Chapter 6
Research on Statistics

Henry Braun

Since its founding in 1947, ETS has supported research in a variety of areas—a fact attested to by the many different chapters comprising this volume. As a private, nonprofit organization known primarily for its products and services related to standardized testing, it comes as no surprise that ETS conducted extensive research in educational measurement and psychometrics, which together provide the scientific foundations for the testing industry. This work is documented in the chapters in this book. At the same time, a good part of educational measurement and perhaps most of psychometrics can be thought of as drawing upon—and providing an impetus for extending—work in theoretical and applied statistics. Indeed, many important developments in statistics are to be found in the reports alluded to above.

One may ask, therefore, if there is a need for a separate chapter on statistics. The short answer is yes. The long answer can be found in the rest of the chapter. A review of the ETS Research Report (RR) series and other archival materials reveals that a great deal of research in both theoretical and applied statistics was carried out at ETS, both by regular staff members and by visitors. Some of the research was motivated by longstanding problems in statistics, such as the Behrens-Fisher problem or the problem of simultaneous inference, and some by issues arising at ETS during the course of business. Much of this work is distinguished by both its depth and generality. Although a good deal of statistics-related research is treated in other chapters, much is not.

The purpose of this chapter, then, is to tell *a* story of statistics research at ETS. It is not *the* story, as it is not complete; rather, it is structured in terms of a number of major domains and, within each domain, a roughly chronological narrative of key highlights. As will be evident, the boundaries between domains are semipermeable so that the various narratives sometimes intermix. Consequently, reference will also

H. Braun (✉)
Boston College, Chestnut Hill, MA, USA
e-mail: braunh@bc.edu

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_6

be made to topic coverage in other chapters. The writing of this chapter was made more challenging by the fact that some important contributions made by ETS researchers or by ETS visitors (supported by ETS) did not appear in the RR series but in other technical report series and/or in the peer-reviewed literature. A good faith effort was made to identify some of these contributions and include them as appropriate.

The chapter begins with a treatment of classic linear models, followed by sections on latent regression, Bayesian methods, and causal inference. It then offers shorter treatments of a number of topics, including missing data, complex samples, and data displays. A final section offers some closing thoughts on the statistical contributions of ETS researchers over the years.

6.1 Linear Models

Linear models, comprising such techniques as regression, analysis of variance, and analysis of covariance, are the workhorses of applied statistics. Whether offering convenient summaries of data patterns, modeling data to make predictions, or even serving as the basis for inferring causal relationships, they are both familiar tools and the source of endless questions and puzzles that have fascinated statisticians for more than a century. Research on problems related to linear models goes back to ETS's earliest days and continues even today.

From the outset, researchers were interested in the strength of the relationship between scores on admissions tests and school performance as measured by grades. The best known example, of course, is the relationship between *SAT*® test scores and performance in the first year of college. The strength of the relationship was evidence of the predictive validity of the test, with predictive validity being one component of the *validity trinity*.[1] From this simple question, many others arose: How did the strength of the relationship change when other predictors (e.g., high school grades) were included in the model? What was the impact of restriction of range on the observed correlations, and to what extent was differential restriction of range the cause of the variation in validity coefficients across schools? What could explain the year-to-year volatility in validity coefficients for a given school, and how could it be controlled? These and other questions that arose over the years provided the impetus for a host of methodological developments that have had an impact on general statistical practice. The work at ETS can be divided roughly into three categories: computation, inference, and prediction.

[1] The validity trinity comprises content validity, criterion-related validity, and predictive validity.

6.1.1 Computation

In his doctoral dissertation, Beaton (1964) developed the sweep operator, which was one of the first computational algorithms to take full advantage of computer architecture to improve statistical calculations with respect to both speed and the size of the problem that could be handled. After coming to ETS, Beaton and his colleagues developed F4STAT, an expandable subroutine library to carry out statistical calculations that put ETS in the forefront of statistical computations. More on F4STAT can be found in Beaton and Barone (Chap. 8, this volume). (It is worth noting that, over the years, the F4STAT system has been expanded and updated to more current versions of FORTRAN and is still in use today.) Beaton et al. (1972) considered the problem of computational accuracy in regression. Much later, Longford, in a series of reports (Longford 1987a, b, 1993), addressed the problem of obtaining maximum likelihood estimates in multilevel models with random effects. Again, accuracy and speed were key concerns. (Other aspects of multilevel models are covered in Sect. 6.2.3). A contribution to robust estimation of regression models was authored by Beaton and Tukey (1974).

6.1.2 Inference

The construction of confidence intervals with specific confidence coefficients is another problem that appears throughout the RR series, with particular attention to the setting of simultaneous confidence intervals when making inferences about multiple parameters, regression planes, and the like. One of the earliest contributions was by Abelson (1953) extending the Neyman-Johnson technique for regression. Aitkin (1973) made further developments. Another famous inference problem, the Behrens-Fisher problem, attracted the attention of Potthoff (1963, 1965), who devised Scheffé-type tests. Beaton (1981) used a type of permutation test approach to offer a way to interpret the coefficients of a least squares fit in the absence of random sampling. This was an important development, as many of the data sets subjected to regression analysis do not have the required pedigree and, yet, standard inferential procedures are applied nonetheless. A. A. von Davier (2003a) treated the problem of comparing regression coefficients in large samples. Related work can be found in Moses and Klockars (2009).

A special case of simultaneous inference arises in analysis of variance (ANOVA) when comparisons among different levels of a factor are of interest and control of the overall error rate is desired. This is known as the problem of multiple comparisons, and many procedures have been devised. Braun and Tukey (1983) proposed a new procedure and evaluated its operating characteristics. Zwick (1993) provided a comprehensive review of multiple comparison procedures. Braun (1994) edited Volume VIII of *The Collected Works of John W. Tukey*, a volume dedicated to Tukey's work in the area of simultaneous inference. Especially noteworthy in this

collection is that Braun, in collaboration with ETS colleagues Kaplan, Sheehan, and Wang, prepared a corrected, complete version of the never-published manuscript (1953) by Tukey titled *The Problem of Multiple Comparisons* (1994), which set the stage for the modern treatment of simultaneous inference. A review of Tukey's contributions to simultaneous inference was presented in Benjamini and Braun (2003).

6.1.3 Prediction

Most of the standardized tests that ETS was and is known for are intended for use in admissions to higher education. A necessary, if not sufficient, justification for their utility is their predictive validity; that is, for example, that scores on the SAT are strongly correlated with first year averages (FYA) in college and, more to the point, that they possess explanatory power above and beyond that available with the use of other quantitative measures, such as high school grades. Another important consideration is that the use of the test does not inappropriately disadvantage specific subpopulations. (A more general discussion of validity can be found in Chap. 16 by Kane and Bridgeman, this volume. See also Kane 2013). Another aspect of test fairness, differential prediction, is discussed in the chapter by Dorans and Puhan (Chap. 4, this volume).

Consequently, the study of prediction equations and, more generally, prediction systems has been a staple of ETS research. Most of the validity studies conducted at ETS were done under the auspices of particular programs and the findings archived in the report series of those programs. At the same time, ETS researchers were continually trying to improve the quality and utility of validity studies through developing new methodologies.

Saunders (1952) investigated the use of the analysis of covariance (ANCOVA) in the study of differential prediction. Rock (1969) attacked a similar problem using the notion of moderator variables. Browne (1969) published a monograph that proposed measures of predictive accuracy, developed estimates of those measures, and evaluated their operating characteristics.

Tucker established ETS's test validity procedures and supervised their implementation until his departure to the University of Illinois. He published some of the earliest ETS work in this area (1957, 1963). His first paper proposed a procedure to simplify the prediction problem with many predictors by constructing a smaller number of composite predictors. The latter paper, titled *Formal Models for a Central Prediction System*, tackled a problem that bedeviled researchers in this area. The problem can be simply stated: Colleges receive applications from students attending many different high schools, each with its own grading standards. Thus, high school grade point averages (HSGPA) are not comparable even when they are reported on a common scale. Consequently, including HSGPA in a single prediction equation without any adjustment necessarily introduces noise in the system and induces bias in the estimated regression coefficients. Standardized test scores, such as the SAT, are on a common scale—a fact that surely contributes to their strong correlation

with FYA. Tucker's monograph discusses three approaches to constructing composite predictors based on placing multiple high school grades on a common scale for purposes of predicting college grades. This work, formally published in Tucker (1963), led to further developments, which were reviewed by Linn (1966) and, later, by Young and Barrett (1992). More recently, Zwick (2013) and Zwick and Himelfarb (2011) conducted further analyses of HSGPA as a predictor of FYA, with a focus on explaining why HSGPA tends to overpredict college performance for students from some demographic subgroups.

Braun and Szatrowski (1984a, b) investigated a complementary prediction problem. When conducting a typical predictive validity study at an institution, the data are drawn from those students who matriculate and obtain a FYA. For schools that use the predictor in the admissions process, especially those that are at least moderately selective, the consequence is a restriction of range for the predictor and an attenuated correlation. Although there are standard corrections for restriction of range, they rest on untestable assumptions. At the same time, unsuccessful applicants to selective institutions likely attend other institutions and obtain FYAs at those institutions. The difficulty is that FYAs from different institutions are not on a common scale and cannot be used to carry out an *ideal validity study* for a single institution in which the prediction equation is estimated on, for example, all applicants.

Using data from the Law School Admissions Council, Braun and Szatrowski (1984a, b) were able to link the FYA grade scales for different law schools to a single, common scale and, hence, carry out institutional validity studies incorporating data from nearly all applicants. The resulting fitted regression planes differed from the standard estimates in expected ways and were in accord with the fitted planes obtained through an Empirical Bayes approach. During the 1980s, there was considerable work on using Empirical Bayes methods to improve the accuracy and stability of prediction equations. (These are discussed in the section on Bayes and Empirical Bayes.)

A longstanding concern with predictive validity studies, especially in the context of college admissions, is the nature of the criterion. In many colleges, freshmen enroll in a wide variety of courses with very different grading standards. Consequently, first year GPAs are rather heterogeneous, which has a complex impact on the observed correlations with predictors. This difficulty was tackled by Ramist et al. (1990). They investigated predictive validity when course-level grades (rather than FYAs) were employed as the criterion. Using this more homogeneous criterion yielded rather different results for the correlations with SAT alone, HSGPA alone, and SAT with HSGPA. Patterns were examined by subject and course rigor, as was variation across the 38 colleges in the study. This approach was further pursued by Lewis et al. (1994) and by Bridgeman et al. (2008).

Over the years, Willingham maintained an interest in investigating the differences between grades and test scores, especially with respect to differential predictive validity (Willingham et al. 2002). Related contributions include Lewis and Willingham (1995) and Haberman (2006). The former showed how restriction of range can affect estimates of *gender bias* in prediction and proposed some strategies

for generating improved estimates. The latter was concerned with the bias in predicting multinomial responses and the use of different penalty functions in reducing that bias.

Over the years, ETS researchers also published volumes that explored aspects of test validity and test use, with some attention to methodological considerations. Willingham (1988) considered issues in testing *handicapped people* (a term now replaced by the term *students with disabilities*) for the SAT and *GRE*® programs. The chapter in that book by Braun et al. (1988) studied the predictive validity for those testing programs for students with different disabilities. Willingham and Cole (1997) examined testing issues in gender-related fairness, with some attention to the implications for predictive validity.

6.1.4 Latent Regression

Latent regression methods were introduced at ETS by Mislevy (1984) for use in the National Assessment of Educational Progress (NAEP) and are further described in Sheehan and Mislevy (1989), Mislevy (1991), and Mislevy et al. (1992). An overview of more recent developments is given in M. von Davier et al. (2006) and M. von Davier and Sinharay (2013). Mislevy's key insight was that NAEP was not intended to, and indeed was prohibited from, reporting scores at the individual level. Instead, scores were to be reported at various levels of aggregation, either by political jurisdiction or by subpopulation of students. By virtue of the matrix sampling design of NAEP, the amount of data available for an individual student is relatively sparse. Consequently, the estimation bias in statistics of interest may be considerable, but can be reduced through application of latent regression techniques. With latent regression models, background information on students is combined with their responses to cognitive items to yield unbiased estimates of score distributions at the subpopulation level—provided that the characteristics used to define the subpopulations are included in the latent regression model. This topic is also dealt with in the chapter by Beaton and Barone (Chap. 8, this volume), especially in Appendix A; the chapter by Kirsch et al. (Chap. 9, this volume) describes assessments of literacy skills in adult populations that use essentially the same methodologies.

In NAEP, the fitting of a latent regression model results in a family of posterior distributions. To generate plausible values, five members of the family are selected at random, and from each a single random draw is made.[2] The plausible values are used to produce estimates of the target population parameters and to estimate the measurement error components of the total variance of the estimates. Note that latent regression models are closely related to empirical Bayes models.

Latent regression models are very complex and, despite more than 25 years of use, many questions remain. In particular, there are attempts to simplify the

[2] In the series of international surveys of adult skills, 10 PV are generated for each respondent.

estimation procedure without increasing the bias. Comparisons of the ETS approach with so-called direct estimation methods were carried out by M. von Davier (2003b). ETS researchers continue to refine the models and the estimation techniques (Li and Oranje 2007; Li et al. 2007; M. von Davier and Sinharay 2010). Goodness-of-fit issues are addressed in Sinharay et al. (2009). In that paper, the authors apply a strategy analogous to Bayesian posterior model checking to evaluate the quality of the fit of a latent regression model and apply the technique to NAEP data.

6.2 Bayesian Methods

Bayesian inference comes in many different flavors, depending on the type of probability formalism that is employed. The main distinction between Bayesian inference and classical, frequentist inference (an amalgam of the approaches of Fisher and Neyman) is that, in the former, distribution parameters of interest are treated as random quantities, rather than as fixed quantities. The Bayesian procedure requires specification of a so-called prior distribution, based on information available before data collection. Once relevant data are collected, they can be combined with the prior distribution to yield a so-called posterior distribution which represents current belief about the likely values of the parameter. This approach can be directly applied to evaluating competing hypotheses, so that one can speak of the posterior probabilities associated with different hypotheses—these are the conditional probabilities of the hypotheses, given prior beliefs and the data collected. As many teachers of elementary (and not so elementary) statistics are aware, these are the kinds of interpretations that many ascribe (incorrectly) to the results of a frequentist analysis.

Over the last 50 years, the Bayesian approach to statistical inference has gained more adherents, particularly as advances in computer hardware/software have made Bayesian calculations more feasible. Both theoretical developments and successful applications have moved Bayesian and quasi-Bayesian methods closer to normative statistical practice. In this respect, a number of ETS researchers have made significant contributions in advancing the Bayesian approach, as well as providing a Bayesian perspective on important statistical issues. This section is organized into three sections: Bayes for classical models, later Bayes, and empirical Bayes.

6.2.1 Bayes for Classical Models

Novick was an early proponent of Bayes methods and a prolific contributor to the Bayesian analysis of classical statistical and psychometric models. Building on earlier work by Bohrer (1964) and Lindley (1969b, c, 1970), Novick and colleagues tackled estimation problems in multiple linear regression with particular attention to applications to predictive validity (Novick et al. 1971, 1972; Novick and Thayer 1969). These studies demonstrated the superior properties of Bayesian regression

estimates when many models were to be estimated. The advantage of *borrowing strength* across multiple contexts anticipated later work by Rubin and others who employed Empirical Bayes methods. Rubin and Stroud (1977) continued this work by treating the problem of Bayesian estimation in unbalanced multivariate analysis of variance (MANOVA) designs.

Birnbaum (1969) presented a Bayesian formulation of the logistic model for test scores, which was followed by Lindley (1969a) and Novick and Thayer (1969), who studied the Bayesian estimation of true scores. Novick et al. (1971) went on to develop a comprehensive Bayesian analysis of the classical test theory model addressing such topics as reliability, validity, and prediction.

During this same period, there were contributions of a more theoretical nature as well. For example, Novick (1964) discussed the differences between the subjective probability approach favored by Savage and the logical probability approach favored by Jefferies, arguing for the relative advantages of the latter. Somewhat later, Rubin (1975) offered an example of where Bayesian and standard frequentist inferences can differ markedly. Rubin (1979a) provided a Bayesian analysis of the bootstrap procedure proposed by Efron, which had already achieved some prominence. Rubin showed that the bootstrap could be represented as a Bayesian procedure—but with a somewhat unusual prior distribution.

6.2.2 Later Bayes

The development of graphical models and associated inference networks found applications in intelligent tutoring systems. The Bayesian formulation is very natural, since prior probabilities on an individual's proficiency profile could be obtained from previous empirical work or simply based on plausible (but not necessarily correct) assumptions about the individual. As the individual attempts problems, data accumulates, the network is updated, and posterior probabilities are calculated. These posterior probabilities can be used to select the next problem in order to optimize some criterion or to maximize the information with respect to a subset of proficiencies.

At ETS, early work on intelligent tutoring systems was carried out by Gitomer and Mislevy under a US Air Force contract to develop a tutoring system for troubleshooting hydraulic systems on F-15s. The system, called HYDRIVE, was one of the first to employ rigorous probability models in the analysis of sequential data. The model is described in Mislevy and Gitomer (1995), building on previous work by Mislevy (1994a, b). Further developments can be found in Almond et al. (2009).

Considerable work in the Bayesian domain concerns issues of either computational efficiency or model validation. Sinharay (2003a, b, 2006) has made contributions to both. In particular, the application of posterior predictive model checking to Bayesian measurement models promises to be an important advance in refining these models. At the same time, ETS researchers have developed Bayesian

formulations of hierarchical models (Johnson and Jenkins 2005) and extensions to testlet theory (Wang et al. 2002).

6.2.3 Empirical Bayes

The term *empirical Bayes* (EB) actually refers to a number of different strategies to eat the Bayesian omelet without breaking the Bayesian eggs; that is, EB is intended to reap the benefits of a Bayesian analysis without initially fully specifying a Bayesian prior. Braun (1988) described some of the different methods that fall under this rubric. We have already noted fully Bayesian approaches to the estimation of prediction equations. Subsequently, Rubin (1980d) proposed an EB strategy to deal with a problem that arose from the use of standardized test scores and school grades in predicting future performance; namely, the prediction equation for a particular institution (e.g., a law school) would often vary considerably from year to year—a phenomenon that caused some concern among admissions officers. Although the causes of this volatility, such as sampling variability and differential restriction of range, were largely understood, they did not lead immediately to a solution.

Rubin's version of EB for estimating many multiple linear regression models (as would be the case in a validity study of 100+ law schools) postulated a multivariate normal prior distribution, but did not specify the parameters of the prior. These were estimated through maximum likelihood along with estimates of the regression coefficients for each institution. In this setting, the resulting EB estimate of the regression model for a particular institution can be represented as a weighted combination of the ordinary least squares (OLS) estimate (based on the data from that institution only) and an overall estimate of the regression (aggregating data across institutions), with the weights proportional to the relative precisions of the two estimates. Rubin showed that, in comparison to the OLS estimate, the EB estimates yielded better prediction for the following year and much lower year-to-year volatility. This work led to changes in the validity study services provided by ETS to client programs.

Braun et al. (1983) extended the EB method to the case where the OLS estimate did not necessarily exist because of insufficient data. This problem can arise in prediction bias studies when the focal group is small and widely scattered among institutions. Later, Braun and Zwick (1993) developed an EB approach to estimating survival curves in a validity study in which the criterion was graduate degree attainment. EB or shrinkage-type estimators are now quite commonly applied in various contexts and are mathematically equivalent to the multilevel models that are used to analyze nested data structures.

6.3 Causal Inference

Causal inference in statistics is concerned with using data to elucidate the causal relationships among different factors. Of course, causal inference holds an important place in the history and philosophy of science. Early statistical contributions centered on the role of randomization and the development of various experimental designs to obtain the needed data most efficiently. In the social sciences, experiments are often not feasible, and various alternative designs and analytic strategies have been devised. The credibility of the causal inferences drawn from those designs has been an area of active research. ETS researchers have made important contributions to both the theoretical and applied aspects of this domain.

With respect to theory, Rubin (1972, 1974b, c), building on earlier work by Neyman, proposed a model for inference from randomized studies that utilized the concept of *potential outcomes*. That is, in comparing two treatments, ordinarily an individual can be exposed to only one of the treatments, so that only one of the two potential outcomes can be observed. Thus, the treatment effect on an individual is inestimable. However, if individuals are randomly allocated to treatments, an unbiased estimate of the average treatment effect can be obtained. He also made explicit the conditions under which causal inferences could be justified.

Later, Rubin (1978a) tackled the role of randomization in Bayesian inference for causality. This was an important development because, until then, many Bayesians argued that randomization was irrelevant to the Bayesian approach. Rubin's argument (in part) was that with a randomized design, Bayesian procedures were not only simpler, but also less sensitive to specification of the prior distribution. He also further explicated the crucial role of the stable unit treatment value assumption (SUTVA) in causal inference. This assumption asserts that the outcome of exposing a unit (e.g., an individual) to a particular treatment does not depend on which other units are exposed to that treatment. Although the SUTVA may be unobjectionable in some settings (e.g., agricultural or industrial experiments), in educational settings it is less plausible and argues for caution in interpreting the results.

Holland and Rubin (1980, 1987) clarified the statistical approach to causal inference. In particular, they emphasized the importance of manipulability; that is, the putative *causal agent* should have at least two possible states. Thus, the investigation of the differential effectiveness of various instructional techniques is a reasonable undertaking since, in principle, students could be exposed to any one of the techniques. On the other hand, an individual characteristic like gender or race cannot be treated as a causal agent, since ordinarily it is not subject to manipulation. (On this point, see also Holland, 2003). They go on to consider these issues in the context of retrospective studies, with consideration of estimating causal effects in various subpopulations defined in different ways.

Lord (1967) posed a problem involving two statisticians who drew radically different conclusions from the same set of data. The essential problem lies in attempting to draw causal conclusions from an analysis of covariance applied to nonexperimental data. The resulting longstanding conundrum, usually known as

Lord's Paradox, engendered much confusion. Holland and Rubin (1983) again teamed up to resolve the paradox, illustrating the power of the application of the Neyman-Rubin model, with careful consideration of the assumptions underlying different causal inferences.

In a much-cited paper, Holland (1986) reviewed the philosophical and epistemological foundations of causal inference and related them to the various statistical approaches that had been proposed to analyze experimental or quasi-experimental data, as well as the related literature on causal modeling. An invitational conference that touched on many of these issues was held at ETS, with the proceedings published in Wainer (1986). Holland (1987) represents a continuation of his work on the foundations of causal inference with a call for the measurement of effects rather than the deduction of causes. Holland (1988) explored the use of path analysis and recursive structural equations in causal inference, while Holland (1993) considered Suppes' theory of causality and related it to the statistical approach based on randomization.

As noted above, observational studies are much more common in the social sciences than are randomized experimental designs. In a typical observational study, units are exposed to treatments through some nonrandom mechanism that is often denoted by the term *self-selection* (whether or not the units actually exercised any discretion in the process). The lack of randomization means that the ordinary estimates of average treatment effects may be biased due to the initial nonequivalence of the groups. If the treatment groups are predetermined, one bias-reducing strategy involves matching units in different treatment groups on a number of observed covariates, with the hope that the resulting matched groups are approximately equivalent on all relevant factors except for the treatments under study. Were that the case, the observed average differences between the matched treatment groups would be approximately unbiased estimates of the treatment effects. Sometimes, an analysis of covariance is conducted instead of matching and, occasionally, both are carried out. These strategies raise some obvious questions. Among the most important are: What are the best ways to implement the matching and how well do they work? ETS researchers have made key contributions to answering both questions.

Rubin (1974b, c, 1980a) investigated various approaches to matching simultaneously on multiple covariates and, later, he considered combined strategies of matching and regression adjustment (1979b). Subsequently, Rosenbaum and Rubin (1985a) investigated the bias due to incomplete matching and suggested strategies for minimizing the number of unmatched treatment cases. Rosenbaum and Rubin (1983b) published a seminal paper on matching using propensity scores. Propensity scores facilitate multifactor matching through construction of a scalar index such that matching on this index typically yields samples that are well-matched on all the factors contributing to the index. Further developments and explications can be found in Rosenbaum and Rubin (1984, 1985b), as well as the now substantial literature that has followed. In 1986, the previously mentioned ETS-sponsored conference (Wainer 1986) examined the topic of inference from self-selected samples. The focus was a presentation by James Heckman on his model-based approach to the problem, with comments and critiques by a number of statisticians. A particular

concern was the sensitivity of the findings to an untestable assumption about the value of a correlation parameter.

More generally, with respect to the question of how well a particular strategy works, one approach is to vary the assumptions and determine (either analytically or through simulation) how much the estimated treatment effects change as a result. In many situations, such sensitivity analyses can yield very useful information. Rosenbaum and Rubin (1983a) pioneered an empirical approach that involved assuming the existence of an unobserved binary covariate that accounts for the residual selection bias and incorporating this variable into the statistical model used for adjustment. By varying the parameters associated with this variable, it is possible to generate a response surface that depicts the sensitivity of the estimated treatment effect as a function of these parameters. The shape of the surface near the *naïve* estimate offers a qualitative sense of the confidence to be placed in its magnitude and direction.

This approach was extended by Montgomery et al. (1986) in the context of longitudinal designs. They showed that if there are multiple observations on the outcome, then under certain stability assumptions it is possible to obtain estimates of the parameters governing the unobserved binary variable and, hence, obtain a point estimate of the treatment effect in the expanded model.

More recently, education policy makers have seized on using indicators derived from student test scores as a basis for holding schools and teachers accountable. Under No Child Left Behind, the principal indicator is the percent of students meeting a state-determined proficiency standard. Because of the many technical problems with such status-based indicators, interest has shifted to indicators related to student progress. Among the most popular are the so-called value-added models (VAM) that attempt to isolate the specific contributions that schools and teachers make to their students' learning. Because neither students nor teachers are randomly allocated to schools (or to each other), this is a problem of causal inference (i.e., attribution of responsibility) from an observational study with a high degree of self-selection. The technical and policy issues were explicated in Braun (2005a, b) and in Braun and Wainer (2007). A comparison of the results of applying different VAMs to the same data was considered in Braun, Qu, and Trapani (2008).

6.4 Missing Data

The problem of missing data is ubiquitous in applied statistics. In a longitudinal study of student achievement, for example, data can be missing because the individual was not present at the administration of a particular assessment. In other cases, relevant data may not have been recorded, recorded but lost, and so on. Obviously, the existence of missing data complicates both the computational and inferential aspects of analysis. Adjusting calculation routines to properly take account of missing values can be challenging. Simple methods, such as deleting cases with missing data or filling in the missing values with some sort of average,

can be wasteful, bias-inducing, or both. Standard inferences can also be suspect when there are missing values if they do not take account of how the data came to be missing. Thus, characterizing the process by which the *missingness* occurs is key to making credible inferences, as well as appropriate uses of the results. Despite the fact that ETS's testing programs and other activities generate oceans of data, problems of missing data are common, and ETS researchers have made fundamental contributions to addressing these problems.

Both Lord (1955) and Gulliksen (1956) tackled specific estimation problems in the presence of missing data. This tradition was continued by Rubin (1974a, 1976b, c). In this last report, concerned with fitting regression models, he considered how patterns of missingness of different potential predictors, along with multiple correlations, can be used to guide the selection of a prediction model. This line of research culminated in the celebrated paper by Dempster et al. (1977) that introduced, and elaborated on, the expectation-maximization (EM) algorithm for obtaining maximum likelihood estimates in the presence of missing data. The EM algorithm is an iterative estimation procedure that converges to the maximum likelihood estimate(s) of model parameters under broad conditions. Since that publication, the EM algorithm has become the tool of choice for a wide range of problems, with many researchers developing further refinements and modifications over the years. An ETS contribution is due to M. von Davier and Sinharay (2007), in which they develop a stochastic EM algorithm that is applied to latent regression problems.

Of course, examples of applications of EM abound. One particular genre involves embedding a complete data problem (for which obtaining maximum likelihood estimates is difficult or computationally intractable) in a larger missing data problem to which EM can be readily applied. Rubin and Szatrowski (1982) employed this strategy to obtain estimates in the case of multivariate normal distributions with patterned covariance matrices. Rubin and Thayer (1982) applied the EM algorithm to estimation problems in factor analysis. A more expository account of the EM algorithm and its applications can be found in Little and Rubin (1983).

With respect to inference, Rubin (1973, 1976b) investigated the conditions under which estimation in the presence of missing data would yield unbiased parameter estimates. The concept of *missing at random* was defined and its implications investigated in both the frequentist and Bayesian traditions. Further work on *ignorable nonresponse* was conducted in the context of sample surveys (see the next section).

6.5 Complex Samples

The problem of missing data, usually termed *nonresponse*, is particularly acute in sample surveys and is the cause of much concern with respect to estimation bias—both of the parameters of interest and their variances. Nonresponse can take many forms, from the complete absence of data to having missing values for certain variables (which may vary from individual to individual). Rubin (1978b) represents an

early contribution using a Bayesian approach to address a prediction problem in which all units had substantial background data recorded but more than a quarter had no data on the dependent variables of interest. The method yields a pseudo-confidence interval for the population average.

Subsequently, Rubin (1980b, c) developed the multiple imputations methodology for dealing with nonresponse. This approach relies on generating posterior distributions for the missing values, based on prior knowledge (if available) and relevant auxiliary data (if available). Random draws from the posterior distribution are then used to obtain estimates of population quantities, as well as estimates of the component of error due to the added uncertainty contributed by the missing data. This work ultimately led to two publications that have had a great impact on the field (Rubin 1987; Rubin et al. 1983). Note that the multiple imputations methodology, combined with latent regression, is central to the estimation strategy in NAEP (Beaton and Barone, Chap. 8, this volume).

A related missing data problem arises in NAEP as the result of differences among states in the proportions of sampled students, either with disabilities or who are English-language learners, who are exempted from sitting for the assessment. Since these differences can be quite substantial, McLaughlin (2000) pointed out that these gaps likely result in biased comparisons between states on NAEP achievement. The suggested solution was to obtain so-called *full-population estimates* based on model assumptions regarding the performance of the excluded students. Braun et al. (2010) attacked the problem by investigating whether the observed differences in exemption rates could be explained by relevant differences in the focal subpopulations. Concluding that was not the case, they devised a new approach to obtaining full-population estimates and developed an agenda to guide further research and policy. Since then, the National Assessment Governing Board has imposed stricter limits on exemption rates.

Of course, missing data is a perennial problem in all surveys. ETS has been involved in a number of international large-scale assessment surveys, including those sponsored by the Organization for Economic Cooperation and Development (e.g., Program for International Student Assessment—PISA, International Adult Literacy Survey – IALS, Program for the International Assessment of Adult Competencies—PIAAC) and by the International Association for the Evaluation of Educational Achievement (e.g., Trends in International Mathematics and Science Study—TIMSS, Progress in International Reading Literacy Study—PIRLS). Different strategies for dealing with missing (or omitted) data have been advanced, especially for the cognitive items. An interesting and informative comparison of different approaches was presented by Rose et al. (2010). In particular, they compared deterministic rules with model-based rules using different item response theory (IRT) models.

6.6 Data Displays

An important tool in the applied statistician's kit is the use of graphical displays, a precept strongly promoted by Tukey in his work on exploratory data analysis. Plotting data in different ways can reveal patterns that are not evident in the usual summaries generated by standard statistical software. Moreover, good displays not only can suggest directions for model improvement, but also may uncover possible data errors.

No one at ETS took this advice more seriously than Wainer. An early effort in this direction can be found in Wainer and Thissen (1981). In subsequent years, he wrote a series of short articles in *The American Statistician* and *Chance* addressing both what to do—and what not to do—in displaying data. See, for example, Wainer (1984, 1993, 1996). During and subsequent to his tenure at ETS, Wainer also was successful in reaching a broader audience through his authorship of a number of well-received books on data display (1997, 2005, 2009).

6.7 Conclusion

This chapter is the result of an attempt to span the range of statistical research conducted at ETS over nearly 70 years, with the proviso that much of that research is covered in other chapters sponsored by this initiative. In the absence of those chapters, this one would have been much, much longer. To cite but one example, Holland and Thayer (1987, 2000) introduced a new approach to smoothing empirical score distributions based on employing a particular class of log-linear models. This innovation was motivated by problems arising in equipercentile equating and led to methods that were much superior to the ones used previously—superior with respect to accuracy, quantification of uncertainty, and asymptotic consistency. This work is described in more detail in Dorans and Puhan (Chap. 4, this volume). In short, only a perusal of many other reports can fully reflect the body of statistical research at ETS.

From ETS's founding, research has been a cornerstone of the organization. In particular, it has always offered a rich environment for statisticians and other quantitatively minded individuals. Its programs and activities generate enormous amounts of data that must be organized, described, and analyzed. Equally important, the various uses proposed for the data often raise challenging issues in computational efficiency, methodology, causality, and even philosophy. To address these issues, ETS has been fortunate to attract and retain (at least for a time) many exceptional individuals, well-trained in statistics and allied disciplines, eager to apply their skills to a wide range of problems, and effective collaborators. That tradition continues with attendant benefits to both ETS and the research community at large.

Acknowledgments The author acknowledges with thanks the superb support offered by Katherine Shields of Boston College. Katherine trolled the ETS ReSEARCHER database at http://search.ets.org/researcher/, made the first attempt to cull relevant reports, and assisted in the organization of the chapter. Thanks also for comments and advice to: Al Beaton, Randy Bennett, Brent Bridgeman, Neil Dorans, Shelby Haberman, Paul Holland, and the editors.

References

Abelson, R. P. (1953). A note on the Neyman-Johnson technique. *Psychometrika, 18*, 213–218. https://doi.org/10.1007/BF02289058

Aitkin, M. A. (1973). Fixed-width confidence intervals in linear regression with applications to the Johnson-Neyman technique. *British Journal of Mathematical and Statistical Psychology, 26*, 261–269. https://doi.org/10.1111/j.2044-8317.1973.tb00521.x

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics, 34*, 491–521.

Beaton, A. E. (1964). *The use of special matrix operators in statistical calculus*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Beaton, A. E. (1981). *Interpreting least squares without sampling assumptions* (Research Report No. RR-81-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01265.x

Beaton, A. E., Rubin, D. B., & Barone, J. L. (1972). The acceptability of regression solutions: Another look at computational accuracy. *Journal of the American Statistical Association, 71*, 158–168. https://doi.org/10.1080/01621459.1976.10481507

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band–spectroscopic data. *Technometrics, 16*, 147–185. https://doi.org/10.1080/00401706.1974.10489171

Benjamini, Y., & Braun, H. I. (2003). John W. Tukey's contributions to multiple comparisons. *Annals of Statistics, 30*, 1576–1594.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258–276. https://doi.org/10.1016/0022-2496(69)90005-4

Bohrer, R. E. (1964). *Bayesian analysis of linear models: Fixed effects* (Research Bulletin No. RB-64-46). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1964.tb00516.x

Braun, H. I. (1988). *Empirical Bayes methods: A tool for exploratory analysis* (Research Report No. RR-88-25). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00281.x

Braun, H. I. (Ed.). (1994). *The collected works of John W. Tukey: Vol. VIII. Multiple comparisons*. New York: Chapman & Hall, Inc..

Braun, H. I. (2005a). *Using student progress to evaluate teachers: A primer on value-added models* (Policy Information Report). Princeton: Educational Testing Service.

Braun, H. I. (2005b). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 19–39). Maple Grove: JAM Press.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika, 48*, 171–181. https://doi.org/10.1007/BF02294013

Braun, H. I., Qu, Y., & Trapani, C. (2008). *Robustness of a value-added assessment of school effectiveness* (Research Report No. RR-08-22). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02108.x

Braun, H. I., Ragosta, M., & Kaplan, B. (1988). Predictive validity. In W. W. Willingham (Ed.), *Testing handicapped people* (pp. 109–132). Boston: Allyn and Bacon.

Braun, H. I., & Szatrowski, T. H. (1984a). The scale-linkage algorithm: Construction of a universal criterion scale for families of institutions. *Journal of Educational Statistics, 9*, 311–330. https://doi.org/10.2307/1164744

Braun, H. I., & Szatrowski, T. H. (1984b). Validity studies based on a universal criterion scale. *Journal of Educational Statistics, 9*, 331–344. https://doi.org/10.2307/1164745

Braun, H. I., & Tukey, J. W. (1983). Multiple comparisons through orderly partitions: The maximum subrange procedure. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 55–65). Hillsdale: Erlbaum.

Braun, H. I., & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 27. Psychometrics* (pp. 867–892). Amsterdam: Elsevier.

Braun, H. I., Zhang, J., & Vezzu, S. (2010). An investigation of bias in reports of the National Aassessment of Educational Progress. *Educational Evaluation and Policy Analysis, 32*, 24–43. https://doi.org/10.3102/0162373709351137

Braun, H. I., & Zwick, R. J. (1993). Empirical Bayes analysis of families of survival curves: Applications to the analysis of degree attainment. *Journal of Educational Statistics, 18*, 285–303.

Bridgeman, B., Pollack, J. M., & Burton, N. W. (2008). Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches. *Journal of College Admission, 199*, 19–25.

Browne, M. W. (1969). *Precision of prediction* (Research Bulletin No. RB-69-69). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00748.x

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *The Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Gulliksen, H. O. (1956). A least squares solution for paired comparisons with incomplete data. *Psychometrika, 21*, 125–134. https://doi.org/10.1007/BF02289093

Haberman, S. J. (2006). Bias in estimation of misclassification rates. *Psychometrika, 71*, 387–394. https://doi.org/10.1007/s11336-004-1145-6

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. https://doi.org/10.1080/01621459.1986.10478354

Holland, P. W. (1987). *Which comes first, cause or effect?* (Research Report No. RR-87-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00212.x

Holland, P. W. (1988). *Causal inference, path analysis and recursive structural equations models* (Research Report No. RR-88-14). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00270.x

Holland, P. W. (1993). *Probabilistic causation without probability* (Research Report No. RR-93-19). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01530.x

Holland, P. W. (2003). *Causation and race* (Research Report No. RR-03-03). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01895.x

Holland, P. W., & Rubin, D. B. (1980). *Causal inference in prospective and retrospective studies*. Washington, DC: Education Resources Information Center.

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

Holland, P. W., & Rubin, D. B. (1987). *Causal inference in retrospective studies* (Research Report No. RR-87-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00211.x

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Report No. RR-87-31). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00235.x

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. https://doi.org/10.3102/10769986025002133

Johnson, M. S., & Jenkins, F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (Research Report No. RR-04-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01965.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000

Lewis, C., McCamley-Jenkins, L., & Ramist, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Research Report No. RR-94-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01600.x

Lewis, C., & Willingham, W. W. (1995). *The effects of sample restriction on gender differences* (Research Report No. RR-95-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01648.x

Li, D., & Oranje, A. (2007). *Estimation of standard error of regression effects in latent regression models using Binder's linearization* (Research Report No. RR-07-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02051.x

Li, D., Oranje, A., & Jiang, Y. (2007). *Parameter recovery and subpopulation proficiency estimation in hierarchical latent regression models* (Research Report No. RR-07-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02069.x

Lindley, D. V. (1969a). *A Bayesian estimate of true score that incorporates prior information* (Research Bulletin No. RB-69-75). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00754.x

Lindley, D. V. (1969b). *A Bayesian solution for some educational prediction problems* (Research Bulletin No. RB-69-57). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00735.x

Lindley, D. V. (1969c). *A Bayesian solution for some educational prediction problems, II* (Research Bulletin No. RB-69-91). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00770.x

Lindley, D. V. (1970). *A Bayesian solution for some educational prediction problems, III* (Research Bulletin No. RB-70-33). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1970.tb00591.x

Linn, R. L. (1966). Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement, 3*, 313–329. https://doi.org/10.1111/j.1745-3984.1966.tb00897.x

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician, 37*, 218–220. https://doi.org/10.1080/00031305.1983.10483106

Longford, N. T. (1987a). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817–827. https://doi.org/10.1093/biomet/74.4.817

Longford, N. T. (1987b). *Fisher scoring algorithm for variance component analysis with hierarchically nested random effects* (Research Report No. RR-87-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00236.x

Longford, N. T. (1993). *Logistic regression with random coefficients* (Research Report No. RR-93-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01531.x

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association, 50*, 870–876. https://doi.org/10.1080/01621459.1955.10501972

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–305. https://doi.org/10.1037/h0025105

McLaughlin, D. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates* (Technical report). Palo Alto: American Institutes for Research.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381. https://doi.org/10.1007/BF02306026

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J. (1994a). *Information-decay pursuit of dynamic parameters in student models* (Research Memorandum No. RM-94-14-ONR). Princeton: Educational Testing Service.

Mislevy, R. J. (1994b). *Virtual representation of IID observations in Bayesian belief networks* (Research Memorandum No. RM-94-13-ONR). Princeton: Educational Testing Service.

Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction, 5*, 253–282. https://doi.org/10.1007/BF01126112

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154. https://doi.org/10.2307/1165166

Montgomery, M. R., Richards, T., & Braun, H. I. (1986). Child health, breast-feeding, and survival in Malaysia: A random-effects logit approach. *Journal of the American Statistical Association, 81*, 297–309. https://doi.org/10.1080/01621459.1986.10478273

Moses, T., & Klockars, A. (2009). *Strategies for testing slope differences* (Research Report No. RR-09-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02189.x

Novick, M. R. (1964). *On Bayesian logical probability* (Research Bulletin No. RB-64-22). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1964.tb00330.x

Novick, M. R., Jackson, P. H., & Thayer, D. T. (1971). Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika, 36*, 261–288. https://doi.org/10.1007/BF02297848

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in m-groups: A cross-validation study. *British Journal of Mathematical and Statistical Psychology, 25*, 33–50. https://doi.org/10.1111/j.2044-8317.1972.tb00476.x

Novick, M. R., & Thayer, D. T. (1969). *A comparison of Bayesian estimates of true score* (Research Bulletin No. RB-69-74). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00753.x

Potthoff, R. F. (1963). *Illustrations of some Scheffe-type tests for some Behrens-Fisher-type regression problems* (Research Bulletin No. RB-63-36). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1963.tb00502.x

Potthoff, R. F. (1965). Some Scheffe-type tests for some Behrens-Fisher-type regression problem. *Journal of the American Statistical Association, 60*, 1163–1190. https://doi.org/10.1080/01621459.1965.10480859

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton: Educational Testing Service.

Rock, D. A. (1969). *The identification and utilization of moderator effects in prediction systems* (Research Bulletin No. RB-69-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00573.x

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Research Report No. RR-10-11). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *The Journal of the Royal Statistical Society, Series B, 45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). *The bias due to incomplete matching* (Research Report No. RR-83-37). Princeton: Educational Testing Service.

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. https://doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics, 41*, 103–116. https://doi.org/10.2307/2530647

Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rubin, D. B. (1972). *Estimating causal effects of treatments in experimental and observational studies* (Research Bulletin No. RB-72-39). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1972.tb00631.x

Rubin, D. B. (1973). *Missing at random: What does it mean?* (Research Bulletin No. RB-73-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1973.tb00198.x

Rubin, D. B. (1974a). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association, 69*, 467–474. https://doi.org/10.1080/01621459.1974.10482976

Rubin, D. B. (1974b). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics, 32*, 109–120. https://doi.org/10.2307/2529342

Rubin, D. B. (1974c). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics, 32*, 121–132. https://doi.org/10.2307/2529343

Rubin, D. B. (1975). *A note on a simple problem in inference* (Research Bulletin No. RB-75-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01061.x

Rubin, D. B. (1976a). Comparing regressions when some predictor values are missing. *Technometrics, 18*, 201–205. https://doi.org/10.1080/00401706.1976.10489425

Rubin, D. B. (1976b). Inference and missing data. *Biometrika, 63*, 581–592.

Rubin, D. B. (1976c). Noniterative least squares estimates, standard errors, and F-tests for any analysis of variance with missing data. *The Journal of the Royal Statistical Society, 38*, 270–274.

Rubin, D. B. (1978a). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34–58. https://doi.org/10.1214/aos/1176344064

Rubin, D. B. (1978b). Multiple imputations in sample surveys—A phenomenological Bayesian approach to nonresponse. *The Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

Rubin, D. B. (1979a). *The Bayesian bootstrap* (Program Statistical Report No. PSRTR-80-03). Princeton: Educational Testing Service.

Rubin, D. B. (1979b). Using multivariate matched sampling and regression to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318–328. https://doi.org/10.2307/2286330

Rubin, D. B. (1980a). Bias reduction using Mahalanobis metric matching. *Biometrics, 36*, 293–298. https://doi.org/10.2307/2529981

Rubin, D. B. (1980b). *Handling nonresponse in sample surveys by multiple imputations*. Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Rubin, D. B. (1980c). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. In *Proceedings of the 42nd session of the International Statistical Institute, 1979* (Book 2, pp. 517–532). The Hague: The International Statistical Institute.

Rubin, D. B. (1980d). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association, 75*, 801–816. https://doi.org/10.1080/01621459.1980.10477553

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley. https://doi.org/10.1002/9780470316696

Rubin, D., & Stroud, T. (1977). The calculation of the posterior distribution of the cell means in a two-way unbalanced MANOVA. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 26*, 60–66. https://doi.org/10.2307/2346868

Rubin, D. B., & Szatrowski, T. H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika, 69*, 657–660. https://doi.org/10.1093/biomet/69.3.657

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47*, 69–76. https://doi.org/10.1007/BF02293851

Rubin, D. B., Madow, W. G., & Olkin, I. (Eds.). (1983). *Incomplete data in sample surveys: Vol. 2. Theory and bibliographies*. New York: Academic Press.

Saunders, D. R. (1952). *The "ruled surface regression" as a starting point in the investigation of "differential predictability"* (Research Memorandum No. RM-52-18). Princeton: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. J. (1989). Information matrices in latent-variable models. *Journal of Educational Statistics, 14*, 335–350.

Sinharay, S. (2003a). *Assessing convergence of the Markov chain Monte Carlo algorithms: A review* (Research Report No. RR-03-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01899.x

Sinharay, S. (2003b). *Practical applications of posterior predictive model checking for assessing fit of the common item response theory models* (Research Report No. RR-03-33). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01925.x

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–33. https://doi.org/10.3102/10769986031001001

Sinharay, S., Guo, Z., von Davier, M., & Veldkamp, B. P. (2009). *Assessing fit of latent regression models* (Research Report No. RR-09-50). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02207.x

Tucker, L. R. (1957). *Computation procedure for transformation of predictor variables to a simplified regression structure* (Research Memorandum No. RM-57-01). Princeton: Educational Testing Service.

Tucker, L. R. (1963). *Formal models for a central prediction system* (Psychometric Monograph No. 10). Richmond: Psychometric Corporation.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Unpublished manuscript.

Tukey, J. W. (1994). The problem of multiple comparisons. In H. I. Braun (Ed.), *The collected works of John Tukey: Vol. VIII. Multiple comparisons* (pp. 1–300). New York: Chapman & Hall.

von Davier, A. A. (2003a). *Large sample tests for comparing regression coefficients in models with normally distributed variables* (Research Report No. RR-03-19). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01911.x

von Davier, M. (2003b). *Comparing conditional and marginal direct estimation of subgroup distributions* (Research Report No. RR-03-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01894.x

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32*, 233–251. https://doi.org/10.3102/1076998607300422

von Davier, M., Sinharay, S., Oranje., A. & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*, 174–193. https://doi.org/10.3102/1076998609346970

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). New York: CRC Press.

Wainer, H. (1984). How to display data badly. *The American Statistician, 38*, 137–147. https://doi.org/10.1080/00031305.1984.10483186

Wainer, H. (Ed.). (1986). *Drawing inferences from self-selected samples*. New York: Springer. https://doi.org/10.1007/978-1-4612-4976-4

Wainer, H. (1993). Graphical answers to scientific questions. *Chance, 6*(4), 48–50. https://doi.org/10.1080/09332480.1993.10542398

Wainer, H. (1996). Depicting error. *The American Statistician, 50*, 101–111. https://doi.org/10.1080/00031305.1996.10474355

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus Books.

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton: Princeton University Press.

Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate and control uncertainty through graphical display*. Princeton: Princeton University Press.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (pp. 191–241). Palo Alto: Annual Reviews. https://doi.org/10.1177/0146621602026001007

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109–128. https://doi.org/10.1177/0146621602026001007

Willingham, W. W. (Ed.). (1988). *Testing handicapped people*. Boston: Allyn and Bacon.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Willingham, W. W., Pollack, J., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*, 1–37. https://doi.org/10.1111/j.1745-3984.2002.tb01133.x

Young, J. W., & Barrett, C. A. (1992). Analyzing high school transcripts to improve prediction of college performance. *Journal of College Admission, 137*, 25–29.

Zwick, R. J. (1993). Pairwise comparison procedures for one-way analysis of variance designs. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 43–71). Hillsdale: Erlbaum.

Zwick, R. J. (2013). *Disentangling the role of high school grades, SAT scores, and SES in predicting college achievement* (Research Report No. RR-13-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02316.x

Zwick, R. J., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement, 48*, 101–121. https://doi.org/10.1111/j.1745-3984.2011.00136.x