools assigned to each cross-validation sample by node

	Estimation sample				
Node	~A	~B	~C	~D	~E
Monitor = L (Node 1)	23	28	26	23	24
Monitor = H & Exemptions = 1M & Cert = AS (Node 2)	38	32	34	36	40
Monitor = H & Exemptions = 1M & Cert = N (Node 3)	14	18	15	17	16
Monitor = H & Exemptions = 0 (Node 4)	43	40	43	43	39
Total	118	118	118	119	119

Note. Sample ~A includes all schools *except* the 30 schools assigned to sample A, etc. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.2 Empirical values for probability that school is associated with LEA, by node and sample

	f estimation sample				
Node	~A	~B	~C	~D	~E
Monitor = L (Node 1)	0.65	0.68	0.69	0.70	0.83
Monitor = H & Exemptions = 1M & Cert = A, S (Node 2)	0.66	0.66	0.71	0.61	0.60
Monitor = H & Exemptions = 1M & Cert = N (Node 3)	0.36	0.33	0.33	0.29	0.19
Monitor = H & Exemptions = 0 (Node 4)	0.30	0.22	0.28	0.21	0.23

Note. Tabled value is the probability that school = LEA|node, sample. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.3 Agreement between observed and predicted classifications

		Predicted school type		
		LEA	Non-LEA	Total
True type	LEA	51	19	70
	Non-LEA	25	53	78
Total		76	72	148

Note. Observed agreement = 0.70; kappa = 0.40; LEA = local education authority

Table 13.4 Adjusted school means for two types of schools: local educational authority (LEA) schools and non-LEA schools

Type	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
LEA	70	70	100	213.44	8.34
Non-LEA	78	0	0	209.96	7.73
Total	148	70	47	211.61	8.19

Table 13.5 Adjusted school means for two types of schools: schools classified at nodes 1 and 2 and schools classified at nodes 3 and 4

Node	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
1 & 2	76	51	67	212.78	8.49
3 & 4	72	19	26	210.37	7.72
Total	148	70	47	211.61	8.19

Note. *LEA* local education authority

Appendix 2

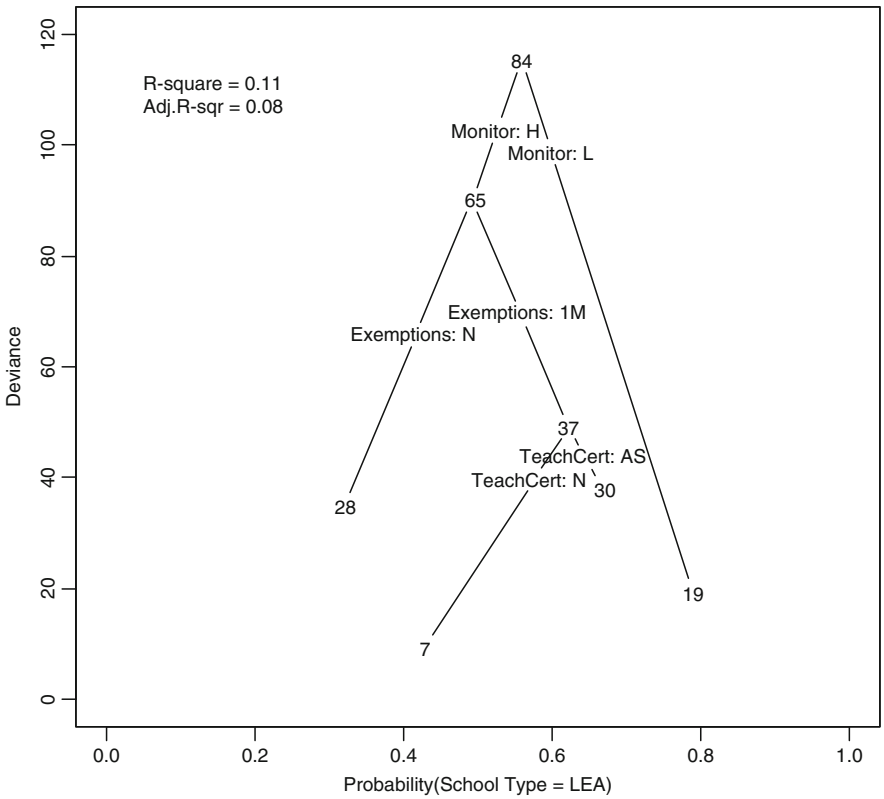

Fig. 13.4 Tree model estimated from all schools with adjusted mean > 211. *Note*. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate; LEA = local education authority

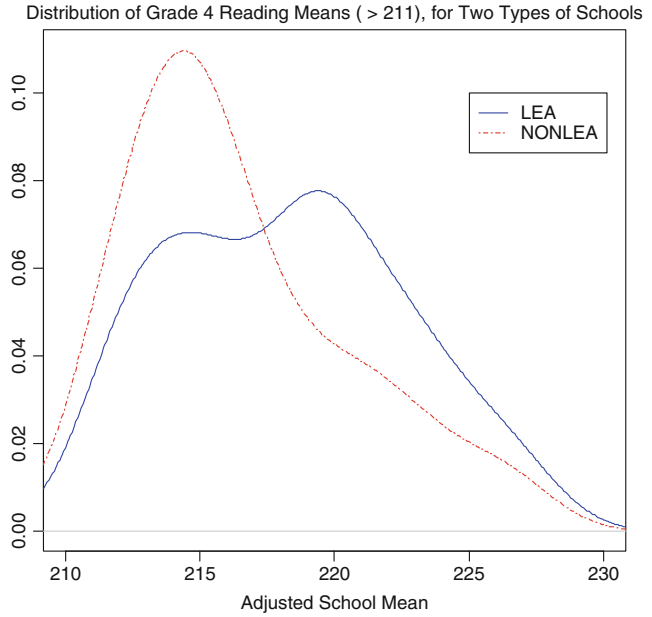

Fig. 13.5 Distribution of adjusted school means for the 47 local education authority (LEA) schools and the 37 non-LEA schools

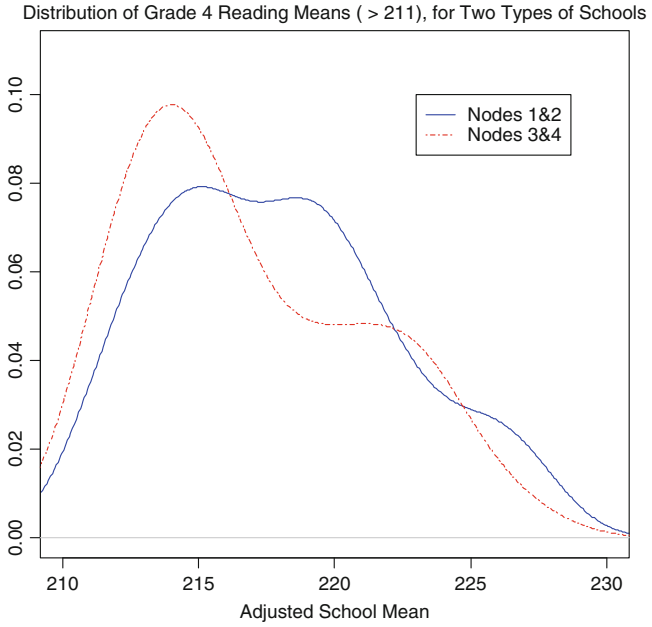

Fig. 13.6 Distribution of adjusted school means for the 49 schools classified at nodes 1 and 2 and the 35 schools classified at nodes 3 and 4

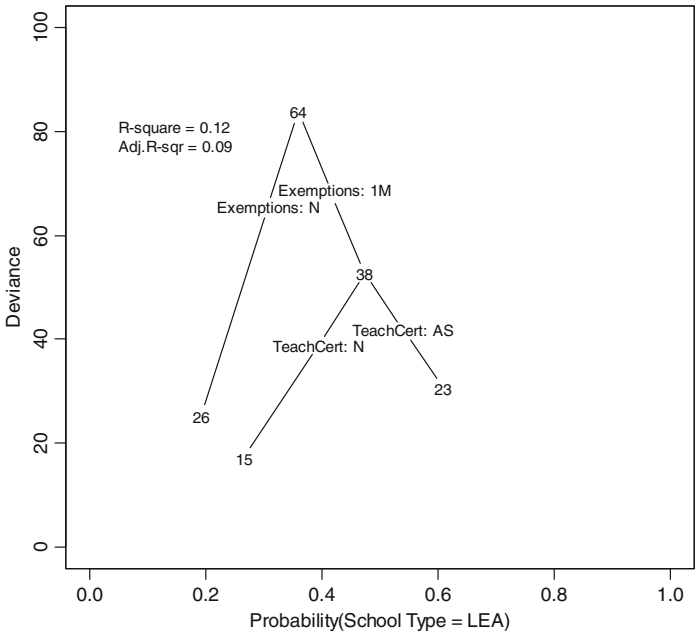

Fig. 13.7 Tree model estimated from all schools with adjusted mean < 211. *Note*. Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate; LEA = local education authority

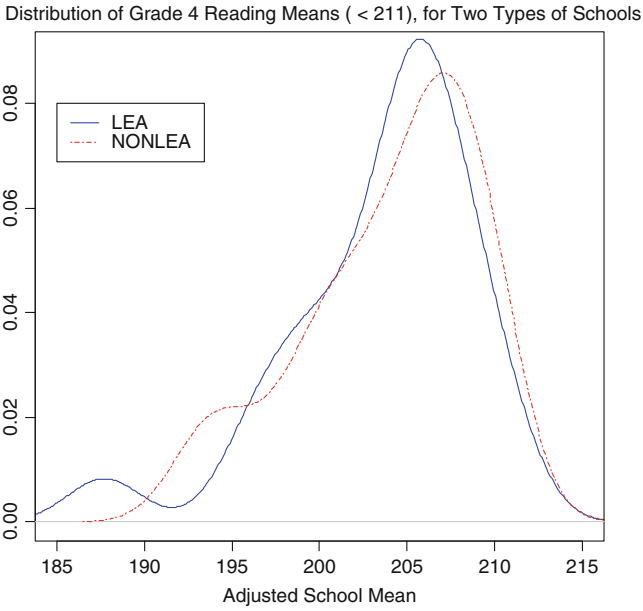

Fig. 13.8 Distribution of adjusted school means for the 23 local education authority (LEA) schools and the 41 non-LEA schools

Table 13.6 Numbers of schools assigned by node

Node	Number of schools
Monitor = L (Node 1)	19
Monitor = H & Exemptions = 1M & TeachCert = AS (Node 2)	30
Monitor = H & Exemptions = 1M & TeachCert = N (Node 3)	7
Monitor = H & Exemptions = 0 (Node 4)	28
Total	84

Note. Mean > 211. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.7 Empirical values for probability by node

Node	Probability
Monitor = L (Node 1)	0.79
Monitor = H & Exemptions = 1M & TeachCert = AS (Node 2)	0.67
Monitor = H & Exemptions = 1M & TeachCert = N (Node 3)	0.43
Monitor = H & Exemptions = 0 (Node 4)	0.32

Note. School = LEA|Node, Sample. Monitor: L = 0–3; Monitor: H = 4–6; Exemptions: N = 0; Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.8 Agreement between observed and predicted classifications

		Predicted school type		
		LEA	Non-LEA	Total
True type	LEA	35	12	47
	Non-LEA	14	23	37
Total		49	35	84

Note. Observed agreement = 0.70; kappa = 0.40; *LEA* local education authority

Table 13.9 Adjusted school means for two types of schools: local education authority (LEA) schools and non-LEA schools

Type	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
LEA	47	47	100	218.27	4.33
Non-LEA	37	0	0	216.50	4.10
Total	84	47	56	217.49	4.30

Table 13.10 Adjusted school means for two types of schools: schools classified at nodes 1 and 2 and schools classified at nodes 3 and 4

Node	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
1 & 2	49	35	71	217.94	4.31
3 & 4	35	12	34	216.87	4.27
Total	84	47	56	217.49	4.30

Note. *LEA* local education authority

Table 13.11 Numbers of schools assigned by node

Node	Number of schools
Exemptions = 1M & TeachCert = AS (Node 1)	23
Exemptions = 1M & TeachCert = N (Node 2)	15
Exemptions = 0 (Node 3)	26
Total	64

Note. Mean < 211. Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.12 Empirical values for probability by node

Node	Probability
Exemptions = 1M & TeachCert = AS (Node 1)	0.61
Exemptions = 1M & TeachCert = N (Node 2)	0.27
Exemptions = 0 (Node 3)	0.19

Note. School = LEA|Node, Sample. Exemptions: 1M = 1 or more; TeachCert: N = no teacher certificate; TeachCert: AS = some or all teacher certificate

Table 13.13 Agreement between observed and predicted classifications

		Predicted school type		
		LEA	Non-LEA	Total
True type	LEA	14	9	23
	Non-LEA	9	32	41
Total		23	41	64

Note. Observed agreement = 0.70; kappa = 0.40; LEA = local education authority

Table 13.14 Adjusted school means for two types of schools: local education authority (LEA) schools and non-LEA schools

Type	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
LEA	23	23	100	203.58	5.25
Non-LEA	41	0	0	204.05	4.95
Total	64	23	35.9	203.88	5.02

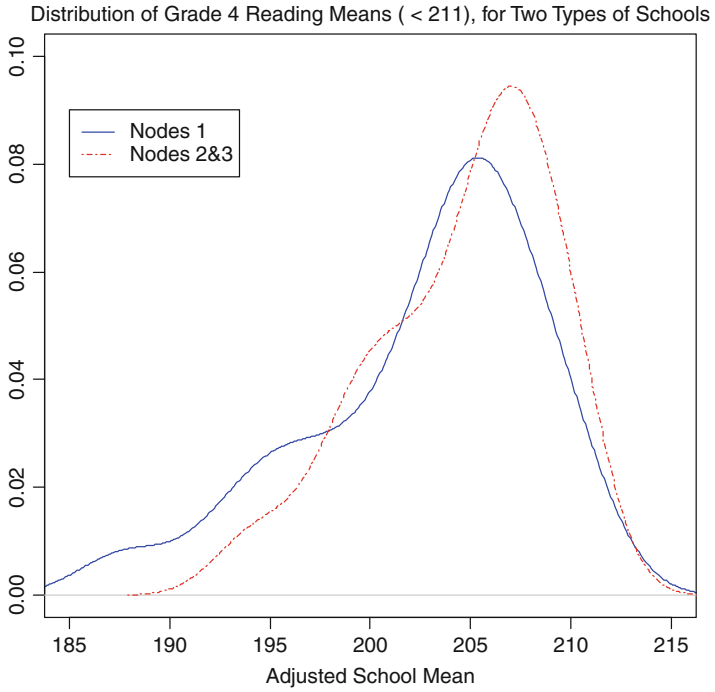

Fig. 13.9 Distribution of adjusted school means for the 23 schools classified at node 1 and the 41 schools classified at nodes 2 and 3

Table 13.15 Adjusted school means for two types of schools: schools classified at node 1 and schools classified at nodes 2 and 3

Node	Total schools	Observed number LEA	Percent LEA (%)	Adjusted mean	SD
1	23	14	61	202.67	5.86
2 & 3	41	9	22	204.56	4.42
	64	23	35.9	203.88	5.02

Note. *LEA* local education authority

Appendix 3

Table 13.16 The absolute differences based on all possible splits on four school characteristics

Group 1	Group 2	Absolute value of difference in adjusted school means across groups
Monitor 0	Monitor 1-6	4.4225
Monitor 0-1	Monitor 2-6	3.8038
Monitor 0-2	Monitor 3-6	3.8320
Monitor 0-3	Monitor 4-6	2.1849
Monitor 0-4	Monitor 5-6	1.2902
Monitor 0-5	Monitor 6	0.7144
Exemptions 0	Exemptions 1-7	0.8126
Exemptions 0-1	Exemptions 2-7	1.3015
Exemptions 0-2	Exemptions 3-7	1.9236
Exemptions 0-3	Exemptions 4-7	1.1150
Exemptions 0-4	Exemptions 5-7	1.3811
Exemptions 0-5	Exemptions 6-7	1.7902
Exemptions 0-6	Exemptions 7	0.2564
Report 0	Report 1-6	2.0043
Report 0-1	Report 2-6	0.1424
Report 0-2	Report 3-6	1.6643
Report 0-3	Report 4-6	0.8846
Report 0-4	Report 5-6	1.0432
Report 0-5	Report 6	2.0790
Teacher Cert 1	Teacher Cert 2-3	2.2946
Teacher Cert 1-2	Teacher Cert 3	2.1062
LEA	Non-LEA	3.4834

Note. *LEA* local education authority

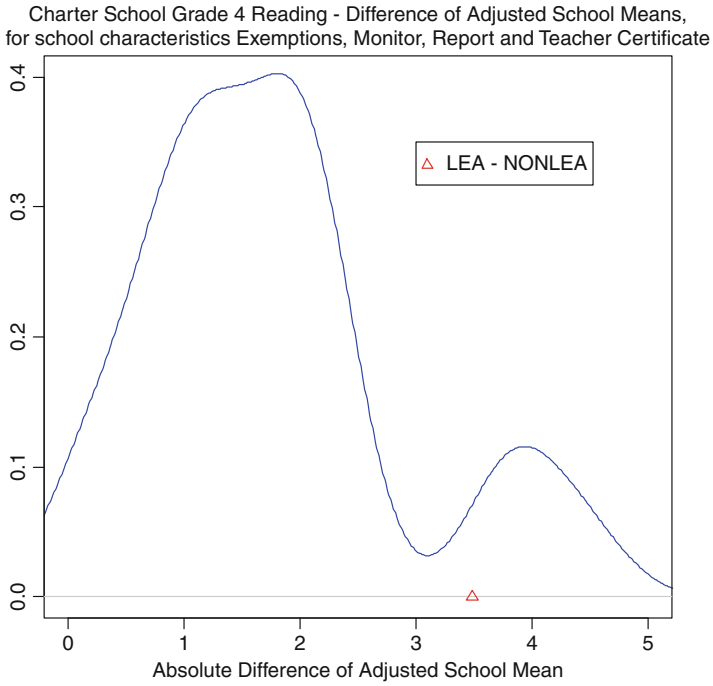

Fig. 13.10 Empirical distribution of the absolute differences based on all possible splits on four school characteristics. *Note*. LEA = local education authority

References

Braun, H., Jenkins, F., & Grigg, W. (2006). *A closer look at charter schools using hierarchical linear modeling (NCES 2006–460)*. Washington, DC: U.S. Government Printing Office.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1988). *Classification and regression trees*. New York, NY: Chapman & Hall.

Carnoy, M., Jacobsen, R., Mishel, L., & Rothstein, R. (2005). *The charter school dust-up: Examining the evidence on enrollment and achievement*. Washington, DC: Economic Policy Institute.

Charter School Achievement Consensus Panel. (2006). *Key issues in studying charter schools and achievement: A review and suggestions for national guidelines* (NCSRP White Paper Series No. 2). Seattle, WA: Center on Reinventing Public Education.

National Center for Education Statistics. (2005). *America's charter schools: Results from the NAEP 2003 pilot study (NCES 2005–456)*. Washington, DC: U.S. Government Printing Office.

Chapter 14
Holland's Advice for the Fourth Generation of Test Theory: Blood Tests Can Be Contests

Neil J. Dorans

14.1 Overview

According to Holland (2008) in *The First Four Generations of Test Theory,* testing as a scientific enterprise is not more than 120 years old. Holland divides this enterprise into four overlapping generations. The first generation, which was influenced by concepts such as error of measurement and correlation that were developed in other fields, focused on test scores and saw developments in the areas of reliability, classical test theory, generalizability theory, and validity. This generation began in the early twentieth century and continues today, but most of its major developments were achieved by 1970. The second generation, which focused on models for item level data, began in the 1940s and peaked in the 1970s but continues into the present as well. The third generation started in the 1970s and continues into today. It is characterized by the application of statistical ideas and sophisticated computational methods to item level models, as well as models of sets of items.

The current fourth generation attempts to bridge the gap between the statistician/psychometrician role and the role of other components of the testing enterprise. It recognizes that testing occurs within a larger complex system and that measurement needs to occur within this larger context. In this paper, we will discuss one of Holland's important contributions to the fourth generation of testing, the notion of tests as both blood tests and contests, and its implications for differential item functioning (DIF), which is a critical statistical procedure for ensuring fair measurement.

While the third generation was marked by statistical and computational advances, the work in this generation was too specialized. It seems as if modeling the item, and indirectly, the test, was the only concern of this generation of model builders. Examinees were needed to produce scores; if unavailable, the model could be used to simulate scores. In fact, simulations were more convenient and less hassle

N.J. Dorans (✉)
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
e-mail: ndorans@ets.org

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_14,

than grappling with unruly uncooperative data. That simulations inform reality is something of a fantasy. Holland (2008) noted that real tests exist in a complex world with test takers, test administrators, test score users, test developers, and legislation and policy issues.

A key feature of Holland's (2008) fourth generation of test theory is that tests are only a part of a testing program. A test is a single instrument, but a testing program is a whole system of test production, administration, scoring, using, and interpreting test results that repeat in annual or other cycles and in many different sites.

Another aspect of the fourth generation of testing is the difference between what Holland (2008) called tests as *blood tests* and tests as *contests*. Users of test results often see tests as measurements in the same way that a blood test is a measurement of some aspect of an individual. In a remark appended to Cattell (1890) work, Galton wrote:

> One of the most important objects of measurement. . .is to obtain a general knowledge of the capacities of a man by sinking shafts, as it were, at a few critical points. In order to ascertain the best points for the purpose, the sets of measures should be compared with an independent estimate of the man's powers. (p. 380)

This is a vintage measurement view of testing. But the contest view should never be forgotten when it is relevant. As Holland (2008) noted, high stakes always make the contest perspective relevant. Test takers often see tests as contests in which they can be winners or losers. They want fairness.

Contest and the blood test views are sometimes in conflict. We address this conflict in the balance of this paper. In Sect. 14.2, we mention some of Holland's major contributions to and influences on the fourth generation of testing. In Sect. 14.3, we apply contest/blood test thinking to the area of DIF. Section 14.4 is a recap of previous sections.

14.2 Briefly, Holland's Contributions to Differential Item Functioning and Equating

Score equating and DIF are fourth generation activities that have been going on for decades. Holland has been active in both. He coauthored four books on these topics: Holland and Rubin (1982), von Davier, Holland, and Thayer (2004), and Dorans, Pommerich, and Holland (2007) about score linking and equating; and Holland and Wainer (1993) on DIF. The difference in number, 3 to 1, and the fact that that DIF is sandwiched in time between the equating books, reveal that Paul was more interested in equating than DIF.

14.2.1 Equating

Early in the 1980s, Paul and I tried to define the notion of score equatability. DIF (Holland & Wainer, 1993; Zieky, this volume, Chap. 8), renorming the SAT®

(Dorans, 2002), 3,000 miles, and a variety of other issues kept us from doing so. In 2000, after Dorans grappled with concording the SAT and ACT (Dorans, Lyu, Pommerich, & Houston, 1997) and Holland chaired the Committee on Equivalency and Linkage of Educational Tests for the National Research Council that produced *Uncommon Measures: Equivalence and Linkage Among Educational Tests* (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), we finally got around to equatability again. The end result was *Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case* (Dorans & Holland, 2000), which made the case for assessing equatability by checking assumptions associated with equating. Holland and Dorans (2006) contained within it the essence of our collaborative effort on score linking.

14.2.2 Differential Item Functioning

Shortly after Holland had completed Alderman and Holland (1981), an early foray into an area that would command much of his attention over the next decade, he introduced Dorans to direct standardization (Mosteller & Tukey, 1977). Dorans adapted this approach and introduced the standardization method, which was soon applied to the SAT program to assess item fairness (Dorans & Kulick, 1983, 1986).

A few years later, as noted by Zieky (this volume, Chap. 8), Holland was drawn deeply into the problem of developing an alternative to the so-called Golden Rule method. Dorans was pulled into this work as well, and a close collaboration on DIF issues occurred over the next several years as Holland spearheaded the implementation of Mantel-Haenszel (MH) and standardization procedures here at ETS. The MH approach became an industry-wide standard.

Holland's book with Wainer (Holland & Wainer, 1993) was the apex of his work with DIF. Holland's work on DIF was mostly reactive, with some notable exceptions. In Hackett, Holland, Pearlman, and Thayer (1987) and Schmitt, Holland, and Dorans (1993), he illustrated how experimentation could be used to advance our substantive understanding of DIF. As he left the DIF domain, Holland (1994) issued a challenge to the DIF community: DIF is a psychometric procedure that is carried out for contest reasons – the public needs to view test items as fair. We take up that challenge in the next section, which examines DIF from contest and measurement perspectives.

14.3 True-Score Estimates Are Really Observed Scores

In this section, I briefly describe the standardization method for DIF assessment and what is sometimes called its true-score version, SIBTEST (Shealy & Stout, 1993). Then I explore the fairness of the SIBTEST procedure.

14.3.1 Standardization

Dorans (1982) reviewed item bias studies that had been conducted on SAT data in the late seventies and concluded that these studies were flawed because either DIF was confounded with lack of model fit or contaminated by impact as a result of *fat matching*, the practice of grouping scores into broad categories of roughly comparable ability. A new method was needed. As noted above, Dorans and Kulick (1983, 1986) developed the standardization approach after consultation with Holland. The formulas in the following section can be found in these articles and in Dorans and Holland (1993).

14.3.1.1 Standardization's Definition of Differential Item Functioning

An item exhibits DIF when the expected performance on an item differs for matched examinees from different groups. Expected performance can be operationalized by nonparametric item-test regressions. Differences in empirical item-test regressions are indicative of DIF.

The first step in the standardization analysis is to use all available data to estimate nonparametric item-test regressions in the reference group and in the focal group. The focal group is the focus of analysis while the reference group serves as a basis for comparison.

Let $\boldsymbol{\varepsilon}_f(Y|X)$ define the empirical item-test regression for the focal group f, and let $\boldsymbol{\varepsilon}_r(Y|X)$ define the empirical item-test regression for the reference group r, where Y is the item score variable and X is the matching variable. The definition of null DIF employed by the standardization approach implies that $\boldsymbol{\varepsilon}_f(Y|X) = \boldsymbol{\varepsilon}_r(Y|X)$.

The most detailed definition of DIF is at the individual score level, m,

$$D_m = \boldsymbol{\varepsilon}_f(Y|X = m) - \boldsymbol{\varepsilon}_r(Y|X = m), \tag{14.1}$$

where $\boldsymbol{\varepsilon}_f(Y|X = m)$ and $\boldsymbol{\varepsilon}_r(Y|X = m)$ are realizations of the item-test regressions at score level m. The D_m are the fundamental measures of DIF according to the standardization method. Plots of these differences, as well as plots of $\boldsymbol{\varepsilon}_f(Y|X)$ and $\boldsymbol{\varepsilon}_r(Y|X)$, provide visual descriptions of DIF in fine detail for binary as well as polytomously scored items. For illustrations of nonparametric item-test regressions and differences for an actual SAT item that exhibits considerable DIF, see Dorans and Kulick (1986).

14.3.1.2 Standardization's Primary Differential Item Functioning Index

While plots describe DIF directly, a need was identified for some numerical index that targets suspect items for close scrutiny, while allowing acceptable items to

pass swiftly through the screening process. Standardization has such an index, *STD EIS-DIF* (Dorans & Kulick, 2006), which uses a weighting function supplied by the standardization group. The standardization group supplies specific weights for each score level that are used in weighting each individual D_m before accumulating the weighted differences across score levels to arrive at a summary item-discrepancy index, *STD EIS-DIF*, which is defined as:

$$\begin{aligned} STD\ EIS\text{-}DIF &= \boldsymbol{\varepsilon}_f(Y) - \hat{\boldsymbol{\varepsilon}}_f(Y) \\ &= \frac{\sum_{m=1}^{M} N_{fm} * \boldsymbol{\varepsilon}_f(Y|X=m)}{\sum_{m=1}^{M} N_{fm}} - \frac{\sum_{m=1}^{M} N_{fm} * \boldsymbol{\varepsilon}_r(Y|X=m)}{\sum_{m=1}^{M} N_{fm}}, \end{aligned} \tag{14.2}$$

where $N_{fm} / \sum_{m=1}^{M} N_{fm}$ is the weighting factor at score level X_m supplied by the standardization group to weight differences in item performance between the focal group $\boldsymbol{\varepsilon}_f(Y|X)$ and the reference group $\boldsymbol{\varepsilon}_r(Y|X)$.

In contrast to impact, in which each group has its relative frequency serve as a weight at each score level, standardization uses a standard or common weight on both $\boldsymbol{\varepsilon}_f(Y|X=m)$ and $\boldsymbol{\varepsilon}_r(Y|X=m)$, namely $N_{fm} / \sum_{m=1}^{M} N_{fm}$. The use of the same weight on both $\boldsymbol{\varepsilon}_f(Y|X=m)$ and $\boldsymbol{\varepsilon}_r(Y|X=m)$ is the essence of the standardization approach.

Use of N_{fm} means that *EISDIF* equals the difference between the observed performance of the focal group on the item and the predicted performance of selected reference group members who are matched in ability to the focal group members. This difference can be derived very simply; see Dorans and Holland (1993).

For standardization, the definition of null-DIF conditions on an observed score,

$$\boldsymbol{\varepsilon}_f(Y|X) = \boldsymbol{\varepsilon}_r(Y|X). \tag{14.3}$$

14.3.2 SIBTEST: A Model-Based Standardization Approach to Differential Item Functioning

Shealy and Stout (1993) introduced a general model-based approach to assessing DIF and other forms of differential functioning. They cite the standardization approach as a progenitor, but claim that SIBTEST was developed independently of standardization. From a theoretical perspective, SIBTEST is elegant. It sets DIF within a general multidimensional model of item and test performance. Unlike most item response theory (IRT) approaches, which posit a peculiar form for the item response model (e.g. a two-parameter logistic model), SIBTEST does not

specify a particular functional form. In this sense, it is a nonparametric IRT model, in principle, in which the null definition of standardization is replaced by

$$\boldsymbol{\varepsilon}_f(Y|T_x) = \boldsymbol{\varepsilon}_r(Y|T_x), \tag{14.4}$$

where T_x represents a true score for X. As such, it employs a measurement invariance definition of null DIF, while standardization employs a prediction invariance definition (Meredith & Millsap, 1992).

Kelley (1927) provided a framework for true-score theory that introduced his formula relating observed test scores, true scores, and reliability. This research led to classical test theory that was eventually first codified by Gulliksen (1950) and later given a sound statistical basis by Lord and Novick (1968). Classical test theory decomposes an observed score for the ith person on occasion o, X_{io}, into a systematic component T_{xi} and an error component E_{xio},

$$X_{io} = T_{xi} + E_{xio}. \tag{14.5}$$

Note that this definition is at the level of the individual, and o in the classical definition could refer to replications of parallel tests. In this representation, T_{xi} is defined as the expected value for a single person i across parallel measurements; expectation is over tests,

$$T_{xi} = \boldsymbol{\varepsilon}_o(X_{io}). \tag{14.6}$$

Holland (Holland & Hoskens, 2003) preferred to think of X_{io} as representing the score of an individual i from a subpopulation in which all individuals have the same true score or ability level. In this case, T_{xi} is defined as the expected value on a single test X_{io} across parallel people from subpopulation o,

$$T_{xo} = \boldsymbol{\varepsilon}_i(X_{io}). \tag{14.7}$$

Because the tests are parallel in one case and the people are parallel in the other case, these two expectations yield the same answer, $T_{xi} = \boldsymbol{\varepsilon}_o(X_{io}) = T_{xo} = \boldsymbol{\varepsilon}_i(X_{io})$, when the tests are parallel and the people are parallel. Hence alternative conceptualizations of the true score exist: one at the level of the individual (across parallel tests), and one at the level of the test (across parallel people). But neither conceptualization can be realized in practice.

To make SIBTEST practical, Shealy and Stout (1993) resorted to Kelley's (1927) equation for estimating true scores from observed scores. In essence, SIBTEST replaces the empirical item-test regression used by standardization with an adjusted regression that employs Kelley's equation. The null definition of DIF for standardization as shown in (14.3) is replaced by this null-DIF hypothesis,

$$\boldsymbol{\varepsilon}_f(Y|X = m) + Adj_{mf} = \boldsymbol{\varepsilon}_r(Y|X = m) + Adj_{mr} \tag{14.8}$$

where

$$Adj_{mf} = \frac{E_f(Y|X = m+1) - E_f(Y|X = m-1)}{\hat{T}_f(X = m+1) - \hat{T}_f(X = m-1)} \left(\frac{\hat{T}_r(X = m) - \hat{T}_f(X = m)}{2} \right) \quad (14.9)$$

and

$$Adj_{mr} = -\frac{E_r(Y|X = m+1) - E_r(Y|X = m-1)}{\hat{T}_r(X = m+1) - \hat{T}_r(X = m-1)} \left(\frac{\hat{T}_r(X = m) - \hat{T}_f(X = m)}{2} \right). \quad (14.10)$$

Kelley's (1927) correction comes into play at this point:

$$\hat{T}_g(X) = rel_g(X) * X + (1 - rel_g(X)) * \boldsymbol{\varepsilon}_g(X). \quad (14.11)$$

This equation produces a subgroup-specific linear transformation of the observed score X. It is not the true score, as defined above, which takes an expectation across parallel people or across parallel test forms. The expectation used by SIBTEST produces a mean for the focal group and a mean for the reference group. In SIBTEST, an observed score on X is treated differently depending on whether it is obtained by the reference group or the focal group. It is regressed to a different mean. This difference in regressed means leads to a higher item-test regression for the lower scoring group and a lower one for the higher scoring group. For example, SIBTEST's effect on Black/White DIF would be to reduce the negative DIF against the Black group on the grounds that DIF indicated by standardization is inflated to the extent that the groups differ on the unreliable observed score matching variable.

Is the use of the differential regression corrections and its effect on the item-test regression defensible from a contest point of view? DIF, after all, is a contest activity. We examine this question in the next subsection.

14.3.2.1 A Dangerous Application

Wainer (2007) cites Kelley's (1927) equation as a contender for the world's most dangerous equation. According to Wainer, a dangerous equation is one that people are ignorant of and has serious implications for a wide variety of applications. Sometimes an equation can be dangerous if it is known and misused. Shealy and Stout (1993) were aware of Kelly's formula, but they misused it. Estimated true scores are not true scores. Instead they are linear transformations of observed scores that are regressed toward a mean to a degree that reflects the uncertainty of the prediction. The use of different transformations for the reference and focal groups is tantamount to using subgroup specific linkings of observed scores that take into account subgroup means and standard deviations.

Shealy and Stout (1993) operate within the classical test theory framework, which is the core of Holland's first generation of testing (Holland, 2008). Whenever the Kelley correction is used in practice, certain problems arise. First, as noted above, TSEs are not true scores. Tucker (1971) addressed this type of distinction in the context of factor scores.

An even more perplexing issue associated with using the use of the Kelley correction is the *which group* question. Each examinee is a member of a large number of groups. In DIF, race/ethnicity and gender are the groups of interest. For example, one test taker is male and White. Hence, he belongs to the group called White males, the group called male, and the group called White, as well as being a member of the total group that includes both gender groups and all ethnic/racial groups and those who choose not to identify themselves. As a White male, he has observed scores that can be regressed to the total group mean, the mean of Whites, and the mean of males or the mean of White males. The observed scores of an Asian American woman, on the other hand, could be regressed to the mean of Asian Americans, the mean of women, or the mean of Asian American women or the overall mean.

If SIBTEST regressed to the overall mean, it would be identical to standardization since the Kelley correction is simply a linear transformation of the observed scores and standardization results are invariant with respect to this linear transformation except for some clumping that might be introduced if scores were rounded. In order for SIBTEST to be different from standardization, it has to regress to different means, namely those of the focal and reference group.

14.3.2.2 SIBTEST True-Score Estimates (TSE): An Example

Consider the following example constructed from data in the public domain on a well-known math test. The SAT is a widely used admissions test with widely published statistical properties. In 2005, the average SAT-Math mean was 520. I chose 2005 because the mean that year is a reportable score; SAT scores ranged from 200 to 800 in steps of 10 (College Board, 2005). Let's assume that the reliability of the test X is the same in both the focal and reference groups.

The leftmost column of Table 14.1 contains labels for each group. Alongside that column are the means for each group. Next come three pairs of columns. Each pair contains TSEs based on Kelley's formula using a common reliability of 0.90 and the difference between the TSEs and the observed score that appears at the top of the each pair of rows. Three observed scores are considered: 420, which is just below the mean score of Black female test takers on SAT-Math; 520, the total group average; and 600, just above the average score for Asian American male test takers.

For an observed score of 420, using of one of the three Black group means (424, 431, or 442) leaves the score basically unchanged (420, 421 or 422), which means a Black examinee with a score of 420 would have an estimated true score close to 420. In contrast, the TSE for an Asian American examinee with a 420 would

Table 14.1 True-score estimate (TSE) and difference between true-score estimate and observed score (OS) as a function of observed score and group mean (mean) for a test score reliability of 0.90

		Reliability = 0.90					
Subgroup		OS = 420		OS = 520		OS = 600	
Name	Mean	TSE	TSE – OS	TSE	TSE – OS	TSE	TSE – OS
Black female	424	420	0	510	***−10***	582	**−18**
Black	431	421	1	511	***−9***	583	**−17**
Black male	442	422	2	512	***−8***	584	**−16**
Female	504	428	***8***	518	−2	590	***−10***
White female	520	430	***10***	520	0	592	***−8***
Total	520	430	***10***	520	0	592	***−8***
White	536	432	***12***	522	2	594	***−6***
Male	538	432	***12***	522	2	594	***−6***
White male	554	433	***13***	523	3	595	***−5***
Asian American female	566	435	**15**	525	***5***	597	−3
Asian American	580	436	**16**	526	***6***	598	−2
Asian American male	595	438	**18**	528	***8***	600	−1

Note. Any difference between OS and TSE in bold could produce a scale score difference of 10–20 points

increase by 15 to 18 points (435, 436, or 438), depending on which of the three Asian American means (566, 580, or 595) were used. If the score of 420 were regressed to the total group mean of 520, all examinees with observed score of 420 would be regressed to 430, an increase of 10 points, regardless of which group they came from.

On the other hand, for a score of 600, regression to the total group mean of 520 would reduce the observed score by about 8 points to 592. If their subgroup specific version of the regression were used, Asian American test takers with 600 would be barely affected, while Black examinees with scores of 600 would be pulled down toward 580.

The average score of 520 would be pulled toward 510 for Black examinees and toward 530 for Asian American examinees.

If the reliability of the test decreases, the TSEs are regressed more toward the mean, and if separate regressions are employed, the estimates are pulled even more toward different means.

In the equal reliability case, the kernel of the SIBTEST correction for the unreliability of the matching variable is captured in the term $\left(T_r(X = m) - T_f(X = m)\right)/2$ which is half the difference in the TSEs for the focal and reference groups at x_m. This term is added to the item-test regression for the focal group and subtracted from the one for the reference group. The ultimate effect is that relative to the observed item-test regressions used by standardization; these adjustments make the item look easier for the lower scoring group and harder for the higher scoring group. Hence a positive DIF item (favors lower-scoring focal group, e.g. Black examinees) under standardization would look even more positive under SIBTEST, while a negative DIF item (favors higher scoring reference group, e.g. White test takers) would look less

negative under SIBTEST. Conversely, a positive DIF item (favors higher-scoring focal group, e.g. Asian American examinees) under standardization would look less positive under SIBTEST while a negative DIF item (favors lower scoring reference group, e.g. White test takers) would look more negative. In effect, SIBTEST would suggest less negative DIF for Black examinees and more negative DIF for Asian American examinees.

14.3.2.3 Which Group?

The use of TSEs in place of observed scores produces results that run counter to the purpose of DIF when examined from a contest perspective. For example, a Black examinee obtains a 600 but instead receives a true score estimate of about 580 (rounded estimate). An Asian American examinee keeps his or her score of 600. A second Black examinee keeps his or her score of 420, but a second Asian American test taker with the 420 gets a 440. A third Asian American test taker with a 520 receives a 530, while a third Black test taker with a 520 receives a lower score of 510. In essence, SIBTEST adjusts your score in the direction of the mean of the group you came from before assessing DIF, making an adjustment that seems to run counter to the intent of the DIF analysis.

Take an Asian American female examinee with a score of 420. Because she is in the Asian American group, she gets boosted past 430 towards 440. As a female, she only gets up to 430. What about the Black female test taker with a score of 600? She gets dropped almost to 580 as a member of the Black group and close to 590 as a female. Which TSE is better? One might argue that conditioning on both gender and race is better than using either one alone. Then the Asian American female test taker with a 420 gets close to 440, while the Black female test taker with a 600 gets close to 580. If we had more useful information, we could condition on that, and in an ideal world, we could reduce our uncertainty about the transformed observed score to an acceptably small level. Along the way, we would have a wide variety of estimates to choose from, none of which is a true score in the sense of an expected score over many parallel people as noted in the next section.

14.3.2.4 Constructing the Perfect True Score Estimate

The example cited above used a reliability of 0.9. While results from SIBTEST and standardization differ here, they don't differ by much. Most of the literature that shows differences between these methods or between MH and SIBTEST involves tests with lower reliabilities. Standardization suffers when the matching variable is unreliable. SIBTEST attempts to fix the unreliability by regressing scores toward the focal and reference group means, respectively.

As noted earlier, the true score of real interest is the expected value of an examinee's performance over many parallel forms of a test or the expected value of many parallel people on the test. The approach employed by SIBTEST of using

subgroup-specific conversions of observed score that regress observe scores toward subgroup means does not achieve this goal.

As noted earlier, classical test theory decomposes an observed score into a systematic component T_{xi} and an error component E_{xio}. Holland (Holland & Hoskens, 2003) preferred to think of X_{io} as representing the score of an individual i from a subpopulation in which all individuals have the same true score or ability level. In their case, T_{xi} is defined as the expected value on a single test X_{io} across parallel people from subpopulation o as shown in (14.7).

SIBTEST uses (14.11) to estimate true score for a given value of X. Equation (14.11) represents a subgroup-specific linear transformation of the observed score X. The expectation used by SIBTEST produces a mean for the focal group and a mean for the reference group. In SIBTEST, an observed score on X is treated differently. Depending on whether the score is obtained by a person in the reference group or the focal group, it is regressed to a different mean. It is not regressed to the true score defined in (14.7), which is an expectation across parallel people.

Hence SIBTEST, as operationalized, fails to achieve what it seeks as a measurement model. In addition, it introduces unfairness into a process that is all about fairness. It replaces an unbiased estimate of individual true score with a least squares estimate that depends on group membership. This is akin to regressing ice skaters' scores, which exhibit some unreliability, towards the mean of ice skaters from their country instead of using their actual ice skating scores.

14.4 Keeping the Contest in Mind

Holland (2008) noted that the fourth generation of testing is characterized by an emerging view that testing should be aware of multiple perspectives, not all of which are compatible. Two important perspectives in high stakes testing are what he calls the contest and blood test perspectives. This paper has described the tests-as-blood-tests perspective as one that is more aligned with the interests of test users, while the tests-as-contests perspective is aligned with the interests of the test taker. To the extent that testing conditions are poor, such as when tests and anchors are unreliable and when matching variables and anchors are unrepresentative of the items and tests being studied, DIF and equating methods aligned with the contest and blood test perspectives will produce different results.

The methods most aligned with the tests-as-blood-tests perspective will replace data with model assumptions. Each method is based on underlying theories and assumptions that are likely to be incorrect when these methods are applied in these undesirable situations. The use of these methods with their strong reliance on measurement models to augment weak data should not be done blindly. Assumptions should be questioned.

The standardization DIF method employs regressions involving observables. It focuses on observed scores and employs regressions that match on an observed score in the same way in both populations of interest. Standardization assesses

whether the item-test regressions are the same across focal and reference groups. It is a contest-oriented method. It has problems, however, when the matching variable is unreliable.

The SIBTEST DIF method appears to be more aligned with measurement models (i.e. blood tests). This method assumes that examinee group differences influence DIF or test form difficulty differences more than can be observed in unreliable test scores. The observed data are pulled toward what is suggested to be appropriate by the measurement model. The degree to which this pulling occurs depends on the extent that these data are unreliable. In the absence of reliable data on the individual, it will presume, for example, that a Black examinee would receive the average score obtained by all Black examinees, and a male examinee would receive the average score obtained by male examinees. SIBTEST regresses observed item data to what would be expected for the focal or reference group on the basis of ample data that show that race and gender are related to item performance. In essence, the SIBTEST method uses a subgroup-specific TSE as a surrogate for the true score that is defined in the classical test theory model.

SIBTEST results differ from standardization results because SIBTEST transforms raw scores differently across the different groups. It starts from the premise that the observed score is not only unreliable but biased against higher scoring groups. Instead of viewing a true score as an expectation over replications of parallel tests or parallel individuals, SIBTEST treats TSE as a prediction problem, introducing bias to reduce mean squared error.

Knowing a person's gender, race, years of schooling, performance on similar tests, and so on should lead to TSEs with smaller mean squared error than test score alone does. But is it sensible to use this information in a process that exists to demonstrate that items behave consistently across subgroups? The author thinks the answer is no.

One of the requirements of test score equating is that equating functions are invariant across test groups. If X and Y are two parallel tests, the linking relationship between them would be invariant across subgroups. In addition, $X = X$ holds in all subgroups because it is parallel to itself. Likewise, the relationship between the true score on X and X is the same in all subgroups.

When SIBTEST employs a subgroup specific transformation of X (or Y) toward a different mean, it implicitly states that the relationship of X to itself is subgroup dependent. Subgroup specific regressions have been rejected as means of equating tests for over 80 years (Kelley, 1927). Why employ these transformations prior to a DIF analysis? SIBTEST use of subgroup specific regressions seems to run counter to the purpose of producing a fair contest.

Poor reliability leads to poor assessment. SIBTEST does not provide a correct solution to the reliability problem. The solution is to marry measurement with contest. The most direct way of doing this is to ensure that the matching variable is reliable enough. Then the observed score approaches the true score. If a score is not reliable enough to support a DIF analysis, it probably is not reliable enough to be reported.

The fourth generation of testing should fully integrate both the contest and blood test perspectives. As Holland (1994) said in the context of DIF:

> . . .tests are not just measuring instruments. . .that they are sometimes contests as well is the main reason that we care about fairness. . .
>
> The measurement view can certainly inform the contest view (and I think that this is important to say to those who only subscribe to the contest view) but neither can replace the other. (p. 29)

The best way to resolve the contest/measurement conflict is not with measurement models that attempt to compensate for poor measurement, but with better measurement. Better measurement should lead to fairer and more useful contests. When the results of statistical procedures based on different perspectives converge, both fairness and measurement are served.

Acknowledgements The author thanks Paul Holland for being the mentor, colleague, and friend who had the most impact on my career. Tim Moses provided valuable advice. Any opinions expressed here are those of the author and not necessarily of Educational Testing Service.

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (ETS Research Rep. No. RR-81-16) Princeton, NJ: ETS.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373–381.

College Board. (2005). *2005 college bound seniors: Total group profile report*. New York, NY: Author.

Dorans, N. J. (1982). *Technical review of item fairness studies: 1975–1979* (ETS Statistical Rep. No. SR-82-90). Princeton, NJ: ETS.

Dorans, N. J. (2002). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 59–84.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (ETS Research Rep. No. RR-83-09). Princeton, NJ: ETS.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel-Haenzel and standardization procedures. *Medical Care, 44 S3,* S107–S114.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *Colleges and Universities, 73*, 24–34.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: 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 (Report of the*

Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Hackett, R. K., Holland, P. W., Pearlman, M., & Thayer, D. T. (1987). *Test construction manipulating scores differences between Black and White examinees: Properties of the resulting tests* (ETS Research Rep. No. RR-87-30). Princeton, NJ: ETS.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/ Donoghue. *Proceedings of the Social Statistics Section of the American Statistical Association, 1994*, 27–29.

Holland, P. W. (2008, March). The first four generations of test theory. Paper presented at the Association of Test Publishers on Innovations in Testing, Dallas, TX.

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/Prager.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*, 123–149.

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

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York, NY: World Book.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289–311.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometika, 58*, 197–239.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika, 36*(4), 427–436.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Wainer, H. (2007). The world's most dangerous equation. *American Scientist, 95*, 249–256.

Author Index

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2,

Subject Index