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

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