Part I
ETS Contributions to Developing Analytic Tools for Educational Measurement

Chapter 2
A Review of Developments and Applications in Item Analysis

Tim Moses

This chapter summarizes contributions ETS researchers have made concerning the applications of, refinements to, and developments in item analysis procedures. The focus is on dichotomously scored items, which allows for a simplified presentation that is consistent with the focus of the developments and which has straightforward applications to polytomously scored items. Item analysis procedures refer to a set of statistical measures used by testing experts to review and revise items, to estimate the characteristics of potential test forms, and to make judgments about the quality of items and assembled test forms. These procedures and statistical measures have been alternatively characterized as conventional item analysis (Lord 1961, 1965a, b), traditional item analysis (Wainer 1989), analyses associated with classical test theory (Embretson and Reise 2000; Hambleton 1989; Tucker 1987; Yen and Fitzpatrick 2006), and simply item analysis (Gulliksen 1950; Livingston and Dorans 2004). This chapter summarizes key concepts of item analysis described in the sources cited. The first section describes item difficulty and discrimination indices. Subsequent sections review discussions about the relationships of item scores and test scores, visual displays of item analysis, and the additional roles item analysis methods have played in various psychometric contexts. The key concepts described in each section are summarized in Table 2.1.

T. Moses (✉)
College Board, New York, NY, USA
e-mail: tmoses@collegeboard.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_2

Table 2.1 Summary key item analysis concepts

Item analysis concept	Motivation	Description of application to item analysis	Description of application(s) to other psychometric questions
Average item score ($\bar{x}_i$) and reference average item score ($\bar{x}_{i,2}$)	Index for summarizing item difficulty	Gulliksen (1950), Horst (1933), Lord and Novick (1968), Thurstone (1925), and Tucker (1987)	DIF (Dorans and Kulick 1986); item context/order (Dorans and Lawrence 1990; Moses et al. 2007)
Delta (Δ_i) and equated delta $\left[\hat{e}_2\left(\Delta_{i,1}\right)\right]$	Index for summarizing item difficulty with reduced susceptibility to score compression due to mostly high scores or mostly low scores	Brigham (1932), Gulliksen (1950), Holland and Thayer (1985), and Tucker (1987)	DIF (Holland and Thayer 1988); IRT comparisons (L. L. Cook et al. 1988)
Point biserial correlation $\left[\hat{r}_{\text{point biserial}}\left(x_i,y\right)\right]$	Index for summarizing item discrimination	Swineford (1936), Gulliksen (1950), and Lord and Novick (1968)	
Biserial correlation $\left[\hat{r}_{\text{biserial}}\left(x_i,y\right)\right]$	Index for summarizing item discrimination with reduced susceptibility to examinee group differences and to dichotomous scoring	Fan (1952), Pearson (1909), Tucker (1987), Turnbull (1946), and Lord and Novick (1968)	
r-Polyreg correlation $\left[\hat{r}_{\text{polyreg}}\left(x_i,y\right)\right]$	Index for summarizing item discrimination with reduced susceptibility to examinee group differences, dichotomous scoring, and the difficulties of estimating the biserial correlation	Lewis et al. (n.d.) and Livingston and Dorans (2004)	
Conditional average item score $\left(\bar{x}_{ik}\right)$ estimated from raw data	Obtain a detailed description of an item's functional relationship (difficulty and discrimination) with the criterion (usually a total test)	Thurstone (1925), Lord (1965a, b, 1970), and Wainer (1989)	DIF (Dorans and Holland 1993); IRT comparisons (Sinharay 2006)
Conditional average item scores $\left(\bar{x}_{ik}\right)$ estimated from raw data on percentile groupings of the total test scores	Obtain a detailed description of an item's functional relationship (difficulty and discrimination) for a total test with reduced susceptibility to sample fluctuations	Turnbull (1946), Tucker (1987), and Wainer (1989)	
Conditional average item scores $\left(\bar{x}_{ik}\right)$ estimated with kernel or other smoothing	Obtain a detailed description of an item's functional relationship (difficulty and discrimination) for a total test with reduced susceptibility to sample fluctuations	Ramsay (1991) and Livingston and Dorans (2004)	DIF (Moses et al. 2010); IRT comparisons (Moses 2016)

Note. DIF differential item functioning, *IRT* item response theory

2.1 Item Analysis Indices

In their discussions of item analysis, ETS researchers Lord and Novick (1968, p. 327) and, two decades later, Wainer (1989, p. 2) regarded items as the building blocks of a test form being assembled. The assembly of a high-quality test form depends on assuring that the individual building blocks are sound. Numerical indices can be used to summarize, evaluate, and compare a set of items, usually with respect to their difficulties and discriminations. Item difficulty and discrimination indices can also be used to check for potential flaws that may warrant item revision prior to item use in test form assembly. The most well-known and utilized difficulty and discrimination indices of item analysis were developed in the early twentieth century (W. W. Cook 1932; Guilford 1936; Horst 1933; Lentz et al. 1932; Long and Sandiford 1935; Pearson 1909; Symonds 1929; Thurstone 1925). Accounts of ETS scientists Tucker (1987, p. ii), Livingston and Dorans (2004) have described how historical item analysis indices have been applied and adapted at ETS from the mid-1940s to the present day.

2.1.1 Item Difficulty Indices

In their descriptions of item analyses, Gulliksen (1950) and Tucker (1987) listed two historical indices of item difficulty that have been the focus of several applications and adaptations at ETS. These item difficulty indices are defined using the following notation:

i is a subscript indexing the $i = 1$ to I items on Test Y,
j is a subscript indexing the $j = 1$ to N examinees taking Test Y,
x_{ij} indicates a score of 0 or 1 on the ith dichotomously scored Item i from examinee j (all N examinees have scores on all I items).

The most well-known item difficulty index is the average item score, or, for dichotomously scored items, the proportion of correct responses, the "p-value" or "P_+" (Gulliksen 1950; Hambleton 1989; Livingston and Dorans 2004; Lord and Novick 1968; Symonds 1929; Thurstone 1925; Tucker 1987; Wainer 1989):

$$\bar{x}_i = \frac{1}{N}\sum_{j}^{N} x_{ij}. \tag{2.1}$$

Estimates of the quantity defined in Eq. 2.1 can be obtained with several alternative formulas.[1] A more complex formula that is the basis of developments described in Sect. 2.2.1 can be obtained based on additional notation, where.

[1] Alternative expressions to the average item score computations shown in Eq. 2.1 are available in other sources. Expressions involving summations with respect to examinees are shown in Gulliksen (1950) and Lord and Novick (1968). More elaborate versions of Eq. 2.1 that address polytomously scored items and tests composed of both dichotomously and polytomously scored items have also been developed (J. Carlson, personal communication, November 6, 2013).

k is a subscript indexing the $k = 0$ to I possible scores of Test Y (y_k),
$\hat{p}_k$ is the observed proportion of examinees obtaining test score y_k,
$\bar{x}_{ik}$ is the average score on Item i for examinees obtaining test score y_k.

With the preceding notation, the average item score as defined in Eq. 2.1 can be obtained as

$$\bar{x}_i = \sum_k \hat{p}_k \, \bar{x}_{ik}.$$

Alternative item difficulty indices that use a transformation based on the inverse of the cumulative distribution function (CDF) of the normal distribution for the $\bar{x}_i$ in Eq. 2.1 have been proposed by ETS scientists (Gulliksen 1950; Horst 1933) and others (Symonds 1929; Thurstone 1925). The transformation based on the inverse of the CDF of the normal distribution is used extensively at ETS is the delta index developed by Brolyer (Brigham 1932; Gulliksen 1950):

$$\hat{\Delta}_i = 13 - 4\Phi^{-1}\left(\bar{x}_i\right), \tag{2.2}$$

where $\Phi^{-1}(p)$ represents the inverse of the standard normal cumulative distribution corresponding to the pth percentile. ETS scientists Gulliksen (1950, p. 368), Fan (1952, p. 1), Holland and Thayer (1985, p. 1), and Wainer (1989, p. 7) have described deltas as having features that differ from those of average item scores:

- The delta provides an increasing expression of an item's difficulty (i.e., is negatively associated with the average item score).
- The increments of the delta index are less compressed for very easy or very difficult items.
- The sets of deltas obtained for a test's items from two different examinee groups are more likely to be linearly related than the corresponding sets of average item scores.

Variations of the item difficulty indices in Eqs. 2.1 and 2.2 have been adapted and used in item analyses at ETS to address examinee group influences on item difficulty indices. These variations have been described both as actual item difficulty parameters (Gulliksen 1950, pp. 368–371) and as adjustments to existing item difficulty estimates (Tucker 1987, p. iii). One adjustment is the use of a linear function to transform the mean and standard deviation of a set of $\hat{\Delta}_i$ values from one examinee group to this set's mean and standard deviation from the examinee group of interest (Gulliksen 1950; Thurstone 1925, 1947; Tucker 1987):

$$\hat{e}_2\left(\hat{\Delta}_{i,1}\right) = \bar{\Delta}_{.,2} + \frac{\hat{\sigma}_{.,2}\left(\Delta\right)}{\hat{\sigma}_{.,1}\left(\Delta\right)}\left(\hat{\Delta}_{i,1} - \bar{\Delta}_{.,1}\right). \tag{2.3}$$

Equation 2.3 shows that the transformation of Group 1's item deltas to the scale of Group 2's deltas, $\hat{e}_2\left(\Delta_{i,1}\right)$, is obtained from the averages, $\overline{\Delta}_{.,1}$ and $\overline{\Delta}_{.,2}$, and standard deviations, $\hat{\sigma}_{.,1}\left(\Delta\right)$ and $\hat{\sigma}_{.,2}\left(\Delta\right)$, of the groups' deltas. The "mean sigma" adjustment in Eq. 2.3 has been exclusively applied to deltas (i.e., "delta equating"; Gulliksen 1950; Tucker 1987, p. ii) due to the higher likelihood of item deltas to reflect linear relationships between the deltas obtained from two examinee groups on the same set of items. Another adjustment uses Eq. 2.1 to estimate the average item scores for an examinee group that did not respond to those items but has available scores and $\hat{p}_k$ estimates on a total test (e.g., Group 2). Using Group 2's $\hat{p}_k$ estimates and the conditional average item scores from Group 1, which actually did respond to the items and also has scores on the same test as Group 2 (Livingston and Dorans 2004; Tucker 1987), the estimated average item score for Item *i* in Group 2 is

$$\overline{x}_{i,2} = \sum_k \hat{p}_{k,2}\, \overline{x}_{ik,1}. \tag{2.4}$$

The Group 2 adjusted or *reference* average item scores produced with Eq. 2.4 can be subsequently used with Eq. 2.2 to obtain delta estimates for Group 2.

Other measures have been considered as item difficulty indices in item analyses at ETS but have not been used as extensively as those in Eqs. 2.1, 2.2, 2.3, and 2.4. The motivation for considering the additional measures was to expand the focus of Eqs. 2.1, 2.2, and 2.3 beyond item difficulty to address the measurement heterogeneity that would presumably be reflected in relatively low correlations with other items, test scores, or assumed underlying traits (Gulliksen 1950, p. 369; Tucker 1948, 1987, p. iii). Different ways to incorporate items' biserial correlations (described in Sect. 2.1.2) have been considered, including the estimation of item–test regressions to identify the test score that predicts an average item score of 0.50 in an item (Gulliksen 1950). Other proposals to address items' measurement heterogeneity were attempts to incorporate heterogeneity indices into difficulty indices, such as by conducting the delta equating of Eq. 2.3 after dividing the items' deltas by the items' biserial correlations (Tucker 1948) and creating alternative item difficulty indices from the parameter estimates of three-parameter item characteristic curves (Tucker 1981). These additional measures did not replace delta equating in historical ETS practice, partly because of the computational and numerical difficulties in estimating biserial correlations (described later and in Tucker 1987, p. iii), accuracy loss due to computational difficulties in estimating item characteristic curves (Tucker 1981), and interpretability challenges (Tucker 1987, p. vi). Variations of the delta statistic in Eq. 2.2 have been proposed based on logistic cumulative functions rather than normal ogives (Holland and Thayer 1985). The potential benefits of logistic cumulative functions include a well-defined standard error estimate, odds ratio interpretations, and smoother and less biased estimation. These benefits have not been considered substantial enough to warrant a change to wide use of logistic cumulative functions, because the difference between the values of the logistic cumulative function and the normal ogive cumulative function is small

(Haley, cited in Birnbaum 1968, p. 399). In other ETS research by Olson, Scheuneman, and Grima (1989), proposals were made to study items' difficulties after exploratory and confirmatory approaches are used to categorize items into sets based on their content, context, and/or task demands.

2.1.2 Item Discrimination Indices

Indices of item discrimination summarize an item's relationship with a trait of interest. In item analysis, the total test score is almost always used as an approximation of the trait of interest. On the basis of the goals of item analysis to evaluate items, items that function well might be distinguished from those with flaws based on whether the item has a positive versus a low or negative association with the total score. One historical index of the item–test relationship applied in item analyses at ETS is the product moment correlation (Pearson 1895; see also Holland 2008; Traub 1997):

$$\hat{r}\left(x_i,y\right)=\frac{\hat{\sigma}\left(x_i,y\right)}{\hat{\sigma}\left(x_i\right)\hat{\sigma}\left(y\right)}, \tag{2.5}$$

where $\hat{\sigma}\left(x_i,y\right)$, $\hat{\sigma}\left(x_i\right)$, and $\hat{\sigma}\left(y\right)$ denote the estimated covariance and standard deviations of the item scores and test scores. For the dichotomously scored items of interest in this chapter, Eq. 2.5 is referred to as a point biserial correlation, which may be computed as

$$\hat{r}_{\text{point biserial}}\left(x_i,y\right)=\frac{\frac{1}{N}\sum_k N_k\overline{x}_{ik}y_k-\overline{x}_i\overline{y}}{\sqrt{\overline{x}_i\left(1-\overline{x}_i\right)}\hat{\sigma}\left(y\right)}, \tag{2.6}$$

where N and N_k denote the sample sizes for the total examinee group and for the subgroup of examinees obtaining total score y_k and $\overline{x}_i$ and $\overline{y}$ are the means of Item i and the test for the total examinee group. As described in Sect. 2.2.1, the point biserial correlation is a useful item discrimination index due to its direct relationship with respect to test score characteristics.

In item analysis applications, ETS researcher Swineford (1936) described how the point biserial correlation can be a "considerably lowered" (p. 472) measure of item discrimination when the item has an extremely high or low difficulty value. The biserial correlation (Pearson 1909) addresses the lowered point biserial correlation based on the assumptions that (a) the observed scores of Item i reflect an artificial dichotomization of a continuous and normally distributed trait (z), (b) y is normally distributed, and (c) the regression of y on z is linear. The biserial correlation can be estimated in terms of the point biserial correlation and is itself an estimate of the product moment correlation of z and y:

$$\hat{r}_{\text{biserial}}\left(x_i,y\right)=\hat{r}_{\text{point biserial}}\left(x_i,y\right)\frac{\sqrt{\overline{x}_i\left(1-\overline{x}_i\right)}}{\varphi\left(\hat{q}_i\right)}\approx\hat{r}_{zy}, \tag{2.7}$$

where $\varphi\left(\hat{q}_i\right)$ is the density of the standard normal distribution at $\hat{q}_i$ and where $\hat{q}_i$ is the assumed and estimated point that dichotomizes z into x_i (Lord and Novick 1968). Arguments have been made for favoring the biserial correlation estimate over the point biserial correlation as a discrimination index because the biserial correlation is not restricted in range due to Item i's dichotomization and because the biserial correlation is considered to be more invariant with respect to examinee group differences (Lord and Novick 1968, p. 343; Swineford 1936).

Despite its apparent advantages over the point biserial correlation (described earlier), ETS researchers and others have noted several drawbacks to the biserial correlation. Some of the potential drawbacks pertain to the computational complexities the $\varphi\left(\hat{q}_i\right)$ in Eq. 2.7 presented for item analyses conducted prior to modern computers (DuBois 1942; Tucker 1987). Theoretical and applied results revealed the additional problem that estimated biserial correlations could exceed 1 (and be lower than −1, for that matter) when the total test scores are not normally distributed (i.e., highly skewed or bimodal) and could also have high standard errors when the population value is very high (Lord and Novick 1968; Tate 1955a, b; Tucker 1987).

Various attempts have been made to address the difficulties of computing the biserial correlation. Prior to modern computers, these attempts usually involved different uses of punch card equipment (DuBois 1942; Tucker 1987). ETS researcher Turnbull (1946) proposed the use of percentile categorizations of the total test scores and least squares regression estimates of the item scores on the categorized total test scores to approximate Eq. 2.7 and also avoid its computational challenges. In other ETS work, lookup tables were constructed using the average item scores of the examinee groups falling below the 27th percentile or above the 73rd percentile on the total test and invoking bivariate normality assumptions (Fan 1952). Attempts to normalize the total test scores resulted in partially improved biserial correlation estimates but did not resolve additional estimation problems due to the discreteness of the test scores (Tucker 1987, pp. ii–iii, v). With the use of modern computers, Lord (1961) used simulations to evaluate estimation alternatives to Eq. 2.7, such as those proposed by Brogden (1949) and Clemens (1958). Other correlations based on maximum likelihood, ad hoc, and two-step (i.e., combined maximum likelihood and ad hoc) estimation methods have also been proposed and shown to have accuracies similar to each other in simulation studies (Olsson, Drasgow, and Dorans 1982).

The biserial correlation estimate eventually developed and utilized at ETS is from Lewis, Thayer, and Livingston (n.d.; see also Livingston and Dorans 2004). Unlike the biserial estimate in Eq. 2.7, the Lewis et al. method can be used with dichotomously or polytomously scored items, produces estimates that cannot exceed 1, and does not rely on bivariate normality assumptions. This correlation has been referred to as an *r*-polyreg correlation, an *r*-polyserial estimated by regression correlation (Livingston and Dorans 2004, p. 14), and an *r*-biserial correlation for dichotomously scored items. The correlation is based on the assumption that the item scores are determined by the examinee's position on an underlying latent continuous variable *z*. The distribution of *z* for candidates with a given criterion score *y* is assumed to be normal with mean $\beta_i y$ and variance 1, implying the following probit regression model:

$$P\left(x_i \le 1 \middle| y\right) = P\left(z \le_i \alpha \middle| y\right) = \varphi\left(a_i - \beta_i y\right), \tag{2.8}$$

where α_i is the value of *z* corresponding to $x_i = 1$, Φ is the standard normal cumulative distribution function, and a_i and β_i are intercept and slope parameters. Using the maximum likelihood estimate of β_i, the *r*-polyreg correlation can be computed as

$$\hat{r}_{\text{polyreg}}\left(x_i, y\right) = \frac{\sqrt{\hat{\beta}_i^2 \hat{\sigma}_y^2}}{\sqrt{\hat{\beta}_i^2 \hat{\sigma}_y^2 + 1}}, \tag{2.9}$$

where $\hat{\sigma}_y$ is the standard deviation of scores on criterion variable *y* and is estimated in the same group of examinees for which the polyserial correlation is to be estimated. In Olsson et al.'s (1982) terminology, the $\hat{r}_{\text{polyreg}}\left(x_i, y\right)$ correlation might be described as a two-step estimator that uses a maximum likelihood estimate of β_i and the traditional estimate of the standard deviation of *y*.

Other measures of item discrimination have been considered at ETS but have been less often used than those in Eqs. 2.5, 2.6, 2.7 and 2.9. In addition to describing relationships between total test scores and items' correct/incorrect responses, ETS researcher Myers (1959) proposed the use of biserial correlations to describe relationships between total test scores and distracter responses and between total test scores and not-reached responses. Product moment correlations are also sometimes used to describe and evaluate an item's relationships with other items (i.e., phi correlations; Lord and Novick 1968). Alternatives to phi correlations have been developed to address the effects of both items' dichotomizations (i.e., tetrachoric correlations; Lord and Novick 1968; Pearson 1909). Tetrachoric correlations have been used less extensively than phi correlations for item analysis at ETS, possibly due to their assumption of bivariate normality and their lack of invariance advantages (Lord and Novick 1968, pp. 347–349). Like phi correlations, tetrachoric correlations may also be infrequently used as item analysis measures because they describe the relationship of only two test items rather than an item and the total test.

2.2 Item and Test Score Relationships

Discussions of the relationships of item and test score characteristics typically arise in response to a perceived need to expand the focus of item indices. For example, in Sect. 2.1.2, item difficulty indices have been noted as failing to account for items' measurement heterogeneity (see also Gulliksen 1950, p. 369). Early summaries and lists of item indices (W. W. Cook 1932; Guilford 1936; Lentz et al. 1932; Long and Sandiford 1935; Pearson 1909; Richardson 1936; Symonds 1929), and many of the refinements and developments of these item indices from ETS, can be described with little coverage of their implications for test score characteristics. Even when test score implications have been covered in historical discussions, this coverage has usually been limited to experiments about how item difficulties relate to one or two characteristics of test scores (Lentz et al. 1932; Richardson 1936) or to "arbitrary indices" (Gulliksen 1950, p. 363) and "arbitrarily defined" laws and propositions (Symonds 1929, p. 482). In reviewing the sources cited earlier, Gulliksen (1950) commented that "the striking characteristic of nearly all the methods described is that no theory is presented showing the relationship between the validity or reliability of the total test and the method of item analysis suggested" (p. 363).

Some ETS contributions to item analysis are based on describing the relationships of item characteristics to test score characteristics. The focus on relationships of items and test score characteristics was a stated priority of Gulliksen's (1950) review of item analysis: "In developing and investigating procedures of item analysis, it would seem appropriate, first, to establish the relationship between certain item parameters and the parameters of the total test" (p. 364). Lord and Novick (1968) described similar priorities in their discussion of item analysis and indices: "In mental test theory, the basic requirement of an item parameter is that it have a definite (preferably a clear and simple) relationship to some interesting total-test-score parameter" (p. 328). The focus of this section's discussion is summarizing how the relationships of item indices and test form characteristics were described and studied by ETS researchers such as Green Jr. (1951), Gulliksen (1950), Livingston and Dorans (2004), Lord and Novick (1968), Sorum (1958), Swineford (1959), Tucker (1987), Turnbull (1946), and Wainer (1989).

2.2.1 Relating Item Indices to Test Score Characteristics

A test with scores computed as the sum of I dichotomously scored items has four characteristics that directly relate to average item scores and point biserial correlations of the items (Gulliksen 1950; Lord and Novick 1968). These characteristics include Test Y's mean (Gulliksen 1950, p. 367, Eq. 5; Lord and Novick 1968, p. 328, Eq. 15.2.3),

$$\overline{y} = \sum_i \overline{x}_i, \tag{2.10}$$

Test *Y*'s variance (Gulliksen 1950, p. 377, Equation 19; Lord and Novick 1968, p. 330, Equations 15.3.5 and 15.3.6),

$$\hat{\sigma}^2(y) = \sum_i \hat{r}_{\text{point biserial}}(x_i, y)\sqrt{\bar{x}_i(1-\bar{x}_i)}\,\hat{\sigma}(y) = \sum_i \hat{\sigma}(x_i, y) \tag{2.11}$$

Test *Y*'s alpha or KR-20 reliability (Cronbach 1951; Gulliksen 1950, pp. 378–379, Eq. 21; Kuder and Richardson 1937; Lord and Novick 1968, p. 331, Eq. 15.3.8),

$$\hat{\mathrm{r}}\mathrm{el}(y) = \left(\frac{I}{I-1}\right)\left\{1 - \frac{\sum_i \bar{x}_i(1-\bar{x}_i)}{\left[\sum_i \hat{r}_{\text{point biserial}}(x_i, y)\sqrt{\bar{x}_i(1-\bar{x}_i)}\right]^2}\right\}, \tag{2.12}$$

and Test *Y*'s validity as indicated by *Y*'s correlation with an external criterion, *W* (Gulliksen 1950, pp. 381–382, Eq. 24; Lord and Novick 1968, p. 332, Eq. 15.4.2),

$$\hat{r}_{wy} = \frac{\sum_i \hat{r}_{\text{point biserial}}(x_i, w)\sqrt{\bar{x}_i(1-\bar{x}_i)}}{\sum_i \hat{r}_{\text{point biserial}}(x_i, y)\sqrt{\bar{x}_i(1-\bar{x}_i)}}. \tag{2.13}$$

Equations 2.10–2.13 have several implications for the characteristics of an assembled test. The mean of an assembled test can be increased or reduced by including easier or more difficult items (Eq. 2.10). The variance and reliability of an assembled test can be increased or reduced by including items with higher or lower item–test correlations (Eqs. 2.11 and 2.12, assuming fixed item variances). The validity of an assembled test can be increased or reduced by including items with lower or higher item–test correlations (Eq. 2.13).

The test form assembly implications of Eqs. 2.10, 2.11, 2.12 and 2.13 have been the focus of additional research at ETS. Empirical evaluations of the predictions of test score variance and reliability from items' variances and correlations with test scores suggest that items' correlations with test scores have stronger influences than items' variances on test score variance and reliability (Swineford 1959). Variations of Eq. 2.12 have been proposed that use an approximated linear relationship to predict test reliability from items' biserial correlations with test scores (Fan, cited in Swineford 1959). The roles of item difficulty and discrimination have been described in further detail for differentiating examinees of average ability (Lord 1950) and for classifying examinees of different abilities (Sorum 1958). Finally, the correlation of a test and an external criterion shown in Eq. 2.13 has been used to develop methods of item selection and test form assembly based on maximizing test validity (Green 1951; Gulliksen 1950; Horst 1936).

2.2.2 Conditional Average Item Scores

In item analyses, the most detailed descriptions of relationships of items and test scores take the form of $\bar{x}_{ik}$, the average item score conditional on the *k*th score of total test Y (i.e., the discussion immediately following Eq. 2.1). ETS researchers have described these conditional average item scores as response curves (Livingston and Dorans 2004, p. 1), functions (Wainer 1989, pp. 19–20), item–test regressions (Lord 1965b, p. 373), and approximations to item characteristic curves (Tucker 1987, p. ii). Conditional average item scores tend to be regarded as one of the most fundamental and useful outputs of item analysis, because the $\bar{x}_{ik}$ are useful as the basis to calculate in item difficulty indices such as the overall average item score (the variation of Eq. 2.1), item difficulties estimated for alternative examinee groups (Eq. 2.4), and item discrimination indices such as the point biserial correlation (Eq. 2.6). Because the $1-\bar{x}_{ik}$ scores are also related to the difficulty and discrimination indices, the percentages of examinees choosing different incorrect (i.e., distracter) options or omitting the item making up the $1-\bar{x}_{ik}$ scores can provide even more information about the item. Item reviews based on conditional average item scores and conditional proportions of examinees choosing distracters and omitting the item involve relatively detailed presentations of individual items rather than tabled listings of all items' difficulty and discrimination indices for an entire test. The greater detail conveyed in conditional average item scores has prompted consideration of the best approaches to estimation and display of results.

The simplest and most direct approach to estimating and presenting $\bar{x}_{ik}$ and $1-\bar{x}_{ik}$ is based on the raw, unaltered conditional averages at each score of the total test. This approach has been considered in very early item analyses (Thurstone 1925) and also in more current psychometric investigations by ETS researchers Dorans and Holland (1993), Dorans and Kulick (1986), and Moses et al. (2010). Practical applications usually reveal that raw conditional average item scores are erratic and difficult to interpret without reference to measures of sampling instabilities (Livingston and Dorans 2004, p. 12).

Altered versions of $\bar{x}_{ik}$ and $1-\bar{x}_{ik}$ have been considered and implemented in operational and research contexts at ETS. Operational applications favored grouping total test scores into five or six percentile categories, with equal or nearly equal numbers of examinees, and reporting conditional average item scores and percentages of examinees choosing incorrect options across these categories (Tucker 1987; Turnbull 1946; Wainer 1989). Other, less practical alterations of the $\bar{x}_{ik}$ were considered in research contexts based on very large samples (N > 100,000), where, rather than categorizing the y_k scores, the $\bar{x}_{ik}$ values were only presented at total test scores with more than 50 examinees (Lord 1965b). Questions remained about how to present $\bar{x}_{ik}$ and $1-\bar{x}_{ik}$ at the uncategorized scores of the total test while also controlling for sampling variability (Wainer 1989, pp. 12–13).

Other research about item analysis has considered alterations of $\bar{x}_{ik}$ and $1-\bar{x}_{ik}$ (Livingston and Dorans 2004; Lord 1965a, b; Ramsay 1991). Most of these alterations involved the application of models and smoothing methods to reveal trends and eliminate irregularities due to sampling fluctuations in $\bar{x}_{ik}$ and $1-\bar{x}_{ik}$. Relatively strong mathematical models such as normal ogive and logistic functions have been found to be undesirable in theoretical discussions (i.e., the average slope of all test items' conditional average item scores does not reflect the normal ogive model; Lord 1965a) and in empirical investigations (Lord 1965b). Eventually,

> the developers of the ETS system chose a more flexible approach—one that allows the estimated response curve to take the shape implied by the data. Nonmonotonic curves, such as those observed with distracters, can be easily fit by this approach. (Livingston and Dorans 2004, p. 2)

This approach utilizes a special version of kernel smoothing (Ramsay 1991) to replace each $\bar{x}_{ik}$ or $1-\bar{x}_{ik}$ value with a weighted average of all $k = 0$ to I values:

$$KS\left(\bar{x}_{ik}\right)=\left(\sum_{l=0}^{I} w_{kl}\right)^{-1}\sum_{l=0}^{I} w_{kl}\bar{x}_{il}. \tag{2.14}$$

The w_{kl} values of Eq. 2.14 are Gaussian weights used in the averaging,

$$w_{kl}=\exp\left[\frac{-1}{2h}\frac{\left(y_l-y_k\right)^2}{\hat{\sigma}^2\left(y\right)}\right]n_l, \tag{2.15}$$

where exp denotes exponentiation, n_l is the sample size at test score y_l, and h is a kernel smoothing bandwidth parameter determining the extent of smoothing (usually set at $1.1N^{-0.2}$; Ramsay 1991). The rationale of the kernel smoothing procedure is to smooth out sampling irregularities by averaging adjacent $\bar{x}_{ik}$ values, but also to track the general trends in $\bar{x}_{ik}$ by giving the largest weights to the $\bar{x}_{ik}$ values at y scores closest to y_k and at y scores with relatively large conditional sample sizes, n_l. As indicated in the preceding Livingston and Dorans (2004) quote, the kernel smoothing in Eqs. 2.14 and 2.15 is also applied to the conditional percentages of examinees omitting and choosing each distracter that contribute to $1-\bar{x}_{ik}$. Standard errors and confidence bands of the raw and kernel-smoothed versions of $\bar{x}_{ik}$ values have been described and evaluated in Lewis and Livingston (2004) and Moses et al. (2010).

2.3 Visual Displays of Item Analysis Results

Presentations of item analysis results have reflected increasingly refined integrations of indices and conditional response information. In this section, the figures and discussions from the previously cited investigations are reviewed to trace the progression of item analysis displays from pre-ETS origins to current ETS practice.

The original item analysis example is Thurstone's (1925) scaling study for items of the Binet–Simon test, an early version of the Stanford–Binet test (Becker 2003; Binet and Simon 1905). The Binet–Simon and Stanford–Binet intelligence tests represent some of the earliest adaptive tests, where examiners use information they have about an examinee's maturity level (i.e., mental age) to determine where to begin testing and then administer only those items that are of appropriate difficulty for that examinee. The use of multiple possible starting points, and subsets of items, results in limited test administration time and maximized information obtained from each item but also presents challenges in determining how items taken by different examinees translate into a coherent scale of score points and of mental age (Becker 2003).

Thurstone (1925) addressed questions about the Binet–Simon test scales by developing and applying the item analysis methods described in this chapter to Burt's (1921) study sample of 2764 examinees' Binet–Simon test and item scores. Some steps of these analyses involved creating graphs of each of the test's 65 items' proportions correct, $\overline{x}_{ik}$, as a function of examinees' chronological ages, y. Then each item's "at par" (p. 444) age, y_k, is found such that 50% of examinees answered the item correctly, $\overline{x}_{ik} = 0.5$. Results of these steps for a subsample of the items were presented and analyzed in terms of plotted $\overline{x}_{ik}$ values (reprinted in Fig. 2.1).

Thurstone's (1925) analyses included additional steps for mapping all 65 items' at par ages to an item difficulty scale for 3.5-year-old examinees:

1. First the proportions correct of the items taken by 3-year-old, 4-year-old, …, 14-year-old examinees were converted into indices similar to the delta index shown in Eq. 2.2. That is, Thurstone's deltas were computed as $\Delta_{ik} = 0 - (1)\Phi^{-1}(\overline{x}_{ik})$, where the i subscript references the item and the k subscript references the age group responding to the item.

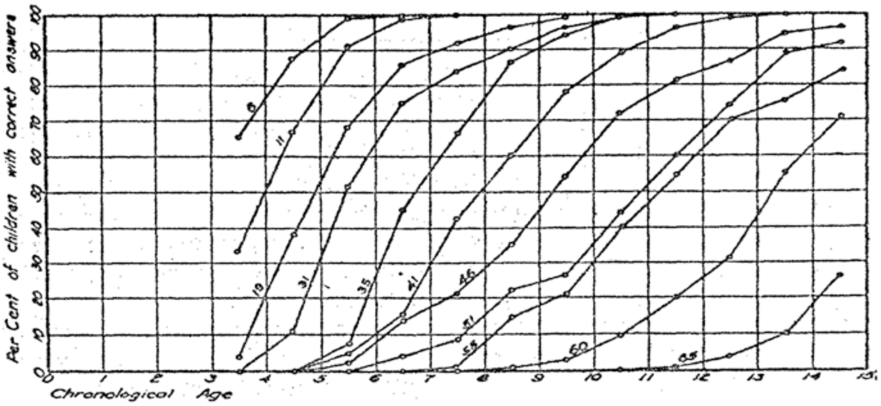


FIG. 5.

Fig. 2.1 Thurstone's (1925) Figure 5, which plots proportions of correct response (vertical axis) to selected items from the Binet–Simon test among children in successive age groups (horizontal axis)

2. For the sets of common items administered to two adjacent age groups (e.g., items administered to 8-year-old examinees and to 7-year-old examinees), the two sets of average item scores, $\bar{x}_{i7}$ and $\bar{x}_{i8}$, were converted into deltas, $\hat{\Delta}_{i7}$ and $\hat{\Delta}_{i8}$.
3. The means and standard deviations of the two sets of deltas from the common items administered to two adjacent age groups (e.g., 7- and 8-year-old examinees) were used with Eq. 2.3 to transform the difficulties of items administered to older examinees to the difficulty scale of items administered to the younger examinees,

$$\hat{e}_7\left(\hat{\Delta}_{i8}\right) = \bar{\Delta}_{.,7} + \frac{\hat{\sigma}_{.,7}\left(\Delta\right)}{\hat{\sigma}_{.,8}\left(\Delta\right)}\left(\hat{\Delta}_{i8} - \bar{\Delta}_{.,8}\right).$$

4. Steps 1–3 were repeated for the two sets of items administered to adjacent age groups from ages 3 to 14 years, with the purpose of developing scale transformations for the item difficulties observed for each age group to the difficulty scale of 3.5-year-old examinees.
5. The transformations obtained in Steps 1–4 for scaling the item difficulties at each age group to the difficulty scale of 3.5-year-old examinees were applied to items' $\hat{\Delta}_{ik}$ and $\bar{x}_{ik}$ estimates nearest to the items' at par ages. For example, with items at an at par age of 7.9, two scale transformations would be averaged, one for converting the item difficulties of 7-year-old examinees to the difficulty scale of 3.5-year-old examinees and another for converting the item difficulties of 8-year-old examinees to the difficulty scale of 3.5-year-old examinees. For items with different at par ages, the scale transformations corresponding to those age groups would be averaged and used to convert to the difficulty scale of 3.5-year-old examinees.

Thurstone (1925) used Steps 1–5 to map all 65 of the Binet–Simon test items to a scale and to interpret items' difficulties for 3.5-year-old examinees (Fig. 2.2). Items 1–7 are located to the left of the horizontal value of 0 in Fig. 2.2, indicating that these items are relatively easy (i.e., have $\bar{x}_{i3.5}$ values greater than 0.5 for the average 3.5-year-old examinee). Items to the right of the horizontal value of 0 in Fig. 2.2 are relatively difficult (i.e., have $\bar{x}_{i3.5}$ values less than 0.5 for the average 3.5-year-old examinee). The items in Fig. 2.2 at horizontal values far above 0 (i.e., greater than the mean item difficulty value of 0 for 3.5-year-old examinees by a given number of standard deviation units) are so difficult that they would not actually be administered to 3.5-year-old examinees. For example, Item 44 was actually administered to examinees 7 years old and older, but this item corresponds to a horizontal value of 5 in Fig. 2.2, implying that its proportion correct is estimated as 0.5 for 3.5-year-old examinees who are 5 standard deviation units more intelligent than the average 3.5-year-old examinee. The presentation in Fig. 2.2 provided empirical evidence that allowed Thurstone (1925) to describe the limitations of assembled forms of Burt–Simon items for measuring the intelligence of examinees at different ability levels and ages: "…the questions are unduly bunched at certain ranges and rather scarce at other ranges" (p. 448). The methods

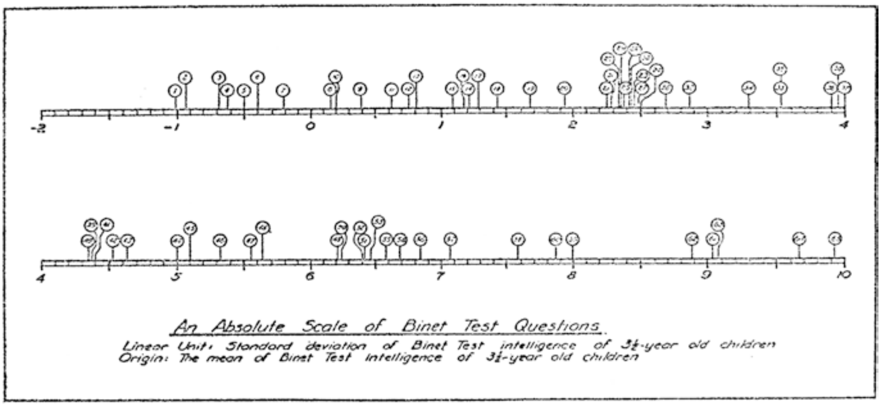


FIG. 6.

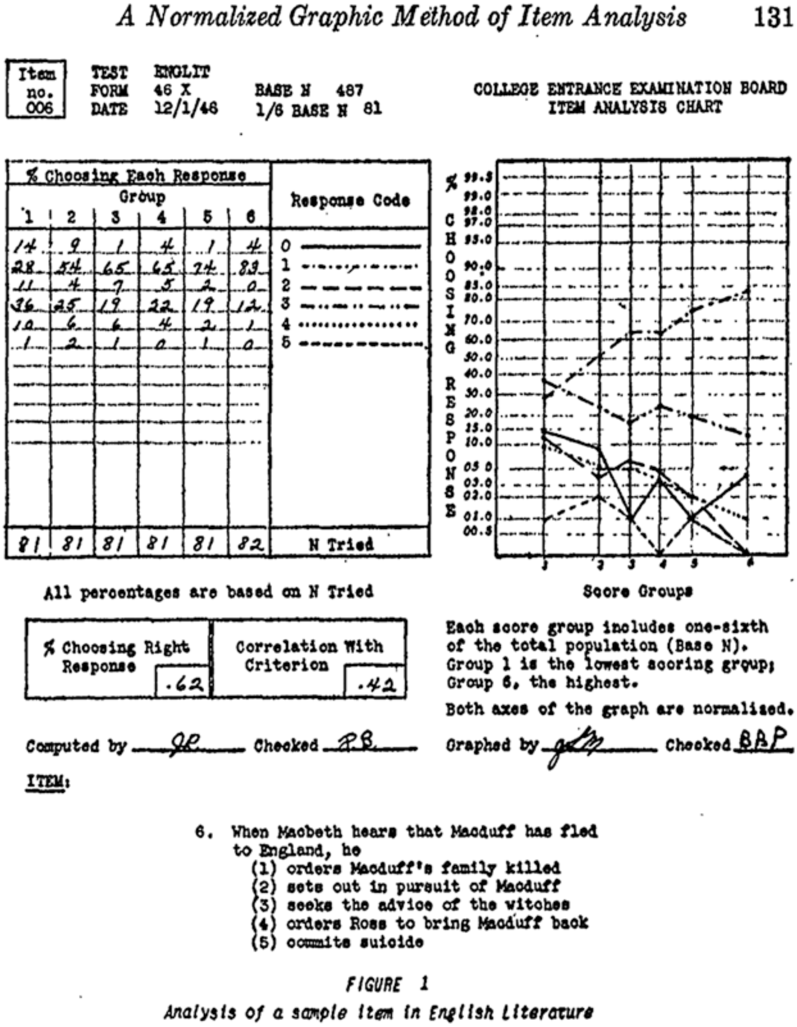
Fig. 2.2 Thurstone's (1925) Figure 6, which represents Binet–Simon test items' average difficulty on an absolute scale

Thurstone (1925) developed, and displayed in Figs. 2.1 and 2.2, were adapted and applied in item analysis procedures used at ETS (Gulliksen 1950, p. 368; Tucker 1987, p. ii).

Turnbull's (1946) presentation of item analysis results for an item from a 1946 College Entrance Examination Board test features an integration of tabular and graphical results, includes difficulty and discrimination indices, and also shows the actual multiple-choice item being analyzed (Fig. 2.3). The graph and table in Fig. 2.3 convey the same information, illustrating the categorization of the total test score into six categories with similar numbers of examinees ($n_k = 81$ or 82). Similar to Thurstone's conditional average item scores (Fig. 2.1), Turnbull's graphical presentation is based on a horizontal axis variable with few categories. The small number of categories limits sampling variability fluctuations in the conditional average item scores, but these categories are labeled in ways that conceal the actual total test scores corresponding to the conditional average item scores. In addition to presenting conditional average item scores, Turnbull's presentation reports conditional percentages of examinees choosing the item's four distracters. Wainer (1989, p. 10) pointed out that the item's correct option is not directly indicated but must be inferred to be the option with conditional scores that monotonically increase with the criterion categories. The item's overall average score (percentage choosing the right response) and biserial correlation, as well as initials of the staff who graphed and checked the results, are also included.

A successor of Turnbull's (1946) item analysis is the ETS version shown in Fig. 2.4 for a 1981 item from the *PSAT/NMSQT*® test (Wainer 1989).[2] The presentation in Fig. 2.4 is completely tabular, with the top table showing conditional sample

[2] In addition to the item analysis issues illustrated in Fig. 2.4 and in Wainer (1989), this particular item was the focus of additional research and discussion, which can be found in Wainer (1983).



A Normalized Graphic Method of Item Analysis **131**

Item no. 006	TEST ENGLIT FORM 46 X DATE 12/1/46	BASE N 487 1/6 BASE N 81	COLLEGE ENTRANCE EXAMINATION BOARD ITEM ANALYSIS CHART

% Choosing Each Response						Response Code
Group						
1	2	3	4	5	6	
14	9	1	4	1	4	0 ———————
28	54	65	65	74	83	1 -·-·-·-·-·-
11	4	7	5	2	0	2 — — — — —
36	25	19	22	19	12	3 —··—··—··—
10	6	6	4	2	1	4 ················
1	2	1	0	1	0	5 - - - - - - - -
81	81	81	81	81	82	N Tried

All percentages are based on N Tried

% Choosing Right Response .62	Correlation With Criterion .42

Each score group includes one-sixth of the total population (Base N). Group 1 is the lowest scoring group; Group 6, the highest.

Both axes of the graph are normalized.

Computed by JR Checked RB Graphed by Checked BAP

ITEM:

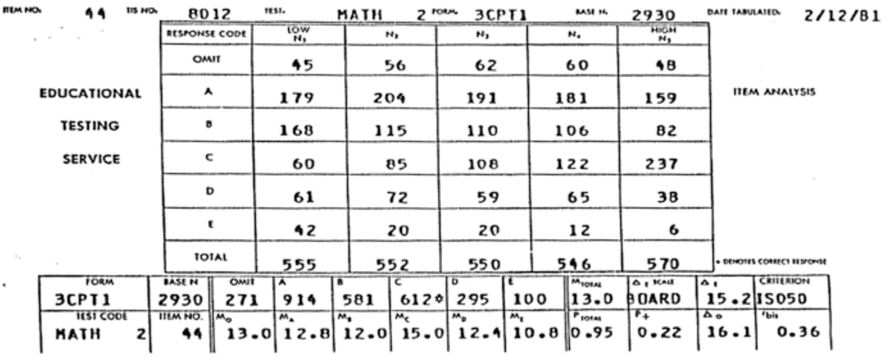
6. When Macbeth hears that Macduff has fled to England, he
   (1) orders Macduff's family killed
   (2) sets out in pursuit of Macduff
   (3) seeks the advice of the witches
   (4) orders Ross to bring Macduff back
   (5) commits suicide

FIGURE 1

Analysis of a sample item in English Literature

Fig. 2.3 Turnbull's (1946) Figure 1, which reports a multiple-choice item's normalized graph (*right*) and table (*left*) for all of its response options for six groupings of the total test score

sizes of examinees choosing the correct option, the distracters, and omitting the item, at five categories of the total test scores (Tucker 1987). The lower table in Fig. 2.4 shows additional overall statistics such as sample sizes and PSAT/NMSQT scores for the group of examinees choosing each option and the group omitting the item, overall average PSAT/NMSQT score for examinees reaching the item (M_{TOTAL}), observed deltas (Δ_O), deltas equated to a common scale using Eq. 2.3 (i.e., "equated deltas," Δ_E), percentage of examinees responding to the item (P_{TOTAL}), percentage of examinees responding correctly to the item (P_+), and the biserial correlation (r_{bis}). The lower table also includes an asterisk with the number of examinees choosing



ITEM NO. 44 TIS NO. 8012 TEST. MATH 2 FORM 3CPT1 BASE N. 2930 DATE TABULATED. 2/12/81

EDUCATIONAL TESTING SERVICE

ITEM ANALYSIS

RESPONSE CODE	LOW N_1	N_2	N_3	N_4	HIGH N_5
OMIT	45	56	62	60	48
A	179	204	191	181	159
B	168	115	110	106	82
C	60	85	108	122	237
D	61	72	59	65	38
E	42	20	20	12	6
TOTAL	555	552	550	546	570

* DENOTES CORRECT RESPONSE

FORM	BASE N	OMIT	A	B	C	D	E	M_{TOTAL}	Δ_E SCALE	Δ_E	CRITERION
3CPT1	2930	271	914	581	612*	295	100	13.0	BOARD	15.2	IS050
TEST CODE	ITEM NO.	M_O	M_A	M_B	M_C	M_D	M_E	P_{TOTAL}	P_+	Δ_O	r_{bis}
MATH 2	44	13.0	12.8	12.0	15.0	12.4	10.8	0.95	0.22	16.1	0.36

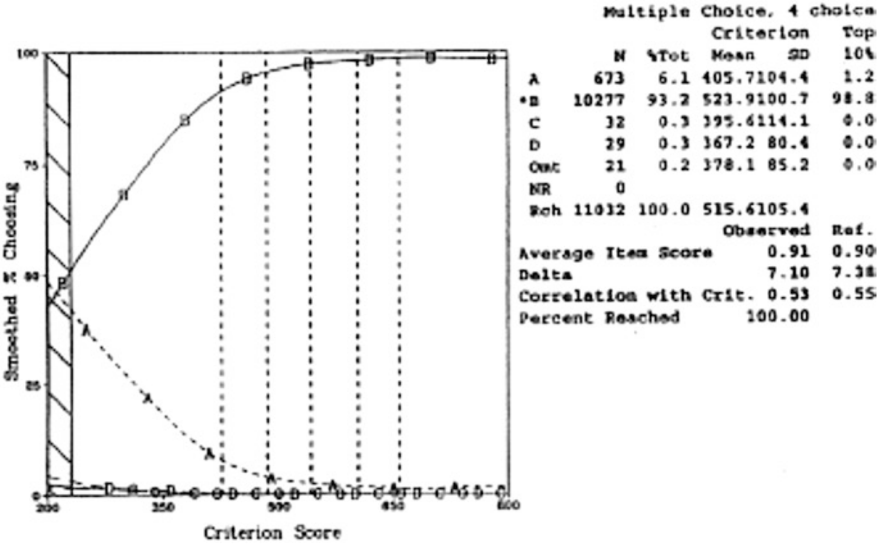
Exhibit 1. Standard ETS Item Analysis Information Strip.

Fig. 2.4 Wainer's (1989) Exhibit 1, which illustrates a tabular display of classical item indices for a PSAT/NMSQT test's multiple-choice item's five responses and omitted responses from 1981

Option C to indicate that Option C is the correct option. Wainer used Turnbull's item presentation (Fig. 2.3) as a basis for critiquing the presentation of Fig. 2.4, suggesting that Fig. 2.4 could be improved by replacing the tabular presentation with a graphical one and also by including the actual item next to the item analysis results.

The most recent versions of item analyses produced at ETS are presented in Livingston and Dorans (2004) and reprinted in Figs. 2.5–2.7. These analysis presentations include graphical presentations of conditional percentages choosing the item's correct option, distracters, omits, and not-reached responses at individual uncategorized criterion scores. The dashed vertical lines represent percentiles of the score distribution where the user can choose which percentiles to show (in this case, the 20th, 40th, 60th, 80th, and 90th percentiles). The figures' presentations also incorporate numerical tables to present overall statistics for the item options and criterion scores as well as observed item difficulty indices, item difficulty indices equated using Eqs. 2.3 and 2.4 (labeled as Ref. in the figures), *r*-biserial correlations ($\hat{r}_{\text{polyreg}}\left(x_i,y\right)$; Eq. 2.9), and percentages of examinees reaching the item. Livingston and Dorans provided instructive discussion of how the item analysis presentations in Figs. 2.5–2.7 can reveal the typical characteristics of relatively easy items (Fig. 2.5), items too difficult for the intended examinee population (Fig. 2.6), and items exhibiting other problems (Fig. 2.7).

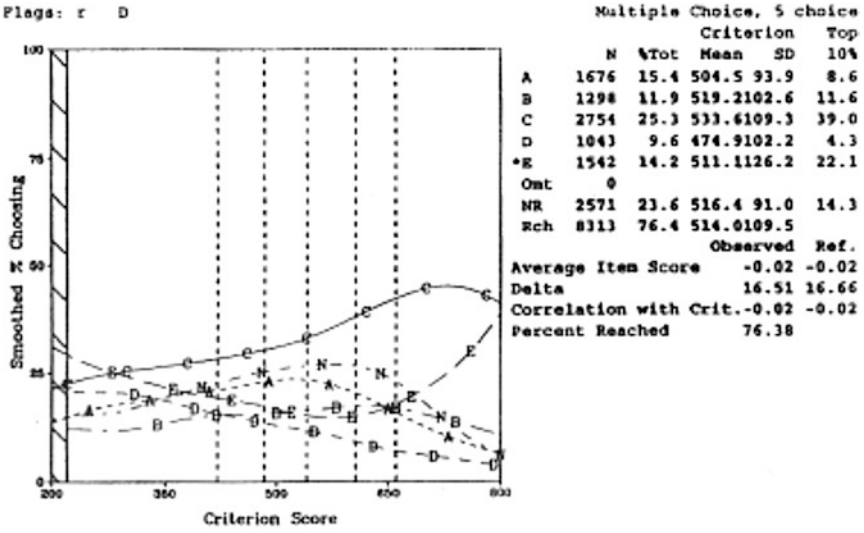
The results of the easy item shown in Fig. 2.5 are distinguished from those of the more difficult items in Figs. 2.6 and 2.7 in that the percentages of examinees choosing the correct option in Fig. 2.5 is 50% or greater for all examinees, and the percentages monotonically increase with the total test score. The items described in Figs. 2.6 and 2.7 exhibit percentages of examinees choosing the correct option that do not obviously rise for most criterion scores (Fig. 2.6) or do not rise more clearly than an intended incorrect option (Fig. 2.7). Livingston and Dorans (2004) interpreted Fig. 2.6 as indicative of an item that is too difficult for the examinees, where examinees do not clearly choose the correct option, Option E, at a higher rate than distracter C, except for the highest total test scores (i.e., the best performing exam-



```
                    Multiple Choice, 4 choice
                              Criterion    Top
          N    %Tot    Mean     SD     10%
A        673    6.1  405.7104.4     1.2
*B     10277   93.2  523.9100.7    98.8
C         32    0.3  395.6114.1     0.0
D         29    0.3  367.2  80.4    0.0
Omt       21    0.2  378.1  85.2    0.0
NR         0
Rch    11032  100.0  515.6105.4
                            Observed   Ref.
Average Item Score             0.91    0.90
Delta                          7.10    7.38
Correlation with Crit.         0.53    0.55
Percent Reached              100.00
```

Figure 1. **An easy item.**

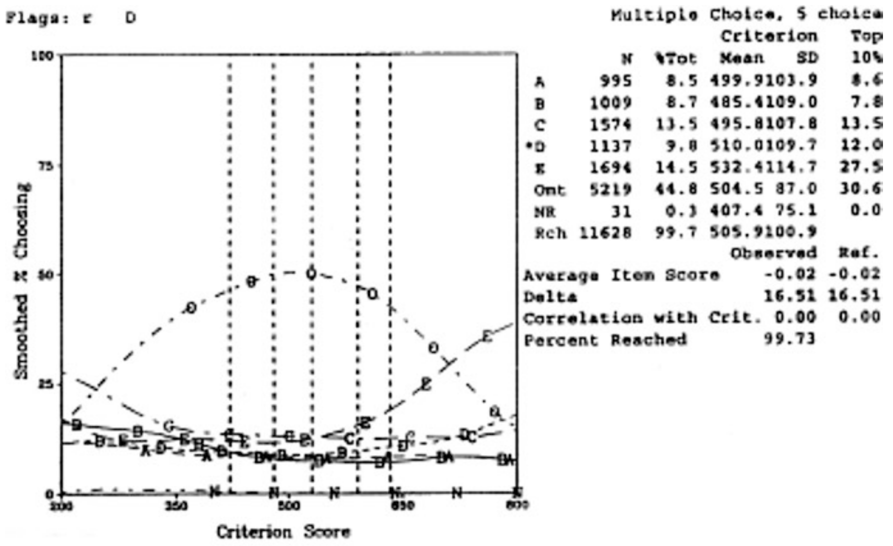
Fig. 2.5 Livingston and Dorans's (2004) Figure 1, which demonstrates classical item analysis results currently used at ETS, for a relatively easy item



```
Flags: r   D
                    Multiple Choice, 5 choice
                              Criterion    Top
          N    %Tot    Mean     SD     10%
A       1676   15.4  504.5  93.9     8.6
B       1298   11.9  519.2102.6    11.6
C       2754   25.3  533.6109.3    39.0
D       1043    9.6  474.9102.2     4.3
*E      1542   14.2  511.1126.2    22.1
Omt        0
NR      2571   23.6  516.4  91.0    14.3
Rch     8313   76.4  514.0109.5
                            Observed   Ref.
Average Item Score            -0.02   -0.02
Delta                         16.51   16.66
Correlation with Crit.-0.02   -0.02
Percent Reached               76.38
```

Figure 5. **An item that is too difficult for the population of examinees.**

Fig. 2.6 Livingston and Dorans's (2004) Figure 5, which demonstrates classical item analysis results currently used at ETS, for a relatively difficult item



```
                 Multiple Choice, 5 choice
                          Criterion    Top
          N   %Tot   Mean     SD     10%
 A      995    8.5  499.9103.9      8.6
 B     1009    8.7  485.4109.0      7.8
 C     1574   13.5  495.8107.8     13.5
*D     1137    9.8  510.0109.7     12.0
 E     1694   14.5  532.4114.7     27.5
 Omt   5219   44.8  504.5 87.0     30.6
 NR      31    0.3  407.4 75.1      0.0
 Rch  11628   99.7  505.9100.9
                         Observed  Ref.
Average Item Score          -0.02 -0.02
Delta                       16.51 16.51
Correlation with Crit.      0.00  0.00
Percent Reached             99.73
```

Figure 7. **An item that does not work for this population.**

Fig. 2.7 Livingston and Dorans's (2004) Figure 7, which demonstrates classical item analysis results currently used at ETS, for a problematic item

inees). Figure 2.7 is interpreted as indicative of an item that functions differently from the skill measured by the test (Livingston and Dorans 2004), where the probability of answering the item correctly is low for examinees at all score levels, where it is impossible to identify the correct answer (D) from the examinee response data, and where the most popular response for most examinees is to omit the item. Figures 2.6 and 2.7 are printed with statistical flags that indicate their problematic results, where the "*r*" flags indicate *r*-biserial correlations that are very low and even negative and the "*D*" flags indicate that high-performing examinees obtaining high percentiles of the criterion scores are more likely to choose one or more incorrect options rather than the correct option.

2.4 Roles of Item Analysis in Psychometric Contexts

2.4.1 Differential Item Functioning, Item Response Theory, and Conditions of Administration

The methods of item analysis described in the previous sections have been used for purposes other than informing item reviews and test form assembly with dichotomously scored multiple-choice items. In this section, ETS researchers' applications of item analysis to psychometric contexts such as differential item functioning

(DIF), item response theory (IRT), and evaluations of item order and context effects are summarized. The applications of item analysis in these areas have produced results that are useful supplements to those produced by the alternative psychometric methods.

2.4.2 Subgroup Comparisons in Differential Item Functioning

Item analysis methods have been applied to compare an item's difficulty for different examinee subgroups. These DIF investigations focus on "unexpected" performance differences for examinee subgroups that are matched in terms of their overall ability or their performance on the total test (Dorans and Holland 1993, p. 37). One DIF procedure developed at ETS is based on evaluating whether two subgroups' conditional average item scores differ from 0 (i.e., standardization; Dorans, and Kulick 1986):

$$\bar{x}_{ik,1} - \bar{x}_{ik,2} \neq 0, \quad k = 0, \ldots, I. \tag{2.16}$$

Another statistical procedure applied to DIF investigations is based on evaluating whether the odds ratios in subgroups for an item i differ from 1 (i.e., the Mantel–Haenszel statistic; Holland and Thayer 1988; Mantel and Haenszel 1959):

$$\frac{\bar{x}_{ik,1} \big/ \left(1 - \bar{x}_{ik,1}\right)}{\bar{x}_{ik,2} \big/ \left(1 - \bar{x}_{ik,2}\right)} \neq 1, \quad k = 0, \ldots, I. \tag{2.17}$$

Most DIF research and investigations focus on averages of Eq. 2.16 with respect to one "standardization" subgroup's total score distribution (Dorans and Holland 1993, pp. 48–49) or averages of Eq. 2.17 with respect to the combined subgroups' test score distributions (Holland and Thayer 1988, p. 134). Summary indices created from Eqs. 2.16 and 2.17 can be interpreted as an item's average difficulty difference for the two matched or standardized subgroups, expressed either in terms of the item's original scale (like Eq. 2.1) or in terms of the delta scale (like Eq. 2.2; Dorans and Holland 1993).

DIF investigations based on averages of Eqs. 2.16 and 2.17 have also been supplemented with more detailed evaluations, such as the subgroups' average item score differences at each of the total test scores indicated in Eq. 2.16. For example, Dorans and Holland (1993) described how the conditional average item score differences in Eq. 2.16 can reveal more detailed aspects of an item's differential functioning, especially when supplemented with conditional comparisons of matched subgroups' percentages choosing the item's distracters or of omitting the item. In ETS practice, conditional evaluations are implemented as comparisons of subgroups' conditional $\bar{x}_{ik}$ and $1 - \bar{x}_{ik}$ values after these values have been estimated with kernel smoothing (Eqs. 2.14 and 2.15). Recent research has shown that evaluations of differences in subgroups' conditional $\bar{x}_{ik}$ values can be biased when estimated with kernel smoothing and that more accurate subgroup comparisons of the conditional $\bar{x}_{ik}$ values can be obtained when estimated with logistic regression or loglinear models (Moses et al. 2010).

2.4.3 *Comparisons and Uses of Item Analysis and Item Response Theory*

Comparisons of item analysis and IRT with respect to methods, assumptions, and results have been an interest of early and contemporary psychometrics (Bock 1997; Embretson and Reise 2000; Hambleton 1989; Lord 1980; Lord and Novick 1968). These comparisons have also motivated considerations for updating and replacing item analysis procedures at ETS. In early years at ETS, potential IRT applications to item analysis were dismissed due to the computational complexities of IRT model estimation (Livingston and Dorans 2004) and also because of the estimation inaccuracies resulting from historical attempts to address the computational complexities (Tucker 1981). Some differences in the approaches' purposes initially slowed the adaptation of IRT to item analysis, as IRT methods were regarded as less oriented to the item analysis goals of item review and revision (Tucker 1987, p. iv). IRT models have also been interpreted to be less flexible in terms of reflecting the shapes of item response curves implied by actual data (Haberman 2009, p. 15; Livingston and Dorans 2004, p. 2).

This section presents a review of ETS contributions describing how IRT compares with item analysis. The contributions are reviewed with respect to the approaches' similarities, the approaches' invariance assumptions, and demonstrations of how item analysis can be used to evaluate IRT model fit. To make the discussions more concrete, the reviews are presented in terms of the following two-parameter normal ogive IRT model:

$$\text{prob}\left(x_i = 1 \middle| \theta, a_i, b_i\right) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt \tag{2.18}$$

where the probability of a correct response to dichotomously scored Item *i* is modeled as a function of an examinee's latent ability, *θ*, Item *i*'s difficulty, b_i, and discrimination, a_i (Lord 1980). Alternative IRT models are reviewed by ETS researchers Lord (1980), Yen and Fitzpatrick (2006), and others (Embretson and Reise 2000; Hambleton 1989).

2.4.3.1 Similarities of Item Response Theory and Item Analysis

Item analysis and IRT appear to have several conceptual similarities. Both approaches can be described as predominantly focused on items and on the implications of items' statistics for assembling test forms with desirable measurement properties (Embretson and Reise 2000; Gulliksen 1950; Wainer 1989; Yen and Fitzpatrick 2006). The approaches have similar historical origins, as the Thurstone (1925) item scaling study that influenced item analysis (Gulliksen 1950; Tucker 1987) has also been described as an antecedent of IRT methods (Bock 1997, pp. 21–23; Thissen and Orlando 2001, pp. 79–83). The kernel smoothing methods used to depict conditional average item scores in item analysis (Eqs. 2.14 and 2.15) were originally developed as an IRT method that is nonparametric with respect to the shapes of its item response functions (Ramsay 1991, 2000).

In Lord and Novick (1968) and Lord (1980), the item difficulty and discrimination parameters of IRT models and item analysis are systematically related, and one can be approximated by a transformation of the other. The following assumptions are made to show the mathematical relationships (though these assumptions are not requirements of IRT models):

- The two-parameter normal ogive model in Eq. 2.18 is correct (i.e., no guessing).
- The regression of x_i on θ is linear with error variances that are normally distributed and homoscedastic.
- Variable θ follows a standard normal distribution.
- The reliability of total score y is high.
- Variable y is linearly related to θ.

With the preceding assumptions, the item discrimination parameter of the IRT model in Eq. 2.18 can be approximated from the item's biserial correlation as

$$a_i \approx \frac{r_{\text{biserial}}\left(x_i, y\right)}{\sqrt{1 - r_{\text{biserial}}\left(x_i, y\right)^2}}. \tag{2.19}$$

With the preceding assumptions, the item difficulty parameter of the IRT model in Eq. 2.18 can be approximated as

$$b_i \approx \frac{l\Delta_i}{r_{\text{biserial}}\left(x_i, y\right)}, \tag{2.20}$$

where $l\Delta_i$ is a linear transformation of the delta (Eq. 2.2). Although IRT does not require the assumptions listed earlier, the relationships in Eqs. 2.19 and 2.20 are used in some IRT estimation software to provide initial estimates in an iterative procedure to estimate $a_i$and b_i (Zimowski et al. 2003).

2.4.3.2 Comparisons and Contrasts in Assumptions of Invariance

One frequently described contrast of item analysis and IRT approaches is with respect to their apparent invariance properties (Embretson and Reise 2000; Hambleton 1989; Yen and Fitzpatrick 2006). A simplified statement of the question of interest is, When a set of items is administered to two not necessarily equal groups of examinees and then item difficulty parameters are estimated in the examinee groups using item analysis and IRT approaches, which approach's parameter estimates are more invariant to examinee group differences? ETS scientists Linda L. Cook, Daniel Eignor, and Hessy Taft (1988) compared the group sensitivities of item analysis deltas and IRT difficulty estimates after estimation and equating using achievement test data, sets of similar examinee groups, and other sets of dissimilar examinee groups. L. L. Cook et al.'s results indicate that equated deltas and IRT models' equated difficulty parameters are similar with respect to their stabilities and their potential for group dependence problems. Both approaches produced inaccurate estimates with very dissimilar examinee groups, results which are consistent with those of equating studies reviewed by ETS scientists L. L. Cook and Petersen (1987) and equating studies conducted by ETS scientists Lawrence and Dorans (1990), Livingston, Dorans, and Nancy Wright (1990), and Schmitt, Cook, Dorans, and Eignor (1990). The empirical results showing that difficulty estimates from item analysis and IRT can exhibit similar levels of group dependence tend to be underemphasized in psychometric discussions, which gives the impression that estimated IRT parameters are more invariant than item analysis indices (Embretson and Reise 2000, pp. 24–25; Hambleton 1989, p. 147; Yen and Fitzpatrick 2006, p. 111).

2.4.3.3 Uses of Item Analysis Fit Evaluations of Item Response Theory Models

Some ETS researchers have suggested the use of item analysis to evaluate IRT model fit (Livingston and Dorans 2004; Wainer 1989). The average item scores conditioned on the observed total test score, $\bar{x}_{ik}$, of interest in item analysis has been used as a benchmark for considering whether the normal ogive or logistic functions assumed in IRT models can be observed in empirical test data (Lord 1965a, b, 1970). One recent application by ETS scientist Sinharay (2006) utilized $\bar{x}_{ik}$ to describe and evaluate the fit of IRT models by considering how well the IRT models' posterior predictions of $\bar{x}_{ik}$ fit the $\bar{x}_{ik}$ values obtained from the raw data. Another recent investigation compared IRT models' $\bar{x}_{ik}$ values to those obtained from loglinear models of test score distributions (Moses 2016).

2.4.4 Item Context and Order Effects

A basic assumption of some item analyses is that items' statistical measures will be consistent if those items are administered in different contexts, locations, or positions (Lord and Novick 1968, p. 327). Although this assumption is necessary for supporting items' administration in adaptive contexts (Wainer 1989), examples in large-scale testing indicate that it is not always tenable (Leary and Dorans 1985; Zwick 1991). Empirical investigations of order and context effects on item statistics have a history of empirical evaluations focused on the changes in IRT estimates across administrations (e.g., Kingston and Dorans 1984). Other evaluations by ETS researchers Dorans and Lawrence (1990) and Moses et al. (2007) have focused on the implications of changes in item statistics on the total test score distributions from randomly equivalent examinee groups. These investigations have a basis in Gulliksen's (1950) attention to how item difficulty affects the distribution of the total test score (Eqs. 2.10 and 2.11). That is, the Dorans and Lawrence (1990) study focused on the changes in total test score means and variances that resulted from changes in the positions of items and intact sections of items. The Moses et al. (2007) study focused on changes in entire test score distributions that resulted from changes in the positions of items and from changes in the positions of intact sets of items that followed written passages.

2.4.5 Analyses of Alternate Item Types and Scores

At ETS, considerable discussion has been devoted to adapting and applying item analysis approaches to items that are not dichotomously scored. Indices of item difficulty and discrimination can be extended, modified, or generalized to account for examinees' assumed guessing tendencies and omissions (Gulliksen 1950; Lord and Novick 1968; Myers 1959). Average item scores (Eq. 2.1), point biserial correlations (Eq. 2.5), *r*-polyreg correlations (Eq. 2.9), and conditional average item scores have been adapted and applied in the analysis of polytomously scored items. Investigations of DIF based on comparing subgroups' average item scores conditioned on total test scores as in Eq. 2.16 have been considered for polytomously scored items by ETS researchers, including Dorans and Schmitt (1993), Moses et al. (2013), and Zwick et al. (1997). At the time of this writing, there is great interest in developing more innovative items that utilize computer delivery and are more interactive in how they engage examinees. With appropriate applications and possible additional refinements, the item analysis methods described in this chapter should have relevance for reviews of innovative item types and for attending to these items' potential adaptive administration contexts, IRT models, and the test forms that might be assembled from them.

Acknowledgments This manuscript was significantly improved from earlier versions thanks to reviews and suggestions from Jim Carlson, Neil Dorans, Skip Livingston and Matthias von Davier, and editorial work from Kim Fryer.

References

Becker, K. A. (2003). *History of the Stanford–Binet intelligence scales: Content and psychometrics* (Stanford–Binet intelligence scales, 5th Ed. Assessment Service Bulletin no. 1). Itasca: Riverside.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du nieveau intellectual anormoux [new methods for the diagnosis of levels of intellectual abnormality]. *L'Année Psychologique, 11*, 191–244. https://doi.org/10.3406/psy.1904.3675.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 374–472). Reading: Addison-Wesley.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x.

Brigham, C. C. (1932). *A study of error*. New York: College Entrance Examination Board.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. *Psychometrika, 14*, 169–182. https://doi.org/10.1007/BF02289151.

Burt, C. (1921). *Mental and scholastic tests*. London: King.

Clemens, W. V. (1958). An index of item-criterion relationship. *Educational and Psychological Measurement, 18*, 167–172. https://doi.org/10.1177/001316445801800118.

Cook, W. W. (1932). *The measurement of general spelling ability involving controlled comparisons between techniques*. Iowa City: University of Iowa Studies in Education.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244. https://doi.org/10.1177/014662168701100302.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31–45. https://doi.org/10.1111/j.1745-3984.1988.tb00289.x.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3*, 245–254. https://doi.org/10.1207/s15324818ame0303_3.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale: Erlbaum.

DuBois, P. H. (1942). A note on the computation of biserial *r* in item validation. *Psychometrika, 7*, 143–146. https://doi.org/10.1007/BF02288074.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale: Erlbaum.

Fan, C.-T. (1952). *Note on construction of an item analysis table for the high-low-27-per-cent group method* (Research Bulletin no. RB-52-13). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1952.tb00227.x

Green, B. F., Jr. (1951). *A note on item selection for maximum validity* (Research Bulletin no. RB-51-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1951.tb00217.x

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. https://doi.org/10.1037/13240-000.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (Research Report No. RR-09-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02172.x

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). Washington, DC: American Council on Education.

Holland, P. W. (2008, March). *The first four generations of test theory*. Paper presented at the ATP Innovations in Testing Conference, Dallas, TX.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00128.x

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

Horst, P. (1933). The difficulty of a multiple choice test item. *Journal of Educational Psychology, 24*, 229–232. https://doi.org/10.1037/h0073588.

Horst, P. (1936). Item selection by means of a maximizing function. *Psychometrika, 1*, 229–244. https://doi.org/10.1007/BF02287875.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147–154. https://doi.org/10.1177/014662168400800202.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*, 19–36. https://doi.org/10.1207/s15324818ame0301_3.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413. https://doi.org/10.3102/00346543055003387.

Lentz, T. F., Hirshstein, B., & Finch, J. H. (1932). Evaluation of methods of evaluating test items. *Journal of Educational Psychology, 23*, 344–350. https://doi.org/10.1037/h0073805.

Lewis, C., & Livingston, S. A. (2004). *Confidence bands for a response probability function estimated by weighted moving average smoothing.* Unpublished manuscript.

Lewis, C., Thayer, D., & Livingston, S. A. (n.d.). *A regression-based polyserial correlation coefficient.* Unpublished manuscript.

Livingston, S. A., & Dorans, N. J. (2004). *A graphical approach to item analysis* (Research Report No. RR-04-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01937.x

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95. https://doi.org/10.1207/s15324818ame0301_6.

Long, J. A., & Sandiford, P. (1935). The validation of test items. *Bulletin of the Department of Educational Research, Ontario College of Education, 3*, 1–126.

Lord, F. M. (1950). *Properties of test scores expressed as functions of the item parameters* (Research Bulletin no. RB-50-56). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00919.x

Lord, F. M. (1961). *Biserial estimates of correlation* (Research Bulletin no. RB-61-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00105.x

Lord, F.M. (1965a). A note on the normal ogive or logistic curve in item analysis. *Psychometrika, 30*, 371–372. https://doi.org/10.1007/BF02289500

Lord, F.M. (1965b). An empirical study of item-test regression. *Psychometrika, 30*, 373–376. https://doi.org/10.1007/BF02289501

Lord, F.M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model. *Psychometrika, 35*, 43–50. https://doi.org/10.1007/BF02290592

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

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

Moses, T. (2016). Estimating observed score distributions with loglinear models. In W. J. van der Linder & R. K. Hambleton (Eds.), *Handbook of item response theory* (2nd ed., pp. 71–85). Boca Raton: CRC Press.

Moses, T., Yang, W., & Wilson, C. (2007). Using kernel equating to check the statistical equivalence of nearly identical test editions. *Journal of Educational Measurement, 44*, 157–178. https://doi.org/10.1111/j.1745-3984.2007.00032.x.

Moses, T., Miao, J., & Dorans, N. J. (2010). A comparison of strategies for estimating conditional DIF. *Journal of Educational and Behavioral Statistics, 6*, 726–743. https://doi.org/10.3102/1076998610379135.

Moses, T., Liu, J., Tan, A., Deng, W., & Dorans, N. J. (2013). *Constructed response DIF evaluations for mixed format tests* (Research Report No. RR-13-33) Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02340.x

Myers, C. T. (1959). *An evaluation of the "not-reached" response as a pseudo-distracter* (Research Memorandum No. RM-59-06). Princeton: Educational Testing Service.

Olson, J. F., Scheuneman, J., & Grima, A. (1989). *Statistical approaches to the study of item difficulty* (Research Report No. RR-89-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00136.x

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*, 337–347. https://doi.org/10.1007/BF02294164.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society, 186*, 343–414. https://doi.org/10.1098/rsta.1895.0010.

Pearson, K. (1909). On a new method for determining the correlation between a measured character a, and a character B. *Biometrika, 7*, 96–105. https://doi.org/10.1093/biomet/7.1-2.96.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630. https://doi.org/10.1007/BF02294494.

Ramsay, J. O. (2000). *TESTGRAF: A program for the graphical analysis of multiple-choice test and questionnaire data* [Computer software and manual]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

Richardson, M. W. (1936). Notes on the rationale of item analysis. *Psychometrika, 1*, 69–76. https://doi.org/10.1007/BF02287926.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3*, 53–71. https://doi.org/10.1207/s15324818ame0301_5.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*, 429–449. https://doi.org/10.1348/000711005X66888.

Sorum, M. (1958). *Optimum item difficulty for a multiple-choice test* (Research memorandum no. RM-58-06). Princeton: Educational Testing Service.

Swineford, F. (1936). Biserial *r* versus Pearson *r* as measures of test-item validity. *Journal of Educational Psychology, 27*, 471–472. https://doi.org/10.1037/h0052118.

Swineford, F. (1959, February). Some relations between test scores and item statistics. *Journal of Educational Psychology, 50*(1), 26–30. https://doi.org/10.1037/h0046332.

Symonds, P. M. (1929). Choice of items for a test on the basis of difficulty. *Journal of Educational Psychology, 20*, 481–493. https://doi.org/10.1037/h0075650.

Tate, R. F. (1955a). Applications of correlation models for biserial data. *Journal of the American Statistical Association, 50*, 1078–1095. https://doi.org/10.1080/01621459.1955.10501293.

Tate, R. F. (1955b). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika, 42*, 205–216. https://doi.org/10.1093/biomet/42.1-2.205.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah: Erlbaum.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451. https://doi.org/10.1037/h0073357.

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist, 3*, 103–104. https://doi.org/10.1037/h0057821.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x.

Tucker, L. R. (1948). A method for scaling ability test items taking item unreliability into account. *American Psychologist, 3*, 309–310.

Tucker, L. R. (1981). *A simulation–Monte Carlo study of item difficulty measures delta and D.6* (Research Report No. RR-81-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01239.x

Tucker, L. R. (1987). *Developments in classical item analysis methods* (Research Report No. RR-87-46). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00250.x

Turnbull, W. W. (1946). A normalized graphic method of item analysis. *Journal of Educational Psychology, 37*, 129–141. https://doi.org/10.1037/h0053589.

Wainer, H. (1983). Pyramid power: Searching for an error in test scoring with 830,000 helpers. *American Statistician, 37*, 87–91. https://doi.org/10.1080/00031305.1983.10483095.

Wainer, H. (1989, Summer). The future of item analysis. *Journal of Educational Measurement, 26*, 191–208.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport: American Council on Education and Praeger.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG [computer software]*. Lincolnwood: Scientific Software International.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP Reading proficiency. *Educational Measurement: Issues and Practice, 10*, 10–16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (Research Report No. RR-97-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1997.tb01726.x

Chapter 3
Psychometric Contributions: Focus on Test Scores

Tim Moses

This chapter is an overview of ETS psychometric contributions focused on test scores, in which issues about items and examinees are described to the extent that they inform research about test scores. Comprising this overview are Sect. 3.1 Test Scores as Measurements and Sect. 3.2 Test Scores as Predictors in Correlational and Regression Relationships. The discussions in these sections show that these two areas are not completely independent. As a consequence, additional contributions are the focus in Sect. 3.3 Integrating Developments About Test Scores as Measurements and Test Scores as Predictors. For each of these sections, some of the most important historical developments that predate and provide context for the contributions of ETS researchers are described.

3.1 Test Scores as Measurements

3.1.1 Foundational Developments for the Use of Test Scores as Measurements, Pre-ETS

By the time ETS officially began in 1947, the fundamental concepts of the classical theory of test scores had already been established. These original developments are usually traced to Charles Spearman's work in the early 1900s (Gulliksen 1950; Mislevy 1993), though Edgeworth's work in the late 1800s is one noteworthy predecessor (Holland 2008). Historical reviews describe how the major ideas of

A version of this chapter was originally published in 2013 by Educational Testing Service as a research report in the ETS R&D Scientific and Policy Contributions Series.

T. Moses (✉)
College Board, New York, NY, USA
e-mail: tmoses@collegeboard.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_3

classical test theory, such as conceptions of test score averages and errors, were borrowed from nineteenth century astronomers and were probably even informed by Galileo's work in the seventeenth century (Traub 1997).

To summarize, the fundamental concepts of classical test theory are that an observed test score for examinee *p* on a particular form produced for test *X*, X'_p, can be viewed as the sum of two independent components: the examinee's true score that is assumed to be stable across all parallel forms of *X*, T_{Xp}, and a random error that is a function of the examinee and is specific to test form *X′*, $E_{X'p}$,

$$X_p^{'} = T_{Xp} + E_{X'p} \tag{3.1}$$

Classical test theory traditionally deals with the hypothetical scenario where examinee *p* takes an infinite number of parallel test forms (i.e., forms composed of different items but constructed to have identical measurement properties, *X′*, *X″*, *X‴*, …). As the examinee takes the infinite number of test administrations, the examinee is assumed to never tire from the repeated testing, does not remember any of the content in the test forms, and does not remember prior performances on the hypothetical test administrations. Under this scenario, classical test theory asserts that means of observed scores and errors for examinee *p* across all the *X′*, *X″*, *X‴*… forms are

$$\mu\left(X_p^{'}\right) = T_{Xp} \; and \; \mu\left(E_{X'p}\right) = 0, \tag{3.2}$$

and the conditional variance for examinee *p* across the forms is

$$\sigma^2_{X_p|T_{X_p}} = \sigma^2_{E_{X_p}} \tag{3.3}$$

The variance of the observed score turns out to be the sum of the true score variance and the error variance,

$$\sigma^2_X = \sigma^2_{T_X} + \sigma^2_{E_X{}'} \tag{3.4}$$

where the covariance of the true scores and errors, $\sigma_{T_X,E_{X_2}}$, is assumed to be zero. Research involving classical test theory often focuses on $\sigma^2_{T_X}$ and $\sigma^2_{E_X}$, meaning that considerable efforts have been devoted to developing approaches for estimating these quantities. The reliability of a test score can be summarized as a ratio of those variances,

$$rel\left(X\right) = \frac{\sigma^2_{T_X}}{\sigma^2_X} = 1 - \frac{\sigma^2_{E_X}}{\sigma^2_X} \tag{3.5}$$

Reliability indicates the measurement precision of a test form for the previously described hypothetical situation involving administrations of an infinite number of parallel forms given to an examinee group.

3.1.2 Overview of ETS Contributions

Viewed in terms of the historical developments summarized in the previous section, many psychometric contributions at ETS can be described as increasingly refined extensions of classical test theory. The subsections in Sect. 3.1 summarize some of the ETS contributions that add sophistication to classical test theory concepts. The summarized contributions have themselves been well captured in other ETS contributions that provide culminating and progressively more rigorous formalizations of classical test theory, including Gulliksen's (1950) *Theory of Mental Tests*, Lord and Novick's (1968) *Statistical Theories of Mental Test Scores*, and Novick's (1965) *The Axioms and Principal Results of Classical Test Theory*. In addition to reviewing and making specific contributions to classical test theory, the culminating formalizations address other more general issues such as different conceptualizations of observed score, true score, and error relationships (Gulliksen 1950), derivations of classical test theory resulting from statistical concepts of sampling, replications and experimental units (Novick 1965), and latent, platonic, and other interpretations of true scores (Lord and Novick 1968). The following subsections of this paper summarize ETS contributions about specific aspects of classical test theory. Applications of these contributions to improvements in the psychometric (measurement) quality of ETS tests are also described.

3.1.3 ETS Contributions About $\sigma_{E_X|T_{X_P}}$

The finding that σ_{E_X} (i.e., the standard error of measurement) may not indicate the actual measurement error for all examinees across all T_{Xp} values is an important, yet often forgotten contribution of early ETS researchers. The belief that classical test theory assumes that $\sigma^2_{E_X|T_{X_p}}$ is constant for all T_{Xp} values has been described as a common misconception (Haertel 2006), and appears to have informed misleading statements about the disadvantages of classical test theory relative to item response theory (e.g., Embretson and Reise 2000, p. 16).

In fact, the variability of the size of tests' conditional standard errors has been the focus of empirical study where actual tests were divided into two halves of equivalent difficulty and length (i.e., tau equivalent, described in Sect. 3.1.5.1), the standard deviation of the differences between the half test scores of examinees grouped by their total scores were computed, and a polynomial regression was fit to the estimated conditional standard errors on the total test scores and graphed (Mollenkopf 1949). By relating the coefficients of the polynomial regression to empirical test score distributions, Mollenkopf showed that conditional standard errors are usually larger near the center of the score distribution than at the tail and may only be expected to be constant for normally distributed and symmetric test-score distributions.

Another contribution to conditional standard error estimation involves assuming a binomial error model for number-correct scores (Lord 1955b, 1957a). If a test is regarded as a random sample of n dichotomously scored items, then the total score for an examinee with a particular true score, $T_{xp,}$ may be modeled as the sum of n draws from a binomial distribution with the probability of success on each draw equal to the average of their scores on the n items. The variance of the number-correct score under this model is binomial,

$$T_{Xp}\left(1-\frac{T_{Xp}}{n}\right). \tag{3.6}$$

The sample estimate of the conditional standard error can be computed by substituting observed scores for true scores and incorporating a correction for the use of the sample estimate of error variance,

$$\sqrt{\frac{X_p\left(n-X_p\right)}{n-1}}. \tag{3.7}$$

It is an estimator of the variance expected across hypothetical repeated measurements for each separate examinee where each measurement employs an independent sample of n items from an infinite population of such items. As such, it is appropriate for absolute or score-focused interpretations for each examinee.

An adjustment to Lord's (1955b, 1957a) conditional standard error for making relative interpretations of examinees' scores in relation to other examinees rather than with respect to absolute true score values was provided by Keats (1957). Noting that averaging Lord's $\frac{X_p\left(n-X_p\right)}{n-1}$ quantity produces the square of the overall standard error of measurement for the Kuder-Richardson Formula 21, $\sigma^2_{Xp}\left[1-rel_{21}\left(X\right)\right]$ (described in Sect. 3.1.5.2), Keats proposed a correction that utilizes the Kuder-Richardson Formula 21 reliability, $rel_{21}(X)$, and any other reliability estimate of interest, $\widehat{rel}\left(X\right)$. The conditional standard error estimate based on Keats' correction,

$$\sqrt{\frac{X_p\left(n-X_p\right)\left[1-\widehat{rel}\left(X\right)\right]}{\left(n-1\right)\left[1-rel_{21}\left(X\right)\right]}}, \tag{3.8}$$

produces a single standard error estimate for each observed score that is appropriate for tests consisting of equally weighted, dichotomously scored items.

3.1.4 *Intervals for True Score Inference*

One application of interest of standard errors of measurement in Sect. 3.1.3 is to true-score estimation, such as in creating confidence intervals for estimates of the true scores of examinees. Tolerance intervals around estimated true scores are attempts to locate the true score at a specified percentage of confidence (Gulliksen 1950). The confidence intervals around true scores formed from overall or conditional standard errors would be most accurate when errors are normally distributed (Gulliksen 1950, p. 17). These relatively early applications of error estimates to true score estimation are questionable, due in part to empirical investigations that suggest that measurement errors are more likely to be binomially distributed rather than normally distributed (Lord 1958a).

For number-correct or proportion-correct scores, two models that do not invoke normality assumptions are the beta-binomial strong true-score model (Lord 1965) and the four-parameter beta model (Keats and Lord 1962). The beta-binomial model builds on the binomial error model described in Sect. 3.1.3. If the observed test score of examinee *p* is obtained by a random sample of *n* items from some item domain, the mean item score is the probability of a correct response to each such randomly chosen item. This fact implies the binomial error model, that the observed score of examinee *p* follows a binomial distribution for the sum of *n* tries with the probability related to the mean for each trial (i.e., the average item score). The four-parameter beta-binomial model is a more general extension of the binomial error model, modeling the true-score distribution as a beta distribution linearly rescaled from the (0,1) interval to the (a,b) interval, $0 \leq a < b \leq 1$. Estimation for two-parameter and four-parameter beta-binomial models can be accomplished by the method of moments (Hanson 1991; Keats and Lord 1962, 1968, Chapter 23). The beta-binomial and four-parameter beta models have had widespread applicability, including not only the construction of tolerance intervals of specified percentages for the true scores of an examinee group (Haertel 2006; Lord and Stocking 1976), but also providing regression-based estimates of true scores (Lord and Novick 1968), and providing estimates of consistency and accuracy when examinees are classified at specific scores on a test (Livingston and Lewis 1995).

3.1.5 *Studying Test Score Measurement Properties With Respect to Multiple Test Forms and Measures*

3.1.5.1 Alternative Classical Test Theory Models

When the measurement properties of the scores of multiple tests are studied, approaches based on the classical test theory model and variations of this model typically begin by invoking assumptions that aspects of the test scores are identical. Strictly parallel test forms have four properties: They are built from identical test specifications, their observed score distributions are identical when administered to any (indefinitely large) population of examinees, they have equal covariances with one another (if there are more than two tests), and they have identical covariances with any other measure of the same or a different construct. Situations with multiple tests that have similar measurement properties but are not necessarily strictly parallel have been defined, and the definitions have been traced to ETS authors (Haertel 2006). In particular, Lord and Novick (1968, p. 48) developed a stronger definition of strictly parallel tests by adding to the requirement of equal covariances that the equality must hold for every subpopulation for which the test is to be used (also in Novick 1965). Test forms can be tau equivalent when each examinee's true score is constant across the forms while the error variances are unequal (Lord and Novick, p. 50). Test forms can be essentially tau equivalent when an examinee's true scores on the forms differ by an additive constant (Lord and Novick, p. 50). Finally, Haertel credits Jöreskog (1971b) for defining a weaker form of parallelism by dropping the requirement of equal true-score variances (i.e., congeneric test forms). That is, congeneric test forms have true scores that are perfectly and linearly related but with possibly unequal means and variances. Although Jöreskog is credited for the official definition of congeneric test form, Angoff (1953) and Kristof (1971) were clearly aware of this model when developing their reliability estimates summarized below.

3.1.5.2 Reliability Estimation

The interest in reliability estimation is often in assessing the measurement precision of a single test form. This estimation is traditionally accomplished by invoking classical test theory assumptions about two or more measures related to the form in question. The scenario in which reliability is interpreted as a measure of score precision when an infinite number of parallel test forms are administered to the same examinees under equivalent administration conditions (see Sect. 3.2.1) is mostly regarded as a hypothetical thought experiment rather than a way to estimate reliability empirically. In practice, reliability estimates are most often obtained as *internal consistency estimates*. This means the only form administered is the one for which reliability is evaluated and variances and covariances of multiple parts constructed from the individual items or half tests on the administered form are obtained while invoking classical test theory assumptions that these submeasures are parallel, tau equivalent, or congeneric.

Many of the popular reliability measures obtained as internal consistency estimates were derived by non-ETS researchers. One of these measures is the Spearman-Brown estimate for a test (X) divided into two strictly parallel halves (X_1 and X_2),

$$\frac{2\rho_{X1,X2}}{1+\rho_{X1,X2}}, \tag{3.9}$$

where $\rho_{X1,X2} = \frac{\sigma_{X1,X2}}{\sigma_{X1}\sigma_{X2}}$ is the correlation of X_1 and X_2 (Brown 1910; Spearman 1910). Coefficient alpha (Cronbach 1951) can be calculated by dividing a test into $i = 1, 2, \ldots, n$ parts assumed to be parallel,

$$\frac{n}{n-1}\left(\frac{\sigma_X^2 - \sum_i \sigma_{X,i}^2}{\sigma_X^2}\right) = \frac{n}{n-1}\left(1 - \frac{\sum_i \sigma_{X,i}^2}{\sigma_X^2}\right). \tag{3.10}$$

Coefficient alpha is known to be a general reliability estimate that produces previously proposed reliability estimates in special cases. For n parts that are all dichotomously scored items, coefficient alpha can be expressed as the Kuder-Richardson Formula 20 reliability (Kuder and Richardson 1937) in terms of the proportion of correct responses on the ith part, $\mu(X_i)$,

$$\frac{n}{n-1}\left(1 - \frac{\sum_i \mu(X_i)\left[1-\mu(X_i)\right]}{\sigma_X^2}\right). \tag{3.11}$$

The Kuder-Richardson Formula 21 (rel_{21} (X)) from Eq. 3.8 in Sect. 3.1.2) can be obtained as a simplification of Eq. 3.11, by replacing each $\mu(X_i)$ for the dichotomously scored items with the mean score on all the items, $\mu(X)$, resulting in

$$\frac{n}{n-1}\left(1 - \frac{\mu(X)\left[n-\mu(X)\right]}{n\sigma_X^2}\right). \tag{3.12}$$

Some ETS contributions to reliability estimation have been made in interpretive analyses of the above reliability approaches. The two Kuder-Richardson formulas have been compared and shown to give close results in practice (Lord 1959b), with the Kuder-Richardson Formula 21 estimate shown by Ledyard R Tucker (1949) always to be less than or equal to the Kuder-Richardson Formula 20 estimate. Cronbach (1951) described his coefficient alpha measure as equal to the mean of all possible split-half reliability estimates, and this feature has been pointed out as eliminating a source of error associated with the arbitrary choice of the split (Lord 1956). Lord (1955b) pointed out that the Kuder-Richardson Formula 21 reliability estimate requires an assumption that all item intercorrelations are equal and went on to show that an average of his binomial estimate of the squared standard errors of measurement can be used in the $1 - \frac{\sigma_{E_X}^2}{\sigma_X^2}$ reliability estimate in Eq. 3.5 to produce the Kuder-Richardson Formula 21 reliability estimate (i.e., the squared values in Eq. 3.7 can be averaged over examinees to estimate $\sigma_{E_X}^2$). Other ETS researchers have pointed out that if the part tests are not essentially tau equivalent, then coefficient alpha is a lower bound to the internal consistency reliability (Novick and Lewis 1967). The worry that internal consistency reliability estimates depend on how closely the parts are to parallel has prompted recommendations for constructing the parts, such as by grouping a test form's items based on their percent-correct score and biserial item-test correlations (Gulliksen 1950). Statistical sampling theory for coefficient alpha was developed by Kristof (1963b; and independently by Feldt 1965). If the coefficient alpha reliability is calculated for a test divided into n strictly parallel parts using a sample of N examinees, then a statistic based on coefficient alpha is distributed as a central F with $N - 1$ and $(n - 1)(N - 1)$ degrees of freedom. This result is exact only under the assumption that part-test scores follow a multivariate normal distribution with equal variances and with equal covariances (the compound symmetry assumption). Kristof (1970) presented a method for testing the significance of point estimates and for constructing confidence intervals for alpha calculated from the division of a test into $n = 2$ parts with unequal variances, under the assumption that the two part-test scores follow a bivariate normal distribution.

The ETS contributions to conditional error variance estimation from Sect. 3.1.2 have been cited as contributors to generalizability (G) theory. G theory uses analysis of variance concepts of experimental design and variance components to reproduce reliability estimates, such as coefficient alpha, and to extend these reliability estimates to address multiple sources of error variance and reliability estimates for specific administration situations (Brennan 1997; Cronbach et al. 1972). A description of the discussion of relative and absolute error variance and of applications of Lord's (1955b, 1957a) binomial error model results (see Sect. 3.1.2) suggested that these ETS contributions were progenitors to G theory:

> The issues Lord was grappling with had a clear influence on the development of G theory. According to Cronbach (personal communication, 1996), about 1957, Lord visited the Cronbach team in Urbana. Their discussions suggested that the error in Lord's formulation of the binomial error model (which treated one person at a time—that is, a completely nested design) could not be the same error as that in classical theory for a crossed design (Lord basically acknowledges this in his 1962 article.) This insight was eventually captured in the distinction between relative and absolute error in G theory, and it illustrated that errors of measurement are influenced by the choice of design. Lord's binomial error model is probably best known as a simple way to estimate conditional SEMs and as an important precursor to strong true score theory, but it is also associated with important insights that became an integral part of G theory. (Brennan 1997, p. 16)

Other ETS contributions have been made by deriving internal consistency reliability estimates based on scores from a test's parts that are not strictly parallel. This situation would seem advantageous because some of the more stringent assumptions required to achieve strictly parallel test forms can be relaxed. However, situations in which the part tests are not strictly parallel pose additional estimation challenges in that the two-part tests, which are likely to differ in difficulty, length, and so on, result in four unknown variances (the true score and error variances of the two parts) that must be estimated from three pieces of information (the variances and the covariance of the part scores). Angoff (1953; also Feldt 1975) addressed this challenge of reliability estimation by assuming that the part tests follow a congeneric model, so that even though the respective lengths of the part tests (i.e., true-score coefficients) cannot be directly estimated, the relative true-score variances and relative error variances of the parts can be estimated as functions of the difference in the effective test lengths of the parts. That is, if one part is longer or shorter than the other part by factor *j*, the proportional true scores of the first and second part differ by *j*, the proportional true-score variances differ by j^2, and the proportional error variances differ by *j*. These results suggest the following reliability coefficient referred to as the Angoff-Feldt coefficient (see Haertel 2006),

$$\frac{4\sigma\left(X_1,X_2\right)}{\sigma_X^2-\frac{\left[\sigma_{X,1}^2-\sigma_{X,2}^2\right]^2}{\sigma_X^2}} \tag{3.13}$$

Angoff also used his results to produce reliability estimates for a whole test, *X*, and an internal part, X_1,

$$\begin{aligned} rel\left(X\right)&=\frac{\rho_{X,X1}\sigma_X-\sigma_{X1}}{\rho_{X,X1}\left(\sigma_X-\rho_{X,X1}\sigma_{X1}\right)}\ and \\ rel\left(X_1\right)&=\frac{\rho_{X,X1}\left(\rho_{X,X1}\sigma_X-\sigma_{X1}\right)}{\sigma_X-\rho_{X,X1}\sigma_{X1}}, \end{aligned} \tag{3.14}$$

and for a whole test *X*, and an external part not contained in *X*, *Y*,

$$\begin{aligned} rel\left(X\right)&=\frac{\rho_{X,Y}\left(\sigma_X+\rho_{X,Y}\sigma_Y\right)}{\sigma_Y+\rho_{X,Y}\sigma_X}\ and \\ rel\left(Y\right)&=\frac{\rho_{X,Y}\left(\sigma_Y+\rho_{X,Y}\sigma_X\right)}{\sigma_X+\rho_{X,Y}\sigma_Y}. \end{aligned} \tag{3.15}$$

The same assumptions later used by Angoff and Feldt were employed in an earlier work by Horst (1951a) to generalize the Spearman-Brown split-half formula to produce a reliability estimate for part tests of unequal but known lengths. Reviews of alternative approaches to reliability estimation when the two-part test lengths are unknown have recommended the Angoff-Feldt estimate in most cases (Feldt 2002).

Kristof made additional contributions to reliability estimation by applying classical test theory models and assumptions (see Sect. 3.1.5.1) to tests divided into more than two parts. He demonstrated that improved statistical precision in reliability estimates could be obtained from dividing a test into more than two tau-equivalent parts (Kristof 1963b). By formulating test length as a parameter in a model for a population covariance matrix of two or more tests, Kristof (1971) described the estimation of test length and showed how to formulate confidence intervals for the relative test lengths. Finally, Kristof (1974) provided a solution to the problem of three congeneric parts of unknown length, where the reliability estimation problem is considered to be just identified, in that there are exactly as many variances and covariances as parameters to be estimated. Kristof's solution was shown to be at least as accurate as coefficient alpha and also gives stable results across alternative partitions. Kristof also addressed the problem of dividing a test into more than three parts of unknown effective test length where the solution is over-determined. Kristof's solution is obtained via maximum-likelihood and numerical methods.

3.1.5.3 Factor Analysis

Some well-known approaches to assessing the measurement properties of multiple tests are those based on factor-analysis models. Factor-analysis models are conceptually like multivariate versions of the classical test theory results in Sect. 3.1.1. Let $\boldsymbol{X}$ denote a q-by-1 column vector with the scores of q tests, $\boldsymbol{\mu}$ denote the q-by-1 vector of means for the q test forms in $\boldsymbol{X}$, $\boldsymbol{\Theta}$ denote a k-by-1 element vector of scores on k common factors, $k < q$, $\boldsymbol{\lambda}$ denote a q-by-k matrix of constants called factor loadings, and finally, let $\mathbf{v}$ denote a q-by-1 row vector of unique factors corresponding to the elements of $\boldsymbol{X}$. With these definitions, the factor-analytic model can be expressed as.

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\lambda}\boldsymbol{\Theta} + \boldsymbol{\nu}, \tag{3.16}$$

and the covariance matrix of $\boldsymbol{X}$, $\boldsymbol{\Sigma}$, can be decomposed into a sum of q-by-q covariance matrices attributable to the common factors ($\boldsymbol{\lambda}\boldsymbol{\Psi}\boldsymbol{\lambda}'$, where $\boldsymbol{\Psi}$ is a k-by-k covariance matrix of the common factors, $\boldsymbol{\Theta}$) and $\boldsymbol{D}^2$ is a diagonal covariance matrix among the uncorrelated unique factors, $\mathbf{v}$,

$$\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\Psi}\boldsymbol{\lambda}' + \boldsymbol{D}^2. \tag{3.17}$$

The overall goal of factor analyses described in Eqs. 3.16 and 3.17 is to meaningfully explain the relationships among multiple test forms and other variables with a small number of common factors (i.e., $k << q$, meaning "k much less than q"). Since Spearman's (1904a) original factor analysis, motivations have been expressed for factor-analysis models that account for observed variables' intercorrelations using one, or very few, common factors. Spearman's conclusions from his factor analysis of scores from tests of abilities in a range of educational subjects (classics, French, English, Math, music, and musical pitch discrimination) and other scores from measures of sensory discrimination to light, sound, and weight were an important basis for describing a range of intellectual abilities in terms of a single, common, general factor:

> We reach the profoundly important conclusion that there really exists a something that we may provisionally term "General Sensory Discrimination" and similarly a "General Intelligence," and further that the functional correspondence between these two is not appreciably less than absolute. (Spearman 1904a, p. 272)

The predominant view regarding factor analysis is as a tool for describing the measurement properties of one or more tests in terms of factors hypothesized to underlie observed variables that comprise the test(s) (Cudeck and MacCallum 2007; Harman 1967; Lord and Novick 1968). Factor analysis models can be viewed as multivariate variations of the classical test theory model described in Sect. 3.1. In this sense, factor analysis informs a "psychometric school" of inquiry, which views a "…battery of tests as a selection from a large domain of tests that could be developed for the same psychological phenomenon and focused on the factors in this domain" (Jöreskog 2007, p. 47). Similar to the classical test theory assumptions, the means of **v** are assumed to be zero, and the variables' covariance matrix, $\boldsymbol{D}^2$, is diagonal, meaning that the unique factors are assumed to be uncorrelated. Somewhat different from the classical test theory model, the unique factors in **v** are not exactly error variables, but instead are the sum of the error factors and specific factors of the q variables. That is, the **v** factors are understood to reflect unreliability (error factors) as well as actual measurement differences (specific factors). The assumption that the **v** factors are uncorrelated implies that the observed covariances between the observed variables are attributable to common factors and loadings, $\boldsymbol{\lambda\Theta}$. The common factors are also somewhat different from the true scores of the variables because the factor-analysis model implies that the true scores reflect common factors as well as specific factors in **v**.

Many developments in factor analysis are attempts to formulate subjective aspects of model selection into mathematical, statistical, and computational solutions. ETS researchers have contributed several solutions pertaining to these interests, which are reviewed in Harman (1967) and in Lord and Novick (1968). In particular, iterative methods have been contrasted and developed for approximating the factor analysis model in observed data by Browne (1969) and Jöreskog (1965, 1967, 1969a; Jöreskog and Lawley 1968), including maximum likelihood, image factor analysis, and alpha factor analysis. An initially obtained factor solution is not uniquely defined, but can be transformed (i.e., rotated) in ways that result in different interpretations of how the factors relate to the observed variables and reproduce the variables' intercorrelations. Contributions by ETS scientists such as Pinzka, Saunders, and Jennrich include the development of different rotation methods that either allow the common factors to be correlated (oblique) or force the factors to remain orthogonal (Browne 1967, 1972a, b; Green 1952; Pinzka and Saunders 1954; Saunders 1953a). The most popular rules for selecting the appropriate number of common factors, k, are based on the values and graphical patterns of factors' eigenvalues, rules that have been evaluated and supported by simulation studies (Browne 1968; Linn 1968; Tucker et al. 1969). Methods for estimating statistical standard errors of estimated factor loadings have been derived (Jennrich 1973; Jennrich and Thayer 1973). Other noteworthy ETS contributions include mathematical or objective formalizations of interpretability in factor analysis (i.e., Thurstone's simple structure, Tucker 1955; Tucker and Finkbeiner 1981), correlation-like measures of the congruence or strength of association among common factors (Tucker 1951), and methods for postulating and simulating data that reflect a factor analysis model in terms of the variables common (major) factors and that also depart from the factor analysis model in terms of several intercorrelated unique (minor) factors (Tucker et al. 1969).

An especially important ETS contribution is the development and naming of confirmatory factor analysis, a method now used throughout the social sciences to address a range of research problems. This method involves fitting and comparing factor-analysis models with factorial structures, constraints, and values specified a priori and estimated using maximum-likelihood methods (Jöreskog 1969b; Jöreskog and Lawley 1968). Confirmatory factor analysis contrasts with the exploratory factor-analysis approaches described in the preceding paragraphs in that confirmatory factor-analysis models are understood to have been specified a priori with respect to the data. In addition, the investigator has much more control over the models and factorial structures that can be considered in confirmatory factor analysis than in exploratory factor analysis. Example applications of confirmatory factor analyses are investigations of the invariance of a factor-analysis solution across subgroups (Jöreskog 1971a) and evaluating test scores with respect to psychometric models (Jöreskog 1969a). These developments expanded factor analyses towards structural-equation modeling, where factors of the observed variables are not only estimated but are themselves used as predictors and outcomes in further analyses (Jöreskog 2007). The LISREL computer program, initially produced by Jöreskog at ETS, was one of the first programs made available to investigators for implementing maximum-likelihood estimation algorithms for confirmatory factor analysis and structural equation models (Jöreskog and van Thillo 1972).

3.1.6 Applications to Psychometric Test Assembly and Interpretation

The ETS contributions to the study of measurement properties of test scores reviewed in the previous sections can be described as relatively general contributions to classical test theory models and related factor-analysis models. Another set of developments has been more focused on applications of measurement theory concepts to the development, use, and evaluation of psychometric tests. These application developments are primarily concerned with building test forms with high measurement precision (i.e., high reliability and low standard errors of measurement).

The basic idea that longer tests are more reliable than shorter tests had been established before ETS (Brown 1910, Spearman 1910; described in Gulliksen 1950 and Mislevy 1993, 1997). ETS researchers developed more refined statements about test length, measurement precision, and scoring systems that maximize reliability. One example of these efforts was establishing that, like reliability, a test's overall standard error of measurement is also directly related to test length, both in theoretical predictions (Lord 1957a) and also in empirical verifications (Lord 1959b). Other research utilized factor-analysis methods to show how reliability for a test of dichotomous items can be maximized by weighting those items by their standardized component loadings on the first principal component (Lord 1958) and how the reliability of a composite can be maximized by weighting the scores for the composite's test battery according to the first principal axis of the correlations and reliabilities of the tests (Green 1950). Finally, conditions for maximizing the reliability of a composite were established, allowing for the battery of tests to have variable lengths and showing that summing the tests after they have been scaled to have equal standard errors of measurement would maximize composite reliability (Woodbury and Lord 1956).

An important limitation of many reliability estimation methods is that they pertain to overall or average score precision. Livingston and Lewis (1995) developed a method for score-specific reliability estimates rather than overall reliability, as score-specific reliability would be of interest for evaluating precision at one or more cut scores. The Livingston and Lewis method is based on taking a test with items not necessarily equally weighted or dichotomously scored and replacing this test with an idealized test consistent with some number of identical dichotomous items. An effective test length of the idealized test is calculated from the mean, variance, and reliability of the original test to produce equal reliability in the idealized test. Scores on the original test are linearly transformed to proportion-correct scores on the idealized test, and the four parameter beta-binomial model described previously is applied. The resulting analyses produce estimates of classification consistency when the same cut scores are used to classify examinees on a hypothetically administered alternate form and estimates of classification accuracy to describe the precision of the cut-score classifications in terms of the assumed true-score distribution.

Statistical procedures have been a longstanding interest for assessing whether two or more test forms are parallel or identical in some aspect of their measurement (i.e., the models in Sect. 3.1.5.1). The statistical procedures are based on evaluating the extent to which two or more test forms satisfy different measurement models when accounting for the estimation error due to inferring from the examinee sample at hand to a hypothetical population of examinees (e.g., Gulliksen 1950, Chapter 14; Jöreskog 2007). ETS researchers have proposed and developed several statistical procedures to assess multiple tests' measurement properties. Kristof (1969) presented iteratively computed maximum-likelihood estimation versions of the procedures described in Gulliksen for assessing whether tests are strictly parallel to also assess if tests are essentially tau equivalent. Procedures for assessing the equivalence of the true scores of tests based on whether their estimated true-score correlation equals 1 have been derived as a likelihood ratio significance test (Lord 1957b) and as F-ratio tests (Kristof 1973). Another F test was developed to assess if two tests differ only with respect to measurement errors, units, and origins of measurement (Lord 1973). A likelihood ratio test was derived for comparing two or more coefficient alpha estimates obtained from dividing two tests each into two part tests with equivalent error variances using a single sample of examinees (Kristof 1964). Different maximum likelihood and chi-square procedures have been developed for assessing whether tests have equivalent overall standard errors of measurement, assuming these tests are parallel (Green 1950), or that they are essentially tau equivalent (Kristof 1963a). Comprehensive likelihood ratio tests for evaluating the fit of different test theory models, including congeneric models, have been formulated within the framework of confirmatory factor-analysis models (Jöreskog 1969a).

3.2 Test Scores as Predictors in Correlational and Regression Relationships

This section describes the ETS contributions to the psychometric study of test scores that are focused on scores' correlations and regression-based predictions to criteria that are not necessarily parallel to the tests. The study of tests with respect to their relationships with criteria that are not necessarily alternate test forms means that test validity issues arise throughout this section and are treated primarily in methodological and psychometric terms. Although correlation and regression issues can be described as if they are parts of classical test theory (e.g., Traub 1997), they are treated as distinct from classical test theory's measurement concepts here because (a) the criteria with which the tests are to be related are often focused on observed scores rather than on explicit measurement models and (b) classical measurement concepts have specific implications for regression and correlation analyses, which are addressed in the next section. Section 3.1.1 reviews the basic correlational and regression developments established prior to ETS. Section 3.2.2 reviews ETS psychometric contributions involving correlation and regression analyses.

3.2.1 Foundational Developments for the Use of Test Scores as Predictors, Pre-ETS

The simple correlation describes the relationship of variables *X* and *Y* in terms of the standardized covariance of these variables, $\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$, and has been traced to the late 1800s work of Galton, Edgeworth, and Pearson (Holland 2008; Traub 1997). The *X,Y* correlation plays a central role in linear regression, the major concepts of which have been credited to the early nineteenth century work of Legendre, Gauss, and Laplace (Holland 2007). The correlation and regression methods establish a predictive relationship of *Y*'s conditional mean to a linear function of *X*,

$$Y = \mu\left(Y|X\right) + \varepsilon = \mu_Y + \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}\left(X - \mu_X\right) + \varepsilon. \tag{3.18}$$

The prediction error, ε, in Eq. 3.18 describes the imprecision of the linear regression function as well as an *X,Y* correlation that is imperfect (i.e., less than 1). Prediction error is different from the measurement errors of *X* and *Y* that reflect unreliability, E_X and E_Y, (Sect. 3.1). The linear regression function in Eq. 3.18 is based on least-squares estimation because using this method results in the smallest possible value of $\sigma_{\varepsilon}^2 = \sigma_Y^2\left[1-\rho_{X,Y}^2\right]$. The multivariate version of Eq. 3.18 is based on predicting the conditional mean of Y from a combination of a set of q observable predictor variables,

$$\boldsymbol{Y} = \mathbf{X}\beta + \varepsilon = \widehat{Y} + \varepsilon, \tag{3.19}$$

where $\boldsymbol{Y}$ is an N-by-1 column vector of the N Y values in the data, $\widehat{Y} = \mathbf{X}\beta$ is an N-by-1 column vector of predicted values ($\widehat{Y}$), $\boldsymbol{X}$ is an N-by-q matrix of values on the predictor variables, $\boldsymbol{\beta}$ is a q-by-1 column vector of the regression slopes of the predictor variables (i.e., scaled semipartial correlations of Y and each X with the relationships to the other Xs partialed out of each X), and $\boldsymbol{\varepsilon}$ is an N-by-1 column vector of the prediction errors. The squared multiple correlation of Y and $\widehat{Y}$ predicted from the Xs in Eqs. 3.18 and 3.19 can be computed given the $\boldsymbol{\beta}$ parameters (or estimated using estimated parameters, $\widehat{\boldsymbol{\beta}}$) as,

$$\rho_{\widehat{Y},Y}^2 = \frac{\sum_{i=1}^{N}\left(\mathbf{X}_i\boldsymbol{\beta}\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{X}_i\boldsymbol{\beta}\right)^2}{\mathbf{Y}^t\mathbf{Y} - \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{Y}_i\right)^2} = 1 - \frac{\sigma_{\varepsilon}^2}{\sigma_Y^2} \tag{3.20}$$

Early applications of correlation and regression concepts dealt with issues such as prediction in astronomy (Holland 2008; Traub 1997) and obtaining estimates of correlations that account for restrictions in the ranges and standard deviations of X and Y (Pearson 1903).

3.2.2 ETS Contributions to the Methodology of Correlations and Regressions and Their Application to the Study of Test Scores as Predictors

The following two subsections summarize ETS contributions about the sample-based aspects of estimated correlations and regressions. Important situations where relationships of tests to other tests and to criteria are of interest involve missing or incomplete data from subsamples of a single population and the feasibility of accounting for incomplete data of samples when those samples reflect distinct populations with preexisting differences. The third subsection deals with ETS contributions that focus directly on detecting group differences in the relationships of tests and what these group differences imply about test validity. The final section describes contributions pertaining to test construction such as determining testing time, weighting subsections, scoring items, and test length so as to maximize test validity.

3.2.2.1 Relationships of Tests in a Population's Subsamples With Partially Missing Data

Some contributions by ETS scientists, such as Gulliksen, Lord, Rubin, Thayer, Horst, and Moses**,** to test-score relationships have established the use of regressions for estimating test data and test correlations when subsamples in a dataset have partially missing data on the test(s) or the criterion. One situation of interest involves examinee subsamples, *R* and *S*, which are missing data on one of two tests, *X* and *Y*, but which have complete data on a third test, *A*. To address the missing data in this situation, regressions of each test onto test *A* can be used to estimate the means and standard deviations of *X* and *Y* for the subsamples with the missing data (Gulliksen 1950; Lord 1955a, c). For example, if group *P* takes tests *X* and *A* and subsample *S* takes only *A*, the mean and variance of the missing *X* scores of *S* can be estimated by applying the *A*-to-*X* regression of subsample *R* to the *A* scores of *S* using the sample statistics in

$$\mu_{X,S} = \mu_{X,R} - \rho_{X,A,R}\frac{\sigma_{X,R}}{\sigma_{A,R}}\left(\mu_{A,R} - \mu_{A,S}\right), \tag{3.21}$$

and

$$\sigma^2_{X,S} = \sigma^2_{X,R} - \left[\rho_{X,A,R}\frac{\sigma_{X,R}}{\sigma_{A,R}}\right]^2\left[\sigma^2_{A,R} - \sigma^2_{A,S}\right]. \tag{3.22}$$

For the more general situation involving a group of standard tests given to an examinee group and one of several new tests administered to random subsamples in the overall group, correlations among all the new and standard tests can be estimated by establishing plausible values for the new tests' partial correlations of the new and standard tests and then using the intercorrelations of the standard tests to "uncondition" the partial correlations and obtain the complete set of simple correlations (Rubin and Thayer 1978, p. 5). Finally, for predicting an external criterion from a battery of tests, it is possible to identify the minimum correlation of an experimental test with the external criterion required to increase the multiple correlation of the battery with that criterion by a specified amount without knowing the correlation of the experimental test with the criterion (Horst 1951c). The fundamental assumption for all of the above methods and situations is that subsamples are randomly selected from a common population, so that other subsamples' correlations of their missing test with other tests and criteria can serve as reasonable estimates of the correlations for the subsamples with missing data.

Regressions and correlations have been regarded as optimal methods for addressing missing test score data in subsamples because under some assumed mathematical model (e.g., normally distributed bivariate or trivariate distributions), regression and correlation estimates maximize the fit of the complete and estimated missing data with the assumed model (Lord 1955a, c; Rubin and Thayer 1978). Thus regressions and correlations can sometimes be special cases of more general maximum-likelihood estimation algorithms for addressing missing data (e.g., the EM algorithm; Dempster et al. 1977). Similar to Lord's (1954b) establishment of linear regression estimates as maximum likelihood estimators for partially missing data, nonlinear regressions estimated with the usual regression methods have been shown to produce results nearly identical to those obtained by using the EM algorithm to estimate the same nonlinear regression models (Moses et al. 2011). It should be noted that the maximum-likelihood results apply to situations involving partially missing data and not necessarily to other situations where a regression equation estimated entirely in one subsample is applied to a completely different, second subsample that results in loss of prediction efficiency (i.e., a larger $\hat{\sigma}^2(\varepsilon)$ for that second subsample; Lord 1950a).

3.2.2.2 Using Test Scores to Adjust Groups for Preexisting Differences

In practice, correlations and regressions are often used to serve interests such as assessing tests taken by subsamples that are likely due to pre-existing population differences that may not be completely explained by *X* or by the study being conducted. This situation can occur in quasi-experimental designs, observational studies, a testing program's routine test administrations, and analyses of selected groups. The possibilities by which preexisting group differences can occur imply that research situations involving preexisting group differences are more likely than subsamples that are randomly drawn from the same population and that have partially missing data (the situation of interest in Sect. 3.2.2.1). The use of correlation and regression for studying test scores and criteria based on examinees with preexisting group differences that have been matched with respect to other test scores has prompted both methodological proposals and discussions about the adequacy of correlation and regression methods for addressing such situations by ETS scientists such as Linn, Charles Werts, Nancy Wright, Dorans, Holland, Rosenbaum, and O'Connor.

Some problems of assessing the relationships among tests taken by groups with preexisting group differences involve a restricted or selected group that has been chosen based either on their criterion performance (explicit selection) or on some third variable (incidental selection, Gulliksen 1950). Selected groups would exhibit performance on tests and criteria that have restricted ranges and standard deviations, thereby affecting these groups' estimated correlations and regression equations. Gulliksen applied Pearson's (1903) ideas to obtain a estimated correlation, prediction error variance, or regression coefficients of the selected group after correcting these estimates for the range-restricted scores of the selected group on *X* and/or *Y*. These corrections for range restrictions are realized by using the *X* and/or *Y* standard deviations from an unselected group in place of those from the selected group.

Concerns have been raised about the adequacy of Gulliksen's (1950) corrections for the statistics of self-selected groups. In particular, the corrections may be inaccurate if the assumed regression model is incorrect (i.e., is actually nonlinear or if the error variance, $\sigma^2(\varepsilon)$, is not constant), or if the corrections are based on a purported selection variable that is not the actual variable used to select the groups (Linn 1967; Lord and Novick 1968). Cautions have been expressed for using the corrections involving selected and unselected groups when those two groups have very different standard deviations (Lord and Novick 1968). The issue of accurately modeling the selection process used to establish the selected group is obviously relevant when trying to obtain accurate prediction estimates (Linn 1983; Linn and Werts 1971; Wright and Dorans 1993).

The use of regressions to predict criterion *Y*'s scores from groups matched on *X* is another area where questions have been raised about applications for groups with preexisting differences. In these covariance analyses (i.e., ANCOVAs), the covariance-adjusted means of the two groups on *Y* are compared, where the adjustment is obtained by applying an *X*-to-*Y* regression using both groups' data to estimate the regression slope ($\rho_{X,Y,R+S}\dfrac{\sigma_{Y,R+S}}{\sigma_{X,R+S}}$) and each group's means ($\mu_{Y,R}$, $\mu_{Y,S}$, $\mu_{X,R}$ and $\mu_{X,S}$) in the estimation and comparison of the groups' intercepts,

$$\mu_{Y,R}-\mu_{Y,S}-\rho_{X,Y,R+S}\frac{\sigma_{Y,R+S}}{\sigma_{X,R+S}}\left(\mu_{X,R}-\mu_{X,S}\right). \tag{3.23}$$

The application of the covariance analyses of Eq. 3.23 to adjust the *Y* means for preexisting group differences by matching the groups on *X* has been criticized for producing results that can, under some circumstances, contradict analyses of average difference scores, $\mu_{Y,R}-\mu_{Y,S}-(\mu_{X,R}-\mu_{X,S})$, (Lord 1967). In addition, covariance analyses have been described as inadequate for providing an appropriate adjustment for the preexisting group differences that are confounded with the study groups and not completely due to *X* (Lord 1969). Attempts have been made to resolve the problems of covariance analysis for groups with preexisting differences. For instance, Novick (1983) elaborated on the importance of making appropriate assumptions about the subpopulation to which individuals are exchangeable members, Holland and Rubin (1983) advised investigators to make their untestable assumptions about causal inferences explicit, and Linn and Werts (1973) emphasized research designs that provide sufficient information about the measurement errors of the variables. Analysis strategies have also been recommended to account for and explain the preexisting group differences with more than one variable using multiple regression (O'Connor 1973), Mahalanobis distances (Rubin 1980), a combination of Mahalanobis distances and regression (Rubin 1979), and propensity-score matching methods (Rosenbaum and Rubin 1984, 1985).

3.2.2.3 Detecting Group Differences in Test and Criterion Regressions

Some ETS scientists such as Schultz, Wilks, Cleary, Frederiksen, and Melville have developed and applied statistical methods for comparing the regression functions of groups. Developments for statistically comparing regression lines of groups tend to be presented in terms of investigations in which the assessment of differences in regressions of groups is the primary focus. Although these developments can additionally be described as informing the developments in the previous section (e.g., establishing the most accurate regressions to match groups from the same population or different populations), these developments tend to describe the applications of matching groups and adjusting test scores as secondary interests. To the extent that groups are found to differ with respect to X,Y correlations, the slopes and/or intercepts of their $Y|X$ regressions and so on, other ETS developments interpret these differences as reflecting important psychometric characteristics of the test(s). Thus these developments are statistical, terminological, and applicative.

Several statistical strategies have been developed for an investigation with the primary focus of determining whether regressions differ by groups. Some statistical significance procedures are based on directly comparing aspects of groups' regression functions to address sequential questions. For example, some strategies center on assessing differences in the regression slopes of two groups and, if the slope differences are likely to be zero, assessing the intercept differences of the groups based on the groups' parallel regression lines using a common slope (Schultz and Wilks 1950). More expansive and general sequential tests involve likelihood ratio and F-ratio tests to sequentially test three hypotheses: first, whether the prediction error variances of the groups are equal; then, whether the regression slopes of the groups are equal (assuming equal error variances), and finally, whether the regression intercepts of the groups are equal (assuming equal error variances and regression slopes; Gulliksen and Wilks 1950). Significance procedures have also been described to consider how the correlation from the estimated regression model in Eq. 3.18, based only on X, might be improved by incorporating a group membership variable, G, as a moderator (i.e., moderated multiple regression; Saunders 1953b),

$$\begin{bmatrix} Y_1 \\ Y_1 \\ . \\ . \\ Y_N \end{bmatrix} = \begin{bmatrix} 1_1 & X_1 & G_1 & X_1G_1 \\ 1_2 & X_2 & G_2 & X_2G_2 \\ . & & & \\ . & & & \\ 1_N & X_N & G_N & X_NG_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_X \\ \beta_G \\ \beta_{XG} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_1 \\ . \\ . \\ e_N \end{bmatrix}. \tag{3.24}$$

Other statistical procedures for assessing group differences include extensions of the Johnson-Neyman procedure for establishing regions of predictor-variable values in which groups significantly differ in their expected criterion scores (Potthoff 1964) and iterative, exploratory procedures for allowing the regression weights of individuals to emerge in ways that maximize prediction accuracy (Cleary 1966a).

The previously described statistical procedures for assessing group differences in regressions have psychometric implications for the tests used as predictors in those regressions. These implications have sometimes been described in terms of test use in which differential predictability investigations have been encouraged that determine the subgroups for which a test is most highly correlated with a criterion and, therefore, most accurate as a predictor of it (Frederiksen and Melville 1954). Other investigators have made particularly enduring arguments that if subgroups are found for which the predictions of a test for a criterion in a total group's regression are inaccurate, the use of that test as a predictor in the total group regression is biased for that subgroup (Cleary 1966b). The statistical techniques in this section, such as moderated multiple regression (Saunders 1953b) for assessing differential predictability and Cleary's test bias,[1] help to define appropriate and valid uses for tests.

3.2.2.4 Using Test Correlations and Regressions as Bases for Test Construction

Interest in test validity has prompted early ETS developments concerned with constructing, scoring, and administering tests in ways that maximized tests' correlations with an external criterion). In terms of test construction, ETS authors such as Gulliksen, Lord, Novick, Horst, Green, and Plumlee have proposed simple, mathematically tractable versions of the correlation between a test and criterion that might be maximized based on item selection (Gulliksen 1950; Horst 1936). Although the correlations to be maximized are different, the Gulliksen and Horst methods led to similar recommendations that maximum test validity can be approximated by selecting items based on the ratio of correlations of items with the criterion and with the total test (Green 1954). Another aspect of test construction addressed in terms of validity implications is the extent to which multiple-choice tests lead to validity reductions relative to open-ended tests (i.e., tests with items that do not present examinees with a set of correct and incorrect options) because of the probability of chance success in multiple-choice items (Plumlee 1954). Validity implications have also been described in terms of the decrement in validity that results when items are administered and scored as the sum of the correct responses of examinees rather than through formulas designed to discourage guessing and to correct examinee scores for random guessing (Lord 1963).

For situations in which a battery of tests are administered under fixed total testing time, several ETS contributions have considered how to determine the length of

[1] Although the summary of Cleary's (1966b) work in this chapter uses the *test bias* phrase actually used by Cleary, it should be acknowledged that more current descriptions of Cleary's regression applications favor different phrases such as prediction bias, overprediction, and underprediction (e.g., Bridgeman et al. 2008). The emphasis of current descriptions on prediction accuracy allows for distinctions to be made between tests that are not necessarily biased but that may be used in ways that result in biased predictions.

each test in ways that maximize the multiple correlation of the battery with an external criterion. These developments have origins in Horst (1951b), but have been extended to a more general and sophisticated solution by Woodbury and Novick (1968). Further extensions deal with computing the composite scores of the battery as the sum of the scores of the unweighted tests in the battery rather than based on the regression weights (Jackson and Novick 1970). These methods have been extensively applied and compared to suggest situations in which validity gains might be worthwhile for composites formed from optimal lengths and regression weights (Novick and Thayer 1969).

3.3 Integrating Developments About Test Scores as Measurements and Test Scores as Predictors

The focus of this section is on ETS contributions that integrate and simultaneously apply measurement developments in Sect. 3.1 and the correlational and regression developments in Sect. 3.2. As previously stated, describing measurement and correlational concepts as if they are completely independent is an oversimplification. Some of the reliability estimates in Sect. 3.1 explicitly incorporate test correlations. In Sect. 3.2, a review of algorithms by Novick and colleagues for determining the lengths of tests in a battery that maximize validity utilize classical test theory assumptions and test reliabilities, but ultimately produce regression and multiple correlation results based on the observed test and criterion scores (Jackson and Novick 1970; Novick and Thayer 1969; Woodbury and Novick 1968). The results by Novick and his colleagues are consistent with other results that have shown that observed-score regressions such as Eq. 3.18 can serve as optimal predictors of the true scores of a criterion (Holland and Hoskens 2003). What distinguishes this section's developments is that measurement, correlational, and regression concepts are integrated in ways that lead to fundamentally unique results.

Integrations of measurement concepts into correlations and regressions build upon historical developments that predate ETS. Spearman's (1904b, 1910) use of classical test theory assumptions to derive an X, Y correlation disattenuated for X and Y's measurement errors (assumed to be independent) is one major influence,

$$\frac{\rho_{X,Y}}{\sqrt{rel(X)rel(Y)}}. \tag{3.25}$$

Kelley's (1923, 1947) regression estimate of the true scores of a variable from its observed scores is another influence,

$$\hat{T}_{Xp} = rel(X)X_p + \left[1 - rel(X)\right]\mu(X) \tag{3.26}$$

Equations 3.25 and 3.26 suggest that some types of analyses that utilize observed scores to compute correlations and regressions can be inaccurate due to measurement errors of *Y*, *X,* or the combination of *Y*, *X,* and additional predictor variables (Moses 2012). Examples of analyses that can be rendered inaccurate when *X* is unreliable are covariance analyses that match groups based on *X* (Linn and Werts 1973) and differential prediction studies that evaluate *X's* bias (Linn and Werts 1971). Lord (1960a) developed an approach for addressing unreliable *X* scores in covariance analyses. In Lord's formulations, the standard covariance analysis model described in Eq. 3.23 is altered to produce an estimate of the covariance results that might be obtained based on a perfectly reliable *X*,

$$\mu_{Y,R} - \mu_{Y,S} - \widehat{\beta}_{T_X}\left(\mu_{X,R} - \mu_{X,S}\right), \tag{3.27}$$

where $\widehat{\beta}_{T_X}$ is estimated as slope disattenuated for the unreliability of *X* based on the classical test theory assumption of *X* having measurement errors independent of measurement errors for *Y*,

$$\widehat{\beta}_{T_X} = \frac{N_R \sigma_{X,Y,R} + N_S \sigma_{X,Y,S}}{N_R rel_R\left(X\right)\sigma^2_{X,R} + N_S rel_S\left(X\right)\sigma^2_{X,S}}\left[1 - \frac{k\left(k-w\right)}{\left(N_R + N_S\right)w^2}\right], \tag{3.28}$$

where

$$k = \frac{N_R \sigma^2_{X,R} + N_S \sigma^2_{X,S}}{N_R + N_S}, w = \frac{N_R rel_R\left(X\right)\sigma^2_{X,R} + N_S rel_S\left(X\right)\sigma^2_{X,S}}{N_R + N_S},$$

and the bracketed term in Eq. 3.28 is a correction for sampling bias. Large sample procedures are used to obtain a sample estimate of the slope in Eq. 3.28 and produce a statistical significance procedure for evaluating Eq. 3.27.

Another ETS contribution integrating measurement, correlation, and regression is in the study of change (Lord 1962a). Regression procedures are described as valuable for estimating the changes of individuals on a measure obtained in a second time period, *Y*, while controlling for the initial statuses of the individuals in a first time period, *X*, *Y* – *X*. Noting that measurement errors can both deflate and inflate regression coefficients with respect to true differences, Lord proposed a multiple regression application to estimate true change from the observed measures, making assumptions that the measurement errors of *X* and *Y* are independent and have the same distributions,

$$\widehat{T}_Y - \widehat{T}_X = \mu\left(Y\right) + \widehat{\beta}_{Y|X}\left[Y - \mu\left(Y\right)\right] - \mu\left(X\right) - \widehat{\beta}_{X|Y}\left[X - \mu\left(X\right)\right], \tag{3.29}$$

where the regression coefficients incorporate disattenuation for the unreliabilities of *X* and *Y*,

$$\widehat{\beta}_{Y|X} = \frac{rel(Y) - \rho_{X,Y}^2 - [1 - rel(X)]\rho_{X,Y}\sigma_X / \sigma_Y}{1 - \rho_{X,Y}^2}, \tag{3.30}$$

$$\widehat{\beta}_{X|Y} = \frac{rel(X) - \rho_{X,Y}^2 - [1 - rel(Y)]\rho_{X,Y}\sigma_Y / \sigma_X}{1 - \rho_{X,Y}^2}. \tag{3.31}$$

Lord also showed that the reliability of the observed change can be estimated as follows (related to the Lord-McNemar estimate of true change, Haertel 2006),

$$rel(Y - X) = \frac{rel(Y)\sigma_Y^2 + rel(X)\sigma_X^2 - 2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_Y^2 + \sigma_X^2 - 2\rho_{X,Y}\sigma_X\sigma_Y}. \tag{3.32}$$

Another ETS contribution, by Shelby Haberman, considers the question of whether subscores should be reported. This question integrates correlational and measurement concepts to determine if the true scores of subscore *X* are better estimated in regressions on the observed scores of the subscore (such as Eq. 3.26), the observed scores of total test *Y*, or a combination of the *X* and *Y* observed scores (Haberman 2008). Extending the results of Lord and Novick (1968) and Holland and Hoskens (2003), versions of the prediction error variance for an *X*-to-*Y* regression, $\sigma_\varepsilon^2 = \sigma_Y^2 \left[1 - \rho_{X,Y}^2\right]$, are produced for the prediction in Eq. 3.26 of the subscore's true score from its observed score,

$$rel(X)\sigma_X^2 [1 - rel(X)], \tag{3.33}$$

and for the prediction from the observed total score, *Y*,

$$rel(X)\sigma_X^2 \left[1 - \rho_{T_X,Y}^2\right] \tag{3.34}$$

The prediction error variance for the regression of the true scores of *X* on both *X* and *Y* is obtained in extensions of Eqs. 3.33 and 3.34,

$$rel(X)\sigma_X^2 [1 - rel(X)]\left[1 - \rho_{Y,T_X.X}^2\right] \tag{3.35}$$

where $\rho_{Y,T_X.X}$ is the partial correlation of the true score of *X* and the observed score of *Y* given the observed score of *X*. Estimates of the correlations in Eqs. 3.34 and 3.35 are obtained somewhat like the disattenuated correlation in Eq. 3.25, but with modifications to account for subscore *X* being contained within total score *Y* (i.e., violations of the classical test theory assumptions of *X* and *Y* having independent measurement errors).

Comparisons of the prediction error variances from Eqs. 3.33, 3.34, and 3.35 produce an indication for when the observed subscore has value for reporting (i.e., when Eq. 3.33 is less than Eqs. 3.34 and 3.35, such as when the subscore has high reliability and a moderate correlation with the total test score). Comparisons of Eqs. 3.33, 3.34 and 3.35 can also suggest when the total test score is a more accurate reflection of the true subscore (i.e., when Eq. 3.34 is less than Eq. 3.33, such as when the subscore has low reliability and/or a high correlation with the total test score). Haberman's (2008) applications to real data from testing programs suggested that the use of the observed scores of the total test is generally more precise than the use of the observed scores of the subscore and also is usually not appreciably worse than the combination of the observed scores of the subscore and the total test.

The final ETS contributions summarized in this section involve true-score estimation methods that are more complex than Kelley's (1923, 1947) linear regression (Eq. 3.26). Some of these more complex true-score regression estimates are based on the tau equivalent classical test theory model, in which frequency distributions are obtained from two or more tests assumed to be tau equivalent and these tests' distributions are used to infer several moments of the tests' true-score and error distributions (i.e., means, variances, skewness, kurtosis, and conditional versions of these; Lord 1959a). Other true-score regression estimates are based on invoking binomial assumptions about a single test's errors and beta distribution assumptions about that test's true scores (Keats and Lord 1962; Lord 1965). These developments imply regressions of true scores on observed scores that are not necessarily linear, though linearity does result when the true scores follow a beta distribution and the observed scores follow a negative hypergeometric distribution. The regressions reflect relationships among true scores and errors that are more complex than assumed in classical test theory, in which the errors are not independent of the true scores and for which attention cannot be restricted only to means, variances, and covariances. Suggested applications for these developments include estimating classification consistency and accuracy (Livingston and Lewis 1995), smoothing observed test score distributions (Hanson and Brennan 1990; Kolen and Brennan 2004), producing interval estimates for true scores (Lord and Novick 1968), predicting test norms (Lord 1962b), and predicting the bivariate distribution of two tests assumed to be parallel (Lord and Novick 1968).

3.4 Discussion

The purpose of this chapter was to summarize more than 60 years of ETS psychometric contributions pertaining to test scores. These contributions were organized into a section about the measurement properties of tests and developments of classical test theory, another section about the use of tests as predictors in correlational and regression relationships, and a third section based on integrating and applying measurement theories and correlational and regression analyses to address test-score issues. Work described in the third section on the integrations of measurement and correlational concepts and their consequent applications, is especially relevant to the operational work of psychometricians on ETS testing programs. Various integrations and applications are used when psychometricians assess a testing program's alternate test forms with respect to their measurement and prediction properties, equate alternate test forms (Angoff 1971; Kolen and Brennan 2004), and employ adaptations of Cleary's (1966b) test bias[2] approach to evaluate the invariance of test equating functions (Dorans and Holland 2000; Myers 1975). Other applications are used to help testing programs face increasing demand for changes that might be supported with psychometric methods based on the fundamental measurement and regression issues about test scores covered in this chapter.

One unfortunate aspect of this undertaking is the large number of ETS psychometric contributions that were not covered. These contributions are difficult to describe in terms of having a clear and singular focus on scores or other issues, but they might be accurately described as studies of the interaction of items and test scores. The view of test scores as a sum of items suggests several ways in which an item's characteristics influence test-score characteristics. Some ETS contributions treat item and score issues almost equally and interactively in describing their relationships, having origins in Gulliksen's (1950) descriptions of how item statistics influence test score means, standard deviations, reliability, and validity. ETS researchers such as Swineford, Lord, and Novick have clarified Gulliksen's descriptions through empirically estimated regression functions that predict test score standard deviations and reliabilities from correlations of items and test scores, through item difficulty statistics (Swineford 1959), and through mathematical functions derived to describe the influence of items with given difficulty levels on the moments of test-score distributions (Lord 1960b; Lord and Novick 1968). Other mathematical functions describe the relationships of the common factor of the items to the discrimination, standard error of measurement, and expected scores of the test (Lord 1950b). Using item response theory (IRT) methods that focus primarily on items rather than scores, ETS researchers (see the chapter on ETS contributions to IRT in this volume) have explained the implications of IRT item models for test-score characteristics, showing how observed test score distributions can be estimated from IRT models (Lord and Wingersky 1984) and showing how classical test theory results can be directly obtained from some IRT models (Holland and Hoskens 2003).

The above contributions are not the only ones dealing with interactions between scores, items, and/or fairness. Similarly, advances such as differential item functioning (DIF) can be potentially described with respect to items, examinees, and item-examinee interactions. Developments such as IRT and its application to adaptive testing can be described in terms of items and using item parameters to estimate examinees' abilities as the examinees interact with and respond to the items. ETS

[2] Although the summary of Cleary's (1966b) work in this chapter uses the *test bias* phrase actually used by Cleary, it should be acknowledged that more current descriptions of Cleary's regression applications favor different phrases such as prediction bias, overprediction, and underprediction (e.g., Bridgeman et al. 2008). The emphasis of current descriptions on prediction accuracy allows for distinctions to be made between tests that are not necessarily biased but that may be used in ways that result in biased predictions.

contributions to DIF and to IRT are just two of several additional areas of psychometrics summarized in other chapters (Carlson and von Davier, Chap. 5, this volume; Dorans, Chap. 7, this volume).

References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*, 1–14. https://doi.org/10.1007/BF02289023

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14–20. https://doi.org/10.1111/j.1745-3992.1997.tb00604.x

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.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Browne, M. W. (1967). On oblique procrustes rotation. *Psychometrika, 32*, 125–132. https://doi.org/10.1007/BF02289420

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33,* 267–334. https://doi.org/10.1007/BF02289327

Browne, M. W. (1969) Fitting the factor analysis model. *Psychometrika, 34*, 375. https://doi.org/10.1007/BF02289365

Browne, M. W. (1972a). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25,* 207–212. https://doi.org/10.1111/j.2044-8317.1972.tb00492.x

Browne, M. W. (1972b). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25,* 115–120. https://doi.org/10.1111/j.2044-8317.1972.tb00482.x

Cleary, T. A. (1966a). An individual differences model for multiple regression. *Psychometrika, 31*, 215–224. https://doi.org/10.1007/BF02289508

Cleary, T. A. (1966b). *Test bias: Validity of the Scholastic Aptitude Test for Negro and White students in integrated colleges* (Research Bulletin No. RB-66-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1966.tb00529.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cudeck, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah: Erlbaum.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–22.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale: Erlbaum.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370. https://doi.org/10.1007/BF02289499

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika, 40,* 557–561. https://doi.org/10.1007/BF02291556

Feldt, L. S. (2002). Reliability estimation when a test is split into two parts of unknown effective length. *Applied Measurement in Education, 15,* 295–308. https://doi.org/10.1207/S15324818AME1503_4

Frederiksen, N., & Melville, S.D. (1954). Differential predictability in the use of test scores. *Educational and Psychological Measurement, 14*, 647–656. https://doi.org/10.1177/001316445401400040

Green, B. F., Jr. (1950). A test of the equality of standard errors of measurement. *Psychometrika, 15*, 251–257. https://doi.org/10.1007/BF02289041

Green, B. F., Jr. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika, 17*, 429–440. https://doi.org/10.1007/BF02288918

Green, B. F., Jr. (1954). A note on item selection for maximum validity. *Educational and Psychological Measurement, 14*, 161–164. https://doi.org.10.1177/001316445401400116

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. https://doi.org/10.1037/13240-000

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114. https://doi.org/10.1007/BF02289195

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33,* 204–229. https://doi.org/10.3102/1076998607302636

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport: American Council on Education and Praeger.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Research Report No. 91–5). Iowa City: American College Testing Program.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345–359. https://doi.org/10.1111/j.1745-3984.1990.tb00753.x

Harman, H. H. (1967). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_2

Holland, P. W. (2008, March). *The first four generations of test theory*. Presentation at the ATP Innovations in Testing Conference, Dallas, TX.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Applications to true-score prediction from a possibly nonparallel test. *Psychometrika, 68,* 123–149. https://doi.org/10.1007/BF02296657

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 Frederic M. Lord* (pp. 3–25). Hillsdale: Erlbaum.

Horst, P. (1936). Item selection by means of a maximizing function. *Psychometrika, 1,* 229–244. https://doi.org/10.1007/BF02287875

Horst, P. (1951a). Estimating total test reliability from parts of unequal length. *Educational and Psychological Measurement, 11,* 368–371. https://doi.org/10.1177/001316445101100306

Horst, P. (1951b). Optimal test length for maximum battery validity. *Psychometrika, 16,* 189–202. https://doi.org/10.1007/BF02289114

Horst, P. (1951c). The relationship between the validity of a single test and its contribution to the predictive efficiency of a test battery. *Psychometrika, 16*, 57–66. https://doi.org/10.1007/BF02313427

Jackson, P. H., & Novick, M. R. (1970). Maximizing the validity of a unit-weight composite as a function of relative component lengths with a fixed total testing time. *Psychometrika, 35*, 333–347. https://doi.org/10.1007/BF02310793

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, *38,* 593–604. https://doi.org/10.1007/BF02291497

Jennrich, R. I., & Thayer, D. T. (1973). A note on Lawley's formulas for standard errors in maximum likelihood factor analysis. *Psychometrika*, *38,* 571–592. https://doi.org/10.1007/BF02291495

Jöreskog, K. G. (1965). *Image factor analysis* (Research Bulletin No RB-65-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1965.tb00134.x

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G. (1969a). Efficient estimation in image factor analysis. *Psychometrika*, *34,* 51–75. https://doi.org/10.1007/BF02290173

Jöreskog, K. G. (1969b). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K.G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika*, *36,* 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K.G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36,* 109–133. https://doi.org/10.1007/BF02291393

Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 47–77). Mahwah: Erlbaum.

Jöreskog, K. G., & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology, 21*, 85–96. https://doi.org/10.1111/j.2044-8317.1968.tb00399.x

Jöreskog, K. G., & van Thillo, M. (1972). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Research Bulletin No. RB-72-56). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1972.tb00827.x

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, *22,* 29–41. https://doi.org/10.1007/BF02289207

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, 27*,* 59–72. https://doi.org/10.1007/BF02289665

Kelley, T. L. (1923). *Statistical methods*. New York: Macmillan.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Kristof, W. (1963a). Statistical inferences about the error variance. Psychometrika, 28, 129–143. https://doi.org/10.1007/BF02289611

Kristof, W. (1963b). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28,* 221–238. https://doi.org/10.1007/BF02289571

Kristof, W. (1964). Testing differences between reliability coefficients. *British Journal of Statistical Psychology, 17,* 105–111. https://doi.org/10.1111/j.2044-8317.1964.tb00253.x

Kristof, W. (1969). Estimation of true score and error variance for tests under various equivalence assumptions. *Psychometrika, 34*, 489–507. https://doi.org/10.1007/BF02290603

Kristof, W. (1970). On the sampling theory of reliability estimation. *Journal of Mathematical Psychology, 7,* 371–377. https://doi.org/10.1016/0022-2496(70)90054-4

Kristof, W. (1971). On the theory of a set of tests which differ only in length. *Psychometrika, 36,* 207–225. https://doi.org/10.1007/BF02297843

Kristof, W. (1973). Testing a linear relation between true scores of two measures. *Psychometrika, 38*, 101–111. https://doi.org/10.1007/BF02291178

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39,* 491–499. https://doi.org/10.1007/BF02291670

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160. https://doi.org/10.1007/BF02288391

Linn, R. L. (1967). *Range restriction problems in the validation of a guidance test battery* (Research Bulletin No. RB-67-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1967.tb00149.x

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika, 33,* 33–71. https://doi.org/10.1007/BF02289675

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale: Erlbaum.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4. https://doi.org/10.1007/BF02289675

Linn, R. L., & Werts, C. E. (1973). Errors of inference due to errors of measurement. *Educational and Psychological Measurement, 33*, 531–543. https://doi.org/10.1177/001316447303300301

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Lord, F. M. (1950a). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. RB-50-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00478.x

Lord, F. M. (1950b). *Properties of test scores expressed as functions of the item parameters* (Research Bulletin No. RB-50-56). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00919.x

Lord, F. M. (1955a). Equating test scores—A maximum likelihood solution. *Psychometrika, 20*, 193–200. https://doi.org/10.1007/BF02289016

Lord, F. M. (1955b). *Estimating test reliability* (Research Bulletin No. RB-55-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1955.tb00054.x

Lord, F. (1955c). Estimation of parameters from incomplete data. *Journal of the American Statistical Association, 50*, 870–876. https://doi.org/10.2307/2281171

Lord, F. M. (1956). Sampling error due to choice of split in split-half reliability coefficients. *Journal of Experimental Education, 24,* 245–249. https://doi.org/10.1080/00220973.1956.11010545

Lord, F. M. (1957a). Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement, 17*, 510–521. https://doi.org/10.1177/001316445701700407

Lord, F. M. (1957b). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika, 22*, 207–220. https://doi.org/10.1007/BF02289122

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291–296. https://doi.org/10.1007/BF02289779

Lord, F. M. (1959a). Statistical inferences about true scores. *Psychometrika, 24*, 1–17. https://doi.org/10.1007/BF02289759

Lord, F. M. (1959b). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement, 19,* 233–239. https://doi.org/10.1177/001316445901900208

Lord, F. M. (1960). An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika, 25*, 91–104. https://doi.org/10.1007/BF02288936

Lord, F. M. (1960a). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307–321. https://doi.org/10.1080/01621459.1960.10482065

Lord, F. M. (1960b). Use of true-score theory to predict moments of univariate and bivariate observed score distributions. *Psychometrika, 25*, 325–342. https://doi.org/10.1007/BF02289751

Lord, F. M. (1962a). *Elementary models for measuring change*. (Research Memorandum No. RM-62-05). Princeton: Educational Testing Service.

Lord. F. M. (1962b). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*, 259–267. https://doi.org/10.1177/001316446202200202

Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 23*, 663–672. https://doi.org/10.1177/001316446302300403

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 30,* 239–270. https://doi.org/10.1007/BF02289490

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–305.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72*, 336–337. https://doi.org/10.1037/h0028108

Lord, F. M. (1973). Testing if two measuring procedures measure the same dimension. *Psychological Bulletin, 79*, 71–72. https://doi.org/10.1037/h0033760

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lord, F. M., & Stocking, M. (1976). An interval estimate for making statistical inferences about true score. *Psychometrika, 41,* 79–87. https://doi.org/10.1007/BF02291699

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461. https://doi.org/10.1177/014662168400800409

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale: Erlbaum.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14,* 189–229. https://doi.org/10.1007/BF02289153

Moses, T. (2012). Relationships of measurement error and prediction error in observed-score regression. *Journal of Educational Measurement, 49,* 380–398. https://doi.org/10.1111/j.1745-3984.2012.00182.x

Moses, T., Deng, W., & Zhang, Y.-L. (2011). Two approaches for using multiple anchors in NEAT equating. *Applied Psychological Measurement, 35*, 362–379. https://doi.org/10.1177/0146621611405510

Myers, C. T. (1975). *Test fairness: A comment on fairness in statistical analysis* (Research Bulletin No. RB-75-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01051.x

Novick, M. R. (1965). *The axioms and principal results of classical test theory* (Research Bulletin No. RB-65-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1965.tb00132.x

Novick, M. R. (1983). The centrality of Lord's paradox and exchangeability for all statistical inference. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 41–53). Hillsdale: Erlbaum.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1–13. https://doi.org/10.1007/BF02289400

Novick, M. R., & Thayer, D. T. (1969). *Some applications of procedures for allocating testing time* (Research Bulletin No. RB-69-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1969.tb00161.x

O'Connor, E. F. (1973). *Unraveling Lord's paradox: The appropriate use of multiple regression analysis in quasi-experimental research* (Research Bulletin No. RB-73-53). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1973.tb00839.x

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophicali. Transactions 200-A,* 1–66. London: Royal Society

Pinzka, C., & Saunders, D. R. (1954). *Analytic rotation to simple structure: II. Extension to an oblique solution* (Research Bulletin No. RB-54-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1954.tb00487.x

Plumlee, L. B. (1954). Predicted and observed effect of chance on multiple-choice test validity. *Psychometrika, 19,* 65–70. https://doi.org/10.1007/BF02288994

Potthoff, R. F. (1964). On the Johnson-Neyman technique and some extensions thereof. *Psychometrika, 29*, 241–256. https://doi.org/10.1007/BF02289721

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. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39,* 33–8. https://doi.org/10.1080/00031305.1985.10479383

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

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

Rubin, D. B., & Thayer, D. (1978). Relating tests given to different samples. *Psychometrika, 43,* 1–10. https://doi.org/10.1007/BF02294084

Saunders, D. R. (1953a). *An analytic method for rotation to orthogonal simple structure* (Research Bulletin No. RB-53-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1953.tb00890.x

Saunders, D. R. (1953b). *Moderator variables in prediction, with special reference to freshman engineering grades and the strong vocational interest blank* (Research Bulletin No. RB-53-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1953.tb00238.x

Schultz, D. G., & Wilks, S. S. (1950). *A method for adjusting for lack of equivalence in groups* (Research Bulletin No. RB-50-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00682.x

Spearman, C. (1904a). General intelligence objectively determined and measured. *American Journal of Psychology, 15,* 201–293. https://doi.org/10.2307/1412107

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101. https://doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Swineford, F. (1959). Some relations between test scores and item statistics. *Journal of Educational Psychology, 50*, 26–30. https://doi.org/10.1037/h0046332

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Tucker, L. R. (1949). A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika, 14,* 117–119. https://doi.org/10.1007/BF02289147

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Tucker, L. R. (1955). The objective definition of simple structure in linear factor analysis. *Psychometrika, 20,* 209–225. https://doi.org/10.1007/BF02289018

Tucker, L. R., & Finkbeiner, C.T. (1981). *Transformation of factors by artificial personal probability functions* (Research Report No. RR-81-58). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01285.x

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421–459. https://doi.org/10.1007/BF02290601

Woodbury, M. A., & Lord, F. M. (1956). The most reliable composite with a specified true score. *British Journal of Statistical Psychology, 9*, 21–28. https://doi.org/10.1111/j.2044-8317.1956.tb00165.x

Woodbury, M. A., & Novick, M. R. (1968). Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. *Journal of Mathematical Psychology, 5*, 242–259. https://doi.org/10.1016/0022-2496(68)90074-6

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01515.x

Chapter 4
Contributions to Score Linking Theory and Practice

Neil J. Dorans and Gautam Puhan

Test score equating is essential for testing programs that use multiple editions of the same test and for which scores on different editions are expected to have the same meaning. Different editions may be built to a common blueprint and be designed to measure the same constructs, but they almost invariably differ somewhat in their psychometric properties. If one edition were more difficult than another, test takers would tend to receive lower scores on the harder form. Score equating seeks to eliminate the effects on scores of these unintended differences in test form difficulty. Score equating is necessary to be fair to test takers.

ETS statisticians and psychometricians have contributed indirectly or directly to the wealth of material in the chapters on score equating or on score linking that have appeared in the four editions of *Educational Measurement*. ETS's extensive involvement with the score equating chapters of these editions of *Educational Measurement* highlights the impact that ETS has had in this important area of psychometrics.

At the time of publication, each of the four editions of *Educational Measurement* represented the state of the art in domains that are essential to the purview of the National Council on Measurement in Education. Experts in each domain wrote a chapter in each edition. Harold Gulliksen was one of the key contributors to the Flanagan (1951) chapter on units, scores, and norms that appeared in the first edition. Several of the issues and problems raised in that first edition are still current, which shows their persistence. Angoff (1971), in the second edition, provided a comprehensive introduction to scales, norms, and test equating. Petersen et al. (1989) introduced new material developed since the Angoff chapter. Holland and Dorans (2006) included a brief review of the history of test score linking. In addition to test equating, Holland and Dorans (2006) discussed other ways that scores on different tests are connected or linked together.

N.J. Dorans (✉) • G. Puhan
Educational Testing Service, Princeton, NJ, USA
e-mail: ndorans@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_4

The purpose of this chapter is to document ETS's involvement with score linking theory and practice. This chapter is not meant to be a book on score equating and score linking.[1] Several books on equating exist; some of these have been authored by ETS staff, as is noted in the last section of this chapter. We do not attempt to summarize all extant research and development pertaining to score equating or score linking. We focus on efforts conducted by ETS staff. We do not attempt to pass judgment on research or synthesize it. Instead, we attempt to describe it in enough detail to pique the interest of the reader and help point him or her in the right direction for further exploration on his or her own. We presume that the reader is familiar enough with the field so as not to be intimidated by the vocabulary that has evolved over the years in this area of specialization so central to ETS's mission to foster fairness and quality.

The particular approach to tackling this documentation task is to cluster studies around different aspects of score linking. Section 4.1 lists several examples of score linking to provide a motivation for the extent of research on score linking. Section 4.2 summarizes published efforts that provide conceptual frameworks of score linking or examples of scale aligning. Section 4.3 deals with data collection designs and data preparation issues. In Sect. 4.4, the focus is on the various procedures that have been developed to link or equate scores. Research describing processes for evaluating the quality of equating results is the focus of Sect. 4.5. Studies that focus on comparing different methods are described in Sect. 4.6. Section 4.7 is a brief chronological summary of the material covered in Sects. 4.2, 4.3, 4.4, 4.5 and 4.6. Section 4.8 contains a summary of the various books and chapters that ETS authors have contributed on the topic of score linking. Section 4.9 contains a concluding comment.

4.1 Why Score Linking Is Important

Two critical ingredients are needed to produce test scores: the test and those who take the test, the test takers. Test scores depend on the blueprint or specifications used to produce the test. The specifications describe the construct that the test is supposed to measure, how the items or components of the test contribute to the measurement of this construct (or constructs), the relative difficulty of these items for the target population of test takers, and how the items and test are scored. The definition of the target population of test takers includes who qualifies as a member of that population and is preferably accompanied by an explanation of why the test

[1] The term *linking* is often used in an IRT context to refer to procedures for aligning item parameter and proficiency metrics from one calibration to another, such as those described by M. von Davier and A. A. von Davier (2007). We do not consider this type of IRT linking in this chapter; it is treated in the chapter by Carlson and von Davier (Chap. 5, this volume). We do, however, address IRT true-score linking in Sect. 4.6.4 and IRT preequating in Sect. 4.4.4.

is appropriate for these test takers and examples of appropriate and inappropriate use.

Whenever scores from two different tests are going to be compared, there is a need to link the scales of the two test scores. The goal of scale aligning is to transform the scores from two different tests onto a common scale. The types of linkages that result depend on whether the test scores being linked measure different constructs or similar constructs, whether the tests are similar or dissimilar in difficulty, and whether the tests are built to similar or different test specifications. We give several practical examples in the following.

When two or more tests that measure different constructs are administered to a common population, the scores for each test may be transformed to have a common distribution for the target population of test takers (i.e., the reference population). The data are responses from (a) administering all the tests to the same sample of test takers or (b) administering the tests to separate, randomly equivalent samples of test takers from the same population. In this way, all of the tests are taken by equivalent groups of test takers from the reference population. One way to define comparable scores is in terms of comparable percentiles in the reference population.

Even though the scales on the different tests are made comparable in this narrow sense, the tests do measure different constructs. The recentering of the *SAT*® *I* test scale is an example of this type of scale aligning (Dorans 2002a, b). The scales for the SAT Verbal (SAT-V) and SAT Mathematical (SAT-M) scores were redefined so as to give the scaled scores on the SAT-V and SAT-M the same distribution in a reference population of students tested in 1990. The recentered score scales enable a student whose SAT-M score is higher than his or her SAT-V score to conclude that he or she did in fact perform better on the mathematical portion than on the verbal portion, at least in relation to the students tested in 1990.

Tests of skill subjects (e.g., reading) that are targeted for different school grades may be viewed as tests of similar constructs that are intended to differ in difficulty—those for the lower grades being easier than those for the higher grades. It is often desired to put scores from such tests onto a common overall scale so that progress in a given subject, such as mathematics or reading, can be tracked over time. A topic such as mathematics or reading, when considered over a range of school grades, has several subtopics or dimensions. At different grades, potentially different dimensions of these subjects are relevant and tested. For this reason, the constructs being measured by the tests for different grade levels may differ somewhat, but the tests are often similar in reliability.

Sometimes tests that measure the same construct have similar levels of difficulty but differ in reliability (e.g., length). The classic case is scaling the scores of a short form of a test onto the scale of its full or long form.

Sometimes tests to be linked all measure similar constructs, but they are constructed according to different specifications. In most cases, they are similar in test length and reliability. In addition, they often have similar uses and may be taken by the same test takers for the same purpose. Score linking adds value to the scores on both tests by expressing them as if they were scores on the other test. Many colleges and universities accept scores on either the ACT or SAT for the purpose of admissions decisions, and they often have more experience interpreting the results from one of these tests than the other.

Test equating is a necessary part of any testing program that produces new test forms and for which the uses of these tests require the meaning of the score scale be maintained over time. Although they measure the same constructs and are usually built to the same test specifications or test blueprint, different editions or forms of a test almost always differ somewhat in their statistical properties. For example, one form may be harder than another, so without adjustments, test takers would be expected to receive lower scores on this harder form. A primary goal of test equating for testing programs is to eliminate the effects on scores of these unintended differences in test form difficulty. The purpose of equating test scores is to allow the scores from each test to be used interchangeably, as if they had come from the same test. This purpose puts strong requirements on the tests and on the method of score linking. Most of the research described in the following pages focused on this particular form of scale aligning, known as *score equating*.

In the remaining sections of this chapter, we focus on score linking issues for tests that measure characteristics at the level of the individual test taker. Large-scale assessments, which are surveys of groups of test takers, are described in Beaton and Barone (Chap. 8, this volume) and Kirsh et al. (Chap. 9, this volume).

4.2 Conceptual Frameworks for Score Linking

Holland and Dorans (2006) provided a framework for classes of score linking that built on and clarified earlier work found in Mislevy (1992) and Linn (1993). Holland and Dorans (2006) made distinctions between different types of linkages and emphasized that these distinctions are related to how linked scores are used and interpreted. A link between scores on two tests is a transformation from a score on one test to a score on another test. There are different types of links, and the major difference between these types is not procedural but interpretative. Each type of score linking uses either equivalent groups of test takers or common items for linkage purposes. It is essential to understand why these types differ because they can be confused in practice, which can lead to violations of the standards that guide professional practice. Section 4.2.1 describes frameworks used for score linking. Section 4.2.2 contains a discussion of score equating frameworks.

4.2.1 Score Linking Frameworks

Lord (1964a, b) published one of the early articles to focus on the distinction between test forms that are actually or rigorously parallel and test forms that are nominally parallel—those that are built to be parallel but fall short for some reason.

This distinction occurs in most frameworks on score equating. Lord (1980) later went on to say that equating was either unnecessary (rigorously parallel forms) or impossible (everything else).

Mislevy (1992) provided one of the first extensive treatments of different aspects of what he called linking of educational assessments: *equating, calibration, projection, statistical moderation,* and *social moderation.*

Dorans (1999) made distinctions between three types of linkages or score correspondences when evaluating linkages among SAT scores and ACT scores. These were equating, scaling, and prediction. Later, in a special issue of *Applied Psychological Measurement*, edited by Pommerich and Dorans (2004), he used the terms *equating*, *concordance*, and *expectation* to refer to these three types of linkings and provided means for determining which one was most appropriate for a given set of test scores (Dorans 2004b). This framework was elaborated on by Holland and Dorans (2006), who made distinctions between *score equating, scale aligning*, and *predicting*, noting that scale aligning was a broad category that could be further subdivided into subcategories on the basis of differences in the construct assessed, test difficulty, test reliability, and population ability.

Many of the types of score linking cited by Mislevy (1992) and Dorans (1999, 2004b) could be found in the broad area of scale aligning, including concordance, vertical linking, and calibration. This framework was adapted for the public health domain by Dorans (2007) and served as the backbone for the volume on linking and aligning scores and scales by Dorans et al. (2007).

4.2.2 Equating Frameworks

Dorans et al. (2010a) provided an overview of the particular type of score linking called score equating from a perspective of best practices. After defining equating as a special form of score linking, the authors described the most common data collection designs used in the equating of test scores, some common observed-score equating functions, common data-processing practices that occur prior to computations of equating functions, and how to evaluate an equating function.

A.A. von Davier (2003, 2008) and A.A. von Davier and Kong (2005), building on the unified statistical treatment of score equating, known as *kernel equating*, that was introduced by Holland and Thayer (1989) and developed further by A.A. von Davier et al. (2004b), described a new unified framework for linear equating in a nonequivalent groups anchor test design. They employed a common parameterization to show that three linear methods, Tucker, Levine observed score, and chained,[2] can be viewed as special cases of a general linear function. The concept of a method function was introduced to distinguish among the possible forms that a linear equating function might take, in general, and among the three equating methods, in particular. This approach included a general formula for the standard error of equating

[2] These equating methods are described in Sect. 4.4.

for all linear equating functions in the nonequivalent groups anchor test design and advocated the use of the standard error of equating difference (SEED) to investigate if the observed differences in the equating functions are statistically significant.

A.A. von Davier (2013) provided a conceptual framework that encompassed traditional observed-score equating methods, kernel equating methods, and item response theory (IRT) observed-score equating, all of which produce one equating function between two test scores, along with local equating or local linking, which can produce a different linking function between two test scores given a score on a third variable (Wiberg et al. 2014). The notion of multiple conversions between two test scores is a source of controversy (Dorans 2013; Gonzalez and von Davier 2013; Holland 2013; M. von Davier et al. 2013).

4.3 Data Collection Designs and Data Preparation

Data collection and preparation are prerequisites to score linking.

4.3.1 Data Collection

Numerous data collection designs have been used for score linking. To obtain unbiased estimates of test form difficulty differences, all score equating methods must control for differential ability of the test-taker groups employed in the linking process. Data collection procedures should be guided by a concern for obtaining equivalent groups, either directly or indirectly. Often, two different, nonstrictly parallel tests are given to two different groups of test takers of unequal ability. Assuming that the samples are large enough to ignore sampling error, differences in the distributions of the resulting scores can be due to one or both of two factors. One factor is the relative difficulty of the two tests, and the other is the relative ability of the two groups of test takers on these tests. Differences in difficulty are what test score equating is supposed to take care of; difference in ability of the groups is a confounding factor that needs to be eliminated before the equating process can take place.

In practice, two distinct approaches address the separation of test difficulty and group ability differences. The first approach is to use a common population of test takers so that there are no ability differences. The other approach is to use an anchor measure of the construct being assessed by the tests to be equated. Ideally, the data should come from a large representative sample of motivated test takers that is divided in half either randomly or randomly within strata to achieve equivalent groups. Each half of this sample is administered either the new form or the old form of a test. It is typical to assume that all samples are random samples from populations of interest, even though, in practice, this may be only an approximation. When the same test takers take both tests, we achieve direct control over differential test-taker ability. In practice, it is more common to use two equivalent samples of test takers from a common population instead of identical test takers.

The second approach assumes that performance on a set of common items or an anchor measure can quantify the ability differences between two distinct, but not necessarily equivalent, samples of test takers. The use of an anchor measure can lead to more flexible data collection designs than those that require common test takers. However, the use of anchor measures requires users to make various assumptions that are not needed when the test takers taking the tests are either the same or from equivalent samples. When there are ability differences between new and old form samples, the various statistical adjustments for ability differences often produce different results because the methods make different assumptions about the relationships of the anchor test score to the scores to be equated. In addition, assumptions are made about the invariance of item characteristics across different locations within the test.

Some studies have attempted to link scores on tests in the absence of either common test material or equivalent groups of test takers. Dorans and Middleton (2012) used the term *presumed linking* to describe these situations. These studies are not discussed here.

It is generally considered good practice to have the anchor test be a mini-version of the total tests being equated. That means it should have the same difficulty and similar content. Often an external anchor is not available, and internal anchors are used. In this case, context effects become a possible issue. To minimize these effects, anchor (or common) items are often placed in the same location within each test. When an anchor test is used, the items should be evaluated via procedures for assessing whether items are functioning in the same way in both the old and new form samples. All items on both total tests are evaluated to see if they are performing as expected. If they are not, it is often a sign of a quality-control problem. More information can be found in Holland and Dorans (2006).

When there are large score differences on the anchor test between samples of test takers given the two different test forms to be equated, equating based on the nonequivalent-groups anchor test design can often become problematic. Accumulation of potentially biased equating results can occur over a chain of prior equatings and lead to a shift in the meaning of numbers on the scores scale.

In practice, the true equating function is never known, so it is wise to look at several procedures that make different assumptions or that use different data. Given the potential impact of the final score conversion on all participants in an assessment process, it is important to check as many factors that can cause problems as possible. Considering multiple conversions is one way to do this.

Whereas many sources, such as Holland and Dorans (2006), have focused on the structure of data collection designs, the amount of data collected has a substantial effect on the usefulness of the resulting equatings. Because it is desirable for the statistical uncertainty associated with test equating to be much smaller than the other sources of variation in test results, it is important that the results of test equating be based on samples that are large enough to ensure this. This fact should always be kept in mind when selecting a data collection design. Section 4.4 describes procedures that have been developed to deal with the threats associated with small samples.

4.3.2 Data Preparation Activities

Prior to equating and other forms of linking, several steps can be taken to improve the quality of the data. These best practices of data preparation often deal with sample selection, smoothing score distributions, excluding outliers, repeaters, and so on. These issues are the focus of the next four parts of this section.

4.3.2.1 Sample Selection

Before conducting the equating analyses, testing programs often filter the data based on certain heuristics. For example, a testing program may choose to exclude test takers who do not attempt a certain number of items on the test. Other programs might exclude test takers based, for example, on repeater status. ETS researchers have conducted studies to examine the effect of such sample selection practices on equating results. Liang et al. (2009) examined whether nonnative speakers of the language in which the test is administered should be excluded and found that this may not be an issue as long as the proportion of nonnative speakers does not change markedly across administrations. Puhan (2009b, 2011c) studied the impact of repeaters in the equating samples and found in the data he examined that inclusion or exclusion of repeaters had very little impact on the final equating results. Similarly, Yang et al. (2011) examined the effect of repeaters on score equating and found no significant effects of repeater performance on score equating for the exam being studied. However, Kim and Walker (2009a, b) found in their study that when the repeater subgroup was subdivided based on the particular form test takers took previously, subgroup equating functions substantially differed from the total-group equating function.

4.3.2.2 Weighted Samples

Dorans (1990c) edited a special issue of *Applied Measurement in Education* that focused on the topic of equating with samples matched on the anchor test score (Dorans 1990a). The studies in that special issue used simulations that varied in the way in which real data were manipulated to produce simulated samples of test takers. These and related studies are described in Sect. 4.6.3.

Other authors used demographic data to achieve a form of matching. Livingston (2014a) proposed the demographically adjusted groups procedure, which uses demographic information about the test takers to transform the groups taking the two different test forms into groups of equal ability by weighting the test takers unequally. Results indicated that although this procedure adjusts for group differences, it does not reduce the ability difference between the new and old form samples enough to warrant use.

Qian et al. (2013) used techniques for weighting observations to yield a weighted sample distribution that is consistent with the target population distribution to achieve true-score equatings that are more invariant across administrations than those obtained with unweighted samples.

Haberman (2015) used adjustment by minimum discriminant information to link test forms in the case of a nonequivalent-groups design in which there are no satisfactory common items. This approach employs background information other than scores on individual test takers in each administration so that weighted samples of test takers form pseudo-equivalent groups in the sense that they resemble samples from equivalent groups.

4.3.2.3 Smoothing

Irregularities in score distributions can produce irregularities in the equipercentile equating adjustment that might not generalize to different groups of test takers because the methods developed for continuous data are applied to discrete data. Therefore it is generally advisable to presmooth the raw-score frequencies in some way prior to equipercentile equating.

The idea of smoothing score distributions prior to equating goes far back to the 1950s. Karon and Cliff (1957) proposed the Cureton–Tukey procedure as a means for reducing sampling error by mathematically smoothing the sample score data before equating. However, the differences among the linear equating method, the equipercentile equating method with no smoothing of the data, and the equipercentile equating method after smoothing by the Cureton–Tukey method were not statistically significant. Nevertheless, this was an important idea, and although Karon and Cliff's results did not show the benefits of smoothing, currently most testing programs using equipercentile equating use some form of pre- or postsmoothing to obtain more stable equating results.

Ever since the smoothing method using loglinear models was adapted by ETS researchers in the 1980s (for details, see Holland and Thayer 1987; Rosenbaum and Thayer 1987) smoothing has been an important component of the equating process. The new millennium saw a renewed interest in smoothing research. Macros using the statistical analysis software SAS loglinear modeling routines were developed at ETS to facilitate research on smoothing (Moses and von Davier 2006, 2013; Moses et al. 2004). A series of studies were conducted to assess selection strategies (e.g., strategies based on likelihood ratio tests, equated score difference tests, Akaike information criterion (AIC) for univariate and bivariate loglinear smoothing models and their effects on equating function accuracy (Moses 2008a, 2009; Moses and Holland 2008, 2009a, b, c, 2010a, b).

Studies also included comparisons of traditional equipercentile equating with various degrees of presmoothing and kernel equating (Moses and Holland 2007) and smoothing approaches for composite scores (Moses 2014) as well as studies that compared smoothing with pseudo-Bayes probability estimates (Moses and Oh 2009).

There has also been an interest in smoothing in the context of systematic irregularities in the score distributions that are due to scoring practice and scaling issues (e.g., formula scoring, impossible scores) rather than random irregularities (J. Liu et al. 2009b; Puhan et al. 2008b, 2010).

4.3.2.4 Small Samples and Smoothing

Presmoothing the data before conducting an equipercentile equating has been shown to reduce error in small-sample equating. For example, Livingston and Feryok (1987) and Livingston (1993b) worked with small samples and found that presmoothing substantially improved the equating results obtained from small samples. Puhan (2011a, b), based on the results of an empirical study, however, concluded that although presmoothing can reduce random equating error, it is not likely to reduce equating bias caused by using an unrepresentative small sample and presented other alternatives to the small-sample equating problem that focused more on improving data collection (see Sect. 4.4.5).

4.4 Score Equating and Score Linking Procedures

Many procedures for equating tests have been developed by ETS researchers. In this section, we consider equating procedures such as linear, equipercentile equating, kernel equating, and IRT true-score linking.[3] Equating procedures developed to equate new forms under special circumstances (e.g., preequating and small-sample equating procedures) are also considered in this section.

[3] We have chosen to use the term *linking* instead of *equating* when it comes to describing the IRT true-score approach that is in wide use. This linking procedure defines the true-score equating that exists between true scores on Test X and true scores on Test Y, which are perfectly related to each other, as both are monotonic transformations of the same IRT proficiency estimate. Typically, this true-score equating is applied to observed scores as if they were true scores. This application produces an observed-score linking that is not likely to yield equated scores, however, as defined by Lord (1980) or Holland and Dorans (2006); hence our deliberate use of linking instead of equating.

4.4.1 Early Equating Procedures

Starting in the 1950s, ETS researchers have made substantial contributions to the equating literature by proposing new methods for equating, procedures for improving existing equating methods, and procedures for evaluating equating results.

Lord (1950) provided a definition of comparability wherein the score scales of two equally reliable tests are considered comparable with respect to a certain group of test takers if the score distributions of the two tests are identical for this group. He provided the basic formulas for equating means and standard deviations (in six different scenarios) to achieve comparability of score scales. Tucker (1951) emphasized the need to establish a formal system within which to consider scaling error due to sampling. Using simple examples, he illustrated possible ways of defining the scaling error confidence range and setting a range for the probability of occurrence of scaling errors due to sampling that would be considered within normal operations. Techniques were developed to investigate whether regressions differ by groups. Schultz and Wilks (1950) presented a technique to adjust for the lack of equivalence in two samples. This technique focused on the intercept differences from the two group regressions of total score onto anchor score obtained under the constraint that the two regressions had the same slope. Koutsopoulos (1961) presented a linear practice effect solution for a counterbalanced case of equating, in which two equally random groups (alpha and beta) take two forms, X and Y, of a test, alpha in the order X, Y and beta in the order Y, X. Gulliksen (1968) presented a variety of solutions for determining the equivalence of two measures, ranging from a criterion for strict interchangeability of scores to factor methods for comparing multifactor batteries of measures and multidimensional scaling. Boldt (1972) laid out an alternative approach to linking scores that involved a principle for choosing objective functions whose optimization would lead to a selection of conversion constants for equating.

Angoff (1953) presented a method of equating test forms of the American Council on Education (ACE) examination by using a miniature version of the full test as an external anchor to equate the test forms. Fan and Swineford (1954) and Swineford and Fan (1957) introduced a method based on item difficulty estimates to equate scores administered under the nonequivalent anchor test design, which the authors claimed produced highly satisfactory results, especially when the two groups taking the two forms were quite different in ability.

Assuming that the new and old forms are equally reliable, Lord (1954, 1955) derived maximum likelihood estimates of the population mean and standard deviation, which were then substituted into the basic formula for linear equating.

Levine (1955) developed two linear equating procedures for the common-item nonequivalent population design. Levine observed-score equating relates observed scores on a new form to the scale of observed scores on an old form. Levine true-score equating equates true scores. Approximately a half-century later, A.A. von Davier et al. (2007) introduced an equipercentile version of the Levine linear observed-score equating function, which is based on assumptions about true scores.

Based on theoretical and empirical results, Chen (2012) showed that linear IRT observed-score linking and Levine observed-score equating for the anchor test design are closely related despite being based on different methodologies. Chen and Livingston (2013) presented a new equating method for the nonequivalent groups with anchor test design: poststratification equating based on true anchor scores. The linear version of this method is shown to be equivalent, under certain conditions, to Levine observed-score equating.

4.4.2 True-Score Linking

As noted in the previous section, Levine (1955) also developed the so-called Levine true-score equating procedure that equates true scores.

Lord (1975) compared equating methods based on item characteristic curve (ICC) theory, which he later called item response theory (IRT) in Lord (1980), with nonlinear conventional methods and pointed out the effectiveness of ICC-based methods for increasing stability of the equating near the extremes of the data, reducing scale drift, and preequating. Lord also included a chapter on IRT preequating. (A review of research related to IRT true-score linking appears in Sect. 4.6.4.)

4.4.3 Kernel Equating and Linking With Continuous Exponential Families

As noted earlier, Holland and Thayer (1989) introduced the kernel method of equating score distributions. This new method included both linear and standard equipercentile methods as special cases and could be applied under most equating data collection designs.

Within the Kernel equating framework, Chen and Holland (2010) developed a new curvilinear equating for the nonequivalent groups with anchor test (NEAT) design which they called curvilinear Levine observed score equating.

In the context of equivalent-groups design, Haberman (2008a) introduced a new way to continuize discrete distribution functions using exponential families of functions. Application of this linking method was also considered for the single-group design (Haberman 2008b) and the nonequivalent anchor test design (Haberman and Yan 2011). For the nonequivalent groups with anchor test design, this linking method produced very similar results to kernel equating and equipercentile equating with loglinear presmoothing.

4.4.4 Preequating

Preequating has been tried for several ETS programs over the years. Most notably, the computer-adaptive testing algorithm employed for the *GRE*® test, the *TOEFL*® test, and GMAT examination in the 1990s could be viewed as an application of IRT preequating. Since the end of the twentieth century, IRT preequating has been used for the *CLEP*® examination and with the GRE revised General Test introduced in 2011. This section describes observed-score preequating procedures. (The results of several studies that used IRT preequating can be found in Sect. 4.6.5.)

In the 1980s, section preequating was used with the GMAT examination. A preequating procedure was developed for use with small-volume tests, most notably the *PRAXIS*® assessments. This approach is described in Sect. 4.4.5. Holland and Wightman (1982) described a preliminary investigation of a linear section preequating procedure. In this statistical procedure, data collected from equivalent groups via the nonscored variable or experimental section(s) of a test were combined across tests to produce statistics needed for linear preequating of a form composed of these sections. Thayer (1983) described the maximum likelihood estimation procedure used for estimating the joint covariance matrix for sections of tests given to distinct samples of test takers, which was at the heart of the section preequating approach.

Holland and Thayer (1981) applied this procedure to the GRE test and obtained encouraging results. Holland and Thayer (1984, 1985) extended the theory behind section preequating to allow for practice effects on both the old and new forms and, in the process, provided a unified account of the procedure. Wightman and Wightman (1988) examined the effectiveness of this approach when there is only one variable or experimental section of the test, which entailed using different missing data techniques to estimate correlations between sections.

After a long interlude, section preequating with a single variable section was studied again. Guo and Puhan (2014) introduced a method for both linear and nonlinear preequating. Simulations and a real-data application showed the proposed method to be fairly simple and accurate. Zu and Puhan (2014) examined an observed-score preequating procedure based on empirical item response curves, building on work done by Livingston in the early 1980s. The procedure worked reasonably well in the score range that contained the middle 90th percentile of the data, performing as well as the IRT true-score equating procedure.

4.4.5 Small-Sample Procedures

In addition to proposing new methods for test equating in general, ETS researchers have focused on equating under special circumstances, such as equating with very small samples. Because equating with very small samples tends to be less stable, researchers have proposed new approaches that aim to produce more stable equating results under small-sample conditions. For example, Kim et al. (2006, 2007, 2008c, 2011) proposed the synthetic linking function (which is a weighted average of the small-sample equating and the identity function) for small samples and conducted several empirical studies to examine its effectiveness in small-sample conditions. Similarly, the circle-arc equating method, which constrains the equating curve to pass through two prespecified endpoints and an empirically determined middle point, was also proposed for equating with small samples (Livingston and Kim 2008, 2009, 2010a, b) and evaluated in empirical studies by Kim and Livingston (2009, 2010). Finally, Livingston and Lewis (2009) proposed the empirical Bayes approach for equating with small samples whereby prior information comes from equatings of other test forms, with an appropriate adjustment for possible differences in test length. Kim et al. (2008d, 2009) conducted resampling studies to evaluate the effectiveness of the empirical Bayes approach with small samples and found that this approach tends to improve equating accuracy when the sample size is 25 or fewer, provided the prior equatings are accurate.

The studies summarized in the previous paragraph tried to incorporate modifications to existing equating methods to improve equating under small-sample conditions. Their efficacy depends on the correctness of the strong assumptions that they employ to affect their proposed solutions (e.g., the appropriateness of the circle arc or the identity equatings).

Puhan (2011a, b) presented other alternatives to the small-sample equating problem that focused more on improving data collection. One approach would be to implement an equating design whereby data conducive to improved equatings can be collected to help with the small-sample equating problem. An example of such a design developed at ETS is the single-group nearly equivalent test design, or the SiGNET design (Grant 2011), which introduces a new form in stages rather than all at once. The SiGNET design has two primary merits. First, it facilitates the use of a single-group equating design that has the least random equating error of all designs, and second, it allows for the accumulation of data to equate the new form with a larger sample. Puhan et al. (2008a, 2009) conducted a resampling study to compare equatings under the SiGNET and common-item equating designs and found lower equating error for the SiGNET design than for the common-item equating design in very small sample size conditions (e.g., $N = 10$).

4.5 Evaluating Equatings

In this part, we address several topics in the evaluation of links formed by scale alignment or by equatings. Section 4.5.1 describes research on assessing the sampling error of linking functions. In Sect. 4.5.2, we summarize research dealing with measures of the effect size for assessing the invariance of equating and scale-aligning functions over subpopulations of a larger population. Section 4.5.3 is concerned with research that deals with scale continuity.

4.5.1 Sampling Stability of Linking Functions

All data based linking functions are statistical estimates, and they are therefore subject to sampling variability. If a different sample had been taken from the target population, the estimated linking function would have been different. A measure of statistical stability gives an indication of the uncertainty in an estimate that is due to the sample selected. In Sect. 4.5.1.1, we discuss the standard error of equating (SEE). Because the same methods are also used for concordances, battery scaling, vertical scaling, calibration, and some forms of anchor scaling, the SEE is a relevant measure of statistical accuracy for these cases of test score linking as well as for equating.

In Sects. 4.5.1.1 and 4.5.1.2, we concentrate on the basic ideas and large-sample methods for estimating standard error. These estimates of the SEE and related measures are based on the delta method. This means that they are justified as standard error estimates only for large samples and may not be valid in small samples.

4.5.1.1 The Standard Error of Equating

Concern about the sampling error associated with different data collection designs for equating has occupied ETS researchers since the 1950s (e.g., Karon 1956; Lord 1950). The SEE is the oldest measure of the statistical stability of estimated linking functions. The SEE is defined as the conditional standard deviation of the sampling distribution of the equated score for a given raw score over replications of the equating process under similar conditions. We may use the SEE for several purposes. It gives a direct measure of how consistently the equating or linking function is estimated. Using the approximate normality of the estimate, the SEE can be used to form confidence intervals. In addition, comparing the SEE for various data collection designs can indicate the relative advantage some designs have over others for particular sample sizes and other design factors. This can aid in the choice of a data collection design for a specific purpose.

The SEE can provide us with statistical caveats about the instability of linkings based on small samples. As the size of the sample(s) increases, the SEE will decrease. With small samples, there is always the possibility that the estimated linking function is a poor representation of the population linking function.

The earliest work on the SEE is found in Lord (1950) and reproduced in Angoff (1971). These papers were concerned with linear-linking methods and assumed normal distributions of scores. Zu and Yuan (2012) examined estimates for linear equating methods under conditions of nonnormality for the nonequivalent-groups design. Lord (1982b) derived the SEE for the equivalent- and single-group designs for the equipercentile function using linear interpolation for continuization of the linking functions. However, these SEE calculations for the equipercentile function did not take into account the effect of presmoothing, which can produce reductions in the SEE in many cases, as demonstrated by Livingston (1993a). Liou and Cheng

(1995) gave an extensive discussion (including estimation procedures) of the SEE for various versions of the equipercentile function that included the effect of presmoothing. Holland et al. (1989) and Liou et al. (1996, 1997) discussed the SEE for kernel equating for the nonequivalent-groups anchor test design.

A.A. von Davier et al. (2004b) provided a system of statistical accuracy measures for kernel equating for several data collection designs. Their results account for four factors that affect the SEE: (a) the sample sizes; (b) the effect of presmoothing; (c) the data collection design; and (d) the form of the final equating function, including the method of continuization. In addition to the SEE and the SEED (described in Sect. 4.5.1.2), they recommend the use of percent relative error to summarize how closely the moments of the equated score distribution match the target score distribution that it is striving to match. A.A. von Davier and Kong (2005) gave a similar analysis for linear equating in the non-equivalent-groups design.

Lord (1981) derived the asymptotic standard error of a true-score equating by IRT for the anchor test design and illustrated the effect of anchor test length on this SEE. Y. Liu et al. (2008) compared a Markov chain Monte Carlo (MCMC) method and a bootstrap method in the estimation of standard errors of IRT true-score linking. Grouped jackknifing was used by Haberman et al. (2009) to evaluate the stability of equating procedures with respect to sampling error and with respect to changes in anchor selection with illustrations involving the two-parameter logistic (2PL) IRT model.

4.5.1.2 The Standard Error of Equating Difference Between Two Linking Functions

Those who conduct equatings are often interested in the stability of differences between linking functions. A.A. von Davier et al. (2004b) were the first to explicitly consider the standard error of the distribution of the difference between two estimated linking functions, which they called the SEED. For kernel equating methods, using loglinear models to presmooth the data, the same tools used for computing the SEE can be used for the SEED for many interesting comparisons of kernel equating functions. Moses and Zhang (2010, 2011) extended the notion of the SEED to comparisons between kernel linear and traditional linear and equipercentile equating functions, as well.

An important use of the SEED is to compare the linear and nonlinear versions of kernel equating. von Davier et al. (2004b) combined the SEED with a graphical display of the plot of the difference between the two equating functions. In addition to the difference, they added a band of ±2SEED to put a rough bound on how far the two equating functions could differ due to sampling variability. When the difference curve is outside of this band for a substantial number of values of the X-scores, this is evidence that the differences between the two equating functions exceed what might be expected simply due to sampling error. The ±2SEED band is narrower for larger sample sizes and wider for smaller sample sizes.

Duong and von Davier (2012) illustrated the flexibility of the observed-score equating framework and the availability of the SEED in allowing practitioners to compare statistically the equating results from different weighting schemes for distinctive subgroups of the target population.

In the special situation where we wish to compare an estimated equating function to another nonrandom function, for example, the identity function, the SEE plays the role of the SEED. Dorans and Lawrence (1988, 1990) used the SEE to create error bands around the difference plot to determine whether the equating between two section orders of a test was close enough to the identity. Moses (2008a, 2009) examined a variety of approaches for selecting equating functions for the equivalent-groups design and recommended that the likelihood ratio tests of loglinear models and the equated score difference tests be used together to assess equating function differences overall and also at score levels. He also encouraged a consideration of the magnitude of equated score differences with respect to score reporting practices.

In addition to the statistical significance of the difference between the two linking functions (the SEED), it is also useful to examine whether this difference has any important consequences for reported scores. This issue was addressed by Dorans and Feigenbaum (1994) in their notion of a difference that matters (DTM). They called a difference in reported score points a DTM if the testing program considered it to be a difference worth worrying about. This, of course, depends on the test and its uses. If the DTM that is selected is smaller than 2 times an appropriate SEE or SEED, then the sample size may not be sufficient for the purposes that the equating is intended to support.

4.5.2 Measures of the Subpopulation Sensitivity of Score Linking Functions

Neither the SEE nor the SEED gives any information about how different the estimated linking function would be if the data were sampled from other populations of test takers. Methods for checking the sensitivity of linking functions to the population on which they are computed (i.e., subpopulation invariance checks) serve as diagnostics for evaluating links between tests (especially those that are intended to be test equatings). The most common way that population invariance checks are made is on subpopulations of test takers within the larger population from which the samples are drawn. Subgroups such as male and female are often easily identifiable in the data. Other subgroups are those based on ethnicity, region of the country, and so on. In general, it is a good idea to select subgroups that are known to differ in their performance on the tests in question.

Angoff and Cowell (1986) examined the population sensitivity of linear conversions for the GRE Quantitative test (GRE-Q) and the specially constituted GRE Verbal-plus-Quantitative test (GREV+Q) using equivalent groups of approximately

13,000 taking each form. The data clearly supported the assumption of population invariance for GRE-Q but not quite so clearly for GREV+Q.

Dorans and Holland (2000a, b) developed general indices of population invariance/sensitivity of linking functions for the equivalent groups and single-group designs. To study population invariance, they assumed that the target population is partitioned into mutually exclusive and exhaustive subpopulations. A.A. von Davier et al. (2004a) extended that work to the nonequivalent-groups anchor test design that involves two populations, both of which are partitioned into similar subpopulations.

Moses (2006, 2008b) extended the framework of kernel equating to include the standard errors of indices described in Dorans and Holland (2000a, b). The accuracies of the derived standard errors were evaluated with respect to empirical standard errors.

Dorans (2004a) edited a special issue of the *Journal of Educational Measurement*, titled "Assessing the Population Sensitivity of Equating Functions," that examined whether equating or linking functions relating test scores achieved population invariance. A. A. von Davier et al. (2004a) extended the work on subpopulation invariance done by Dorans and Holland (2000a, b) for the single-population case to the two-population case, in which the data are collected on an anchor test as well as the tests to be equated. Yang (2004) examined whether the multiple-choice (MC) to composite linking functions of the *Advanced Placement*® examinations remain invariant over subgroups by region. Dorans (2004c) examined population invariance across gender groups and placed his investigation within a larger fairness context by introducing score equity analysis as another facet of fair assessment, a complement to differential item functioning and differential prediction.

A.A. von Davier and Liu (2007) edited a special issue of *Applied Psychological Measurement*, titled "Population Invariance," that built on and extended prior research on population invariance and examined the use of population invariance measures in a wide variety of practical contexts. A.A. von Davier and Wilson (2008) examined IRT models applied to Advanced Placement exams with both MC and constructed-response (CR) components. M. Liu and Holland (2008) used Law School Admission Test (LSAT) data to extend the application of population invariance methods to subpopulations defined by geographic region, whether test takers applied to law school, and their law school admission status. Yang and Gao (2008) investigated the population invariance of the one-parameter IRT model used with the testlet-based computerized exams that are part of CLEP. Dorans et al. (2008) examined the role that the choice of anchor test plays in achieving population invariance of linear equatings across male and female subpopulations and test administrations.

Rijmen et al. (2009) compared two methods for obtaining the standard errors of two population invariance measures of equating functions. The results indicated little difference between the standard errors found by the delta method and the grouped jackknife method.

Dorans and Liu (2009) provided an extensive illustration of the application of score equity assessment (SEA), a quality-control process built around the use of population invariance indices, to the SAT-M exam. Moses et al. (2009, 2010b) developed a SAS macro that produces Dorans and Liu's (2009) prototypical SEA analyses, including various tabular and graphical analyses of the differences between scaled score conversions from one or more subgroups and the scaled score conversion based on a total group. J. Liu and Dorans (2013) described how SEA can be used as a tool to assess a critical aspect of construct continuity, the equivalence of scores, whenever planned changes are introduced to testing programs. They also described how SEA can be used as a quality-control check to evaluate whether tests developed to a static set of specifications remain within acceptable tolerance levels with respect to equitability.

Kim et al. (2012) illustrated the use of subpopulation invariance with operational data indices to assess whether changes to the test specifications affected the equatability of a redesigned test to the current test enough to change the meaning of points on the score scale. Liang et al. (2009), also reported in Sinharay et al. (2011b), used SEA to examine the sensitivity of equating procedures to increasing numbers of nonnative speakers in equating samples.

4.5.3 Consistency of Scale Score Meaning

In an ideal world, measurement is flawless, and score scales are properly defined and well maintained. Shifts in performance on a test reflect shifts in the ability of test-taker populations, and any variability in the raw-to-scale conversions across editions of a test is minor and due to random sampling error. In an ideal world, many things need to mesh. Reality differs from the ideal in several ways that may contribute to scale inconsistency, which, in turn, may contribute to the appearance or actual existence of scale drift. Among these sources of scale inconsistency are inconsistent or poorly defined test-construction practices, population changes, estimation error associated with small samples of test takers, accumulation of errors over a long sequence of test administrations, inadequate anchor tests, and equating model misfit. Research into scale continuity has become more prevalent in the twenty-first century. Haberman and Dorans (2011) made distinctions among different sources of variation that may contribute to score-scale inconsistency. In the process of delineating these potential sources of scale inconsistency, they indicated practices that are likely either to contribute to inconsistency or to attenuate it.

Haberman (2010) examined the limits placed on scale accuracy by sample size, number of administrations, and number of forms to be equated. He demonstrated analytically that a testing program with a fixed yearly volume is likely to experience more substantial scale drift with many small-volume administrations than with fewer large volume administrations. As a consequence, the comparability of scores across different examinations is likely to be compromised from many small-volume administrations. This loss of comparability has implications for some modes of continuous testing. Guo (2010) investigated the asymptotic accumulative SEE for linear equating methods under the nonequivalent groups with anchor test design. This tool measures the magnitude of equating errors that have accumulated over a series of equatings.

Lee and Haberman (2013) demonstrated how to use harmonic regression to assess scale stability. Lee and von Davier (2013) presented an approach for score-scale monitoring and assessment of scale drift that used quality-control charts and time series techniques for continuous monitoring, adjustment of customary variations, identification of abrupt shifts, and assessment of autocorrelation.

With respect to the SAT scales established in the early 1940s, Modu and Stern (1975) indicated that the reported score scale had drifted by almost 14 points for the verbal section and 17 points for the mathematics section between 1963 and 1973. Petersen et al. (1983) examined scale drift for the verbal and mathematics portions of the SAT and concluded that for reasonably parallel tests, linear equating was adequate, but for tests that differed somewhat in content and length, 3PL IRT-based methods lead to greater stability of equating results. McHale and Ninneman (1994) assessed the stability of the SAT scale from 1973 to 1984 and found that the SAT-V score scale showed little drift. Furthermore, the results from the Mathematics scale were inconsistent, and therefore the stability of this scale could not be determined.

With respect to the revised SAT scales introduced in 1995, Guo et al. (2012) examined the stability of the SAT Reasoning Test score scales from 2005 to 2010. A 2005 old form was administered along with a 2010 new form. Critical Reading and Mathematics score scales experienced, at most, a moderate upward scale drift that might be explained by an accumulation of random equating errors. The Writing score scale experienced a significant upward scale drift, which might reflect more than random error.

Scale stability depends on the number of items or sets of items used to link tests across administrations. J. Liu et al. (2014) examined the effects of using one, two, or three anchor tests on scale stability of the SAT from 1995 to 2003. Equating based on one old form produced persistent scale drift and also showed increased variability in score means and standard deviations over time. In contrast, equating back to two or three old forms produced much more stable conversions and had less variation.

Guo et al. (2013) advocated the use of the conditional standard error of measurement when assessing scale deficiencies as measured by gaps and clumps, which were defined in Dorans et al. (2010b).

Using data from a teacher certification program, Puhan (2007, 2009a) examined scale drift for parallel equating chains and a single long chain. Results of the study indicated that although some drift was observed, the effect on pass or fail status of test takers was not large.

Cook (1988) explored several alternatives to the scaling procedures traditionally used for the College Board Achievement Tests. The author explored additional scaling covariates that might improve scaling results for tests that did not correlate highly with the SAT Reasoning Test, possible respecification of the sample of students used to scale the tests, and possible respecification of the hypothetical scaling population.

4.6 Comparative Studies

As new methods or modifications to existing methods for data preparation and analysis continued to be developed at ETS, studies were conducted to evaluate the new approaches. These studies were diverse and included comparisons between newly developed methods and existing methods, chained versus poststratification methods, comparisons of equatings using different types of anchor tests, and so on. In this section we attempt to summarize this research in a manner that parallels the structure employed in Sects. 4.3 and 4.4. In Sect. 4.6.1, we address research that focused on data collection issues, including comparisons of equivalent-groups equating and anchor test equating and comparisons of the various anchor test equating procedures. Section 4.6.2 contains research pertaining to anchor test properties. In Sect. 4.6.3, we consider research that focused on different types of samples of test takers. Next, in Sect. 4.6.4, we consider research that focused on IRT equating. IRT preequating is considered in Sect. 4.6.5. Then some additional topics are addressed. Section 4.6.6 considers equating tests with CR components. Equating of subscores is considered in Sect. 4.6.7, whereas Sect. 4.6.8 considers equating in the presence of multidimensional data. Because several of the studies addressed in Sect. 4.6 used simulated data, we close with a caveat about the strengths and limitations of relying on simulated data in Sect. 4.6.9.

4.6.1 Different Data Collection Designs and Different Methods

Comparisons between different equating methods (e.g., chained vs. poststratification methods) and different equating designs (e.g., equivalent groups vs. nonequivalent groups with anchor test design) have been of interest for many ETS researchers. (Comparisons that focused on IRT linking are discussed in Sect. 4.6.4.)

Kingston and Holland (1986) compared alternative equating methods for the GRE General Test. They compared the equivalent-groups design with two other designs (i.e., nonequivalent groups with an external anchor test and equivalent groups with a preoperational section) and found that the equivalent groups with preoperational section design produced fairly poor results compared to the other designs.

After Holland and Thayer introduced kernel equating in 1989, Livingston (1993b) conducted a study to compare kernel equating with traditional equating methods and concluded that kernel equating and equipercentile equating based on smoothed score distributions produce very similar results, except at the low end of the score scale, where the kernel results were slightly more accurate. However, much of the research work at ETS comparing kernel equating with traditional equating methods happened after A.A. von Davier et al. (2004b) was published. For example, A.A. von Davier et al. (2006) examined how closely the kernel equating (KE) method approximated the results of other observed-score equating methods under the common-item equating design and found that the results from kernal equating (KE) and the other methods were quite similar. Similarly, results from a study by Mao et al. (2006) indicated that the differences between KE and the traditional equating methods are very small (for most parts of the score scale) for both the equivalent-groups and common-item equating design. J. Liu and Low (2007, 2008) compared kernel equating with analogous traditional equating methods and concluded that KE results are comparable to the results of other methods. Similarly, Grant et al. (2009) compared KE with traditional equating methods, such as Tucker, Levine, chained linear, and chained equipercentile methods, and concluded that the differences between KE and traditional equivalents were quite small. Finally, Lee and von Davier (2008) compared equating results based on different kernel functions and indicated that the equated scores based on different kernel functions do not vary much, except for extreme scores.

There has been renewed interest in chained equating (CE) versus poststratification equating (PSE) research in the new millennium. For example, Guo and Oh (2009) evaluated the frequency estimation (FE) equating method, a PSE method, under different conditions. Based on their results, they recommended FE equating when neither the two forms nor the observed conditional distributions are very different. Puhan (2010a, b) compared Tucker, chained linear, and Levine observed equating under conditions where the new and old form samples were either similar in ability or not and where the tests were built to the same set of content specifications and concluded that, for most conditions, chained linear equating produced fairly accurate equating results. Predictions from both PSE and CE assumptions were compared using data from a special study that used a fairly novel approach (Holland et al. 2006, 2008). This research used real data to simulate tests built to the same set of content specifications and found that that both CE and PSE make very similar predictions but that those of CE are slightly more accurate than those of PSE, especially where the linking function is nonlinear. In a somewhat similar vein as the preceding studies, Puhan (2012) compared Tucker and chained linear equating in two scenarios. In the first scenario, known as rater comparability scoring and equating, chained linear equating produced more accurate results. Note that although rater comparability scoring typically results in a single-group equating design, the study evaluated a special case in which the rater comparability scoring data were used under a common-item equating design. In the second situation, which used a common-item equating design where the new and old form samples were randomly equivalent, Tucker equating produced more accurate results. Oh and Moses (2012) investigated differences between uni- and bidirectional approaches to chained equipercentile equating and concluded that although the bidirectional results were slightly less erratic and smoother, both methods, in general, produce very similar results.

4.6.2 The Role of the Anchor

Studies have examined the effect of different types of anchor tests on test equating, including anchor tests that are different in content and statistical characteristics. For example, Echternacht (1971) compared two approaches (i.e., using common items or scores from the GRE Verbal and Quantitative measures as the anchor) for equating the GRE Advanced tests. Results showed that both approaches produce equating results that are somewhat different from each other. DeMauro (1992) examined the possibility of equating the *TWE*® test by using TOEFL as an anchor and concluded that using TOEFL as an anchor to equate the TWE is not appropriate.

Ricker and von Davier (2007) examined the effects of external anchor test length on equating results for the common-item equating design. Their results indicated that bias tends to increase in the conversions as the anchor test length decreases, although FE and kernel poststratification equating are less sensitive to this change than other equating methods, such as chained equipercentile equating. Zu and Liu (2009, 2010) compared the effect of discrete and passage-based anchor items on common-item equating results and concluded that anchor tests that tend to have more passage-based items than discrete items result in larger equating errors, especially when the new and old samples differ in ability. Liao (2013) evaluated the effect of speededness on common-item equating and concluded that including an item set toward the end of the test in the anchor affects the equating in the anticipated direction, favoring the group for which the test is less speeded.

Moses and Kim (2007) evaluated the impact of unequal reliability on test equating methods in the common-item equating design and noted that unequal and/or low reliability inflates equating function variability and alters equating functions when there is an ability difference between the new and old form samples.

Sinharay and Holland (2006a, b) questioned conventional wisdom that an anchor test used in equating should be a statistical miniature version of the tests to be equated. They found that anchor tests with a spread of item difficulties less than that of a total test (i.e., a midi test) seem to perform as well as a mini test (i.e., a miniature version of the full test), thereby suggesting that the requirement of the anchor test to mimic the statistical characteristics of the total test may not be optimal. Sinharay et al. (2012) also demonstrated theoretically that the mini test may not be the optimal anchor test with respect to the anchor test–total test correlation. Finally, several empirical studies by J. Liu et al. (2009a, 2011a, b) also found that the midi anchor performed as well or better than the mini anchor across most of the score scale, except the top and bottom, which is where inclusion or exclusion of easy or hard items might be expected to have an effect.

For decades, new editions of the SAT were equated back to two past forms using the nonequivalent-groups anchor test design (Holland and Dorans 2006). Successive new test forms were linked back to different pairs of old forms. In 1994, the SAT equatings began to link new forms back to four old forms. The rationale for this new scheme was that with more links to past forms, it is easier to detect a poor past conversion function, and it makes the final new conversion function less reliant on any particular older equating function. Guo et al. (2011) used SAT data collected from 44 administrations to investigate the effect of accumulated equating error in equating conversions and the effect of the use of multiple links in equating. It was observed that the single-link equating conversions drifted further away from the operational four-link conversions as equating results accumulated over time. In addition, the single-link conversions exhibited an instability that was not obvious for the operational data. A statistical random walk model was offered to explain the mechanism of scale drift in equating caused by random equating error. J. Liu et al. (2014) tried to find a balance point where the needs for equating, control of item/form exposure, and pretesting could be satisfied. Three equating scenarios were examined using real data: equating to one old form, equating to two old forms, or equating to three old forms. Equating based on one old form produced persistent score drift and showed increased variability in score means and standard deviations over time. In contrast, equating back to two or three old forms produced much more stable conversions and less variation in means and standard deviations. Overall, equating based on multiple linking designs produced more consistent results and seemed to limit scale drift.

Moses et al. (2010a, 2011) studied three different ways of using two anchors that link the same old and new form tests in the common-item equating design. The overall results of this study suggested that when using two anchors, the poststratification approach works better than the imputation and propensity score matching approaches. Poststratification also produced more accurate SEEDs, quantities that are useful for evaluating competing equating and scaling functions.

4.6.3 Matched-Sample Equating

Equating based on samples with identical anchor score distributions was viewed as a potential solution to the variability seen across equating methods when equating samples of test takers were not equivalent (Dorans 1990c). Cook et al. (1988) discussed the need to equate achievement tests using samples of students who take the new and old forms at comparable points in the school year. Stocking et al. (1988) compared equating results obtained using representative and matched samples and concluded that matching equating samples on the basis of a fallible measure of ability is not advisable for any equating method, except possibly the Tucker equating method. Lawrence and Dorans (1988) compared equating results obtained using a representative old-form sample and an old-form sample matched to the new-form sample (matched sample) and found that results for the five studied equating methods tended to converge under the matched sample condition.

Lawrence and Dorans (1990), using the verbal anchor to create differences from the reference or base population and the pseudo-populations, demonstrated that the poststratification methods did best and the true-score methods did slightly worse than the chained method when the same verbal anchor was used for equating. Eignor et al. (1990a, b) used an IRT model to simulate data and found that the weakest results were obtained for poststratification on the basis of the verbal anchor and that the true-score methods were slightly better than the chained method. Livingston et al. (1990) used SAT-M scores to create differences in populations and examined the equating of SAT-V scores via multiple methods. The poststratification method produced the poorest results. They also compared equating results obtained using representative and matched samples and found that the results for all equating methods in the matched samples were similar to those for the Tucker and FE methods in the representative samples. In a follow-up study, Dorans and Wright (1993) compared equating results obtained using representative samples, samples matched on the basis of the equating set, and samples matched on the basis of a selection variable (i.e., a variable along which subpopulations differ) and indicated that matching on the selection variable improves accuracy over matching on the equating test for all methods. Finally, a study by Schmitt et al. (1990) indicated that matching on an anchor test score provides greater agreement among the results of the various equating procedures studied than were obtained under representative sampling.

4.6.4 Item Response Theory True-Score Linking

IRT true-score linking[4] was first used with TOEFL in 1979. Research on IRT-based linking methods received considerable attention in the 1980s to examine their applicability to other testing programs. ETS researchers have focused on a wide variety of research topics, including studies comparing non-IRT observed-score and IRT-based linking methods (including IRT true-score linking and IRT observed-score equating methods), studies comparing different IRT linking methods, studies examining the consequences of violation of assumptions on IRT equating, and so on. These studies are summarized here.

Marco et al. (1983a) examined the adequacy of various linear and curvilinear (observed-score methods) and ICC (one- and three-parameter logistic) equating models when certain sample and test characteristics were systematically varied. They found the 3PL model to be most consistently accurate. Using TOEFL data, Hicks (1983, 1984) evaluated three IRT variants and three conventional equating methods (Tucker, Levine and equipercentile) in terms of scale stability and found that the true-score IRT linking based on scaling by fixing the *b* parameters produces the least discrepant results. Lord and Wingersky (1983, 1984) compared IRT true-score linking with equipercentile equating using observed scores and concluded that the two methods yield almost identical results.

Douglass et al. (1985) studied the extent to which three approximations to the 3PL model could be used in item parameter estimation and equating. Although

[4] Several of the earlier studies cited in this section used the phrase IRT equating to describe the application of an IRT true-score equating function to linking two sets of observed scores. We are using the word linking because this procedure does not ensure that the linked scores are interchangeable in the sense described by Lord (1980) and Holland and Dorans (2006).

these approximations yielded accurate results (based on their circular equating criteria), the authors recommended further research before these methods are used operationally. Boldt (1993) compared linking based on the 3PL IRT model and a modified Rasch model (common nonzero lower asymptote) and concluded that the 3PL model should not be used if sample sizes are small. Tang et al. (1993) compared the performance of the computer programs LOGIST and BILOG (see Carlson and von Davier, Chap. 5, this volume, for more on these programs) on TOEFL 3PL IRT-based linking. The results indicated that the BILOG estimates were closer to the true parameter values in small-sample conditions. In a simulation study, Y. Li (2012) examined the effect of drifted (i.e., items performing differently than the remaining anchor items) polytomous anchor items on the test characteristic curve (TCC) linking and IRT true-score linking. Results indicated that drifted polytomous items have a relatively large impact on the linking results and that, in general, excluding drifted polytomous items from the anchor results in an improvement in equating results.

Kingston et al. (1985) compared IRT linking to conventional equating of the GMAT and concluded that violation of local independence had a negligible effect on the linking results. Cook and Eignor (1985) indicated that it was feasible to use IRT to link the four College Board Achievement tests used in their study. Similarly, McKinley and Kingston (1987) investigated the use of IRT linking for the GRE Subject Test in Mathematics and indicated that IRT linking was feasible for this test. McKinley and Schaefer (1989) conducted a simulation study to evaluate the feasibility of using IRT linking to reduce test form overlap of the GRE Subject Test in Mathematics. They compared double-part IRT true-score linking (i.e., linking to two old forms) with 20-item common-item blocks to triple-part linking (i.e., linking to three old forms) with 10-item common-item blocks. On the basis of the results of their study, they suggested using more than two links.

Cook and Petersen (1987) summarized a series of ETS articles and papers produced in the 1980s that examined how equating is affected by sampling errors, sample characteristics, and the nature of anchor items, among other factors. This summary added greatly to our understanding of the uses of IRT and conventional equating methods in suboptimal situations encountered in practice. Cook and Eignor (1989, 1991) wrote articles and instructional modules that provided a basis for understanding the process of score equating through the use of IRT. They discussed the merits of different IRT equating approaches.

A.A. von Davier and Wilson (2005, 2007) used data from the *Advanced Placement Program*® examinations to investigate the assumptions made by IRT true-score linking method and discussed the approaches for checking whether these assumptions are met for a particular data set. They provided a step-by-step check of how well the assumptions of IRT true-score linking are met. They also compared equating results obtained using IRT as well as traditional methods and showed that IRT and chained equipercentile equating results were close for most of the score range.

D. Li et al. (2012) compared the IRT true-score equating to chained equipercentile equating and observed that the sample variances for the chained equipercentile equating were much smaller than the variances for the IRT true-score equating, except at low scores.

4.6.5 Item Response Theory Preequating Research

In the early 1980s, IRT was evaluated for its potential in preequating tests developed from item pools. Bejar and Wingersky (1981) conducted a feasibility study for preequating the TWE and concluded that the procedure did not exhibit problems beyond those already associated with using IRT on this exam. Eignor (1985) examined the extent to which item parameters estimated on SAT-V and SAT-M pretest data could be used for equating purposes. The preequating results were mixed; three of the four equatings examined were marginally acceptable at best. Hypotheses for these results were posited by the author. Eignor and Stocking (1986) studied these hypotheses in a follow-up investigation and concluded that there was a problem either with the SAT-M data or the way in which LOGIST calibrated items under the 3PL model. Further hypotheses were generated. Stocking and Eignor (1986) investigated these results further and concluded that difference in ability across samples and multidimensionality may have accounted for the lack of item parameter invariance that undermined the preequating effort. While the SAT rejected the use of preequating on the basis of this research, during the 1990s, other testing programs moved to test administration and scoring designs, such as computer-adaptive testing, that relied on even more restrictive invariance assumptions than those that did not hold in the SAT studies.

Gao et al. (2012) investigated whether IRT true-score preequating results based on a Rasch model agreed with equating results based on observed operational data (postequating) for CLEP. The findings varied from subject to subject. Differences among the equating results were attributed to the manner of pretesting, contextual/order effects, or the violations of IRT assumptions. Davey and Lee (2011) examined the potential effect of item position on item parameter and ability estimates for the GRE revised General Test, which would use preequating to link scores obtained via its two-stage testing model. In an effort to mitigate the impact of position effects, they recommended that questions be pretested in random locations throughout the test. They also recommended considering the impact of speededness in the design of the revised test because multistage tests are more subject to speededness compared to linear forms of the same length and testing time.

4.6.6 Equating Tests With Constructed-Response Items

Large-scale testing programs often include CR as well as MC items on their tests. Livingston (2014b) listed some characteristics of CR tests (i.e., small number of tasks and possible raw scores, tasks that are easy to remember and require judgment for scoring) that cause problems when equating scores obtained from CR tests. Through the years, ETS researchers have tried to come up with innovative solutions to equating CR tests effectively.

When a CR test form is reused, raw scores from the two administrations of the form may not be comparable due to two different sets of raters among other reasons. The solution to this problem requires a rescoring, at the new administration, of test-taker responses from a previous administration. The scores from this "rescoring" are used as an anchor for equating, and this process is referred to as rater comparability scoring and equating (Puhan 2013b). Puhan (2013a, b) challenged conventional wisdom and showed theoretically and empirically that the choice of target population weights (for poststratification equating) has a predictable impact on final equating results obtained under the rater comparability scoring and equating scenario. The same author also indicated that chained linear equating produces more accurate equating results than Tucker equating under this equating scenario (Puhan 2012).

Kim et al. (2008a, b, 2010a, b) have compared various designs for equating CR-only tests, such as using an anchor test containing either common CR items or rescored common CR items or an external MC test and an equivalent-groups design incorporating rescored CR items (no anchor test). Results of their studies showed that the use of CR items without rescoring results in much larger bias than the other designs. Similarly, they have compared various designs for equating tests containing both MC and CR items such as using an anchor test containing only MC items, both MC and CR items, both MC and rescored CR items, and an equivalent-groups design incorporating rescored CR items (no anchor test). Results of their studies indicated that using either MC items alone or a mixed anchor without CR item rescoring results in much larger bias than the other two designs and that the equivalent-groups design with rescoring results in the smallest bias. Walker and Kim (2010) examined the use of an all-MC anchor for linking mixed-format tests containing both MC and CR items in a nonequivalent-groups design. They concluded that a MC-only anchor could effectively link two such test forms if either the MC or CR portion of the test measured the same knowledge and skills and if the relationship between the MC portion and the total test remained constant across the new and reference linking groups.

Because subpopulation invariance is considered a desirable property for equating relationships, Kim and Walker (2009b, 2012a) examined the appropriateness of the anchor composition in a mixed-format test, which includes both MC and CR items, using subpopulation invariance indices. They found that the mixed anchor was a better choice than the MC-only anchor to achieve subpopulation invariance between males and females. Muraki et al. (2000) provided an excellent summary describing issues and developments in linking performance assessments and included comparisons of common linking designs (single group, equivalent groups, nonequivalent groups) and linking methodologies (traditional and IRT).

Myford et al. (1995) pilot-tested a quality-control procedure for monitoring and adjusting for differences in reader performance and discussed steps that might enable different administrations of the TWE to be equated. Tan et al. (2010)

compared equating results using different sample sizes and equating designs (i.e., single group vs. common-item equating designs) to examine the possibility of reducing the rescoring sample. Similarly, Kim and Moses (2013) conducted a study to evaluate the conditions under which single scoring for CR items is as effective as double scoring in a licensure testing context. Results of their study indicated that under the conditions they examined, the use of single scoring would reduce scoring time and cost without increasing classification inconsistency. Y. Li and Brown (2013) conducted a rater comparability scoring and equating study and concluded that raters maintained the same scoring standards across administrations for the CRs in the *TOEFL iBT*® test Speaking and Writing sections. They recommended that the TOEFL iBT program use this procedure as a tool to periodically monitor Speaking and Writing scoring.

Some testing programs require all test takers to complete the same common portion of a test but offer a choice of essays in another portion of the test. Obviously there can be a fairness issue if the different essays vary in difficulty. ETS researchers have come up with innovative procedures whereby the scores on the alternate questions can be adjusted based on the estimated total group mean and standard deviation or score distribution on each alternate question (Cowell 1972; Rosenbaum 1985). According to Livingston (1988), these procedures tend to make larger adjustments when the scores to be adjusted are less correlated with scores on the common portion. He therefore suggested an adjustment procedure that makes smaller adjustments when the correlation between the scores to be adjusted and the scores on the common portion is low. Allen et al. (1993) examined Livingston's proposal, which they demonstrate to be consistent with certain missing data assumptions, and compared its adjustments to those from procedures that make different kinds of assumptions about the missing data that occur with essay choice.

In an experimental study, Wang et al. (1995) asked students to identify which items within three pairs of MC items they would prefer to answer, and the students were required to answer both items in each of the three pairs. The authors concluded that allowing choice will only produce fair tests when it is not necessary to allow choice. Although this study used tests with MC items only and involved small numbers of items and test takers, it attempted to answer via an experiment a question similar to what the other, earlier discussed studies attempted to answer, namely, making adjustments for test-taker choice among questions.

The same authors attempted to equate tests that allowed choice of questions by using existing IRT models and the assumption that the ICCs for the items obtained from test takers who chose to answer them are the same as the ICCs that would be obtained from the test takers who did not answer them (Wainer et al. 1991, 1994). Wainer and Thissen (1994) discussed several issues pertaining to tests that allow a choice to test takers. They provided examples where equating such tests is impossible and where allowing choice does not necessarily elicit the test takers' best performance.

4.6.7 Subscores

The demand for subscores has been increasing for a number of reasons, including the desire of candidates who fail the test to know their strengths and weaknesses in different content areas and because of mandates by legislatures to report subscores. Furthermore, states and academic institutions such as colleges and universities want a profile of performance for their graduates to better evaluate their training and focus on areas that need remediation. However, for subscores to be reported operationally, they should be comparable across the different forms of a test. One way to achieve comparability is to equate the subscores.

Sinharay and Haberman (2011a, b) proposed several approaches for equating augmented subscores (i.e., a linear combination of a subscore and the total score) under the nonequivalent groups with anchor test design. These approaches only differ in the way the anchor score is defined (e.g., using subscore, total score or augmented subscore as the anchor). They concluded that these approaches performed quite accurately under most practical situations, although using the total score or augmented subscore as the anchor performed slightly better than using only the subscore as the anchor. Puhan and Liang (2011a, b) considered equating subscores using internal common items or total scaled scores as the anchor and concluded that using total scaled scores as the anchor is preferable, especially when the internal common items are small.

4.6.8 Multidimensionality and Equating

The call for CR items and subscores on MC tests reflects a shared belief that a total score based on MC items underrepresents the construct of interest. This suggests that more than one dimension may exist in the data.

ETS researchers such as Cook et al. (1985) examined the relationship between violations of the assumption of unidimensionality and the quality of IRT true-score equating. Dorans and Kingston (Dorans and Kingston 1985; Kingston and Dorans 1982) examined the consequences of violations of unidimensionality assumptions on IRT equating and noted that although violations of unidimensionality may have an impact on equating, the effect may not be substantial. Using data from the LSAT, Camilli et al. (1995) examined the effect of multidimensionality on equating and concluded that violations of unidimensionality may not have a substantial impact on estimated item parameters and true-score equating tables. Dorans et al. (2014) did a comparative study where they varied content structure and correlation between underlying dimensions to examine their effect on latent-score and observed-score linking results. They demonstrated analytically and with simulated data that score equating is possible with multidimensional tests, provided the tests are parallel in content structure.

4.6.9 A Caveat on Comparative Studies

Sinharay and Holland (2008, 2010a, b) demonstrated that the equating method with explicit or implicit assumptions most consistent with the model used to generate the data performs best with those simulated data. When they compared three equating methods—the FE equipercentile equating method, the chained equipercentile equating method, and the IRT observed-score equating method—each one worked best in data consistent with its assumptions. The chained equipercentile equating method was never the worst performer. These studies by Sinharay and Holland provide a valuable lens from which to view the simulation studies summarized in Sect. 4.6 whether they used data simulated from a model or real test data to construct simulated scenarios: The results of the simulation follow from the design of the simulation. As Dorans (2014) noted, simulation studies may be helpful in studying the strengths and weakness of methods but cannot be used as a substitute for analysis of real data.

4.7 The Ebb and Flow of Equating Research at ETS

In this section, we provide a high-level summary of the ebb and flow of equating research reported in Sects. 4.2, 4.3, 4.5, and 4.6. We divide the period from 1947, the birth of ETS, through 2015 into four periods: (a) before 1970, (b) 1970s to mid-1980s, (c) mid-1980s to 2000, and (d) 2001–2015.

4.7.1 Prior to 1970

As might be expected, much of the early research on equating was procedural as many methods were introduced, including those named after Tucker and Levine (Sect. 4.4.1). Lord attended to the SEE (Sect. 4.5.1.1). There were early efforts to smooth data from small samples (Sect. 4.3.2.3). With the exception of work done by Lord in 1964, distinctions between equating and other forms of what is now called score linking did not seem to be made (Sect. 4.2.1).

4.7.2 The Year 1970 to the Mid-1980s

Equating research took on new importance in the late 1970s and early 1980s as test disclosure legislation led to the creation of many more test forms in a testing program than had been needed in the predisclosure period. This required novel data collection designs and led to the investigation of preequating approaches. Lord introduced his equating requirements (Sect. 4.2.1) and concurrently introduced IRT score linking methods, which became the subject of much research (Sects. 4.4.2 and 4.6.4). Lord estimated the SEE for IRT (Sect. 4.5.1.1). IRT preequating research was prevalent and generally discouraging (Sect. 4.6.5). Holland and his colleagues introduced section preequating (section 4.4.4) as another preequating solution to the problems posed by the test disclosure legislation.

4.7.3 The Mid-1980s to 2000

Equating research was more dormant in this period, as first differential item functioning and then computer-adaptive testing garnered much of the research funding at ETS. While some work was motivated by practice, such as matched-sample equating research (Sect. 4.6.3) and continued investigations of IRT score linking (Sect. 4.6.4), there were developments of theoretical import. Most notable among these were the development of kernel equating by Holland and his colleagues (Sects. 4.4.3 and 4.6.1), which led to much research about its use in estimating standard errors (Sect. 4.5.1.1). Claims made by some that scores from a variety of sources could be used interchangeably led to the development of cogent frameworks for distinguishing between different kinds of score linkings (Sect. 4.2.1). The role of dimensionality in equating was studied (Sect. 4.6.8).

4.7.4 The Years 2002–2015

The twenty-first century witnessed a surge of equating research. The kernel equating method and its use in estimating standard errors was studied extensively (Sects. 4.4.3, 4.5.1, 4.5.2, and 4.6.1). A new equating method was proposed by Haberman (Sect. 4.4.3).

Data collection and preparation received renewed interest in the areas of sample selection (Sect. 4.3.2.1) and weighting of samples (Sect. 4.3.2.2). A considerable amount of work was done on smoothing (Sect. 4.3.2.3), mostly by Moses and Holland and their colleagues. Livingston and Puhan and their colleagues devoted much attention to developing small-sample equating methods (Sect. 4.4.5).

CE was the focus of many comparative investigations (Sect. 4.6.1). The anchor continued to receive attention (Sect. 4.6.2). Equating subscores became an important issue as there were more and more calls to extract information from less and less (Sect. 4.6.7). The comparability problems faced by reliance on subjectively scored CR items began to be addressed (Sect. 4.6.6). The role of dimensionality in equating was examined again (Sect. 4.6.8).

Holland and Dorans provided a detailed framework for classes of linking (Sect. 4.2.1) as a further response to calls for linkages among scores from a variety of sources. Central to that framework was the litmus test of population invariance, which led to an area of research that uses equating to assess the fairness of test scores across subgroups (Sect. 4.5.2).

4.8 Books and Chapters

Books and chapters can be viewed as evidence that the authors are perceived as possessing expertise that is worth sharing with the profession. We conclude this chapter by citing the various books and chapters that have been authored by ETS staff in the area of score linking, and then we allude to work in related fields and forecast our expectation that ETS will continue to work the issues in this area.

An early treatment of score equating appeared in Gulliksen (1950), who described, among other things, Ledyard R Tucker's proposed use of an anchor test to adjust for differences in the abilities of samples. Tucker proposed this approach to deal with score equating problems with the SAT that occurred when the SAT started to be administered more than once a year to test takers applying to college. Books that dealt exclusively with score equating did not appear for more than 30 years, until the volume edited by ETS researchers Holland and Rubin (1982) was published. The 1980s was the first decade in which much progress was made in score equating research, spearheaded in large part by Paul Holland and his colleagues.

During the 1990s, ETS turned its attention first toward differential item functioning (Dorans, Chap. 7, this volume) and then toward CR and computer-adaptive testing. The latter two directions posed particular challenges to ensuring comparability of measurements, leaning more on strong assumptions than on an empirical basis. After a relatively dormant period in the 1990s, score equating research blossomed in the twenty-first century. Holland and his colleagues played major roles in this rebirth. The Dorans and Holland (2000a, b) article on the population sensitivity of score linking functions marked the beginning of a renaissance of effort on score equating research at ETS.

With the exception of early chapters by Angoff (1967, 1971), most chapters on equating prior to 2000 appeared between 1981 and 1990. Several appeared in the aforementioned Holland and Rubin (1982). Angoff (1981) provided a summary of procedures in use at ETS up until that time. Braun and Holland (1982) provided a formal mathematical framework to examine several observed-score equating procedures used at ETS at that time. Cowell (1982) presented an early application of IRT true-score linking, which was also described in a chapter by Lord (1982a). Holland and Wightman (1982) described a preliminary investigation of a linear section pre-equating procedure. Petersen et al. (1982) summarized the linear equating portion of a massive simulation study that examined linear and curvilinear methods of anchor test equating, ranging from widely used methods to rather obscure methods. Some anchors were external (did not count toward the score), whereas others were internal. They examined different types of content for the internal anchor. Anchors varied in difficulty. In addition, equating samples were randomly equivalent, similar, or dissimilar in ability. Rock (1982) explored how equating could be represented from the perspective of confirmatory factor analysis. Rubin (1982) commented on the chapter by Braun and Holland, whereas Rubin and Szatrowski (1982) critiqued the preequating chapter.

ETS researchers contributed chapters related to equating and linking in edited volumes other than Holland and Rubin's (1982). Angoff (1981) discussed equating and equity in a volume on new directions in testing and measurement circa 1980. Marco (1981) discussed the efforts of test disclosure on score equating in a volume on coaching, disclosure, and ethnic bias. Marco et al. (1983b) published the curvilinear equating analogue to their linear equating chapter that appeared in Holland and Rubin (1982) in a volume on latent trait theory and computer-adaptive testing. Cook and Eignor (1983) addressed the practical considerations associated with using IRT to equate or link test scores in a volume on IRT. Dorans (1990b) produced a chapter on scaling and equating in a volume on computer-adaptive testing edited by Wainer et al. (1990). Angoff and Cook (1988) linked scores across languages by relating the SAT to the College Board *PAA*™ test in a chapter on access and assessment for Hispanic students.

Since 2000, ETS authors have produced several books on the topics of score equating and score linking, including two quite different books, the theory-oriented unified statistical treatment of score equating by A.A. von Davier et al. (2004b) and an introduction to the basic concepts of equating by Livingston (2004). A.A. von Davier et al. (2004b) focused on a single method of test equating (i.e., kernel equating) in a unifying way that introduces several new ideas of general use in test equating. Livingston (2004) is a lively and straightforward account of many of the major issues and techniques. Livingston (2014b) is an updated version of his 2004 publication.

In addition to these two equating books were two edited volumes, one by Dorans et al. (2007) and one by A.A. von Davier (2011c). ETS authors contributed several chapters to both of these volumes.

There were six integrated parts to the volume *Linking and Aligning Scores and Scales* by Dorans et al. (2007). The first part set the stage for the remainder of the volume. Holland (2007) noted that linking scores or scales from different tests has a history about as long as the field of psychometrics itself. His chapter included a typology of linking methods that distinguishes among predicting, scaling, and equating. In the second part of the book, Cook (2007) considered some of the daunting challenges facing practitioners and discussed three major stumbling blocks encountered when attempting to equate scores on tests under difficult conditions: characteristics of the tests to be equated, characteristics of the groups used for equating, and characteristics of the anchor tests. A. A. von Davier (2007) addressed potential future directions for improving equating practices and included a brief introduction to kernel equating and issues surrounding assessment of the population sensitivity of equating functions. Educational testing programs in a state of transition were considered in the third part of the volume. J. Liu and Walker (2007) addressed score linking issues associated with content changes to a test. Eignor (2007) discussed linkings between test scores obtained under different modes of administration, noting why scores from computer-adaptive tests and paper-and-pencil tests cannot be considered equated. Concordances between tests built for a common purpose but in different ways were discussed by Dorans and Walker (2007) in a whimsical chapter that was part of the fourth part of the volume, which dealt with concordances. Yen (2007) examined the role of vertical scaling in the pre–No Child Left Behind (NCLB) era and the NCLB era in the fifth part, which was dedicated to vertical scaling. The sixth part dealt with relating the results obtained by surveys of educational achievement that provide aggregate results to tests designed to assess individual test takers. Braun and Qian (2007) modified and evaluated a procedure developed to link state standards to the National Assessment of Educational Progress scale and illustrated its use. In the book's postscript, Dorans et al. (2007) peered into the future and speculated about the likelihood that more and more linkages of dubious merit would be sought.

The A.A. von Davier (2011c) volume titled *Statistical Models for Test Equating, Scaling and Linking*, which received the American Educational Research Association 2013 best publication award, covered a wide domain of topics. Several chapters in the book addressed score linking and equating issues. In the introductory chapter of the book, A.A. von Davier (2011a) described the equating process as a feature of complex statistical models used for measuring abilities in standardized assessments and proposed a framework for observed-score equating methods. Dorans et al. (2011) emphasized the practical aspects of the equating process, the need for a solid data collection design for equating, and the challenges involved in applying specific equating procedures. Carlson (2011) addressed how to link vertically the results of tests that are constructed to intentionally differ in difficulty and content and that are taken by groups of test takers who differ in ability. Holland and Strawderman (2011) described a procedure that might be considered for averaging equating conversions that come from linkings to multiple old forms. Livingston and Kim (2011) addressed different approaches to dealing with the problems associated with equating test scores in small samples. Haberman (2011b) described the use of exponential families for continuizing test score distributions. Lee and von Davier (2011) discussed how various continuous variables with distributions (normal, logistic, and uniform) can be used as kernels to continuize test score distributions. Chen et al. (2011) described new hybrid models within the kernel equating framework, including a nonlinear version of Levine linear equating. Sinharay et al. (2011a) presented a detailed investigation of the untestable assumptions behind two popular nonlinear equating methods used with a nonequivalent-groups design. Rijmen et al. (2011) applied the SEE difference developed by A.A. von Davier et al. (2004b) to the full vector of equated raw scores and constructed a test for testing linear hypotheses about the equating results. D. Li et al. (2011) proposed the use of time series methods for monitoring the stability of reported scores over a long sequence of administrations.

ETS researchers contributed chapters related to equating and linking in edited volumes other than Dorans et al. (2007) and A. A. von Davier (2011c). Dorans (2000) produced a chapter on scaling and equating in a volume on computer-adaptive testing edited by Wainer et al. (2000). In a chapter in a volume dedicated to examining the adaptation of tests from one language to another, Cook and Schmitt-Cascallar (2005) reviewed different approaches to establishing score linkages on tests that are administered in different languages to different populations and critiqued three attempts to link the English-language SAT to the Spanish-language PAA over a 25-year period, including Angoff and Cook (1988) and Cascallar and Dorans (2005). In volume 26 of the *Handbook of Statistics*, dedicated to psychometrics and edited by Rao and Sinharay (2007), Holland et al. (2007) provided an introduction to test score equating, its data collection procedures, and methods used for equating. They also presented sound practices in the choice and evaluation of equating designs and functions and discussed challenges often encountered in practice.

Dorans and Sinharay (2011) edited a volume dedicated to feting the career of Paul Holland, titled *Looking Back*, in which the introductory chapter by Haberman (2011a) listed score equating as but one of Holland's many contributions. Three chapters on score equating were included in that volume. These three authors joined Holland and other ETS researchers in promoting the rebirth of equating research at ETS. Moses (2011) focused on one of Holland's far-reaching applications: his application of loglinear models as a smoothing method for equipercentile equating. Sinharay (2011) discussed the results of several studies that compared the performances of the poststratification equipercentile and chained equipercentile equating methods. Holland was involved in several of these studies. In a book chapter, A. A. von Davier (2011b) focused on the statistical methods available for equating test forms from standardized educational assessments that report scores at the individual level.

4.9 Concluding Comment

Lord (1980) stated that score equating is either not needed or impossible. Scores will be compared, however. As noted by Dorans and Holland (2000a),

> The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science. Psychological and educational measurement is no exception to this rule. Test equating techniques are those statistical and psychometric methods used to adjust scores obtained on different tests measuring the same construct so that they are comparable. (p. 281)

Procedures will attempt to facilitate these comparisons.

As in any scientific endeavor, instrument preparation and data collection are critical. With large equivalent groups of motivated test takers taking essentially parallel forms, the ideal of "no need to equate" is within reach. Score equating methods converge. As samples get small or contain unmotivated test takers or test takers with preknowledge of the test material, or as test takers take un-pretested tests that differ in content and difficulty, equating will be elusive. Researchers in the past have suggested solutions for suboptimal conditions. They will continue to do so in the future. We hope this compilation of studies will be valuable for future researchers who grapple with the inevitable less-than-ideal circumstances they will face when linking score scales or attempting to produce interchangeable scores via score equating.

References

Allen, N. L., Holland, P. W., & Thayer, D. T. (1993). *The optional essay problem and the hypothesis of equal difficulty* (Research Report No. RR-93-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01551.x

Angoff, W. H. (1953). *Equating of the ACE psychological examinations for high school students* (Research Bulletin No. RB-53-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1953.tb00887.x

Angoff, W. H. (1967). Technical problems of obtaining equivalent scores on tests. In W. A. Mehrens & R. L. Ebel (Eds.), *Principles of educational and psychological measurement: A book of selected readings* (pp. 84–86). Chicago: Rand McNally.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1981). Equating and equity. *New Directions for Testing and Measurement, 9*, 15–20.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores on the "Prueba de Apitud Academica" and the "Scholastic Aptitude Test"* (College Board Report No. 88-2). New York: College Board. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00259.x

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327–345. https://doi.org/10.1111/j.1745-3984.1986.tb00253.x

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (Research Report No. RR-81-35). Princeton: Educational Testing Service.

Boldt, R. F. (1972). *Anchored scaling and equating: Old conceptual problems and new methods* (Research Bulletin No. RB-72-28). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1972.tb01025.x

Boldt, R. F. (1993). *Simulated equating using several item response curves* (Research Report No. RR-93-57). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01568.x

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_17

Camilli, G., Wang, M.-M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79–96. https://doi.org/10.1111/j.1745-3984.1995.tb00457.x

Carlson, J. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_4

Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration involving methodological alternatives. *International Journal of Testing, 5*, 337–356. https://doi.org/10.1207/s15327574ijt0504_1

Chen, H. (2012). A comparison between linear IRT observed score equating and Levine observed score equating under the generalized kernel equating framework. *Journal of Educational Measurement, 49*, 269–284. https://doi.org/10.1111/j.1745-3984.2012.00175.x

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika, 75*, 542–557. https://doi.org/10.1007/s11336-010-9171-7

Chen, H., & Livingston, S. A. (2013). *Poststratification equating based on true anchor scores and its relationship to Levine observed score equating* (Research Report No. RR-13-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02318.x

Chen, H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 89–107). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_12

Cook, L. L. (1988). *Achievement test scaling* (Research Report No. RR-88-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00290.x

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73–88). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_5

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1985). *An investigation of the feasibility of applying item response theory to equate achievement tests* (Research Report No. RR-85-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00116.x

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*, 161–173. https://doi.org/10.1016/0883-0355(89)90004-9

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37–45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244. https://doi.org/10.1177/014662168701100302

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. Hambleton, P. F. Meranda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 171–192). Mahwah: Erlbaum.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (Research Report No. RR-85-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00115.x

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (Research Report No. RR-88-52). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00308.x

Cowell, W. R. (1972). *A technique for equating essay question scores* (Statistical Report No. SR-72-70). Princeton: Educational Testing Service.

Cowell, W. R. (1982). Item-response-theory pre-equating in the TOEFL testing program. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 149–161). New York: Academic Press.

Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised General Test* (Research Report No. RR-11-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02262.x

DeMauro, G. E. (1992). *An investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics* (Research Report No. RR-92-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01457.x

Dorans, N. J. (1990a). The equating methods and sampling designs. *Applied Measurement in Education, 3*, 3–17. https://doi.org/10.1207/s15324818ame0301_2

Dorans, N. J. (1990b). Scaling and equating. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137–160). Hillsdale: Erlbaum.

Dorans, N. J. (Ed.). (1990c). Selecting samples for equating: To match or not to match [Special issue]. *Applied Measurement in Education, 3*(1).

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (College Board Report No. 99-1). New York: College Board. https://doi.org/10.1002/j.2333-8504.1999.tb01800.x

Dorans, N. J. (2000). Scaling and equating. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 135–158). Hillsdale: Erlbaum.

Dorans, N. J. (2002a). *The recentering of SAT scales and its effects on score distributions and score interpretations* (College Board Research Report No. 2002-11). New York: College Board. https://doi.org/10.1002/j.2333-8504.2002.tb01871.x

Dorans, N. J. (2002b). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 59–84. https://doi.org/10.1111/j.1745-3984.2002.tb01135.x

Dorans, N. J. (Ed.). (2004a). Assessing the population sensitivity of equating functions. [Special issue]. *Journal of Educational Measurement, 41*(1).

Dorans, N. J. (2004b). Equating, concordance and expectation. *Applied Psychological Measurement, 28*, 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J. (2004c). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*(S1), S85–S94. https://doi.org/10.1007/s11136-006-9155-3

Dorans, N. J. (2013). On attempting to do what Lord said was impossible: Commentary on van der Linden's conceptual issues in observed-score equating. *Journal of Educational Measurement, 50*, 304–314. https://doi.org/10.1111/jedm.12017

Dorans, N. J. (2014). *Simulate to understand models, not nature* (Research Report No. RR-14-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12013

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and *PSAT/NMSQT*®. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum No. RM-94-10). Princeton: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000a). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., & Holland, P. W. (2000b). *Population invariance and the equatability of tests: Basic theory and the linear case* (Research Report No. RR-00-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2000.tb01842.x

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249–262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Dorans, N. J., & Lawrence, I. M. (1988). *Checking the equivalence of nearly identical test forms* (Research Report No. RR-88-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00262.x

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3*, 245–254. https://doi.org/10.1207/s15324818ame0303_3

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (Research Report No. RR-09-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02165.x

Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement, 49*, 1–18. https://doi.org/10.1111/j.1745-3984.2011.00157.x

Dorans, N. J., & Sinharay, S. (Eds.). (2011). *Looking back: Proceedings of a conference in honor of Paul W. Holland*. New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179–198). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_10

Dorans, N. J., & Wright, N. K. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01515.x

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

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81–97. https://doi.org/10.1177/0146621607311580

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010a). *Principles and practices of test score equating* (Research Report No. RR-10-29). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02236.x

Dorans, N. J., Liang, L., & Puhan, G. (2010b). *Aligning scales of certification tests* (Research Report No. RR-10-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02214.x

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Towards best practices. In A. A. von Davier (Ed.), *Statistical models for scaling, equating and linking* (pp. 21–42). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_2

Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (Research Report No. RR-14-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12041

Douglass, J. B., Marco, G. L., & Wingersky, M. S. (1985). *An evaluation of three approximate item response theory models for equating test scores* (Research Report No. RR-85-46). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00131.x

Duong, M., & von Davier, A. A. (2012). Observed-score equating with a heterogeneous target population. *International Journal of Testing, 12*, 224–251. https://doi.org/10.1080/15305058.2011.620725

Echternacht, G. (1971). *Alternate methods of equating GRE advanced tests* (GRE Board Professional Report No. GREB No. 69-2P). Princeton: Educational Testing Service.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of the pre-equating of the SAT Verbal and Mathematical sections* (Research Report No. RR-85-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00095.x

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_8

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of the possible causes of the inadequacy of IRT pre-equating* (Research Report No. RR-86-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00169.x

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990a). *The effect on observed- and true-score equating procedures of matching on a fallible criterion: A simulation with test variation* (Research Report No. RR-90-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1990.tb01361.x

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990b). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3*, 37–55. https://doi.org/10.1207/s15324818ame0301_4

Fan, C. T., & Swineford, F. (1954). *A method of score conversion through item statistics* (Research Bulletin No. RB-54-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1954.tb00243.x

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington DC: American Council on Education.

Gao, R., He, W., & Ruan, C. (2012). *Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data* (Research Report No. RR-12-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02294.x

Gonzalez, J., & von Davier, M. (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement, 50*, 315–320. https://doi.org/10.1111/jedm.12018

Grant, M. C. (2011). *The single group with nearly equivalent tests (SiGNET) design for equating very small volume multiple-choice tests* (Research Report No. RR-11-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02267.x

Grant, M. C., Zhang, Y., & Damiano, M. (2009). *An evaluation of kernel equating: Parallel equating with classical methods in the SAT Subject tests program* (Research Report No. RR-09-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02163.x

Gulliksen, H. (1950). *Theory of mental test scores*. New York: Wiley. https://doi.org/10.1037/13240-000

Gulliksen, H. (1968). Methods for determining equivalence of measures. *Psychological Bulletin, 70*, 534–544. https://doi.org/10.1037/h0026721

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika, 75*, 438–453. https://doi.org/10.1007/s11336-010-9160-x

Guo, H., & Oh, H.-J. (2009). *A study of frequency estimation equipercentile equating when there are large ability differences* (Research Report No. RR-09-45). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02202.x

Guo, H., & Puhan, G. (2014). Section pre-equating under the equivalent groups design without IRT. *Journal of Educational Measurement, 51*, 301–317. https://doi.org/10.1111/jedm.12049

Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (Research Report No. RR-11-46). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02282.x

Guo, H., Liu, J., Curley, E., Dorans, N., & Feigenbaum, M. (2012). *The stability of the score scale for the SAT Reasoning Test from 2005–2012* (Research Report No. RR-12-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02297.x

Guo, H., Puhan, G., & Walker, M. E. (2013). *A criterion to evaluate the individual raw-to-scale equating conversions* (Research Report No. RR-13-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02312.x

Haberman, S. J. (2008a). *Continuous exponential families: An equating tool* (Research Report No. RR-08-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02091.x

Haberman, S. J. (2008b). *Linking with continuous exponential families: Single-group designs* (Research Report No. RR-08-61). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02147.x

Haberman, S. (2010). *Limits on the accuracy of linking* (Research Report No. RR-10-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02229.x

Haberman, S. J. (2011a). The contributions of Paul Holland. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 3–17). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_1

Haberman, S. J. (2011b). Using exponential families for equating. In A. A. von Davier (Ed.), *Statistical models for scaling, equating and linking* (pp. 125–140). New York: Springer.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*, 254–273. https://doi.org/10.3102/1076998615574772

Haberman, S. J., & Dorans, N. J. (2011). *Sources of scale inconsistency* (Research Report No. RR-11-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02246.x

Haberman, S. J., & Yan, D. (2011). *Use of continuous exponential families to link forms via anchor tests* (Research Report No. RR-11-11), Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02247.x

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02196.x

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. *Applied Psychological Measurement, 7*, 255–266. https://doi.org/10.1177/014662168300700302

Hicks, M. M. (1984). *A comparative study of methods of equating TOEFL test scores* (Research Report No. RR-84-20). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00060.x

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_2

Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement, 50*, 286–294. https://doi.org/10.1111/jedm.12015

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.

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

Holland, P. W., & Strawderman, W. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 89–107). New York: Springer.

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating the graduate record examination* (Program Statistics Research Technical Report No. 81-51). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01278.x

Holland, P. W., & Thayer, D. T. (1984). *Section pre-equating in the presence of practice effects* (Research Report No. RR-84-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00047.x

Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational Statistics, 10*, 109–120. https://doi.org/10.2307/1164838

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

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Program Statistics Research Technical Report No. 89-84).Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01278.x

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating. A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Testing equating* (pp. 217–297). New York: Academic Press.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (Research Report No. RR-89-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00332.x

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design* (Research Report No. RR-06-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02023.x

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26, Psychometrics* (pp. 169–203). Amsterdam: Elsevier.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*(1), 17–43. https://doi.org/10.1111/j.1745-3984.2007.00050.x

Karon, B. P. (1956). The stability of equated test scores. *Journal of Experimental Education, 24*, 181–195. https://doi.org/10.1080/00220973.1956.11010539

Karon, B. P., & Cliff, R. H. (1957). *The Cureton–Tukey method of equating test scores* (Research Bulletin No. RB-57-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1957.tb00072.x

Kim, S., & Livingston, S. A. (2009). *Methods of linking with small samples in a common-item design: An empirical comparison* (Research Report No. RR-09-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02195.x

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*, 286–298. https://doi.org/10.1111/j.1745-3984.2010.00114.x

Kim, S., & Moses, T. P. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing, 13*, 314–328. https://doi.org/10.1080/15305058.2013.776050

Kim, S., & Walker, M. E. (2009a). *Effect of repeaters on score equating in a large scale licensure test* (Research Report No. RR-09-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02184.x

Kim, S., & Walker, M. E. (2009b). *Evaluating subpopulation invariance of linking functions to determine the anchor composition for a mixed-format test* (Research Report No. RR-09-36). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02193.x

Kim, S., & Walker, M. E. (2012a). Determining the anchor composition for a mixed-format test: Evaluation of subpopulation invariance of linking functions. *Applied Measurement in Education, 25*, 178–195. https://doi.org/10.1080/08957347.2010.524720

Kim, S., & Walker, M. E. (2012b). Examining repeater effects on chained equipercentile equating with common anchor items. *Applied Measurement in Education, 25*, 41–57. https://doi.org/10.1080/08957347.2012.635481

Kim, S., von Davier, A. A., & Haberman, S. J. (2006). *An alternative to equating with small samples in the non-equivalent groups anchor test design* (Research Report No. RR-06-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02033.x

Kim, S., von Davier, A. A., & Haberman, S. (2007). *Investigating the effectiveness of a synthetic linking function on small sample equating* (Research Report No. RR-07-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02079.x

Kim, S., Walker, M. E., & McHale, F. (2008a). *Comparisons among designs for equating constructed response tests* (Research Report No. RR-08-53). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02139.x

Kim, S., Walker, M. E., & McHale, F. (2008b). *Equating of mixed-format tests in large-scale assessments* (Research Report No. RR-08-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02112.x

Kim, S., von Davier, A. A., & Haberman, S. (2008c). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*, 325–342. https://doi.org/10.1111/j.1745-3984.2008.00068.x

Kim, S., Livingston, S. A., & Lewis, C. (2008d). *Investigating the effectiveness of collateral information on small-sample equating.* (Research Report No. RR-08-52). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02138.x

Kim, S., Livingston, S. A., & Lewis, C. (2009). *Evaluating sources of collateral information on small-sample equating* (Research Report No. RR-09-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02171.x

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*, 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigation the effectiveness of equating designs for constructed response tests in large scale assessment. *Journal of Educational Measurement, 47*, 186–201. https://doi.org/10.1111/j.1745-3984.2010.00108.x

Kim, S., von Davier, A. A., & Haberman, S. (2011). Practical application of a synthetic linking function on small sample equating. *Applied Measurement in Education, 24*, 95–114. http://dx.doi.org/10.1080/08957347.2011.554601

Kim, S., Walker, M. E., & Larkin, K. (2012). Examining possible construct changes to a licensure test by evaluating equating requirements. *International Journal of Testing, 12*, 365–381. https://doi.org/10.1080/15305058.2011.645974

Kingston, N. M., & Dorans, N. J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE aptitude test* (Research Report No. RR-82-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01298.x

Kingston, N. M., & Holland, P. W. (1986). *Alternative methods for equating the GRE general test* (Research Report No. RR-86-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00171.x

Kingston, N. M., Leary, L. F., & Wightman, L. E. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (Research Report No. RR-85-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00119.x

Koutsopoulos, C. J. (1961). *A linear practice effect solution for the counterbalanced case of equating* (Research Bulletin No. RB-61-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00287.x

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (Research Report No. RR-88-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00279.x

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*, 19–36. https://doi.org/10.1207/s15324818ame0301_3

Lee, Y.-H., & Haberman, S. H. (2013). Harmonic regression and scale stability. *Psychometrika, 78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & von Davier, A. A. (2008). *Comparing alternative kernels for the kernel method of test equating: Gaussian, logistic, and uniform kernels* (Research Report No. RR-08-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02098.x

Lee, Y.-H., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 159–173). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_10

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*, 557–575. https://doi.org/10.1007/s11336-013-9317-5

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. RB-55-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1955.tb00266.x

Li, D., Li, S., & von Davier, A. A. (2011). Applying time-series analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_20

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement, 49*, 167–189. https://doi.org/10.1111/j.1745-3984.2012.00167.x

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report No. RR-12-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02291.x

Li, Y., & Brown, T. (2013). *A trend-scoring study for the TOEFL iBT Speaking and Writing Sections* (Research Memorandum No. RM-13-05). Princeton: Educational Testing Service.

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (Research Report No. RR-09-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02162.x

Liao, C. (2013). *An evaluation of differential speededness and its impact on the common item equating of a reading test* (Research Memorandum No. RM-13-02). Princeton: Educational Testing Service.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102. https://doi.org/10.1207/s15324818ame0601_5

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*, 259–286. https://doi.org/10.3102/10769986020003259

Liou, M., Cheng, P. E., & Johnson, E. G. (1996). *Standard errors of the kernel equating methods under the common-item design* (Research Report No. RR-96-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01689.x

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*, 349–369. https://doi.org/10.1177/01466216970214005

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*, 15–22. https://doi.org/10.1111/emip.12001

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32*, 27–44. https://doi.org/10.1177/0146621607311576.

Liu, J., & Low, A. C. (2007). *An exploration of kernel equating using SAT data: Equating to a similar population and to a distant population* (Research Report No. RR-07-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02059.x

Liu, J., & Low, A. (2008). A comparison of the kernel equating method with the traditional equating methods using SAT data. *Journal of Educational Measurement, 45*, 309–323. https://doi.org/10.1111/j.1745-3984.2008.00067.x

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_7

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2009a). *The effect of different types of anchor test on observed score equating* (Research Report No. RR-09-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02198.x

Liu, J., Moses, T. P., & Low, A. C. (2009b). *Evaluation of the effects of loglinear smoothing models on equating functions in the presence of structured data irregularities* (Research Report No. RR-09-22). Princeton*:* Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02179.x

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011a). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361. https://doi.org/10.1177/0013164410375571

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2011b). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Journal of Educational Measurement, 48*, 361–379. https://doi.org/10.1111/j.1745-3984.2011.00150.x

Liu, J., Guo, H., & Dorans, N. J. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design*

(Research Report No. RR-14-13). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12014

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32*, 27–44. https://doi.org/10.1177/0146621607311576.

Liu, Y., Shultz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33*, 257–278. https://doi.org/10.3102/1076998607306076.

Livingston, S. A. (1988). *Adjusting scores on examinations offering a choice of essay questions* (Research Report No. RR-88-64). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00320.x

Livingston, S. A. (1993a). *An empirical tryout of kernel equating* (Research Report No. RR-93-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01544.x

Livingston, S. A. (1993b). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23–39. https://doi.org/10.1111/j.1745-3984.1993.tb00420.x.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton: Educational Testing Service.

Livingston, S. A. (2014a). *Demographically adjusted groups for equating test scores* (Research Report No. RR-14-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12030

Livingston, S. A. (2014b). *Equating test scores (without IRT)* (2nd ed.). Princeton: Educational Testing Service.

Livingston, S. A., & Feryok, N. J. (1987). *Univariate versus bivariate smoothing in frequency estimation equating* (Research Report No. RR-87-36). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00240.x

Livingston, S. A., & Kim, S. (2008). *Small sample equating by the circle-arc method* (Research Report No. RR-08-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02125.x

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

Livingston, S. A., & Kim, S. (2010a). *An empirical comparison of methods for equating with randomly equivalent groups of 50 to 400 test takers* (Research Report No. RR-10-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02212.x

Livingston, S. A., & Kim, S. (2010b). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*, 175–185. https://doi.org/10.1111/j.1745-3984.2010.00107.x

Livingston, S. A., & Kim, S. (2011). New approaches to equating with small samples. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 109–122). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_7

Livingston, S. A., & Lewis, C. (2009). *Small-sample equating with prior information* (Research Report No. RR-09-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02182.x

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95. https://doi.org/10.1207/s15324818ame0301_6

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin No. RB-50-48). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00673.x

Lord, F. M. (1954). *Equating test scores: The maximum likelihood solution for a common item equating problem* (Research Bulletin No. RB-54-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1954.tb00040.x

Lord, F. M. (1955). Equating test scores: A maximum likelihood solution. *Psychometrika, 20*, 193–200. https://doi.org/10.1007/BF02289016

Lord, F. M. (1964a). Nominally and rigorously parallel test forms. *Psychometrika, 29*, 335–345. https://doi.org/10.1007/BF02289600

Lord, F. M. (1964b). *Rigorously and nonrigorously parallel test forms* (Research Bulletin No. RB-64-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1964.tb00323.x

Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory* (Research Bulletin No. RB-75-13). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01052.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum Associates.

Lord, F. M. (1981). *Standard error of an equating by item response theory* (Research Report No. RR-81-49). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01276.x

Lord, F. M. (1982a). Item response theory and equating: A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 141–148). New York: Academic Press. https://doi.org/10.2307/1164642

Lord, F. M. (1982b). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*, 165–174. https://doi.org/10.2307/1164642

Lord, F. M., & Wingersky, M. S. (1983). *Comparison of IRT observed-score and true-score equatings* (Research Report No. RR-83-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1983.tb00026.x

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed score equatings. *Applied Psychological Measurement, 8*, 453–461. https://doi.org/10.1177/014662168400800409

Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS data* (Research Report No. RR-06-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02036.x

Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.), *New directions for testing and measurement: Issues in testing—coaching, disclosure, and ethnic bias* (pp. 105–122). San Francisco: Jossey-Bass.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983a). *A large-scale evaluation of linear and curvilinear score equating models* (Research Memorandum No. RM-83-02). Princeton: Educational Testing Service.

Marco, G. L., Stewart, E. E., & Petersen, N. S. (1983b). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 147–177). New York: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50018-4

McHale, F. J., & Ninneman, A. M. (1994). *The stability of the score scale for the Scholastic Aptitude Test from 1973 to 1984* (Statistical Report No. SR-94-27). Princeton: Educational Testing Service.

McKinley, R. L., & Kingston, N. M. (1987). *Exploring the use of IRT equating for the GRE Subject Test in Mathematics* (Research Report No. RR-87-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00225.x

McKinley, R. L., & Schaefer, G. (1989). *Reducing test form overlap of the GRE Subject Test in Mathematics using IRT triple-part equating* (Research Report No. RR-89-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00334.x

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Report). Princeton: Educational Testing Service.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT-Verbal score scale* (Research Bulletin No. RB-75-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01048.x

Moses, T. P. (2006). *Using the kernel method of test equating for estimating the standard errors of population invariance measures* (Research Report No. RR-06-20). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02026.x

Moses, T. P. (2008a). *An evaluation of statistical strategies for making equating function selections* (Research Report No. RR-08-60). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02146.x

Moses, T. P. (2008b). Using the kernel method of test equating for estimating the standard errors of population invariance measures. *Journal of Educational and Behavioral Statistics, 33*, 137–157. https://doi.org/10.3102/1076998607302634

Moses, T. P. (2009). A comparison of statistical significance tests for selecting equating functions. *Applied Psychological Measurement, 33*, 285–306. https://doi.org/10.1177/0146621608321757

Moses, T. P. (2011). Log-linear models as smooth operators: Holland's statistical applications and their practical uses. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 185–202). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_10

Moses, T. P. (2014). Alternative smoothing and scaling strategies for weighted composite scores. *Educational and Psychological Measurement, 74*, 516–536. https://doi.org/10.1177/0013164413507725

Moses, T. P., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (Research Report No. RR-07-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02057.x

Moses, T. P., & Holland, P. W. (2008). *The influence of strategies for selecting loglinear smoothing models on equating functions* (Research Report No. RR-08-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02111.x

Moses, T. P., & Holland, P. W. (2009a). *Alternative loglinear smoothing models and their effect on equating function accuracy* (Research Report No. RR-09-48). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02205.x

Moses, T. P., & Holland, P. W. (2009b). *Selection strategies for bivariate loglinear smoothing models and their effects on NEAT equating functions* (Research Report No. RR-09-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02161.x

Moses, T. P., & Holland, P. W. (2009c). Selection strategies for univariate loglinear smoothing models and their effects on equating function accuracy. *Journal of Educational Measurement, 46*, 159–176. https://doi.org/10.1111/j.1745-3984.2009.00075.x

Moses, T. P., & Holland, P. W. (2010a). A comparison of statistical selection strategies for univariate and bivariate loglinear smoothing models. *British Journal of Mathematical and Statistical Psychology, 63*, 557–574. https://doi.org/10.1348/000711009X478580

Moses, T. P., & Holland, P. W. (2010b). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47*(1), 76–91. https://doi.org/10.1111/j.1745-3984.2009.00100.x

Moses, T. P., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (Research Report No. RR-07-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02058.x

Moses, T. P., & Oh, H. (2009). *Pseudo Bayes estimates for test score distributions and chained equipercentile equating* (Research Report No. RR-09-47). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02204.x

Moses, T. P., & von Davier, A. A. (2006). *A SAS macro for loglinear smoothing: Applications and implications* (Research Report No. RR-06-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02011.x

Moses, T. P., & von Davier, A. A. (2013). A SAS IML macro for loglinear smoothing. *Applied Psychological Measurement, 35*, 250–251. https://doi.org/10.1177/0146621610369909

Moses, T. P., & Zhang, W. (2010). *Research on standard errors of equating differences* (Research Report No. RR-10-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02232.x

Moses, T. P., & Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions, and simulations. *Journal of Educational and Behavioral Statistics, 36*, 779–803. https://doi.org/10.3102/1076998610396892

Moses, T. P., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (Research Report No. RR-04-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01954.x

Moses, T. P., Liu, J., & Dorans, N. J. (2009). *Systematized score equity assessment in SAS* (Research Memorandum No. RM-09-08). Princeton: Educational Testing Service.

Moses, T. P., Deng, W., & Zhang, Y.-L. (2010a). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (Research Report No. RR-10-23). Princeton*:* Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02230.x

Moses, T. P., Liu, J., & Dorans, N. J. (2010b). Systemized SEA in SAS. *Applied Psychological Measurement, 34*, 552–553. https://doi.org/10.1177/0146621610369909

Moses, T. P., Deng, W., & Zhang, Y.-L. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement, 35*, 362–379. https://doi.org/10.1177/0146621611405510

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337. https://doi.org/10.1177/01466210022031787

Myford, C., Marr, D. B., & Linacre, J. M. (1995). *Reader calibration and its potential role in equating for the Test of Written English* (Research Report No. RR-95-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01674.x

Oh, H., & Moses, T. P. (2012). Comparison of the one- and bi-direction chained equipercentile equating. *Journal of Educational Measurement, 49*, 399–418. https://doi.org/10.1111/j.1745-3984.2012.00183.x

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156. https://doi.org/10.2307/1164922

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Pommerich, M., & Dorans, N. J. (Eds.). (2004). Concordance. [Special issue]. *Applied Psychological Measurement, 28*(4).

Puhan, G. (2007). *Scale drift in equating on a test that employs cut scores* (Research Report No. RR-07-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02076.x

Puhan, G. (2009a). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education, 22*, 79–103. https://doi.org/10.1080/08957340802558391

Puhan, G. (2009b). *What effect does the inclusion or exclusion of repeaters have on test equating?* (Research Report No. RR-09-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02176.x

Puhan, G. (2010a). *Chained versus post stratification equating: An evaluation using empirical data* (Research Report No. RR-10-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02213.x

Puhan, G. (2010b). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*, 54–75. https://doi.org/10.1111/j.1745-3984.2009.00099.x

Puhan, G. (2011a). *Can smoothing help when equating with unrepresentative small samples* (Research Report No. RR-11-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02245.x

Puhan, G. (2011b). Futility of log linear smoothing when equating with unrepresentative small samples. *Journal of Educational Measurement, 48*, 274–292. https://doi.org/10.1111/j.1745-3984.2011.00147.x

Puhan, G. (2011c). Impact of inclusion or exclusion of repeaters on test equating. *International Journal of Testing, 11*, 215–230. https://doi.org/10.1080/15305058.2011.555575

Puhan, G. (2012). Tucker versus chained linear equating in two equating situations—Rater comparability scoring and randomly equivalent groups with an anchor. *Journal of Educational Measurement, 49*, 313–330. https://doi.org/10.1111/j.1745-3984.2012.00177.x.

Puhan, G. (2013a). *Choice of target population weights in rater comparability scoring and equating* (Research Report No. RR-13-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02310.x

Puhan, G. (2013b). Rater comparability scoring and equating: Does choice of target population weights matter in this context? *Journal of Educational Measurement, 50*, 374–380. https://doi.org/10.1111/jedm.12023

Puhan, G., & Liang, L. (2011a). Equating subscores under the non-equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice, 30*(1), 23–35. https://doi.org/10.1111/j.1745-3992.2010.00197.x

Puhan, G., & Liang, L. (2011b). *Equating subscores using total scaled scores as the anchor test* (Research Report No. RR-11-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02243.x

Puhan, G., Moses, T. P., Grant, M., & McHale, F. (2008a). *An alternative data collection design for equating with very small samples* (Research Report No. RR-08-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02097.x

Puhan, G., von Davier, A. A., & Gupta, S. (2008b). *Impossible scores resulting in zero frequencies in the anchor test: Impact on smoothing and equating* (Research Report No. RR-08-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02096.x

Puhan, G., Moses, T. P., Grant, M., & McHale, F. (2009). Small sample equating using a single group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement, 46*, 344–362. https://doi.org/10.1111/j.1745-3984.2009.00085.x

Puhan, G., von Davier, A. A., & Gupta, S. (2010). A brief report on how impossible scores impact smoothing and equating. *Educational and Psychological Measurement, 70*, 953–960. https://doi.org/10.1177/0013164410382731

Qian, J., von Davier, A. A., & Jiang, Y. (2013). *Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating* (Research Report No. RR-13-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02346.x

Rao, C. R., & Sinharay, S. (Eds.). (2007). *Psychometrics* (Handbook of statistics, Vol. 26). Amsterdam: Elsevier.

Ricker, K. L., & von Davier, A. A. (2007). *The impact of anchor test lengths on equating results in a nonequivalent groups design* (Research Report No. RR-07-44). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02086.x

Rijmen, F., Manalo, J., & von Davier, A. A. (2009). Asymptotic and sampling-based standard errors for two population invariance measures in the linear equating case. *Applied Psychological Measurement, 33*, 222–237. https://doi.org/10.1177/0146621608323927

Rijmen, F., Qu, Y., & von Davier, A. A. (2011). Hypothesis testing of equating differences in the kernel equating framework. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 317–326). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_19

Rock, D. R. (1982). Equating using the confirmatory factor analysis model. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 247–257). New York: Academic Press.

Rosenbaum, P. R. (1985). *A generalization of adjustment, with an application to the scaling of essay scores* (Research Report No. RR-85-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00087.x

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40*, 43–49. https://doi.org/10.1111/j.2044-8317.1987.tb00866.x

Rubin, D. B. (1982). Discussion of "Partial orders and partial exchangeability in test theory." In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 339–341). New York: Academic Press.

Rubin, D. B., & Szatrowski, T. (1982). Discussion of "Section pre-equating: A preliminary investigation." In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 301–306). New York: Academic Press.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3*, 53–71. https://doi.org/10.1207/s15324818ame0301_5

Schultz, D. G., & Wilks, S. S. (1950). *A method for adjusting for lack of equivalence in groups* (Research Bulletin No. RB-50-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00682.x

Sinharay, S. (2011). Chain equipercentile equating and frequency estimation equipercentile equating: Comparisons based on real and stimulated data. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 203–219). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_11

Sinharay, S., & Haberman, S. J. (2011a). Equating of augmented subscores. *Journal of Educational Measurement, 48*, 122–145. https://doi.org/10.1111/j.1745-3984.2011.00137.x

Sinharay, S., & Haberman, S. J. (2011b). *Equating of subscores and weighted averages under the NEAT design* (Research Report No. RR-11-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02237.x

Sinharay, S., & Holland, P. W. (2006a). *Choice of anchor test in equating* (Research Report No. RR-06-35). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02040.x

Sinharay, S., & Holland, P. W. (2006b). *The correlation between the scores of a test and an anchor test* (Research Report No. RR-06-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02010.x

Sinharay, S., & Holland, P. W. (2008). *The missing data assumption of the NEAT design and their implications for test equating* (Research Report No. RR-09-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02173.x

Sinharay, S., & Holland, P. W. (2010a). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*, 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x

Sinharay, S., & Holland, P. W. (2010b). The missing data assumption of the NEAT design and their implications for test equating. *Psychometrika, 75*, 309–327. https://doi.org/10.1007/s11336-010-9156-6

Sinharay, S., Holland, P. W., & von Davier, A. A. (2011a). Evaluating the missing data assumptions of the chain and poststratification equating methods. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 281–296). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_17

Sinharay, S., Dorans, N. J., & Liang, L. (2011b). First language of examinees and fairness assessment procedures. *Educational Measurement: Issues and Practice, 30*, 25–35. https://doi.org/10.1111/j.1745-3992.2011.00202.x

Sinharay, S., Haberman, S., Holland, P. W., & Lewis, C. (2012). *A note on the choice of an anchor test in equating* (Research Report No. RR-12-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02296.x

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT pre-equating* (Research Report No. RR-86-49). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00204.x

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (Research Report No. RR-88-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00297.x

Swineford, F., & Fan, C. T. (1957). A method of score conversion through item statistics. *Psychometrika, 22*, 185–188. https://doi.org/10.1007/BF02289053

Tan, X., Ricker, K., & Puhan, G. (2010). *Single versus double scoring of trend responses in trend score equating with constructed response tests* (Research Report No. RR-10-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02219.x

Tang, L. K., Way, W. D., & Carey, P. A. (1993). *The effect of small calibration sample sizes on TOEFL IRT-based equating* (Research Report No. RR-93-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01570.x

Thayer, D. T. (1983). Maximum likelihood estimation of the joint covariance matrix for sections of tests given to distinct samples with application to test equating. *Psychometrika, 48*, 293–297. https://doi.org/10.1007/BF02294023

Tucker, L. (1951). *Notes on the nature of gamble in test score scaling* (Research Bulletin No. RB-51-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1951.tb00226.x

von Davier, A. A. (2003). *Notes on linear equating methods for the non-equivalent-groups design* (Research Report No. RR-03-24). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01916.x

von Davier, A. A. (2007). Potential solutions to practical equating issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 89–106). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_6

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Journal of Educational and Behavioral Statistics, 33*, 186–203. https://doi.org/10.3102/1076998607302633

von Davier, A. A. (2011a). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 1–17). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A. (2011b). An observed-score equating framework. In N. J. Dorans & S. Sinharay (Eds.), *A festschrift for Paul W. Holland* (pp. 221–237). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_12

von Davier, A. A. (Ed.). (2011c). *Statistical models for test equating, scaling and linking*. New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A. (2013). Observed score equating: An overview. *Psychometrika, 78*, 605–623. https://doi.org/10.1007/s11336-013-9319-3

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the non-equivalent groups design. *Journal of Educational and Behavioral Statistics, 30*, 313–334. https://doi.org/10.3102/10769986030003313

von Davier, A. A., & Liu, M. (Eds.). (2007). Population invariance. [Special issue]. *Applied Psychological Measurement, 32*(1).

von Davier, A. A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement, 67*, 940–957. https://doi.org/10.1177/0013164407301543

von Davier, A. A., & Wilson, C. (2008). Investigation the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11–26. https://doi.org/10.1177/0146621607311560

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods of observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15–32. https://doi.org/10.1111/j.1745-3984.2004.tb01156.x von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. New York: Springer. https://doi.org/10.1007/b97446

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data* (Research Report No. RR-06-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02008.x

von Davier, A.A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed score equating function using the methods of kernel equating* (Research Report No. RR-07-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02056.x

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linkage and scale transformations. *Methodology, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115.

von Davier, M., Gonzalez, J., & von Davier, A. A. (2013). Local equating using the Rasch model, the OPLM, and the 2PL IRT model—or—What is it anyway if the model captures everything there is to know about the test takers? *Journal of Educational Measurement, 50*, 295–303. https://doi.org/10.1111/jedm.12016

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159–195. https://doi.org/10.3102/00346543064001159

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R., Sternberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale: Lawrence Erlbaum Associates.

Wainer, H., Wang, X.-B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by the examinees* (Research Report No. RR-91-57). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01424.x

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinee choice? *Journal of Educational Measurement, 31*, 183–199. https://doi.org/10.1111/j.1745-3984.1994.tb00442.x

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R., Sternberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale: Erlbaum.

Walker, M. E., & Kim, S. (2010). *Examining two strategies to link mixed-format tests using multiple-choice anchors* (Research Report No. RR-10-18). Princeton: Educational Testing Service.. https://doi.org/10.1002/j.2333-8504.2010.tb02225.x

Wang, X.-B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8*, 211–225. https://doi.org/10.1207/s15324818ame0803_2

Wiberg, M., van der Linden, W., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement, 51*, 57–74. https://doi.org/10.1111/jedm.12034

Wightman, L. E., & Wightman, L. F. (1988). *An empirical investigation of one variable section pre-equating* (Research Report No. RR-88-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00293.x

Yang, W.-L. (2004). Sensitivity of linkings between *AP*® multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33–41. https://doi.org/10.1111/j.1745-3984.2004.tb01157.x

Yang, W.-L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement, 32*, 45–61. https://doi.org/10.1177/0146621607311577

Yang, W.-L., Bontya, A. M., & Moses, T. P. (2011). *Repeater effects on score equating for a graduate admissions exam* (Research Report No. RR-11-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02253.x

Yen, W. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_15

Zu, J., & Liu, J. (2009). *Comparison of the effects of discrete anchor items and passage-based anchor items on observed-score equating results* (Research Report No. RR-09-44). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02201.x

Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement, 47*, 395–412. https://doi.org/10.1111/j.1745-3984.2010.00120.x

Zu, J., & Puhan, G. (2014). Pre-equating with empirical item characteristic curves: An observed-score pre-equating method. *Journal of Educational Measurement, 51*, 281–300. https://doi.org/10.1111/jedm.12047

Zu, J., & Yuan, K. (2012). Standard error of linear observed-score equating for the neat design with nonnormally distributed data. *Journal of Educational Measurement, 49*, 190–213. https://doi.org/10.1111/j.1745-3984.2012.00168.x

Chapter 5
Item Response Theory

James E. Carlson and Matthias von Davier

Item response theory (IRT) models, in their many forms, are undoubtedly the most widely used models in large-scale operational assessment programs. They have grown from negligible usage prior to the 1980s to almost universal usage in large-scale assessment programs, not only in the United States, but in many other countries with active and up-to-date programs of research in the area of psychometrics and educational measurement.

Perhaps the most important feature leading to the dominance of IRT in operational programs is the characteristic of estimating individual item locations (difficulties) and test-taker locations (abilities) separately, but on the same scale, a feature not possible with classical measurement models. This estimation allows for tailoring tests through judicious item selection to achieve precise measurement for individual test takers (e.g., in computerized adaptive testing, CAT) or for defining important cut points on an assessment scale. It also provides mechanisms for placing different test forms on the same scale (linking and equating). Another important characteristic of IRT models is local independence: for a given location of test takers on the scale, the probability of success on any item is independent of that of every other item on that scale. This characteristic is the basis of the likelihood function used to estimate test takers' locations on the scale.

Few would doubt that ETS researchers have contributed more to the general topic of IRT than individuals from any other institution. In this chapter we briefly review most of those contributions, dividing them into sections by decades of publication. Of course, many individuals in the field have changed positions between

This chapter was originally published in 2013 as a research report in the ETS R&D Scientific and Policy Contributions Series.

This work was conducted while M. von Davier was employed with Educational Testing Service.

J.E. Carlson (✉) • M. von Davier
Educational Testing Service, Princeton, NJ, USA
e-mail: jcarlson@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_5

different testing agencies and universities over the years, some having been at ETS during more than one period of time. This chapter includes some contributions made by ETS researchers before taking a position at ETS, and some contributions made by researchers while at ETS, although they have since left. It is also important to note that IRT developments at ETS were not made in isolation. Many contributions were collaborations between ETS researchers and individuals from other institutions, as well as developments that arose from communications with others in the field.

5.1 Some Early Work Leading up to IRT (1940s and 1950s)

Tucker (1946) published a precursor to IRT in which he introduced the term *item characteristic curve*, using the normal ogive model (Green 1980).[1] Green stated:

> Workers in IRT today are inclined to reference Birnbaum in Novick and Lord [sic] when needing historical perspective, but, of course Lord's 1955 monograph, done under Tuck's direction, precedes Birnbaum, and Tuck's 1946 paper precedes practically everybody. He used normal ogives for item characteristic curves, as Lord did later. (p. 4)

Some of the earliest work leading up to a complete specification of IRT was carried out at ETS during the 1950s by Lord and Green. Green was one of the first two psychometric fellows in the joint doctoral program of ETS and Princeton University. Note that the work of Lord and Green was completed prior to Rasch's (1960) publication describing and demonstrating the one-parameter IRT model, although in his preface Rasch mentions modeling data in the mid-1950s, leading to what is now referred to as the Rasch model. Further background on the statistical and psychometric underpinnings of IRT can be found in the work of a variety of authors, both at and outside of ETS (Bock 1997; Green 1980; Lord 1952a, b, 1953).[2]

Lord (1951, 1952a, 1953) discussed test theory in a formal way that can be considered some of the earliest work in IRT. He introduced and defined many of the now common IRT terms such as item characteristic curves (ICCs), test characteristic curves (TCCs), and standard errors conditional on latent ability.[3] He also

[1] Green stated that Tucker was at Princeton and ETS from 1944 to 1960; as head of statistical analysis at ETS, Tucker was responsible for setting up the statistical procedures for test and item analysis, as well as equating.

[2] These journal articles by Green and Lord are based on their Ph.D. dissertations at Princeton University, both presented in 1951.

[3] Lord (1980a, p. 19) attributes the term *local independence* to Lazarsfeld (1950) and mentions that Lazarsfeld used the term *trace line* for a curve like the ICC. Rasch (1960) makes no mention of the earlier works referred to by Lord so we have to assume he was unaware of them or felt they were not relevant to his research direction.

discussed what we now refer to as local independence and the invariance of item parameters (not dependent on the ability distribution of the test takers). His 1953 article is an excellent presentation of the basics of IRT, and he also mentions the relevance of works specifying mathematical forms of ICCs in the 1940s (by Lawley, by Mosier, and by Tucker), and in the 1950s, (by Carroll, by Cronbach & Warrington, and by Lazarsfeld).

The emphasis of Green (1950a, b, 1951a, b, 1952) was on analyzing item response data using latent structure (LS) and latent class (LC) models. Green (1951b) stated:

> Latent Structure Analysis is here defined as a mathematical model for describing the interrelationships of items in a psychological test or questionnaire on the basis of which it is possible to make some inferences about hypothetical fundamental variables assumed to underlie the responses. It is also possible to consider the distribution of respondents on these underlying variables. This study was undertaken to attempt to develop a general procedure for applying a specific variant of the latent structure model, the latent class model, to data. (abstract)

He also showed the relationship of the latent structure model to factor analysis (FA)

> The general model of latent structure analysis is presented, as well as several more specific models. The generalization of these models to continuous manifest data is indicated. It is noted that in one case, the generalization resulted in the fundamental equation of linear multiple factor analysis. (abstract)

The work of Green and Lord is significant for many reasons. An important one is that IRT (previously referred to as latent trait, or LT, theory) was shown by Green to be directly related to the models he developed and discussed. Lord (1952a) showed that if a single latent trait is normally distributed, fitting a linear FA model to the tetrachoric correlations of the items yields a unidimensional normal-ogive model for the item response function.

5.2 More Complete Development of IRT (1960s and 1970s)

During the 1960s and 1970s, Lord (1964, 1965a, b, 1968a, b, 1970) expanded on his earlier work to develop IRT more completely, and also demonstrated its use on operational test scores (including early software to estimate the parameters). Also at this time, Birnbaum (1967) presented the theory of logistic models and Ross (1966) studied how actual item response data fit Birnbaum's model. Samejima (1969)[4] published her development of the graded response (GR) model suitable for polytomous data. The theoretical developments of the 1960s culminated in some of

[4] Samejima produced this work while at ETS. She later developed her GR models more fully while holding university positions.

the most important work on IRT during this period, much of it assembled into Lord and Novick's (1968) *Statistical Theories of Mental Test Scores* (which also includes contributions of Birnbaum: Chapters 17, 18, 19, and 20). Also Samejima's continuing work on graded response models, was further developed (1972) while she held academic positions.

An important aspect of the work at ETS in the 1960s was the development of software, particularly by Wingersky, Lord, and Andersen (Andersen 1972; Lord, 1968a; Lord and Wingersky 1973) enabling practical applications of IRT. The LOGIST computer program (Lord et al. 1976; see also Wingersky 1983) was the standard IRT estimation software used for many years in many other institutions besides ETS. Lord (1975b) also published a report in which he evaluated LOGIST estimates using artificial data. Developments during the 1950s were limited by a lack of such software and computers sufficiently powerful to carry out the estimation of parameters. In his 1968 publication, Lord presented a description and demonstration of the use of maximum likelihood (ML) estimation of the ability and item parameters in the three-parameter logistic (3PL) model, using *SAT*® items. He stated, with respect to ICCs:

> The problems of estimating such a curve for each of a large number of items simultaneously is one of the problems that has delayed practical application of Birnbaum's models since they were first developed in 1957. The first step in the present project (see Appendix B) was to devise methods for estimating three descriptive parameters simultaneously for each item in the Verbal test. (1968a, p. 992)

Lord also discussed and demonstrated many other psychometric concepts, many of which were not put into practice until fairly recently due to the lack of computing power and algorithms. In two publications (1965a, b) he emphasized that ICCs are the functions relating probability of response to the underlying latent trait, not to the total test score, and that the former and not the latter can follow a cumulative normal or logistic function (a point he originally made much earlier, Lord 1953). He also discussed (1968a) optimum weighting in scoring and information functions of items from a Verbal SAT test form, as well as test information, and relative efficiency of tests composed of item sets having different psychometric properties. A very interesting fact is that Lord (1968a, p. 1004) introduced and illustrated multistage tests (MTs), and discussed their increased efficiency relative to "the present Verbal SAT" (p. 1005). What we now refer to as *router* tests in using MTs, Lord called *foretests*. He also introduced *tailor-made tests* in this publication (and in Lord 1968c) and discussed how they would be administered using computers. Tailor-made tests are now, of course, commonly known as computerized adaptive tests (CATs); as suggested above, MTs and CATs were not employed in operational testing programs until fairly recently, but it is fascinating to note how long ago Lord introduced these notions and discussed and demonstrated the potential increase in efficiency of assessments achievable with their use. With respect to CATs Lord stated:

> The detailed strategy for selecting a sequence of items that will yield the most information about the ability of a given examinee has not yet been worked out. It should be possible to work out such a strategy on the basis of a mathematical model such as that used here, however. (1968a, p. 1005)

In this work, Lord also presented a very interesting discussion (1968a, p. 1007) on improving validity by using the methods described and illustrated. Finally, in the appendix, Lord derived the ML estimators (MLEs) of the item parameters and, interestingly points out the fact, well known today, that MLEs of the 3PL lower asymptote or *c* parameter, are often “poorly determined by the data” (p. 1014). As a result, he fixed these parameters for the easier items in carrying out his analyses.

During the 1970s Lord produced a phenomenal number of publications, many of them related to IRT, but many on other psychometric topics. On the topics related to IRT alone, he produced six publications besides those mentioned above; these publications dealt with such diverse topics as individualized testing (1974b), estimating power scores from tests that used improperly timed administration (1973), estimating ability and item parameters with missing responses (1974a), the ability scale (1975c), practical applications of item characteristic curves (1977), and equating methods (1975a). In perusing Lord’s work, including Lord and Novick (1968), the reader should keep in mind that he discussed many item response methods and functions using classical test theory (CTT) as well as what we now call IRT. Other work by Lord includes discussions of item characteristic curves and information functions without, for example, using normal ogive or logistic IRT terminology, but the methodology he presented dealt with the theory of item response data. During this period, Erling Andersen visited ETS and during his stay developed one of the seminal papers on testing goodness of fit for the Rasch model (Andersen 1973). Besides the work of Lord, during this period ETS staff produced many publications dealing with IRT, both methodological and application oriented. Marco (1977), for example, described three studies indicating how IRT can be used to solve three relatively intractable testing problems: designing a multipurpose test, evaluating a multistage test, and equating test forms using pretest statistics. He used data from various College Board testing programs and demonstated the use of the information function and relative efficiency using IRT for preequating. Cook (Hambleton and Cook 1977) coauthored an article on using LT models to analyze educational test data. Hambleton and Cook described a number of different IRT models and functions useful in practical applications, demonstrated their use, and cited computer programs that could be used in estimating the parameters. Kreitzberg et al. (1977) discussed potential advantages of CAT, constraints and operational requirements, psychometric and technical developments that make it practical, and its advantages over conventional paper-and-pencil testing. Waller (1976) described a method of estimating Rasch model parameters eliminating the effects of random guessing, without using a computer, and reported a Monte Carlo study on the performance of the method.

5.3 Broadening the Research and Application of IRT (the 1980s)

During this decade, psychometricians, with leadership from Fred Lord, continued to develop the IRT methodology. Also, of course, computer programs for IRT were further developed. During this time many ETS measurement professionals were engaged in assessing the use of IRT models for scaling dichotomous item response data in operational testing programs. In many programs, IRT linking and equating procedures were compared with conventional methods, to inform programs about whether changing these methods should be considered.

5.3.1 Further Developments and Evaluation of IRT Models

In this section we describe further psychometric developments at ETS, as well as research studies evaluating the models, using both actual test and simulated data.

Lord continued to contribute to IRT methodology with works by himself as well as coauthoring works dealing with unbiased estimators of ability parameters and their parallel forms reliability (1983d), a four-parameter logistic model (Barton and Lord 1981), standard errors of IRT equating (1982), IRT parameter estimation with missing data (1983a), sampling variances and covariances of IRT parameter estimates (Lord and Wingersky 1982), IRT equating (Stocking and Lord 1983), statistical bias in ML estimation of IRT item parameters (1983c), estimating the Rasch model when sample sizes are small (1983b), comparison of equating methods (Lord and Wingersky 1984), reducing sampling error (Wingersky and Lord 1984), conjunctive and disjunctive item response functions (1984), ML and Bayesian parameter estimation in IRT (1986), and confidence bands for item response curves with Pashley (Lord and Pashley 1988).

Although Lord was undoubtedly the most prolific ETS contributor to IRT during this period, other ETS staff members made many contributions to IRT. Holland (1981), for example, wrote on the question, "When are IRT models consistent with observed data?" and Cressie and Holland (1983) examined how to characterize the manifest probabilities in LT models. Holland and Rosenbaum (1986) studied monotone unidimensional latent variable models. They discussed applications and generalizations and provided a numerical example. Holland (1990b) also discussed the *Dutch identity* as a useful tool for studying IRT models and conjectured that a quadratic form based on the identity is a limiting form for log manifest probabilities for all smooth IRT models as test length tends to infinity (but see Zhang and Stout 1997, later in this chapter). Jones discussed the adequacy of LT models (1980) and robustness tools for IRT (1982).

Wainer and several colleagues published articles dealing with standard errors in IRT (Wainer and Thissen 1982), review of estimation in the Rasch model for "longish tests" (Gustafsson et al. 1980), fitting ICCs with spline functions (Winsberg et al. 1984), estimating ability with wrong models and inaccurate parameters (Jones et al. 1984), evaluating simulation results of IRT ability estimation (Thissen and Wainer 1984; Thissen et al. 1984), and confidence envelopes for IRT (Thissen and Wainer 1990). Wainer (1983) also published an article discussing IRT and CAT, which he described as a coming technological revolution. Thissen and Wainer (1985) followed up on Lord's earlier work, discussing the estimation of the *c* parameter in IRT. Wainer and Thissen (1987) used the 1PL, 2PL, and 3PL models to fit simulated data and study accuracy and efficiency of robust estimators of ability. For short tests, simple models and robust estimators best fit the data, and for longer tests more complex models fit well, but using robust estimation with Bayesian priors resulted in substantial shrinkage. Testlet theory was the subject of Wainer and Lewis (1990).

Mislevy has also made numerous contributions to IRT, introducing Bayes modal estimation (1986b) in 1PL, 2PL, and 3PL IRT models, providing details of an expectation-maximization (EM) algorithm using two-stage modal priors, and in a simulation study, demonstrated improvement in estimation. Additionally he wrote on Bayesian treatment of latent variables in sample surveys (Mislevy 1986a). Most significantly, Mislevy (1984) developed the first version of a model that would later become the standard analytic approach for the National Assessment of Educational Progress (NAEP) and virtually all other large scale international survey assessments (see also Beaton and Barone's Chap. 8 and Chap. 9 by Kirsch et al. in this volume on the history of adult literacy assessments at ETS). Mislevy (1987a) also introduced application of empirical Bayes procedures, using auxililary information about test takers, to increase the precision of item parameter estimates. He illustrated the procedures with data from the Profile of American Youth survey. He also wrote (1988) on using auxilliary information about items to estimate Rasch model item difficulty parameters and authored and coauthored other papers, several with Sheehan, dealing with use of auxiliary/collateral information with Bayesian procedures for estimation in IRT models (Mislevy 1988; Mislevy and Sheehan 1989b; Sheehan and Mislevy 1988). Another contribution Mislevy made (1986c) is a comprehensive discussion of FA models for test item data with reference to relationships to IRT models and work on extending currently available models. Mislevy and Sheehan (1989a) discussed consequences of uncertainty in IRT linking and the information matrix in latent variable models. Mislevy and Wu (1988) studied the effects of missing responses and discussed the implications for ability and item parameter estimation relating to alternate test forms, targeted testing, adaptive testing, time limits, and omitted responses. Mislevy also coauthored a book chapter describing a hierarchical IRT model (Mislevy and Bock 1989).

Many other ETS staff members made important contributions. Jones (1984a, b) used asymptotic theory to compute approximations to standard errors of Bayesian and robust estimators studied by Wainer and Thissen. Rosenbaum wrote on testing the local independence assumption (1984) and showed (1985) that the observable distributions of item responses must satisfy certain constraints when two groups of test takers have generally different ability to respond correctly under a unidimensional

IRT model. Dorans (1985) contributed a book chapter on item parameter invariance. Douglass et al. (1985) studied the use of approximations to the 3PL model in item parameter estimation and equating. Methodology for comparing distributions of item responses for two groups was contributed by Rosenbaum (1985). McKinley and Mills (1985) compared goodness of fit statistics in IRT models, and Kingston and Dorans (1985) explored item-ability regressions as a tool for model fit.

Tatsuoka (1986) used IRT in developing a probabilistic model for diagnosing and classifying cognitive errors. While she held a postdoctoral fellowship at ETS, Lynne Steinberg coathored (Thissen and Steinberg 1986) a widely used and cited taxonomy of IRT models, which mentions, among other contributions, that the expressions they use suggest additional, as yet undeveloped, models. One explicitly suggested is basically the two-parameter partial credit (2PPC) model developed by Yen (see Yen and Fitzpatrick 2006) and the equivalent generalized partial credit (GPC) model developed by Muraki (1992a), both some years after the Thissen-Steinberg article. Rosenbaum (1987) developed and applied three nonparametric methods for comparisons of the shapes of two item characteristic surfaces. Stocking (1989) developed two methods of online calibration for CAT tests and compared them in a simulation using item parameters from an operational assessment. She also (1990) conducted a study on calibration using different ability distributions, concluding that the best estimation for applications that are highly dependent on item parameters, such as CAT and test construction, resulted when the calibration sample contained widely dispersed abilities. McKinley (1988) studied six methods of combining item parameter estimates from different samples using real and simulated item response data. He stated, “results support the use of covariance matrix-weighted averaging and a procedure that involves sample-size-weighted averaging of estimated item characteristic curves at the center of the ability distribution.” (abstract). McKinley also (1989a) developed and evaluated with simulated data a confirmatory multidimensional IRT (MIRT) model. Yamamoto (1989) developed HYBRID, a model combining IRT and LC analysis, and used it to “present a structure of cognition by a particular response vector or set of them” (abstract). The software developed by Yamamoto was also used in a paper by Mislevy and Verhelst (1990) that presented an approach to identifying latent groups of test takers. Folk (Folk and Green 1989) coauthored a work on adaptive estimation when the unidimensionality assumption of IRT is violated.

5.3.2 IRT Software Development and Evaluation

With respect to IRT software, Mislevy and Stocking (1987) provided a guide to use of the LOGIST and BILOG computer programs that was very helpful to new users of IRT in applied settings. Mislevy, of course, was one of the developers of BILOG (Mislevy and Bock 1983). Wingersky (1987), the primary developer of LOGIST, developed and evaluated, with real and artificial data, a one-stage version of LOGIST for use when estimates of item parameters but not test-taker abilities are required.

Item parameter estimates were not as good as those from LOGIST, and the one-stage software did not reduce computer costs when there were missing data in the real dataset. Stocking (1989) conducted a study of estimation errors and relationship to properties of the test or item set being calibrated; she recommended improvements to the methods used in the LOGIST and BILOG programs. Yamamoto (1989) produced the HYBIL software for the HYBRID model and mixture IRT we referred to above. Both HYBIL and BILOG utilize marginal ML estimation, whereas LOGIST uses joint ML estimation methods.

5.3.3 Explanation, Evaluation, and Application of IRT Models

During this decade ETS scientists began exploring the use of IRT models with operational test data and producing works explaining IRT models for potential users. Applications of IRT were seen in many ETS testing programs.

Lord's book, *Applications of Item Response Theory to Practical Testing Problems* (1980a), presented much of the current IRT theory in language easily understood by many practitioners. It covered basic concepts, comparison to CTT methods, relative efficiency, optimal number of choices per item, flexilevel tests, multistage tests, tailored testing, mastery testing, estimating ability and item parameters, equating, item bias, omitted responses, and estimating true score distributions. Lord (1980b) also contributed a book chapter on practical issues in tailored testing.

Bejar illustrated use of item characteristic curves in studying dimensionality (1980), and he and Wingersky (1981, 1982) applied IRT to the Test of Standard Written English, concluding that using the 3PL model and IRT preequating "did not appear to present problems" (abstract). Kingston and Dorans (1982) applied IRT to the *GRE*® Aptitude Test, stating that "the most notable finding in the analytical equatings was the sensitivity of the precalibration design to practice effects on analytical items … this might present a problem for any equating design" (abstract). Kingston and Dorans (1982a) used IRT in the analysis of the effect of item position on test taker responding behavior. They also (1982b) compared IRT and conventional methods for equating the GRE Aptitude Test, assessing the reasonableness of the assumptions of item response theory for GRE item types and test taker populations, and finding that the IRT precalibration design was sensitive to practice effects on analytical items. In addition, Kingston and Dorans (1984) studied the effect of item location on IRT equating and adaptive testing, and Dorans and Kingston (1985) studied effects of violation of the unidimensionality assumption on estimation of ability and item parameters and on IRT equating with the GRE Verbal Test, concluding that there were two highly correlated verbal dimensions that had an effect on equating, but that the effect was slight. Kingston et al. (1985) compared IRT to conventional equating of the Graduate Management Admission Test (GMAT) and concluded that violation of local independence of this test had little effect on the equating results (they cautioned that further study was necessary before using other IRT-based procedures with the test). McKinley and Kingston (1987) investigated using IRT equating for the GRE Subject Test in Mathematics and also studied the unidimensionality and model fit assumptions, concluding that the test was reasonably unidimensional and the 3PL model provided reasonable fit to the data.

Cook, Eignor, Petersen and colleagues wrote several explanatory papers and conducted a number of studies of application of IRT on operational program data, studying assumptions of the models, and various aspects of estimation and equating (Cook et al. 1985a, c, 1988a, b; Cook and Eignor 1985, 1989; Eignor 1985; Stocking 1988). Cook et al. (1985b, 1988c) examined effects of curriculum (comparing results for students tested before completing the curriculum with students tested after completing it) on stability of CTT and IRT difficulty parameter estimates, effects on equating, and the dimensionality of the tests. Cook and colleagues (Wingersky et al. 1987), using simulated data based on actual SAT item parameter estimates, studied the effect of anchor item characteristics on IRT true-score equating.

Kreitzberg and Jones (1980) presented results of a study of CAT using the Broad-Range Tailored Test and concluded,“computerized adaptive testing is ready to take the first steps out of the laboratory environment and find its place in the educational community” (abstract). Scheuneman (1980) produced a book chapter on LT theory and item bias. Hicks (1983) compared IRT equating with fixed versus estimated parameters and three “conventional” equating methods using *TOEFL*® test data, concluding that fixing the *b* parameters to pretest values (essentially this is what we now call preequating) is a “very acceptable option.” She followed up (1984) with another study in which she examined controlling for native language and found this adjustment resulted in increased stability for one test section but a decrease in another section. Peterson, Cook, and Stocking (1983) studied several equating methods using SAT data and found that for reasonably parallel tests, linear equating methods perform adequately, but when tests differ somewhat in content and length, methods based on the three-parameter logistic IRT model lead to greater stability of equating results. In a review of research on IRT and conventional equating procedures, Cook and Petersen (1987) discussed how equating methods are affected by sampling error, sample characteristics, and anchor item characteristics, providing much useful information for IRT users.

Cook coauthored a book chapter (Hambleton and Cook 1983) on robustness of IRT models, including effects of test length and sample size on precision of ability estimates. Several ETS staff members contributed chapters to that same edited book on applications of item response theory (Hambleton 1983). Bejar (1983) contributed an introduction to IRT and its assumptions; Wingersky (1983) a chapter on the LOGIST computer program; Cook and Eignor (1983) on practical considerations for using IRT in equating. Tatsuoka coauthored on appropriateness indices (Harnisch and Tatsuoka 1983); and Yen wrote on developing a standardized test with the 3PL model (1983); both Tatsuoka and Yen later joined ETS.

Lord and Wild (1985) compared the contribution of the four verbal item types to measurement accuracy of the GRE General Test, finding that the reading comprehension item type measures something slightly different from what is measured by sentence completion, analogy, or antonym item types. Dorans (1986) used IRT to study the effects of item deletion on equating functions and the score distribution on the SAT, concluding that reequating should be done when an item is dropped. Kingston and Holland (1986) compared equating errors using IRT and several other equating methods, and several equating designs, for equating the GRE General Test, with varying results depending on the specific design and method. Eignor and Stocking (Eignor and Stocking 1986) conducted two studies to investigate whether calibration or linking methods might be reasons for poor equating results on the SAT. In the first study they used actual data, and in the second they used simulations, concluding that a combination of differences in true mean ability and multidimensionality were consistent with the real data. Eignor et al. (1986) studied the potential of a new plotting procedures for assessing fit to the 3PL model using SAT and TOEFL data. Wingersky and Sheehan (1986) also wrote on fit to IRT models, using regressions of item scores onto observed (number correct) scores rather than the previously used method of regressing onto estimated ability.

Bejar (1990), using IRT, studied an approach to psychometric modeling that explicitly incorporates information on the mental models test takers use in solving an item, and concluded that it is not only workable, but also necessary for future developments in psychometrics. Kingston (1986) used full information FA to estimate difficulty and discrimination parameters of a MIRT model for the GMAT, finding there to be dominant first dimensions for both the quantitative and verbal measures. Mislevy (1987b) discussed implications of IRT developments for teacher certification. Mislevy (1989) presented a case for a new test theory combining modern cognitive psychology with modern IRT. Sheehan and Mislevy (1990) wrote on the integration of cognitive theory and IRT and illustrated their ideas using the Survey of Young Adult Literacy data. These ideas seem to be the first appearance of a line of research that continues today. The complexity of these models, built to integrate cognitive theory and IRT, evolved dramatically in the twenty-first century due to rapid increase in computational capabilities of modern computers and developments in understanding problem solving. Lawrence coauthored a paper (Lawrence and Dorans 1988) addressing the sample invariance properties of four equating methods with two types of test-taker samples (matched on anchor test score distributions or taken from different administrations and differing in ability). Results for IRT, Levine, and equipercentile methods differed for the two types of samples, whereas the Tucker observed score method did not. Henning (1989) discussed the appropriateness of the Rasch model for multiple-choice data, in response to an article that questioned such appropriateness. McKinley (1989b) wrote an explanatory article for potential users of IRT. McKinley and Schaeffer (1989) studied an IRT equating method for the GRE designed to reduce the overlap on test forms. Bejar et al. (1989), in a paper on methods used for patient management items in medical licensure testing, outlined recent developments and introduced a procedure that integrates those developments with IRT. Boldt (1989) used LC analysis to study the dimensionality of the TOEFL and assess whether different dimensions were necessary to fit models to diverse groups of test takers. His findings were that a single dimension LT model fits TOEFL data well but “suggests the use of a restrictive assumption of proportionality of item response curves” (p. 123).

In 1983, ETS assumed the primary contract for NAEP, and ETS psychometricians were involved in designing analysis procedures, including the use of an IRT-based latent regression model using ML estimation of population parameters from observed item responses without estimating ability parameters for test takers (e.g., Mislevy 1984, 1991). Asymptotic standard errors and tests of fit, as well as approximate solutions of the integrals involved, were developed in Mislevy's 1984 article. With leadership from Messick (Messick 1985; Messick et al. 1983), a large team of ETS staff developed a complex assessment design involving new analysis procedures for direct estimation of average achievement of groups of students. Zwick (1987) studied whether the NAEP reading data met the unidimensionality assumption underlying the IRT scaling procedures. Mislevy (1991) wrote on making inferences about latent variables from complex samples, using IRT proficiency estimates as an example and illustrating with NAEP reading data. The innovations introduced include the linking of multiple test forms using IRT, a task that would be virtually impossible without IRT-based methods, as well as the intregration of IRT with a regression-based population model that allows the prediction of an ability prior, given background data collected in student questionnaires along with the cogntive NAEP tests.

5.4 Advanced Item Response Modeling: The 1990s

During the 1990s, the use of IRT in operational testing programs expanded considerably. IRT methodology for dichotomous item response data was well developed and widely used by the end of the 1980s. In the early years of the 1990s, models for polytomous item response data were developed and began to be used in operational programs. Muraki (1990) developed and illustrated an IRT model for fitting a polytomous item response theory model to Likert-type data. Muraki (1992a) also developed the GPC model, which has since become one of the most widely used models for polytomous IRT data. Concomitantly, before joining ETS, Yen[5] developed the 2PPC model that is identical to the GPC, differing only in the parameterization incorporated into the model. Muraki (1993) also produced an article detailing the IRT information functions for the GPC model. Chang and Mazzeo (1994) discussed item category response functions (ICRFs) and the item response functions (IRFs), which are weighted sums of the ICRFs, of the partial credit and graded response models. They showed that if two polytomously scored items have the same IRF, they must have the same number of categories that have the same ICRFs. They also discussed theoretical and practical implications. Akkermans and Muraki (1997) studied and described characteristics of the item information and discrimination functions for partial credit items.

[5] Developed in 1991 (as cited in Yen and Fitzpatrick 2006), about the same time as Muraki was developing the GPC model.

In work reminiscent of the earlier work of Green and Lord, Gitomer and Yamamoto (1991) described HYBRID (Yamamoto 1989), a model that incorporates both LT and LC components; these authors, however, defined the latent classes by a cognitive analysis of the understanding that individuals have for a domain. Yamamoto and Everson (1997) also published a book chapter on this topic. Bennett et al. (1991) studied new cognitively sensitive measurement models, analyzing them with the HYBRID model and comparing results to other IRT methodology, using partial-credit data from the GRE General Test. Works by Tatsuoka (1990, 1991) also contributed to the literature relating IRT to cognitive models. The integration of IRT and a person-fit measure as a basis for rule space, as proposed by Tatsuoka, allowed in-depth examinations of items that require multiple skills. Sheehan (1997) developed a tree-based method of proficiency scaling and diagnostic assessment and applied it to developing diagnostic feedback for the SAT I Verbal Reasoning Test. Mislevy and Wilson (1996) presented a version of Wilson's Saltus model, an IRT model that incorporates developmental stages that may involve discontinuities. They also demonstrated its use with simulated data and an example of mixed number subtraction.

The volume *Test Theory for a New Generation of Tests* (Frederiksen et al. 1993) presented several IRT-based models that anticipated a more fully integrated approach providing information about measurement qualities of items as well as about complex latent variables that align with cognitive theory. Examples of these advances are the chapters by Yamamoto and Gitomer (1993) and Mislevy (1993a).

Bradlow (1996) discussed the fact that, for certain values of item parameters and ability, the information about ability for the 3PL model will be negative and has consequences for estimation—a phenomenon that does not occur with the 2PL. Pashley (1991) proposed an alternative to Birnbaum's 3PL model in which the asymptote parameter is a linear component within the logit of the function. Zhang and Stout (1997) showed that Holland's (1990b) conjecture that a quadratic form for log manifest probabilities is a limiting form for all smooth unidimensional IRT models does not always hold; these authors provided counterexamples and suggested that only under strong assumptions can this conjecture be true.

Holland (1990a) published an article on the sampling theory foundations of IRT models. Stocking (1990) discussed determining optimum sampling of test takers for IRT parameter estimation. Chang and Stout (1993) showed that, for dichotomous IRT models, under very general and nonrestrictive nonparametric assumptions, the posterior distribution of test taker ability given dichotomous responses is approximately normal for a long test. Chang (1996) followed up with an article extending this work to polytomous responses, defining a global information function, and he showed the relationship of the latter to other information functions.

Mislevy (1991) published on randomization-based inference about latent variables from complex samples. Mislevy (1993b) also presented formulas for use with Bayesian ability estimates. While at ETS as a postdoctoral fellow, Roberts coauthored works on the use of unfolding[6] (Roberts and Laughlin 1996). A parametric IRT model for unfolding dichotomously or polytomously scored responses, called the graded unfolding model (GUM), was developed; a subsequent recovery simulation showed that reasonably accurate estimates could be obtained. The applicability of the GUM to common attitude testing situations was illustrated with real data on student attitudes toward capital punishment. Roberts et al. (2000) described the generalized GUM (GGUM), which introduced a parameter to the model, allowing for variation in discrimination across items; they demonstrated the use of the model with real data.

Wainer and colleagues wrote further on testlet response theory, contributing to issues of reliability of testlet-based tests (Sireci et al. 1991). These authors also developed, and illustrated using operational data, statistical methodology for detecting differential item functioning (DIF) in testlets (Wainer et al. 1991). Thissen and Wainer (1990) also detailed and illustrated how *confidence envelopes* could be formed for IRT models. Bradlow et al. (1999) developed a Bayesian IRT model for testlets and compared results with those from standard IRT models using a released SAT dataset. They showed that degree of precision bias was a function of testlet effects and the testlet design. Sheehan and Lewis (1992) introduced, and demonstrated with actual program data, a procedure for determining the effect of testlet nonequivalence on the operating characteristics of a computerized mastery test based on testlets.

Lewis and Sheehan (1990) wrote on using Bayesian decision theory to design computerized mastery tests. Contributions to CAT were made in a book, *Computer Adaptive Testing: A Primer*, edited by Wainer et al. (1990a) with chapters by ETS psychometricians: "Introduction and History" (Wainer 1990), "Item Response Theory, Item Calibration and Proficiency Estimation" (Wainer and Mislevy 1990); "Scaling and Equating" (Dorans 1990); "Testing Algorithms" (Thissen and Mislevy 1990); "Validity" (Steinberg et al. 1990); "Item Pools" (Flaugher 1990); and "Future Challenges" (Wainer et al. 1990b). Automated item selection (AIS) using IRT was the topic of two publications (Stocking et al. 1991a, b). Mislevy and Chang (2000) introduced a term to the expression for probability of response vectors to deal with item selection in CAT, and to correct apparent incorrect response pattern probabilities in the context of adaptive testing. Almond and Mislevy (1999) studied graphical modeling methods for making inferences about multifaceted skills and models in an IRT CAT environment, and illustrated in the context of language testing.

In an issue of an early volume of *Applied Measurement in Education*, Eignor et al. (1990) expanded on their previous studies (Cook et al. 1988b) comparing IRT

[6] Unfolding models are proximity IRT models developed for assessments with binary disagree-agree or graded disagree-agree responses. Responses on these assessments are not necessarily cumulative and one cannot assume that higher levels of the latent trait will lead to higher item scores and thus to higher total test scores. Unfolding models predict item scores and total scores on the basis of the distances between the test taker and each item on the latent continuum (Roberts n.d.).

equating with several non-IRT methods and with different sampling designs. In another article in that same issue, Schmitt et al. (1990) reported on the sensitivity of equating results to sampling designs; Lawrence and Dorans (1990) contributed with a study of the effect of matching samples in equating with an anchor test; and Livingston et al. (1990) also contributed on sampling and equating methodolgy to this issue.

Zwick (1990) published an article showing when IRT and Mantel-Haenszel definitions of DIF coincide. Also in the DIF area, Dorans and Holland (1992) produced a widely disseminated and used work on the Mantel-Haenszel (MH) and standardization methodologies, in which they also detailed the relationship of the MH to IRT models. Their methodology, of course, is the mainstay of DIF analyses today, at ETS and at other institutions. Muraki (1999) described a stepwise DIF procedure based on the multiple group PC model. He illustrated the use of the model using NAEP writing trend data and also discussed item parameter drift. Pashley (1992) presented a graphical procedure, based on IRT, to display the location and magnitude of DIF along the ability continuum.

MIRT models, although developed earlier, were further developed and illustrated with operational data during this decade; McKinley coauthored an article (Reckase and McKinley 1991) describing the discrimination parameter for these models. Muraki and Carlson (1995) developed a multidimensional graded response (MGR) IRT model for polytomously scored items, based on Samejima‘s normal ogive GR model. Relationships to the Reckase-McKinley and FA models were discussed, and an example using NAEP reading data was presented and discussed. Zhang and Stout (1999a, b) described models for detecting dimensionality and related them to FA and MIRT.

Lewis coauthored publications (McLeod and Lewis 1999; McLeod et al. 2003) with a discussion of person-fit measures as potential ways of detecting memorization of items in a CAT environment using IRT, and introduced a new method. None of the three methods showed much power to detect memorization. Possible methods of altering a test when the model becomes inappropriate for a test taker were discussed.

5.4.1 IRT Software Development and Evaluation

During this period, Muraki developed the PARSCALE computer program (Muraki and Bock 1993) that has become one of the most widely used IRT programs for polytomous item response data. At ETS it has been incorporated into the GENASYS software used in many operational programs to this day. Muraki (1992b) also developed the RESGEN software, also widely used, for generating simulated polytomous and dichotomous item response data.

Many of the research projects in the literature reviewed here involved development of software for estimation of newly developed or extended models. Some examples involve Yamamoto’s (1989) HYBRID model, the MGR model (Muraki and Carlson 1995) for which Muraki created the POLYFACT software, and the Saltus model (Mislevy and Wilson 1996) for which an EM algorithm-based program was created.

5.4.2 Explanation, Evaluation, and Application of IRT Models

In this decade ETS researchers continued to provide explanations of IRT models for users, to conduct research evaluating the models, and to use them in testing programs in which they had not been previously used. The latter activity is not emphasized in this section as it was for sections on previous decades because of the sheer volume of such work and the fact that it generally involves simply applying IRT to testing programs, whereas in previous decades the research made more of a contribution, with recommendations for practice in general. Although such work in the 1990s contributed to improving the methodology used in specific programs, it provided little information that can be generalized to other programs. This section, therefore covers research that is more generalizable, although illustrations may have used specific program data.

Some of this research provided new information about IRT scaling. Donoghue (1992), for example, described the common misconception that the partial credit and GPC IRT model item category functions are symmetric, helping explain characteristics of items in these models for users of them. He also (1993) studied the information provided by polytomously scored NAEP reading items and made comparisons to information provided by dichotomously scored items, demonstrating how other users can use such information for their own programs. Donoghue and Isham (1998) used simulated data to compare IRT and other methods of detecting item parameter drift. Zwick (1991), illustrating with NAEP reading data, presented a discussion of issues relating to two questions: "What can be learned about the effects of item order and context on invariance of item parameter estimates?" and "Are common-item equating methods appropriate when measuring trends in educational growth?" Camili et al. (1993) studied scale shrinkage in vertical equating, comparing IRT with equipercentile methods using real data from NAEP and another testing program. Using IRT methods, variance decreased from fall to spring testings, and also from lower- to upper-grade levels, whereas variances have been observed to increase across grade levels for equipercentile equating. They discussed possible reasons for scale shrinkage and proposed a more comprehensive, model-based approach to establishing vertical scales. Yamamoto and Everson (1997) estimated IRT parameters using TOEFL data and Yamamoto's extended HYBRID model (1989), which uses a combination of IRT and LC models to characterize when test takers switch from ability-based to random responses. Yamamoto studied effects of time limits on speededness, finding that this model estimated the parameters more accurately than the usual IRT model. Yamamoto and Everson (1995) using three different sets of actual test data, found that the HYBRID model successfully determined the switch point in the three datasets. Liu coauthored (Lane et al.

1995) an article in which mathematics performance-item data were used to study the assumptions of and stability over time of item parameter estimates using the GR model. Sheehan and Mislevy (1994) used a tree-based analysis to examine the relationship of three types of item attributes (constructed-response [CR] vs. multiple choice [MC], surface features, aspects of the solution process) to operating characteristics (using 3PL parameter estimates) of computer-based *PRAXIS*® mathematics items. Mislevy and Wu (1996) built on their previous research (1988) on estimation of ability when there are missing data due to assessment design (alternate forms, adaptive testing, targeted testing), focusing on using Bayesian and direct likelihood methods to estimate ability parameters.

Wainer et al. (1994) examined, in an IRT framework, the comparability of scores on tests in which test takers choose which CR prompts to respond to, and illustrated using the College Board *Advanced Placement*® Test in Chemistry.

Zwick et al. (1995) studied the effect on DIF statistics of fitting a Rasch model to data generated with a 3PL model. The results, attributed to degredation of matching resulting from Rasch model ability estimation, indicated less sensitive DIF detection.

In 1992, special issues of the *Journal of Educational Measurement* and the *Journal of Educational Statistics* were devoted to methodology used by ETS in NAEP, including the NAEP IRT methodology. Beaton and Johnson (1992), and Mislevy et al. (1992b) detailed how IRT is used and combined with the plausible values methodology to estimate proficiencies for NAEP reports. Mislevy et al. (1992a) wrote on how population characteristics are estimated from sparse matrix samples of item responses. Yamamoto and Mazzeo (1992) described IRT scale linking in NAEP.

5.5 IRT Contributions in the Twenty-First Century

5.5.1 Advances in the Development of Explanatory and Multidimensional IRT Models

Multidimensional models and dimensionality considerations continued to be a subject of research at ETS, with many more contributions than in the previous decades. Zhang (2004) proved that, when simple structure obtains, estimation of unidimensional or MIRT models by joint ML yields identical results, but not when marginal ML is used. He also conducted simulations and found that, with small numbers of items, MIRT yielded more accurate item parameter estimates but the unidimensional approach prevailed with larger numbers of items, and that when simple structure does not hold, the correlations among dimensions are overestimated.

A genetic algorithm was used by Zhang (2005b) in the maximization step of an EM algorithm to estimate parameters of a MIRT model with complex, rather than simple, structure. Simulated data suggested that this algorithm is a promising approach to estimation for this model. Zhang (2007) also extended the theory of conditional covariances to the case of polytomous items, providing a theoretical foundation for study of dimensionality. Several estimators of conditional covariance were constructed, including the case of complex incomplete designs such as those used in NAEP. He demonstrated use of the methodology with NAEP reading assessment data, showing that the dimensional structure is consistent with the purposes of reading that define NAEP scales, but that the degree of multidimensionality is weak in those data.

Haberman et al. (2008) showed that MIRT models can be based on ability distributions that are multivariate normal or multivariate polytomous, and showed, using empirical data, that under simple structure the two cases yield comparable results in terms of model fit, parameter estimates, and computing time. They also discussed numerical methods for use with the two cases.

Rijmen wrote two papers dealing with methodology relating to MIRT models, further showing the relationship between IRT and FA models. As discussed in the first section of this chapter, such relationships were shown for more simple models by Bert Green and Fred Lord in the 1950s. In the first (2009) paper, Rijmen showed how an approach to full information ML estimation can be placed into a graphical model framework, allowing for derivation of efficient estimation schemes in a fully automatic fashion. This avoids tedious derivations, and he demonstrated the approach with the bifactor and a MIRT model with a second-order dimension. In the second paper, (2010) Rijmen studied three MIRT models for testlet-based tests, showing that the second-order MIRT model is formally equivalent to the testlet model, which is a bifactor model with factor loadings on the specific dimensions restricted to being proportional to the loadings on the general factor.

M. von Davier and Carstensen (2007) edited a book dealing with multivariate and mixture distribution Rasch models, including extensions and applications of the models. Contributors to this book included: Haberman (2007b) on the interaction model; M. von Davier and Yamamoto (2007) on mixture distributions and hybrid Rasch models; Mislevy and Huang (2007) on measurement models as narrative structures; and Boughton and Yamamoto (2007) on a hybrid model for test speededness.

Antal (2007) presented a coordinate-free approach to MIRT models, emphasizing understanding these models as extensions of the univariate models. Based on earlier work by Rijmen et al. (2003), Rijmen et al. (2013) described how MIRT models can be embedded and understood as special cases of generalized linear and nonlinear mixed models.

Haberman and Sinharay (2010) studied the use of MIRT models in computing subscores, proposing a new statistical approach to examining when MIRT model subscores have added value over total number correct scores and subscores based on CTT. The MIRT-based methods were applied to several operational datasets, and results showed that these methods produce slightly more accurate scores than CTT-based methods.

Rose et al. (2010) studied IRT modeling of nonignorable missing item responses in the context of large-scale international assessments, comparing using CTT and simple IRT models, the usual two treatments (missing item responses as wrong, or as not administered), with two MIRT models. One model used indicator variables as a dimension to designate where missing responses occurred, and the other was a multigroup MIRT model with grouping based on a within-country stratification by the amount of missing data. Using both simulated and operational data, they demonstrated that a simple IRT model ignoring missing data performed relatively well when the amount of missing data was moderate, and the MIRT-based models only outperformed the simple models with larger amounts of missingness, but they yielded estimates of the correlation of missingness with ability estimates and improved the reliability of the latter.

van Rijn and Rijmen (2015) provided an explanation of a "paradox" that in some MIRT models answering an additional item correctly can result in a decrease in the test taker's score on one of the latent variables, previously discussed in the psychometric literature. These authors showed clearly how it occurs and also pointed out that it does not occur in testlet (restricted bifactor) models.

ETS researchers also continued to develop CAT methodology. Yan et al. (2004b) introduced a nonparametric tree-based algorithm for adaptive testing and showed that it may be superior to conventional IRT methods when the IRT assumptions are not met, particularly in the presence of multidimensionality. While at ETS, Weissman coauthored an article (Belov et al. 2008) in which a new CAT algorithm was developed and tested in a simulation using operational test data. Belov et al. showed that their algorithm, compared to another algorithm incorporating content constraints had lower maximum item exposure rates, higher utilization of the item pool, and more robust ability estimates when high (low) ability test takers performed poorly (well) at the beginning of testing.

The second edition of *Computerized Adaptive Testing: A Primer* (Wainer et al. 2000b) was published and, as in the first edition (Wainer et al. 1990a), many chapters were authored or coauthored by ETS researchers (Dorans 2000; Flaugher 2000; Steinberg et al. 2000; Thissen and Mislevy 2000; Wainer 2000; Wainer et al. 2000c; Wainer and Eignor 2000; Wainer and Mislevy 2000). Xu and Douglas (2006) explored the use of nonparametric IRT models in CAT; derivatives of ICCs required by the Fisher information criterion might not exist for these models, so alternatives based on Shannon entropy and Kullback-Leibler information (which do not require derivatives) were proposed. For long tests these methods are equivalent to the maximum Fisher information criterion, and simulations showed them to perform similarly, and much better than random selection of items.

Diagnostic models for assessment including cognitive diagnostic (CD) assessment, as well as providing diagnostic information from common IRT models, continued to be an area of research by ETS staff. Yan et al. (2004a), using a mixed number subtraction dataset, and cognitive research originally developed by Tatsuoka and her colleagues, compared several models for providing diagnostic information on score reports, including IRT and other types of models, and characterized the kinds of problems for which each is suited. They provided a general Bayesian psychometric framework to provide a common language, making it easier to appreciate the differences. M. von Davier (2008a) presented a class of general diagnostic (GD) models that can be estimated by marginal ML algorithms; that allow for both dichotomous and polytomous items, compensatory and noncompensatory models; and subsume many common models including unidimensional and multidimensional Rasch models, 2PL, PC and GPC, facets, and a variety of skill profile models. He demonstrated the model using simulated as well as TOEFL iBT data.

Xu (2007) studied monotonicity properties of the GD model and found that, like the GPC model, monotonicity obtains when slope parameters are restricted to be equal, but does not when this restriction is relaxed, although model fit is improved. She pointed out that trade offs between these two variants of the model should be considerred in practice. M. von Davier (2007) extended the GD model to a hierarchical model and further extended it to the mixture general diagnostic (MGD) model (2008b), which allows for estimation of diagnostic models in multiple known populations as well as discrete unknown, or not directly observed mixtures of populations.

Xu and von Davier (2006) used a MIRT model specified in the GD model framework with NAEP data and verified that the model could satisfactorily recover parameters from a sparse data matrix and could estimate group characteristics for large survey data. Results under both single and multiple group assumptions and comparison with the NAEP model results were also presented. The authors suggested that it is possible to conduct cognitive diagnosis for NAEP proficiency data. Xu and von Davier (2008b) extended the GD model, employing a log-linear model to reduce the number of parameters to be estimated in the latent skill distribution. They extended that model (2008a) to allow comparison of constrained versus nonconstrained parameters across multiple populations, illustrating with NAEP data.

M. von Davier et al. (2008) discussed models for diagnosis that combine features of MIRT, FA, and LC models. Hartz and Roussos (2008)[7] wrote on the fusion model for skills diagnosis, indicating that the development of the model produced advancements in modeling, parameter estimation, model fitting methods, and model fit evaluation procedures. Simulation studies demonstrated the accuracy of the estimation procedure, and effectiveness of model fitting and model fit evaluation procedures. They concluded that the model is a promising tool for skills diagnosis that merits further research and development.

Linking and equating also continue to be important topics of ETS research. In this section the focus is research on IRT-based linking/equating methods. M. von Davier and von Davier (2007, 2011) presented a unified approach to IRT scale linking and transformation. Any linking procedure is viewed as a restriction on the item parameter space, and then rewriting the log-likelihood function together with implementation of a maximization procedure under linear or nonlinear restrictions accomplishes the linking. Xu and von Davier (2008c) developed an IRT linking approach for use with the GD model and applied the proposed approach to NAEP data. Holland and Hoskens (2002) developed an approach viewing CTT as a first-order version of IRT and the latter as detailed elaborations of CTT, deriving general results for the prediction of true scores from observed scores, leading to a new view

[7] While these authors were not ETS staff members, this report was completed under the auspices of the External Diagnostic Research Team, supported by ETS.

of linking tests not designed to be linked. They illustrated the theory using simulated and actual test data. M. von Davier et al. (2011) presented a model that generalizes approaches by Andersen (1985), and Embretson (1991), respectively, to utilize MIRT in a multiple-population longitudinal context to study individual and group-level learning trajectories.

Research on testlets continued to be a focus at ETS, as well as research involving item families. Wang et al. (2002) extended the development of testlet models to tests comprising polytomously scored and/or dichotomously scored items, using a fully Bayesian method. They analyzed data from the Test of Spoken English (TSE) and the North Carolina Test of Computer Skills, concluding that the latter exhibited significant testlet effects, whereas the former did not. Sinharay et al. (2003) used a Bayesian hierarchical model to study item families, showing that the model can take into account the dependence structure built into the families, allowing for calibration of the family rather than the individual items. They introduced the family expected response function (FERF) to summarize the probability of a correct response to an item randomly generated from the family, and suggested a way to estimate the FERF.

Wainer and Wang (2000) conducted a study in which TOEFL data were fitted to an IRT testlet model, and for comparative purposes to a 3PL model. They found that difficulty parameters were estimated well with either model, but discrimination and lower asymptote parameters were biased when conditional independence was incorrectly assumed. Wainer also coauthored book chapters explaining methodology for testlet models (Glas et al. 2000; Wainer et al. 2000a).

Y. Li et al. (2010) used both simulated data and operational program data to compare the parameter estimation, model fit, and estimated information of testlets comprising both dichotomous and polytomous items. The models compared were a standard 2PL/GPC model (ignoring local item dependence within testlets) and a general dichotomous/polytomous testlet model. Results of both the simulation and real data analyses showed little difference in parameter estimation but more difference in fit and information. For the operational data, they also made comparisons to a MIRT model under a simple structure constraint, and this model fit the data better than the other two models.

Roberts et al. (2002) in a continuation of their research on the GGUM, studied the characteristics of marginal ML and expected a posteriori (EAP) estimates of item and test-taker parameter estimates, respectively. They concluded from simulations that accurate estimates could be obtained for items using 750–1000 test takers and for test takers using 15–20 items.

Checking assumptions, including the fit of IRT models to both the items and test takers of a test, is another area of research at ETS during this period. Sinharay and Johnson (2003) studied the fit of IRT models to dichotomous item response data in the framework of Bayesian posterior model checking. Using simulations, they studied a number of discrepancy measures and suggest graphical summaries as having a potential to become a useful psychometric tool. In further work on this model checking (Sinharay 2003, 2005, 2006; Sinharay et al. 2006) they discussed the model-checking technique, and IRT model fit in general, extended some aspects of it, demonstrated it with simulations, and discussed practical applications. Deng coauthored (de la Torre and Deng 2008) an article proposing a modification of the standardized log likelihood of the response vector measure of person fit in IRT models, taking into account test reliability and using resampling methods. Evaluating the method, they found type I error rates were close to the nominal and power was good, resulting in a conclusion that the method is a viable and promising approach.

Based on earlier work during a postdoctoral fellowship at ETS, M. von Davier and Molenaar (2003) presented a person-fit index for dichotomous and polytomous IRT and latent structure models. Sinharay and Lu (2008) studied the correlation between fit statistics and IRT parameter estimates; previous researchers had found such a correlation, which was a concern for practitioners. These authors studied some newer fit statistics not examined in the previous research, and found these new statistics not to be correlated with the item parameters. Haberman (2009b) discussed use of generalized residuals in the study of fit of 1PL and 2PL IRT models, illustrating with operational test data.

Mislevy and Sinharay coauthored an article (Levy et al. 2009) on posterior predictive model checking, a flexible family of model-checking procedures, used as a tool for studying dimensionality in the context of IRT. Factors hypothesized to influence dimensionality and dimensionality assessment are couched in conditional covariance theory and conveyed via geometric representations of multidimensionality. Key findings of a simulation study included support for the hypothesized effects of the manipulated factors with regard to their influence on dimensionality assessment and the superiority of certain discrepancy measures for conducting posterior predictive model checking for dimensionality assessment.

Xu and Jia (2011) studied the effects on item parameter estimation in Rasch and 2PL models of generating data from different ability distributions (normal distribution, several degrees of generalized skew normal distributions), and estimating parameters assuming these different distributions. Using simulations, they found for the Rasch model that the estimates were little affected by the fitting distribution, except for fitting a normal to an extremely skewed generating distribution; whereas for the 2PL this was true for distributions that were not extremely skewed, but there were computational problems (unspecified) that prevented study of extremely skewed distributions.

M. von Davier and Yamamoto (2004) extended the GPC model to enable its use with discrete mixture IRT models with partially missing mixture information. The model includes LC analysis and multigroup IRT models as special cases. An application to large-scale assessment mathematics data, with three school types as groups and 20% of the grouping data missing, was used to demonstrate the model.

M. von Davier and Sinharay (2010) presented an application of a stochastic approximation EM algorithm using a Metropolis-Hastings sampler to estimate the parameters of an item response latent regression (LR) model. These models extend IRT to a two-level latent variable model in which covariates serve as predictors of the conditional distribution of ability. Applications to data from NAEP were presented, and results of the proposed method were compared to results obtained using the current operational procedures.

Haberman (2004) discussed joint and conditional ML estimation for the dichotomous Rasch model, explored conditions for consistency and asymptotic normality, investigated effects of model error, estimated errors of prediction, and developed generalized residuals. The same author (Haberman 2005a) showed that if a parametric model for the ability distribution is not assumed, the 2PL and 3PL (but not 1PL) models have identifiability problems that impose restrictions on possible models for the ability distribution. Haberman (2005b) also showed that LC item response models with small numbers of classes are competitive with IRT models for the 1PL and 2PL cases, showing that computations are relatively simple under these conditions. In another report, Haberman (2006) applied adaptive quadrature to ML estimation for IRT models with normal ability distributions, indicating that this method may achieve significant gains in speed and accuracy over other methods.

Information about the ability variable when an IRT model has a latent class structure was the topic of Haberman (2007a) in another publication. He also discussed reliability estimates and sampling and provided examples. Expressions for bounds on log odds ratios involving pairs of items for unidimensional IRT models in general, and explicit bounds for 1PL and 2Pl models were derived by Haberman, Holland, and Sinharay (2007). The results were illustrated through an example of their use in a study of model-checking procedures. These bounds can provide an elementary basis for assessing goodness of fit of these models. In another publication, Haberman (2008) showed how reliability of an IRT scaled score can be estimated and that it may be obtained even though the IRT model may not be valid.

Zhang (2005a) used simulated data to investigate whether Lord's bias function and weighted likelihood estimation method for IRT ability with known item parameters would be effective in the case of unknown parameters, concluding that they may not be as effective in that case. He also presented algorithms and methods for obtaining the global maximum of a likelihood, or weighted likelihood (WL), function.

Lewis (2001) produced a chapter on expected response functions (ERFs) in which he discussed Bayesian methods for IRT estimation. Zhang and Lu (2007) developed a new corrected weighted likelihood (CWL) function estimator of ability in IRT models based on the asymptotic formula of the WL estimator; they showed via simulation that the new estimator reduces bias in the ML and WL estimators, caused by failure to take into account uncertainty in item parameter estimates. Y.-H. Lee and Zhang (2008) further studied this estimator and Lewis' ERF estimator under various conditions of test length and amount of error in item parameter estimates. They found that the ERF reduced bias in ability estimation under all conditions and the CWL under certain conditions.

Sinharay coedited a volume on psychometrics in the *Handbook of Statistics* (Rao and Sinharay 2007), and contributions included chapters by: M. von Davier et al. (2007) describing recent developments and future directions in NAEP statistical procedures; Haberman and von Davier (2007) on models for cognitively based skills; von Davier and Rost (2007) on mixture distribution IRT models; Johnson et al. (2007) on hierarchical IRT models; Mislevy and Levy (2007) on Bayesian approaches; Holland et al. (2007) on equating, including IRT.

D. Li and Oranje (2007) compared a new method for approximating standard error of regression effects estimates within an IRT-based regression model, with the imputation-based estimator used in NAEP. The method is based on accounting for complex samples and finite populations by Taylor series linearization, and these authors formally defined a general method, and extended it to multiple dimensions. The new method was compared to the NAEP imputation-based method.

Antal and Oranje (2007) described an alternative numerical integration applicable to IRT and emphasized its potential use in estimation of the LR model of NAEP. D. Li, Oranje, and Jiang (2007) discussed parameter recovery and subpopulation proficiency estimation using the hierarchical latent regression (HLR) model and made comparisons with the LR model using simulations. They found the regression effect estimates were similar for the two models, but there were substantial differences in the residual variance estimates and standard errors, especially when there was large variation across clusters because a substantial portion of variance is unexplained in LR.

M. von Davier and Sinharay (2004) discussed stochastic estimation for the LR model, and Sinharay and von Davier (2005) extended a bivariate approach that represented the gold standard for estimation to allow estimation in more than two dimensions. M. von Davier and Sinharay (2007) presented a Robbins-Monro type stochastic approximation algorithm for LR IRT models and applied this approach to NAEP reading and mathematics data.

5.6 IRT Software Development and Evaluation

Wang et al. (2001, 2005) produced SCORIGHT, a program for scoring tests composed of testlets. M. von Davier (2008a) presented stand-alone software for multidimensional discrete latent trait (MDLT) models that is capable of marginal ML estimation for a variety of multidimensional IRT, mixture IRT, and hierarchical IRT models, as well as the GD approach. Haberman (2005b) presented a stand-alone general software for MIRT models. Rijmen (2006) presented a MATLAB toolbox utilizing tools from graphical modeling and Bayesian networks that allows estimation of a range of MIRT models.

5.6.1 Explanation, Evaluation, and Application of IRT Models

For the fourth edition of *Educational Measurement* edited by Brennan, authors Yen and Fitzpatrick (2006) contributed the chapter on IRT, providing a great deal of information useful to both practictioners and researchers. Although other ETS staff were authors or coauthors of chapters in this book, they did not focus on IRT methodology, per se.

Muraki et al. (2000) presented IRT methodology for psychometric procedures in the context of performance assessments, including description and comparison of many IRT and CTT procedures for scaling, linking, and equating. Tang and Eignor (2001), in a simulation, studied whether CTT item statistics could be used as collateral information along with IRT calibration to reduce sample sizes for pretesting TOEFL items, and found that CTT statistics, as the only collateral information, would not do the job.

Rock and Pollack (2002) investigated model-based methods (including IRT-based methods), and more traditional methods of measuring growth in prereading and reading at the kindergarten level, including comparisons between demographic groups. They concluded that the more traditional methods may yield uninformative if not incorrect results.

Scrams et al. (2002) studied use of item variants for continuous linear computer-based testing. Results showed that calibrated difficulty parameters of analogy and antonym items from the GRE General Test were very similar to those based on variant family information, and, using simulations, they showed that precision loss in ability estimation was less than 10% in using parameters estimated from expected response functions based only on variant family information.

A study comparing linear, fixed common item, and concurrent parameter estimation equating methods in capturing growth was conducted and reported by Jodoin et al. (2003). A. A. von Davier and Wilson studied the assumptions made at each step of calibration through IRT true-score equating and methods of checking whether the assumptions are met by a dataset. Operational data from the *AP*® Calculus AB exam were used as an illustration. Rotou et al. (2007) compared the measurement precision, in terms of reliability and conditional standard error of measurement (CSEM), of multistage (MS), CAT, and linear tests, using 1PL, 2PL, and 3PL IRT models. They found the MS tests to be superior to CAT and linear tests for the 1PL and 2PL models, and performance of the MS and CAT to be about the same, but better than the linear for the 3PL case.

Liu et al. (2008) compared the bootstrap and Markov chain Monte Carlo (MCMC) methods of estimation in IRT true-score equating with simulations based on operational testing data. Patterns of standard error estimates for the two methods were similar, but the MCMC produced smaller bias and mean square errors of equating. G. Lee and Fitzpatrick (2008), using operational test data, compared IRT equating by the Stocking-Lord method with and without fixing the *c* parameters. Fixing the *c* parameters had little effect on parameter estimates of the nonanchor items, but a considerable effect at the lower end of the scale for the anchor items. They suggeted that practitioners consider using the fixed-*c* method.

A regression procedure was developed by Haberman (2009a) to simultaneously link a very large number of IRT parameter estimates obtained from a large number of test forms, where each form has been separately calibrated and where forms can be linked on a pairwise basis by means of common items. An application to 2PL and GPC model data was also presented. Xu et al. (2011) presented two methods of using nonparametric IRT models in linking, illustrating with both simulated and operational datasets. In the simulation study, they showed that the proposed methods recover the true linking function when parametric models do not fit the data or when there is a large discrepancy in the populations.

Y. Li (2012), using simulated data, studied the effects, for a test with a small number of polytomous anchor items, of item parameter drift on TCC linking and IRT true-score equating. Results suggest that anchor length, number of items with drifting parameters, and magnitude of the drift affected the linking and equating results. The ability distributions of the groups had little effect on the linking and equating results. In general, excluding drifted polytomous anchor items resulted in an improvement in equating results.

D. Li et al. (2012) conducted a simulation study of IRT equating of six forms of a test, comparing several equating transformation methods and separate versus concurrent item calibration. The characteristic curve methods yielded smaller biases and smaller sampling errors (or accumulation of errors over time) so the former were concluded to be superior to the latter and were recommended in practice.

Livingston (2006) described IRT methodology for item analysis in a book chapter in *Handbook of Test Development* (Downing and Haladyna 2006). In the same publication, Wendler and Walker (2006) discussed IRT methods of scoring, and Davey and Pitoniak (2006) discussed designing CATs, including use of IRT in scoring, calibration, and scaling.

Almond et al. (2007) described Bayesian network models and their application to IRT-based CD modeling. The paper, designed to encourage practitioners to learn to use these models, is aimed at a general educational measurement audience, does not use extensive technical detail, and presents examples.

5.6.2 The Signs of (IRT) Things to Come

The body of work that ETS staff has contributed to in the development and applications of IRT, MIRT, and comprehensive integrated models based on IRT has been documented in multiple published monographs and edited volumes. At the point of writing this chapter, the history is still in the making; there are three more edited volumes that would have not been possible without the contributions of ETS researchers reporting on the use of IRT in various applications. More specifically:

- *Handbook of Item Response Theory* (second edition) contains chapters by Shelby Haberman, John Mazzeo, Robert Mislevy, Tim Moses, Frank Rijmen, Sandip Sinharay, and Matthias von Davier.
- *Computerized Multistage Testing: Theory and Applications* (edited by Duanli Yan, Alina von Davier, & Charlie Lewis, 2014) contains chapters by Isaac Bejar, Brent Bridgeman, Henry Chen, Shelby Haberman, Sooyeon Kim, Ed Kulick,

Yi-Hsuan Lee, Charlie Lewis, Longjuan Liang, Skip Livingston, John Mazzeo, Kevin Meara, Chris Mills, Andreas Oranje, Fred Robin, Manfred Steffen, Peter van Rijn, Alina von Davier, Matthias von Davier, Carolyn Wentzel, Xueli Xu, Kentaro Yamamoto, Duanli Yan, and Rebecca Zwick.

- *Handbook of International Large Scale International Assessment* (edited by Leslie Rutkowski, Matthias von Davier, & David Rutkowski, 2013) contains chapters by Henry Chen, Eugenio Gonzalez, John Mazzeo, Andreas Oranje, Frank Rijmen, Matthias von Davier, Jonathan Weeks, Kentaro Yamamoto, and Lei Ye.

5.7 Conclusion

Over the past six decades, ETS has pushed the envelope of modeling item response data using a variety of latent trait models that are commonly subsumed under the label IRT. Early developments, software tools, and applications allowed insight into the particular advantages of approaches that use item response functions to make inferences about individual differences on latent variables. ETS has not only provided theoretical developments, but has also shown, in large scale applications of IRT, how these methodologies can be used to perform scale linkages in complex assessment designs, and how to enhance reporting of results by providing a common scale and unbiased estimates of individual or group differences.

In the past two decades, IRT, with many contributions from ETS researchers, has become an even more useful tool. One main line of development has connected IRT to cognitive models and integrated measurement and structural modeling. This integration allows for studying questions that cannot be answered by secondary analyses using simple scores derived from IRT- or CTT-based approaches. More specifically, differential functioning of groups of items, the presence or absence of evidence that suggests that multiple diagnostic skill variables can be identified, and comparative assessment of different modeling approaches are part of what the most recent generation of multidimensional explanatory item response models can provide.

ETS will continue to provide cutting edge research and development on future IRT-based methodologies, and continues to play a leading role in the field, as documented by the fact that nine chapters of the *Handbook of Item Response Theory (second edition)* are authored by ETS staff. Also, of course, at any point in time, including the time of publication of this work, there are numerous research projects being conducted by ETS staff, and for which reports are being drafted, reviewed, or submitted for publication. By the timeaa this work is published, there will undoubtedly be additional publications not included herein.

References

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62,* 569–578. https://doi.org/10.1007/BF02294643

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223–237. https://doi.org/10.1177/01466219922031347

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*, 341–359.

Andersen, E. B. (1972). *A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires* (Research Memorandum No. RM-72-06). Princeton: Educational Testing Service.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140. https://doi.org/10.1007/BF02291180

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. https://doi.org/10.1007/BF02294143

Antal, T. (2007). *On multidimensional item response theory: A coordinate-free approach* (Research Report No. RR-07-30). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02072.x

Antal, T., & Oranje, A. (2007). *Adaptive numerical integration for item response theory* (Research Report No. RR-07-06). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02048.x

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report No. RR-81-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 29,* 163–175. https://doi.org/10.1111/j.1745-3984.1992.tb00372.x

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283–296. https://doi.org/10.1111/j.1745-3984.1980.tb00832.x

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1–23). Vancouver: Educational Research Institute of British Columbia.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14,* 237–245. https://doi.org/10.1177/014662169001400302

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (Research Report No. RR-81-35). Princeton: Educational Testing Service.

Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement, 6*, 309–325. https://doi.org/10.1177/014662168200600308

Bejar, I. I., Braun, H. I., & Carlson, S. B. (1989). *Psychometric foundations of testing based on patient management problems* (Research Memorandum No. RM-89-02). Princeton: Educational Testing Service.

Belov, D., Armstrong, R. D., & Weissman, A. (2008). A Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement, 32,* 431–446. https://doi.org/10.1177/0146621607309081

Bennett, R. E., Sebrechts, M. M., & Yamamoto, K. (1991). *Fitting new measurement models to GRE General Test constructed-response item data* (Research Report No. RR-91-60). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01427.x

Birnbaum, A. (1967). *Statistical theory for logistic mental test models with a prior distribution of ability* (Research Bulletin No. RB-67-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1967.tb00363.x

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Boldt, R. F. (1989). Latent structure analysis of the Test of English as a Foreign Language. *Language Testing, 6,* 123–142. https://doi.org/10.1177/026553228900600201

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 147–156). New York: Springer. https://doi.org/10.1007/978-0-387-49839-3_9

Bradlow, E. T. (1996). Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics, 21*, 179–185.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168. https://doi.org/10.1007/BF02294533

Camilli, G., Yamamoto, K., & Wang, M.-M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17,* 379–388. https://doi.org/10.1177/014662169301700407

Chang, H.-H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika, 61,* 445–463. https://doi.org/10.1007/BF02294549

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391–404. https://doi.org/10.1007/BF02296132

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37–52. https://doi.org/10.1007/BF02294469

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1985). *An investigation of the feasibility of applying item response theory to equate achievement tests* (Research Report No. RR-85-31). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00116.x

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*, 161–173.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225–244. https://doi.org/10.1177/014662168701100302

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985a). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (Research Report No. RR-85-30) . Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00115.x

Cook, L. L., Eignor, D. R., & Taft, H. L. (1985b). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* (Research Report No. RR-85-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00123.x

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1985c). *A study of the temporal stability of IRT item parameter estimates* (Research Report No. RR-85-45). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00130.x

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988a). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (Research Report No. RR-88-52). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00308.x

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988b). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13,* 19–43. https://doi.org/10.2307/1164949

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988c). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25,* 31–45. https://doi.org/10.1111/j.1745-3984.1988.tb00289.x

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48*, 129–141. https://doi.org/10.1007/BF02314681

Davey, T., & Pitoniak, M. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543–573). Mahwah: Erlbaum.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45,* 159–177. https://doi.org/10.1111/j.1745-3984.2008.00058.x

Donoghue, J. R. (1992). *On a common misconception concerning the partial credit and generalized partial credit polytomous IRT models* (Research Memorandum No. RM-92-12). Princeton: Educational Testing Service.

Donoghue, J. R. (1993). *An empirical examination of the IRT information in polytomously scored reading items* (Research Report No. RR-93-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01523.x

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22,* 33–51. https://doi.org/10.1177/01466216980221002

Dorans, N. J. (1985). Item parameter invariance: The cornerstone of item response theory. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3, pp. 55–78). Greenwich: JAI Press.

Dorans, N. J. (1986). The impact of item deletion on equating conversions and reported score distributions. *Journal of Educational Measurement, 23,* 245–264. https://doi.org/10.1111/j.1745-3984.1986.tb00250.x

Dorans, N. J. (1990). Scaling and equating. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137–160). Hillsdale: Erlbaum.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 135–158). Mahwah: Erlbaum.

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (Research Report No. RR-92-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE Verbal Scale. *Journal of Educational Measurement, 22,* 249–262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Douglass, J. B., Marco, G. L., & Wingersky, M. S. (1985). *An evaluation of three approximate item response theory models for equating test scores* (Research Report No. RR-85-46). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00131.x

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah: Erlbaum.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT Verbal and Mathematical sections* (Research Report No. RR-85-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00095.x

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating* (Research Report No. RR-86-14). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00169.x

Eignor, D. R., Golub-Smith, M. L., & Wingersky, M. S. (1986). *Application of a new goodness-of-fit plot procedure to SAT and TOEFL item type data* (Research Report No. RR-86-47). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00202.x

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3,* 37–52. https://doi.org/10.1207/s15324818ame0301_4

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515. https://doi.org/10.1007/BF02294487

Flaugher, R. (1990). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 41–63). Hillsdale: Erlbaum.

Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 37–59). Mahwah: Erlbaum.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373–389. https://doi.org/10.1177/014662168901300404

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (1993). *Test theory for a new generation of tests*. Hillsdale: Erlbaum.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189. https://doi.org/10.1111/j.1745-3984.1991.tb00352.x

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Dordrecht: Kluwer. https://doi.org/10.1007/0-306-47531-6_14

Green, B. F., Jr. (1950a). *A proposal for a comparative study of the measurement of attitude* (Research Memorandum no. RM-50-20). Princeton: Educational Testing Service.

Green, B. F., Jr. (1950b). *A proposal for an empirical evaluation of the latent class model of latent structure analysis* (Research Memorandum No. RM-50-26). Princeton: Educational Testing Service.

Green, B. F., Jr. (1951a). A general solution for the latent class model of latent structure analysis. *Psychometrika, 16,* 151–166. https://doi.org/10.1007/BF02289112

Green, B. F., Jr. (1951b). *Latent class analysis: A general solution and an empirical evaluation* (Research Bulletin No. RB-51-15). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00215.x

Green, B. F., Jr. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association, 47,* 71–76. https://doi.org/10.1080/01621459.1952.10501155

Green, B. F., Jr. (1980, April). *Ledyard R Tucker's affair with psychometrics: The first 45 years*. Paper presented at a special symposium in honor of Ledyard R Tucker. Champaign: The University of Illinois.

Gustafsson, J.-E., Morgan, A. M. B., & Wainer, H. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics, 5*, 35–64.

Haberman, S. J. (2004). *Joint and conditional maximum likelihood estimation for the Rasch model of binary responses* (Research Report No. RR-04-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01947.x

Haberman, S. J. (2005a). *Identifiability of parameters in item response models with unconstrained ability distributions* (Research Report No. RR-05-24). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02001.x

Haberman, S. J. (2005b). *Latent-class item response models* (Research Report No. RR-05-28). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02005.x

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (Research Report No. RR-06-29). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02035.x

Haberman, S. J. (2007a). *The information a test provides on an ability parameter* (Research Report No. RR-07-18). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02060.x

Haberman, S. J. (2007b). The interaction model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 201–216). New York: Springer. https://doi.org/10.1007/978-0-387-49839-3_13

Haberman, S. J. (2008). *Reliability of scaled scores* (Research Report No. RR-08-70). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02156.x

Haberman, S. J. (2009a). *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report No. RR-09-40). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Haberman, S. J. (2009b). *Use of generalized residuals to examine goodness of fit of item response models* (Research Report No. RR-09-15). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02172.x

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209–227. https://doi.org/10.1007/s11336-010-9158-4

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Haberman, S. J., Holland, P. W., & Sinharay, S. (2007). Limits on log odds ratios for unidimensional item response theory models. *Psychometrika, 72,* 551–561. https://doi.org/10.1007/s11336-007-9009-0

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. RR-08-45). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Hambleton, R. K. (1983). *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75–96. https://doi.org/10.1111/j.1745-3984.1977.tb00030.x

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31–49). New York: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50010-X

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 104–122). Vancouver: Educational Research Institute of British Columbia.

Hartz, S., & Roussos, L. (2008). *The fusion model for skills diagnosis: Blending theory with practicality* (Research Report No. RR-08-71). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02157.x

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91–97. https://doi.org/10.1111/j.1745-3984.1989.tb00321.x

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. *Applied Psychological Measurement, 7,* 255–266. https://doi.org/10.1177/014662168300700302

Hicks, M. M. (1984). *A comparative study of methods of equating TOEFL test scores* (Research Report No. RR-84-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00060.x

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79–92. https://doi.org/10.1007/BF02293920

Holland, P. (1990a). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601. https://doi.org/10.1007/BF02294609

Holland, P. (1990b). The Dutch identity: A new tool for the study of item response theory models. *Psychometrika, 55,* 5–18. https://doi.org/10.1007/BF02294739

Holland, P. W., & Hoskens, M. (2002). *Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test* (Research Report No. RR-02-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01887.x

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14,* 1523–1543. https://doi.org/10.1214/aos/1176350174

Holland, P., Dorans, N., & Petersen, N. (2007). Equating. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 169–204). Amsterdam: Elsevier.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229–250. https://doi.org/10.1080/00220970309602064

Johnson, M., Sinharay, S., & Bradlow, E. T. (2007). Hierarchical item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 587–605). Amsterdam: Elsevier.

Jones, D. H. (1980). *On the adequacy of latent trait models* (Program Statistics Research Technical Report No. 80–08). Princeton: Educational Testing Service.

Jones, D. H. (1982). *Tools of robustness for item response theory* (Research Report No. RR-82-41). Princeton: Educational Testing Service.

Jones, D. H. (1984a). *Asymptotic properties of the robustified jackknifed estimator* (Research Report No. RR-84-41). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00081.x

Jones, D. H. (1984b). *Bayesian estimators, robust estimators: A comparison and some asymptotic results* (Research Report No. RR-84-42). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00082.x

Jones, D. H., Kaplan, B. A., & Wainer, H. (1984). *Estimating ability with three item response models when the models are wrong and their parameters are inaccurate* (Research Report No. RR-84-26). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00066.x

Kingston, N. M. (1986). *Assessing the dimensionality of the GMAT Verbal and Quantitative measures using full information factor analysis* (Research Report No. RR-86-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00168.x

Kingston, N. M., & Dorans, N. J. (1982a). *The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory* (Research Report No. RR-82-22). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01308.x

Kingston, N. M., & Dorans, N. J. (1982b). *The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test* (Research Report No. RR-82-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01298.x

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147–154. https://doi.org/10.1177/014662168400800202

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288. https://doi.org/10.1177/014662168500900306

Kingston, N. M., & Holland, P. W. (1986). *Alternative methods of equating the GRE General Test* (Research Report No. RR-86-16). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00171.x

Kingston, N. M., Leary, L. F., & Wightman, L. E. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (Research Report No. RR-85-34). Princeton: Educational Testing Service.

Kreitzberg, C. B., & Jones, D. H. (1980). *An empirical study of the broad range tailored test of verbal ability* (Research Report No. RR-80-05). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01195.x

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1977). *Computerized adaptive testing: The concepts and its potentials* (Research Memorandum No. RM-77-03). Princeton: Educational Testing Service.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8,* 313–340. https://doi.org/10.1207/s15324818ame0804_3

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (Research Report No. RR-88-23). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00279.x

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3,* 19–36. https://doi.org/10.1207/s15324818ame0301_3

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. 4. Measurement and prediction* (pp. 362–472). Princeton: Princeton University Press.

Lee, G., & Fitzpatrick, A. R. (2008). A new approach to test score equating using item response theory with fixed c-parameters. *Asia Pacific Education Review, 9,* 248–261. https://doi.org/10.1007/BF03026714

Lee, Y.-H., & Zhang, J. (2008). *Comparing different approaches of bias correction for ability estimation in IRT models* (Research Report No. RR-08-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02099.x

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519–537. https://doi.org/10.1177/0146621608329504

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 163–171). New York: Springer. https://doi.org/10.1007/978-1-4613-0169-1_9

Lewis, C., & Sheehan, K. M. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367–386. https://doi.org/10.1177/014662169001400404

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

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement, 49,* 167–189. https://doi.org/10.1111/j.1745-3984.2012.00167.x

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report No. RR-12-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02291.x

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (Research Report No. RR-10-21). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02228.x

Liu, Y., Schulz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33,* 257–278. https://doi.org/10.3102/1076998607306076

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Mahwah: Erlbaum.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73–95. https://doi.org/10.1207/s15324818ame0301_6

Lord, F. M. (1951). *A theory of test scores and their relation to the trait measured* (Research Bulletin No. RB-51-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00922.x

Lord, F. M. (1952a). *A theory of test scores* (Psychometric Monograph No. 7). Richmond: Psychometric Corporation.

Lord, F. M. (1952b). *The scale proposed for the academic ability test* (Research Memorandum No. RM-52-03). Princeton: Educational Testing Service.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549. https://doi.org/10.1177/001316445301300401

Lord, F. M. (1964). *A strong true score theory, with applications* (Research Bulletin No. RB-64-19). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00922.x

Lord, F. M. (1965a). An empirical study of item-test regression. *Psychometrika, 30,* 373–376. https://doi.org/10.1007/BF02289501

Lord, F. M. (1965b). A note on the normal ogive or logistic curve in item analysis. *Psychometrika, 30,* 371–372. https://doi.org/10.1007/BF02289500

Lord, F. M. (1968a). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020. https://doi.org/10.1177/001316446802800401

Lord, F. M. (1968b). *Some test theory for tailored testing* (Research Bulletin No. RB-68-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1968.tb00562.x

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—A confrontation of Birnbaum's logistic model. *Psychometrika, 35,* 43–50. https://doi.org/10.1007/BF02290592

Lord, F. M. (1973). Power scores estimated by item characteristic curves. *Educational and Psychological Measurement, 33,* 219–224. https://doi.org/10.1177/001316447303300201

Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264. https://doi.org/10.1007/BF02291471

Lord, F. M. (1974b). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. II, pp. 106–126). San Francisco: Freeman.

Lord, F. M. (1975a). *A survey of equating methods based on item characteristic curve theory* (Research Bulletin No. RB-75-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01052.x

Lord, F. M. (1975b). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin No. RB-75-33). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01073.x

Lord, F. M. (1975c). The 'ability' scale in item characteristic curve theory. *Psychometrika, 40,* 205–217. https://doi.org/10.1007/BF02291567

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138. https://doi.org/10.1111/j.1745-3984.1977.tb00032.x

Lord, F. M. (1980a). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Lord, F. M. (1980b). Some how and which for practical tailored testing. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 189–205). New York: Wiley.

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6,* 463–472. https://doi.org/10.1177/014662168200600407

Lord, F. M. (1983a). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48,* 477–482. https://doi.org/10.1007/BF02293689

Lord, F. M. (1983b). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51–61). New York: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50011-1

Lord, F. M. (1983c). Statistical bias in maximum likelihood estimation of item parameters. *Psychometrika, 48,* 425–435. https://doi.org/10.1007/BF02293684

Lord, F. M. (1983d). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233–245. https://doi.org/10.1007/BF02294018

Lord, F. M. (1984). *Conjunctive and disjunctive item response functions* (Research Report No. RR-84-45). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00085.x

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157–162. https://doi.org/10.1111/j.1745-3984.1986.tb00241.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lord, F. M., & Pashley, P. (1988). *Confidence bands for the three-parameter logistic item response curve* (Research Report No. RR-88-67). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00323.x

Lord, F. M., & Wild, C. L. (1985). *Contribution of verbal item types in the GRE General Test to accuracy of measurement of the verbal scores* (Research Report No. RR-85-29). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00114.x

Lord, F. M., & Wingersky, M. S. (1973). *A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. RM-73-02). Princeton: Educational Testing Service.

Lord, F. M., & Wingersky, M. S. (1982). *Sampling variances and covariances of parameter estimates in item response theory* (Research Report No. RR-82-33). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01318.x

Lord, F. M., & Wingersky, M. S. (1983). *Comparison of IRT observed-score and true-score 'equatings'* (Research Report No. RR-83-26). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1983.tb00026.x

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 453–461. https://doi.org/10.1177/014662168400800409

Lord, F. M., Wingersky, M. S., & Wood, R. L. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. RM-76-06). Princeton: Educational Testing Service.

Marco, G. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

McKinley, R. L. (1988). A comparison of six methods for combining multiple IRT item parameter estimates. *Journal of Educational Measurement, 25,* 233–246. https://doi.org/10.1111/j.1745-3984.1988.tb00305.x

McKinley, R. L. (1989a). *Confirmatory analysis of test structure using multidimensional item response theory* (Research Report No. RR-89-31). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00145.x

McKinley, R. L. (1989b). Methods plainly speaking: An introduction to item response theory. *Measurement and Evaluation in Counseling and Development, 22*, 37–57.

McKinley, R. L., & Kingston, N. M. (1987). *Exploring the use of IRT equating for the GRE subject test in mathematics* (Research Report No. RR-87-21). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00225.x

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57. https://doi.org/10.1177/014662168500900105

McKinley, R. L., & Schaeffer, G. A. (1989). *Reducing test form overlap of the GRE Subject Test in Mathematics using IRT triple-part equating* (Research Report No. RR-89-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00334.x

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23,* 147–160. https://doi.org/10.1177/01466219922031275

McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27,* 121–137. https://doi.org/10.1177/0146621602250534

Messick, S. J. (1985). *The 1986 NAEP design: Changes and challenges* (Research Memorandum No. RM-85-02). Princeton: Educational Testing Service.

Messick, S. J., Beaton, A. E., Lord, F. M., Baratz, J. C., Bennett, R. E., Duran, R. P., … Wainer, H. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83–01). Princeton: Educational Testing Service.

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

Mislevy, R. (1985). *Inferences about latent populations from complex samples* (Research Report No. RR-85-41). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00126.x

Mislevy, R. (1986a). *A Bayesian treatment of latent variables in sample surveys* (Research Report No. RR-86-01). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00155.x

Mislevy, R. (1986b). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195. https://doi.org/10.1007/BF02293979

Mislevy, R. (1986c). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3–31. https://doi.org/10.2307/1164846

Mislevy, R. (1987a). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81–91. https://doi.org/10.1177/014662168701100106

Mislevy, R. (1987b). Recent developments in item response theory with implications for teacher certification. *Review of Research in Education, 14,* 239–275. https://doi.org/10.2307/1167313

Mislevy, R. (1988). Exploiting auxiliary information about items in the estimation of Rasch item parameters. *Applied Psychological Measurement, 12,* 281–296. https://doi.org/10.1177/014662168801200306

Mislevy, R. (1989). *Foundations of a new test theory* (Research Report No. RR-89-52). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01336.x

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

Mislevy, R. (1993a). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale: Erlbaum.

Mislevy, R. (1993b). Some formulas for use with Bayesian ability estimates. *Educational and Psychological Measurement, 53,* 315–328. https://doi.org/10.1177/0013164493053002002

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models [Computer software]*. Mooresville: Scientific Software.

Mislevy, R., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57–75). San Diego: Academic Press.

Mislevy, R., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65,* 149–156. https://doi.org/10.1007/BF02294370

Mislevy, R. J., & Huang, C.-W. (2007). Measurement models as narrative structures. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch model: Extensions and applications* (pp.15–35). New York: Springer. https://doi.org/10.1007/978-0-387-49839-3_2

Mislevy, R., & Levy, R. (2007). Bayesian psychometric modeling from an evidence centered design perspective. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 839–866). Amsterdam: Elsevier.

Mislevy, R. J., & Sheehan, K. M. (1989a). Information matrices in latent-variable models. *Journal of Educational Statistics, 14,* 335–350. https://doi.org/10.2307/1164943

Mislevy, R. J., & Sheehan, K. M. (1989b). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661–679. https://doi.org/10.1007/BF02296402

Mislevy, R., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG* (Research Report No. RR-87-43). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00247.x

Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215. https://doi.org/10.1007/BF02295283

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71. https://doi.org/10.1007/BF02296958

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report No. RR-88-48). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00304.x

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Mislevy, R. J., Beaton, A. E., Kaplan, B. A., & Sheehan, K. M. (1992a). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

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

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71. https://doi.org/10.1177/014662169001400106

Muraki, E. (1992a). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E. (1992b). *RESGEN item response generator* (Research Report No. RR-92-07). Princeton: Educational Testing Service.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363. https://doi.org/10.1177/014662169301700403

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36,* 217–232. https://doi.org/10.1111/j.1745-3984.1999.tb00555.x

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT-based test scoring and item analysis for graded items and rating scales* [Computer program]. Chicago: Scientific Software.

Muraki, E., & Carlson, J. E. (1995). Full information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90. https://doi.org/10.1177/014662169501900109

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24,* 325–337. https://doi.org/10.1177/01466210022031787

Pashley, P. J. (1991). *An alternative three-parameter logistic item response model* (Research Report No. RR-91-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01376.x

Pashley, P. J. (1992). *Graphical IRT-based DIF analyses* (Research Report No. RR-92-66). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01497.x

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156.

Rao, C. R., & Sinharay, S. (Eds.). (2007). *Handbook of statistics: Vol. 26. Psychometrics*. Amsterdam: Elsevier.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373. https://doi.org/10.1177/014662169101500407

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam: VU University Medical Center.

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Research Report No. RR-09-03). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02160.x

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2013). A general psychometric approach for educational survey assessments: Flexible statistical models and efficient estimation methods. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis*. London: Chapman & Hall.

Roberts, J. S. (n.d.). *Item response theory models for unfolding*. Retrieved from http://www.psychology.gatech.edu/unfolding/Intro.html

Roberts, J. S., & Laughlin, J. E (1996). A unidimensional item response model for unfolding from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231–255. https://doi.org/10.1177/014662169602000305

Roberts, J. S., Donoghue, J. R., & Laughlin, L. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32. https://doi.org/10.1177/01466216000241001

Roberts, J. S., Donoghue, J. R., & Laughlin, L. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26,* 192–207. https://doi.org/10.1177/01421602026002006

Rock, D. A., & Pollack, J. M. (2002). *A model-based approach to measuring cognitive growth in pre-reading and reading skills during the kindergarten year* (Research Report No. RR-02-18). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01885.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. (1984). *Testing the local independence assumption in item response theory* (Research Report No. RR-84-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00049.x

Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology, 38,* 206–215. https://doi.org/10.1111/j.2044-8317.1985.tb00836.x

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika, 52*, 217–233. https://doi.org/10.1007/BF02294236

Ross, J. (1966). An empirical study of a logistic mental test model. *Psychometrika, 31*, 325–340. https://doi.org/10.1007/BF02289466

Rotou, O., Patsula, L. N., Steffen, M., & Rizavi, S. M. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* (Research Report No. RR-07-04). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02046.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores *Psychometrika*, *34*(4, Whole Pt. 2). https://doi.org/10.1007/BF03372160

Samejima, F. (1972). A general model for free-response data. *Psychometrika*, *37*(1, Whole Pt. 2).

Scheuneman, J. D. (1980). Latent trait theory and item bias. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 140–151). New York: Wiley.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3,* 53–71. https://doi.org/10.1207/s15324818ame0301_5

Scrams, D. J., Mislevy, R. J., & Sheehan, K. M. (2002). *An analysis of similarities in item functioning within antonym and analogy variant families* (Research Report No. RR-02-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01880.x

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34,* 333–352. https://doi.org/10.1111/j.1745-3984.1997.tb00522.x

Sheehan, K. M., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65–76. https://doi.org/10.1177/014662169201600108

Sheehan, K. M., & Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures* (Research Report No. RR-88-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00294.x

Sheehan, K. M., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27,* 255–272 https://doi.org/10.1111/j.1745-3984.1990.tb00747.x

Sheehan, K. M., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills* (Research Report No. RR-94-14). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01587.x

Sinharay, S. (2003). *Practical applications of posterior predictive model checking for assessing fit of 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. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42,* 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59,* 429–449. https://doi.org/10.1348/000711005X66888

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (Research Report No. RR-03-28). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01920.x

Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement, 45,* 1–15. https://doi.org/10.1111/j.1745-3984.2007.00049.x

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BRGOUP program to higher dimensions* (Research Report No. RR-05-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02004.x

Sinharay, S., Johnson, M. S., & Williamson, D. (2003). *An application of a Bayesian hierarchical model for item family calibration* (Research Report No. RR-03-04). Princeton*:* Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01896.x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321. https://doi.org/10.1177/0146621605285517

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187–231). Hillsdale: Erlbaum.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 185–229). Mahwah: Erlbaum.

Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Report No. RR-88-28). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties* (Research Report No. RR-89-05). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00331.x

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55,* 461–475. https://doi.org/10.1007/BF02294761

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210. https://doi.org/10.1177/014662168300700208

Stocking, M. L., Swanson, L., & Pearlman, M. (1991a). *Automatic item selection (AIS) methods in the ETS testing environment* (Research Memorandum No. RM-91-05). Princeton: Educational Testing Service.

Stocking, M. L., Swanson, L., & Pearlman, M. (1991b). *Automated item selection using item response theory* (Research Report No. RR-91-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01375.x

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (Research Report No. RR-01-11). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01853.x

Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. *Behaviormetrika, 13,* 73–86. https://doi.org/10.2333/bhmk.13.19_73

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

Tatsuoka, K. K. (1991). *A theory of IRT-based diagnostic testing* (Office of Naval Research Report). Princeton: Educational Testing Service.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 103–135). Hillsdale: Erlbaum.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–131). Mahwah: Erlbaum.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577. https://doi.org/10.1007/BF02295596

Thissen, D., & Wainer, H. (1984). *The graphical display of simulation results with applications to the comparison of robust IRT estimators of ability* (Research Report No. RR-84-36). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00076.x

Thissen, D., & Wainer, H. (1985). *Some supporting evidence for Lord's guideline for estimating "c" theory* (Research Report No. RR-85-15). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00100.x

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics, 15,* 113–128. https://doi.org/10.2307/1164765

Thissen, D., Wainer, H., & Rubin, D. (1984). *A computer program for simulation evaluation of IRT ability estimators* (Research Report No. RR-84-37). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00077.x

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1–13. https://doi.org/10.1007/BF02288894

van Rijn, P., & Rijmen, F. (2015). On the explaining-away phenomenon in multivariate latent variable models. *British Journal of Mathematical and Statistical Psychology, 68,* 1–22. https://doi.org/10.1111/bmsp.12046

von Davier, A. A., & Wilson, C. (2005). *A didactic approach to the use of IRT true-score equating model* (Research Report No. RR-05-26). Princeton: Educational Testing Service.

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. RR-07-19). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02061.x

von Davier, M. (2008a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61,* 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2008b). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 255–274). Charlotte: Information Age Publishing.

von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M. & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68,* 213–228. https://doi.org/10.1007/BF02294798

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 643–661). Amsterdam: Elsevier. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M., & Sinharay, S. (2004). *Application of the stochastic EM method to latent regression models* (Research Report No. RR-04-34). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01961.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. (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., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformation. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3,* 115–124. https://doi.org/10.1027/1614-2241.3.3.115

von Davier, M., & von Davier, A. A. (2011). A general model for IRT scale linking and scale transformation. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 225–242). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389–406. https://doi.org/10.1177/0146621604268734

von Davier, M., & Yamamoto, K. (2007). Mixture distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99–115). New York: Springer. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay S. (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–174). Cambridge, MA: Hogrefe & Huber.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76,* 318–336. https://doi.org/10.1007/s11336-011-9202-z

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technological revolution in testing. *Journal of College Admission, 28*, 9–16.

Wainer, H. (1990). Introduction and history. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 1–21). Hillsdale: Erlbaum.

Wainer, H. (2000). Introduction and history. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1–21). Mahwah: Erlbaum.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 271–299). Mahwah: Erlbaum.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1–14. https://doi.org/10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Mislevy, R. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, D. R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (pp. 65–102). Hillsdale: Erlbaum.

Wainer, H., & Mislevy, R. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 61–100). Mahwah: Erlbaum.

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412. https://doi.org/10.1007/BF02293705

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339–368. https://doi.org/10.2307/1165054

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203–220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (1990a). *Computer adaptive testing: A primer*. Hillsdale: Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990b). Future challenges. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (pp. 233–270). Hillsdale: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991) Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219. https://doi.org/10.1111/j.1745-3984.1991.tb00354.x

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement, 31*, 183–199. https://doi.org/10.1111/j.1745-3984.1994.tb00442.x

Wainer, H., Bradlow, E. T., & Du, Z. (2000a). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht: Kluwer. https://doi.org/10.1007/0-306-47531-6_13

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (2000b). *Computer adaptive testing: A primer* (2nd ed.). Mahwah: Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000c). Future challenges. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy,

L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 231–269). Mahwah: Erlbaum.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing theory* (Research Bulletin No. RB-76-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1976.tb01094.x

Wang, X., Bradlow, E. T., & Wainer, H. (2001). *User's guide for SCORIGHT (version 1.2): A computer program for scoring tests built of testlets* (Research Report No. RR-01-06). Princeton: Educational Testing Service.

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

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis* (Research Report No. RR-04-49). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01976.x

Wendler, C. L. W., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Mahwah: Erlbaum.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S. (1987). *One-stage LOGIST* (Research Report No. RR-87-45). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00249.x

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364. https://doi.org/10.1177/014662168400800312

Wingersky, M. S., & Sheehan, K. M. (1986). *Using estimated item observed-score regressions to test goodness-of-fit of IRT models* (Research Report No. RR-86-23). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00178.x

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report No. RR-87-24). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00228.x

Winsberg, S., Thissen, D., & Wainer, H. (1984). *Fitting item characteristic curves with spline functions* (Research Report No. RR-84-40). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00080.x

Xu, X. (2007). *Monotone properties of a general diagnostic model* (Research Report No. RR-07-25). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02067.x

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika, 71,* 121–137. https://doi.org/10.1007/s11336-003-1154-5

Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (Research Report No. RR-11-41). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02277.x

Xu, X., & von Davier, M. (2006). *Cognitive diagnostics for NAEP proficiency data* (Research Report No. RR-06-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02014.x

Xu, X., & von Davier, M. (2008a). *Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model* (Research Report No. RR-08-35). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02121.x

Xu, X., & von Davier, M. (2008b). *Fitting the structured general diagnostic model to NAEP* data (Research Report No. RR-08-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02113.x

Xu, X., & von Davier, M. (2008c). Linking for the general diagnostic model. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 1*, 97–111.

Xu, X., Douglas, J., & Lee, Y.-S. (2011). Linking with nonparametric IRT models. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 243–258). New York: Springer.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (Research Report No. RR-89-41). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Research Report No. RR-95-16). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01651.x

Yamamoto, K., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait class models in the social sciences* (pp. 89–99). New York: Waxmann.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 275–295). Hillsdale: Erlbaum.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17,* 155–173. https://doi.org/10.2307/1165167

Yan, D., Almond, R. G., & Mislevy, R. J. (2004a). *A comparison of two models for cognitive diagnosis* (Research Report No. RR-04-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01929.x

Yan, D., Lewis, C., & Stocking, M. L. (2004b). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics, 29,* 293–316. https://doi.org/10.3102/10769986029003293

Yen, W. M. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 123–141). Vancouver: Educational Research Institute of British Columbia.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport: American Council on Education and Praeger Publishers.

Zhang, J. (2004). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (Research Report No. RR-04-44). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01971.x

Zhang, J. (2005a). *Bias correction for the maximum likelihood estimate of ability* (Research Report No. RR-05-15). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01992.x

Zhang, J. (2005b). *Estimating multidimensional item response models with mixed structure* (Research Report No. RR-05-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01981.x

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72,* 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Lu, T. (2007). *Refinement of a bias-correction procedure for the weighted likelihood estimator of ability* (Research Report No. RR-07-23). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02065.x

Zhang, J., & Stout, W. (1997). On Holland's Dutch identity conjecture. *Psychometrika, 62,* 375–392. https://doi.org/10.1007/BF02294557

Zhang, J. & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129–152. https://doi.org/10.1007/BF02294532

Zhang, J. & Stout, W. (1999b). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249. https://doi.org/10.1007/BF02294536

Zwick, R. J. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293–308. https://doi.org/10.1111/j.1745-3984.1987.tb00281.x

Zwick, R. J. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197. https://doi.org/10.2307/1165031

Zwick, R. J. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10–16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x

Zwick, R. J., Thayer, D. T., & Wingersky, M. S. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32,* 341–363. https://doi.org/10.1111/j.1745-3984.1995.tb00471.x

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/10.1177/014662160202600l007

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

Chapter 7
Contributions to the Quantitative Assessment of Item, Test, and Score Fairness

Neil J. Dorans

ETS was founded in 1947 as a not-for-profit organization (Bennett, Chap. 1, this volume). Fairness concerns have been an issue at ETS almost since its inception. William Turnbull (1949, 1951a, b), who in 1970 became the second president of ETS, addressed the Canadian Psychological Association on socioeconomic status and predictive test scores. He made a cogent argument for rejecting the notion that differences in subgroup performance on a test means that a test score is biased. He also advocated the comparison of prediction equations as a means of assessing test fairness. His article was followed by a number of articles by ETS staff on the issue of differential prediction. By the 1980s, under the direction of its third president, Gregory Anrig, ETS established the industry standard for fairness assessment at the item level (Holland and Wainer 1993; Zieky 2011). This century, fairness analyses have begun to focus on relationships between tests that purport to measure the same thing in the same way across different subgroups (Dorans and Liu 2009; Liu and Dorans 2013).

In this chapter, I review quantitative fairness procedures that have been developed and modified by ETS staff over the past decades. While some reference is made to events external to ETS, the focus is on ETS, which has been a leader in fairness assessment. In the first section, Fair Prediction of a Criterion, I consider differential prediction and differential validity, procedures that examine whether test scores predict a criterion, such as performance in college, across different subgroups in a similar manner. The bulk of this review is in the second section, Differential Item Functioning (DIF), which focuses on item-level fairness, or

This chapter was originally published in 2013 by Educational Testing Service as a research report in the ETS R&D Scientific and Policy Contributions Series. ETS staff continue to contribute to the literature on fairness. See Dorans and Cook (2016) for some examples.

N.J. Dorans (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: ndorans@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_7

DIF. Then in the third section, Fair Linking of Test Scores, I consider research pertaining to whether tests built to the same set of specifications produce scores that are related in the same way across different gender and ethnic groups. In the final section, Limitations of Quantitative Fairness Assessment Procedures, limitations of these procedures are mentioned.

7.1 Fair Prediction of a Criterion

Turnbull (1951a) concluded his early ETS treatment of fairness with the following statement: “Fairness, like its amoral brother, validity, resides not in tests or test scores but in the relation to its uses” (p. 4–5).

While several ETS authors had addressed the relative lower performance of minority groups on tests of cognitive ability and its relationship to grades (e.g., Campbell 1964), Cleary (1968) conducted one of the first differential prediction studies. That study has been widely cited and critiqued. A few years after the Cleary article, the field was replete with differential validity studies, which focus on comparing correlation coefficients, and differential prediction studies, which focus on comparing regression functions, in large part because of interest engendered by the Supreme Court decision *Griggs v. Duke Power Co.* in 1971. This decision included the terms *business necessity* and *adverse impact*, both of which affected employment testing. Adverse impact is a substantially different rate of selection in hiring, promotion, transfer, training, or other employment-related decisions for any race, sex, or ethnic group. Business necessity can be used by an employer as a justification for using a selection mechanism that appears to be neutral with respect to sex, race, national origin, or religious group even though it excludes members of one sex, race, national origin, or religious group at a substantially higher rate than members of other groups. The employer must prove that the selection requirement having the adverse impact is job related and consistent with business necessity. In other words, in addition to avoiding the use of race/ethnic/gender explicitly as part of the selection process, the selection instrument had to have demonstrated predictive validity for its use. Ideally, this validity would be the same for all subpopulations.

Linn (1972) considered the implications of the Griggs decision for test makers and users. A main implication was that there would be a need for empirical demonstrations that test scores predict criterion performance, such as how well one does on the job. (In an educational context, test scores may be used with other information to predict the criterion of average course grade). Reliability alone would not be an adequate justification for use of test scores. Linn also noted that for fair prediction to hold, the prediction model must include all the appropriate variables in the model. Otherwise misspecification of the model can give the appearance of statistical bias. The prediction model should include all the predictors needed to predict *Y*, and the functional form used to combine the predictors should be the correct one. The reliabilities of the predictors also were noted to play a role. These limitations with differential validity and differential predictions studies were cogently summarized in four pages by Linn and Werts (1971). One of the quandaries faced by researchers that was not noted in this 1971 study is that some of the variables that contribute to prediction are variables over which a test taker has little control, such as gender, race, parent's level of education and income, and even zip code. Use of variables such as zip code to predict grades in an attempt to eliminate differential prediction would be unfair.

Linn (1975) later noted that differential prediction analyses should be preferred to differential validity studies because differences in predictor or criterion variability can produce differential validity even when the prediction model is fair. Differential prediction analyses examine whether the same prediction models hold across different groups. Fair prediction or selection requires invariance of prediction equations across groups,

$$R\left(Y|\mathbf{X},G=1\right)=R\left(Y|\mathbf{X},G=2\right)=\ldots=R\left(Y|\mathbf{X},G=g\right),$$

where R is the symbol for the function used to predict Y, the criterion score, from $\mathbf{X}$, the predictor. G is a variable indicating subgroup membership.

Petersen and Novick (1976) compared several models for assessing fair selection, including the regression model (Cleary 1968), the constant ratio model (Thorndike 1971), the conditional probability model (Cole 1973), and the constant probability model (Linn 1973) in the lead article in a special issue of the *Journal of Educational Measurement* dedicated to the topic of fair selection. They demonstrated that the regression, or Cleary, model, which is a differential prediction model, was a preferred model from a logical perspective in that it was consistent with its converse (i.e., fair selection of applicants was consistent with fair rejection of applicants). In essence, the Cleary model examines whether the regression of the criterion onto the predictor space is invariant across subpopulations.

Linn (1976) in his discussion of the Petersen and Novick (1976) analyses noted that the quest to achieve fair prediction is hampered by the fact that the criterion in many studies may itself be unfairly measured. Even when the correct equation is correctly specified and the criterion is measured well in the full population, invariance may not hold in subpopulations because of selection effects. Linn (1983) described how predictive bias may be an artifact of selection procedures. Linn used a simple case to illustrate his point. He posited that a single predictor X and linear model were needed to predict Y in the full population P. To paraphrase his argument, assume that a very large sample is drawn from P based on a selection variable U that might depend on X in a linear way. Errors in the prediction of Y from X and U from X are thus also linearly related because of their mutual dependence on X. Linn showed that the sample regression for the selected sample, R $(Y|X, G)$ equals the regression in the full unselected population if the correlation between X and U is zero, or if errors in prediction of Y from X and U from X are uncorrelated.

Myers (1975) criticized the regression model because regression effects can produce differences in intercepts when two groups differ on X and Y and the predictor is unreliable, a point noted by Linn and Werts (1971). Myers argued for a linking or scaling model for assessing fairness. He noted that his approach made sense when

X and *Y* were measures of the same construct, but admitted that scaling test scores to grades or vice versa had issues. This brief report by Myers can be viewed as a remote harbinger of work on the population invariance of score linking functions done by Dorans and Holland (2000), Dorans (2004), Dorans and Liu (2009), and Liu and Dorans (2013).

As can be inferred from the studies above, in particular Linn and Werts (1971) and Linn (1975, 1983), there are many ways in which a differential prediction study can go awry, and even more ways that differential validity studies can be problematic.

7.2 Differential Item Functioning (DIF)

During the 1980s, the focus in the profession shifted to DIF studies. Although interest in item bias studies began in the 1960s (Angoff 1993), it was not until the 1980s that interest in fair assessment at the item level became widespread. During the 1980s, the measurement profession engaged in the development of item level models for a wide array of purposes. DIF procedures developed as part of that shift in attention from the score to the item.

Moving the focus of attention to prediction of item scores, which is what DIF is about, represented a major change from focusing primarily on fairness in a domain, where so many factors could spoil the validity effort, to a domain where analyses could be conducted in a relatively simple, less confounded way. While factors such as multidimensionality can complicate a DIF analysis, as described by Shealy and Stout (1993), they are negligible compared to the many influences that can undermine a differential prediction study, as described in Linn and Werts (1971). In a DIF analysis, the item is evaluated against something designed to measure a particular construct and something that the test producer controls, namely a test score.

Around 100 ETS research bulletins, memoranda, or reports have been produced on the topics of item fairness, DIF, or item bias. The vast majority of these studies were published in the late 1980s and early 1990s. The major emphases of these reports can be sorted into categories and are treated in subsections of this section: Differential Item Functioning Methods, Matching Variable Issues, Study Group Definitions, and Sample Size and Power Issues. The DIF methods section begins with some definitions followed by a review of procedures that were suggested before the term DIF was introduced. Most of the section then describes the following procedures: Mantel-Haenszel (MH), standardization (STAND), item response theory (IRT), and SIBTEST.

7.2.1 Differential Item Functioning (DIF) Methods

Two reviews of DIF methods were conducted by ETS staff: Dorans and Potenza (1994), which was shortened and published as Potenza and Dorans (1995), and Mapuranga et al. (2008), which then superseded Potenza and Dorans. In the last of these reviews, the criteria for classifying DIF methods were (a) definition of null DIF, (b) definition of the studied item score, (c) definition of the matching variable, and (d) the variable used to define groups.

Null DIF is the absence of DIF. One definition of null DIF, observed score null DIF, is that all individuals with the same score on a test of the shared construct measured by that item should have the same proportions answering the item correctly regardless of whether they are from the reference or focal group. The latent variable definition of null DIF can be used to compare the performance of focal and reference subgroups that are matched with respect to a latent variable. An observed difference in average item scores between two groups that may differ in their distributions of scores on the matching variable is referred to as *impact*. With impact, we compare groups that may or may not be comparable with respect to the construct being measured by the item; using DIF, we compare item scores on groups that are comparable with respect to an estimate of their standing on that construct.

The s*tudied item score* refers to the scoring rule used for the items being studied for DIF. Studied items are typically[1] scored as correct/incorrect (i.e., binary) or scored using more than two response categories (i.e., polytomous). The *matching variable* is a variable used in the process of comparing the reference and focal groups (e.g., total test score or subscore) so that comparable groups are formed. In other words, matching is a way of establishing score equivalence between groups that are of interest in DIF analyses. The matching variable can either be an observed score or an estimate of the unobserved latent variable consistent with a specific model for item performance, and can be either a univariate or multivariate variable.

In most DIF analyses, a single focal group is compared to a single reference group where the subgroup-classification variable (gender, race, geographic location, etc.) is referred to as the *grouping variable*. This approach ignores potential interactions between types of subgroups, (e.g., male/female and ethnic/racial). Although it might be better to analyze all grouping variables for DIF simultaneously (for statistical and computational efficiency), most DIF methods compare only two groups at a time. While convention is often the reason for examining two groups at a time, small sample size sometimes makes it a necessity.

The remainder of this section describes briefly the methods that have been developed to assess what has become known as DIF. After reviewing some early work, I turn to the two methods that are still employed operationally here at ETS: the MH method and the STAND method. After briefly discussing IRT methods, I mention

[1] All options can be treated as nominally scored, which could be useful in cases where the focus is on differential functioning on options other than the key (distractors).

the SIBTEST method. Methods that do not fit into any of these categories are addressed in what seems to be the most relevant subsection.

7.2.1.1 Early Developments: The Years Before Differential Item Functioning (DIF) Was Defined at ETS

While most of the focus in the 1960s and 1970s was on the differential prediction issue, several researchers turned their attention to item-level fairness issues. Angoff (1993) discussed several, but not all of these efforts. Cardall and Coffman (1964) and Cleary and Hilton (1966, 1968) defined *item bias*, the phrase that was commonly used before DIF was introduced, as an item-by-subgroup interaction. Analysis of variance was used by both studies of DIF. Identifying individual problem items was not the goal of either study.

Angoff and Sharon (1974) also employed an analysis of variance (ANOVA) method, but by then the transformed item difficulty (TID) or delta-plot method had been adopted for item bias research. Angoff (1972) introduced this approach, which was rooted in Thurstone's absolute scaling model. This method had been employed by Tucker (1951) in a study of academic ability on vocabulary items and by Gulliksen (1964) in a cross-national study of occupation prestige. This method uses an inverse normal transformation to convert item proportion-correct values for two groups to normal deviates that are expect to form an ellipse. Items that deviate from the ellipse exhibit the item difficulty by group interaction that is indicative of what was called item bias. Angoff and Ford (1971, 1973) are the standard references for this approach.

The delta-plot method ignores differences in item discrimination. If items differ in their discriminatory power and the groups under study differ in terms of proficiency, then items will exhibit item-by-group interactions even when there are no differences in item functioning. This point was noted by several scholars including Lord (1977) and affirmed by Angoff (1993). As a consequence, the delta-plot method is rarely used for DIF assessment, except in cases where small samples are involved.

Two procedures may be viewed as precursors of the eventual move to condition directly on total score that was adopted by the STAND (Dorans and Kulick 1983) and MH (Holland and Thayer 1988) DIF approaches. Stricker (1982) recommended a procedure that looks for DIF by examining the partial correlation between group membership and item score with the effect of total test score removed. Scheuneman (1979) proposed a test statistic that looked like a chi-square. This method was shown by Baker (1981) and others to be affected inappropriately by sample size and to possess no known sampling distribution.

The late 1980s and the early 1990s were the halcyon days of DIF research and development at ETS and in the profession. Fairness was of paramount concern, and practical DIF procedures were developed and implemented (Dorans and Holland 1993; Zieky 1993). In October 1989, ETS and the Air Force Human Resources Laboratory sponsored a DIF conference that was held at ETS in October 1989. The papers presented at that conference, along with a few additions, were collected in the volume edited by Holland and Wainer (1993), known informally as the DIF book. It contains some of the major work conducted in this early DIF era, including several chapters about MH and STAND. The chapter by Dorans and Holland (1993) is the source of much of the material in the next two sections, which describe the MH and STAND procedures in some detail because they have been used operationally at ETS since that time. Dorans (1989) is another source that compares and contrasts these two DIF methods.

7.2.1.2 Mantel-Haenszel (MH): Original Implementation at ETS

In their seminal paper, Mantel and Haenszel (1959) introduced a new procedure for the study of matched groups. Holland and Thayer (1986, 1988) adapted the procedure for use in assessing DIF. This adaptation, the MH method, is used at ETS as the primary DIF detection device. The basic data used by the MH method are in the form of *M* 2-by-2 contingency tables or one large three dimensional 2-by-2-by-M table, where *M* is the number of levels of the matching variable.

Under rights-scoring for the items in which responses are coded as either correct or incorrect (including omissions), proportions of rights and wrongs on each item in the target population can be arranged into a contingency table for each item being studied. There are two levels for group: the focal group (f) that is the focus of analysis, and the reference group (r) that serves as a basis for comparison for the focal group. There are also two levels for item response: right (R) or wrong (W), and there are *M* score levels on the matching variable, (e.g., total score). Finally, the item being analyzed is referred to as the studied item. The 2 (groups)-by-2 (item scores)-by-M (score levels) contingency table (see Table 7.1) for each item can be viewed in *2-by-2 slices.*

The null DIF hypothesis for the MH method can be expressed as

$$H_0 : \left[R_{rm} / W_{rm}\right] = \left[R_{fm} / W_{fm}\right] \ \forall m = 1,\ldots,M.$$

In other words, the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and the reference group portions of the population, and this equality holds across all M levels of the matching variable.

Table 7.1 2-by-2-by-M contingency table for an item, viewed in a 2-by-2 Slice

	Item score		
Group	Right	Wrong	Total
Focal group (f)	R_{fm}	W_{fm}	N_{fm}
Reference group (r)	R_{rm}	W_{rm}	N_{rm}
Total group (t)	R_{tm}	W_{tm}	N_{tm}

In their original work, Mantel and Haenszel (1959) developed a chi-square test of the null hypothesis against a particular alternative hypothesis known as the constant odds ratio hypothesis,

$$\mathrm{H_a} : \left[\mathrm{R}_{rm} / \mathrm{W}\right] = \alpha\left[\mathrm{R}_{fm} / \mathrm{W}_{fm}\right] \forall \mathrm{m} = 1, \ldots, \mathrm{M}, \mathit{and}\, \alpha \neq 1.$$

Note that when α = 1, the alternative hypothesis reduces to the null DIF hypothesis. The parameter α is called the *common odds ratio* in the M 2-by-2 tables because under $\mathrm{H_a}$, the value of α is the odds ratio common for all m.

Holland and Thayer (1988) reported that the MH approach is the test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds-ratio hypothesis.

Mantel and Haenszel (1959) also provided an estimate of the constant odds – ratio that ranges from 0 to ∞, for which a value of 1 can be taken to indicate null DIF. This odds-ratio metric is not particularly meaningful to test developers who are used to working with numbers on an item difficulty scale. In general, odds are converted to ln[odds-ratio] because the latter is symmetric around zero and easier to interpret.

At ETS, test developers use item difficulty estimates in the *delta metric*, which has a mean of 13 and a standard deviation of 4. Large values of delta correspond to difficult items, while easy items have small values of delta. Holland and Thayer (1985) converted the estimate of the common odds ratio, α_{MH}, into a difference in deltas via:

$$\mathit{MH}\,\mathrm{D} - \mathit{DIF} = -2.35\ln\left[\alpha \mathit{MH}\right].$$

Note that positive values of MH D-DIF favor the focal group, while negative values favor the reference group. An expression for the standard error of for MH D-DIF was provided in Dorans and Holland (1993).

7.2.1.3 Subsequent Developments With the Mantel-Haenszel (MH) Approach

Subsequent to the operational implementation of the MH approach to DIF detection by ETS in the late 1980s (Zieky 1993, 2011), there was a substantial amount of DIF research conducted by ETS staff through the early 1990s. Some of this research was presented in Holland and Wainer (1993); other presentations appeared in journal articles and ETS Research Reports. This section contains a partial sampling of research conducted primarily on the MH approach.

Donoghue et al. (1993) varied six factors in an IRT-based simulation of DIF in an effort to better understand the properties of the MH and STAND (to be described in the next section) effect sizes and their standard errors. The six factors varied were level of the IRT discrimination parameter, the number of DIF items in the matching variable, the amount of DIF on the studied item, the difficulty of the studied item, whether the studied item was included in the matching variable, and the number of items in the matching variable. Donoghue et al. found that both the MH and STAND methods had problems detecting IRT DIF in items with nonzero lower asymptotes. Their two major findings were the need to have enough items in the matching variable to ensure reliable matching for either method, and the need to include the studied item in the matching variable in MH analysis. This study thus provided support for the analytical argument for inclusion of the studied item that had been made by Holland and Thayer (1986). As will be seen later, Zwick et al. (1993a), Zwick (1990), Lewis (1993), and Tan et al. (2010) also addressed the question of inclusion of the studied item.

Longford et al. (1993) demonstrated how to use a random-effect or variance-component model to aggregate DIF results for groups of items. In particular they showed how to combine DIF estimates from several administrations to obtain variance components for administration differences for DIF within an item. In their examples, they demonstrated how to use their models to improve estimations within an administration, and how to combine evidence across items in randomized DIF studies. Subsequently, ETS researchers have employed Bayesian methods with the goal of pooling data across administrations to yield more stable DIF estimates within an administration. These approaches are discussed in the section on sample size and power issues.

Allen and Holland (1993) used a missing data framework to address the missing data problem in DIF analyses where “no response” to the self-reported group identification question is large, a common problem in applied settings. They showed how MH and STAND statistics can be affected by different assumptions about nonresponses.

Zwick and her colleagues examined DIF in the context of computer adaptive testing (CAT) in which tests are tailored to the individual test taker on the basis of his or her response to previous items. Zwick et al. (1993b) described in great detail a simulation study in which they examined the performance of MH and STAND procedures that had been modified for use with data collected adaptively. The modification to the DIF procedures involved replacing the standard number-right matching variable with a matching variable based on IRT, which was obtained by converting a maximum likelihood estimate of ability to an expected number-right true score on all items in the reference pool. Examinees whose expected true scores fell in the same one-unit intervals were considered to be matched. They found that DIF statistics computed in this way for CAT were similar to those obtained with the traditional matching variable of performance on the total test. In addition they found that pretest DIF statistics were generally well behaved, but the MH DIF statistics tended to have larger standard errors for the pretest items than for the CAT items.

Zwick et al. (1994) addressed the effect of using alternative matching methods for pretest items. Using a more elegant matching procedure did not lead to a reduction of the MH standard errors and produced DIF measures that were nearly identical to those from the earlier study. Further investigation showed that the MH standard errors tended to be larger when items were administered to examinees with a wide ability range, whereas the opposite was true of the standard errors of the

STAND DIF statistic. As reported in Zwick (1994), there may be a theoretical explanation for this phenomenon.

CAT can be thought of as a very complex form of item sampling. The sampling procedure used by the National Assessment of Educational Progress (NAEP) is another form of complex sampling. Allen and Donoghue (1996) used a simulation study to examine the effect of complex sampling of items on the measurement of DIF using the MH DIF procedure. Data were generated using a three-parameter logistic (3PL) IRT model according to the balanced incomplete block design. The length of each block of items and the number of DIF items in the matching variable were varied, as was the difficulty, discrimination, and presence of DIF in the studied item. Block, booklet, pooled booklet, and other approaches to matching on more than the block, were compared to a complete data analysis using the transformed log-odds on the delta scale. The pooled booklet approach was recommended for use when items are selected for examinees according to a balanced incomplete block (BIB) data collection design.

Zwick et al. (1993a) noted that some forms of performance assessment may in fact be more likely to tap construct-irrelevant factors than multiple-choice items are. The assessment of DIF can be used to investigate the effect on subpopulations of the introduction of performance tasks. Two extensions of the MH procedure were explored: the test of conditional association proposed by Mantel (1963) and the generalized statistic proposed by Mantel and Haenszel (1959). Simulation results showed that, for both inferential procedures, the studied item should be included in the matching variable, as in the dichotomous case. Descriptive statistics that index the magnitude of DIF, including that proposed by Dorans and Schmitt (1991; described below) were also investigated.

7.2.1.4 Standardization (STAND)

Dorans (1982) reviewed item bias studies that had been conducted on data from the *SAT*® exam in the late 1970s, and concluded that these studies were flawed because either DIF was confounded with lack of model fit or it was contaminated by impact as a result of *fat matching*, the practice of grouping scores into broad categories of roughly comparable ability. A new method was needed. Dorans and Kulick (1983, 1986) developed the STAND approach after consultation with Holland. The formulas in the following section can be found in these articles and in Dorans and Holland (1993) and Dorans and Kulick (2006).

Standardization's (STAND's) Definition of Differential Item Functioning (DIF)

An item exhibits DIF when the expected performance on an item differs for matched examinees from different groups. Expected performance can be estimated by nonparametric item-test regressions. Differences in empirical item-test regressions are indicative of DIF.

The first step in the STAND analysis is to use all available item response data in the target population of interest to estimate nonparametric item-test regressions in the reference group and in the focal group. Let $E_f(Y|X)$ define the empirical item-test regression for the focal group f, and let $E_r(Y|X)$ define the empirical item-test regression for the reference group r, where Y is the item-score variable and X is the matching variable. For STAND, the definition of null-DIF conditions on an observed score is $E_r(Y|X)=E_r(Y|X)$ Plots of difference in empirical item-test regressions, focal minus reference, provide visual descriptions of DIF in fine detail for binary as well as polytomously scored items. For illustrations of nonparametric item-test regressions and differences for an actual SAT item that exhibits considerable DIF, see Dorans and Kulick (1986).

Standardization's (STAND's) Primary Differential Item Functioning (DIF) Index

While plots described DIF directly, there was a need for some numerical index that targets suspect items for close scrutiny while allowing acceptable items to pass swiftly through the screening process. For each score level, the focal group supplies specific weights that are used for each individual D_m before accumulating the weighted differences across score levels to arrive at a summary item-discrepancy index, $STD-EISDIF$, which is defined as:

$$STD-EISDIF=\frac{\sum_{m=1}^{M}N_{fm}*E_f\left(Y|X=m\right)}{\sum_{m=1}^{M}N_{fm}}-\frac{\sum_{m=1}^{M}N_{fm}*E_r\left(Y|X=m\right)}{\sum_{m=1}^{M}N_{fm}}$$

or simplified

$$STD-EISDIF=\frac{\sum_{m=1}^{M}N_{fm}\left(*E_f\left(Y|X=m\right)-E_r\left(Y|X=m\right)\right)}{\sum_{m=1}^{M}N_{fm}}$$

where $N_{fm}/\sum_{m=1}^{M}N_{fm}$ is the weighting factor at score level X_m supplied by the focal group to weight differences in expected item performance observed in the focal group $E_f(Y|X)$ and expected item performance observed in the reference group $E_r(Y|X)$.

In contrast to impact, in which each group has its relative frequency serve as a weight at each score level, STAND uses a standard or common weight on both $E_f(Y|X)$ and $E_r(Y|X)$, namely $N_{fm}/\sum_{m=1}^{M}N_{fm}$. The use of the same weight on both $E_f(Y|X)$ and $E_r(Y|X)$ is the essence of the STAND approach. Use of N_{fm} means that $STD-EISDIF$ equals the difference between the observed performance of the focal group on the item and the predicted performance of selected reference group members who are matched in ability to the focal group members. This difference can be derived very simply; see Dorans and Holland (1993).

Extensions to Standardization (STAND)

The generalization of the STAND methodology to all response options including omission and not reached is straightforward and is known as standardized distractor analysis (Dorans and Schmitt 1993; Dorans et al. 1988, 1992). It is as simple as replacing the keyed response with the option of interest in all calculations. For example, a standardized response-rate analysis on Option A would entail computing the proportions choosing A in both the focal and reference groups. The next step is to compute differences between these proportions at each score level. Then these individual score-level differences are summarized across score levels by applying some standardized weighting function to these differences to obtain $STD-DIF(A)$, the standardized difference in response rates to Option A. In a similar fashion one can compute standardized differences in response rates for Options B, C, D, and E, and for nonresponses as well. This procedure is used routinely at ETS.

Application of the STAND methodology to counts of examinees at each level of the matching variable who did not reach the item results in a standardized not-reached difference. For items at the end of a separately timed section of a test, these standardized differences provide measurement of the differential speededness of a test. *Differential speededness* refers to the existence of differential response rates between focal group members and matched reference group members to items appearing at the end of a section. Schmitt et al. (1993) reported that excluding examinees who do not reach an item from the calculation of the DIF statistic for that item partially compensates for the effects of item location on the DIF estimate.

Dorans and Schmitt (1991) proposed an extended version of STAND for ordered polytomous data. This extension has been used operationally with NAEP data since the early 1990s. This approach, called standardized mean difference (*SMD*) by Zwick et al. (1993a), provides an average DIF value for describing DIF on items with ordered categories. At each matching score level, there exist distributions of ordered item scores, I, for both the focal group (e.g., females) and the reference group (e.g., males). The expected item scores for each group at each matching score level can be computed by using the frequencies to obtain a weighted average of the score levels. The difference between these expected items scores for the focal and reference groups, $STD-EISDIF$, is the DIF statistic. Zwick and Thayer (1996) provide standard errors for *SMD* (or $STD-EISDIF$).

7.2.1.5 Item Response Theory (IRT)

DIF procedures differ with respect to whether the matching variable is explicitly an observed score (Dorans and Holland 1993) or implicitly a latent variable (Thissen et al. 1993). Observed score DIF and DIF procedures based on latent variables do not measure the same thing, and both are not likely to measure what they strive to measure, which is DIF with respect to the construct that the item purports to measure. The observed score procedures condition on an observed score, typically the score reported to a test taker, which contains measurement error and clearly differs from a pure measure of the construct of interest, especially for test scores of inadequate reliability. The latent variable approaches in essence condition on an unobservable that the test is purportedly measuring. As such they employ what Meredith and Millsap (1992) would call a measurement invariance definition of null DIF, while methods like MH and STAND employ a prediction invariance definition, which may be viewed as inferior to measurement invariance from a theoretical perspective. On the other hand, procedures that purport to assess measurement invariance employ a set of assumptions; in essence they are assessing measurement invariance under the constraint that the model they assume to be true is in fact true.

The observed score methods deal with the fact that an unobservable is unknowable by replacing the null hypothesis of measurement invariance (i.e., the items measure the construct of interest in the same way in the focal and reference groups with a prediction invariance assumption and use the data directly to assess whether expected item score is a function of observed total score in the same way across groups). The latent variable approaches retain the measurement invariance hypothesis and use the data to estimate and compare functional forms of the measurement model relating item score to a latent variable in the focal and reference groups. The assumptions embodied in these functional forms may or may not be correct, however, and model misfit might be misconstrued as a violation of measurement invariance, as noted by Dorans and Kulick (1983). For example applying the Rasch (1960) model to data fit by the two-parameter logistic (2PL) model would flag items with lower IRT slopes as having DIF favoring the lower scoring group, while items with higher slopes would favor the higher scoring group.

Lord (1977, 1980) described early efforts to assess DIF from a latent trait variable perspective. Lord recommended a statistical significance test on the joint difference between the IRT difficulty and discrimination parameters between the two groups under consideration. Thissen et al. (1993) discussed Lord's procedure and described the properties of four other procedures that used IRT. All these methods used statistical significance testing. They also demonstrated how the IRT methods can be used to assess differential distractor functioning. Thissen et al. remains a very informative introduction and review of IRT methods circa 1990.

Pashley (1992) suggested a method for producing simultaneous confidence bands for the difference between item response curves. After these bands have been plotted, the size and regions of DIF can be easily identified. Wainer (1993) provided an IRT-based effect size of amount of DIF that is based on the STAND weighting system that allows one to weight difference in the item response functions (IRF) in a manner that is proportional to the density of the ability distribution.

Zwick et al. (1994) and Zwick et al. (1995) applied the Rasch model to data simulated according to the 3PL model. They found that the DIF statistics based on the Rasch model were highly correlated with the DIF values associated with the generated data, but that they tended to be smaller in magnitude. Hence the Rasch model did not detect DIF as well, which was attributed to degradation in the accuracy of matching. Expected true scores from the Rasch-based computer-adaptive test tended to be biased downward, particularly for lower-ability examinees. If the Rasch model had been used to generate the data, different results would probably have been obtained.

Wainer et al. (1991) developed a procedure for examining DIF in collections of related items, such as those associated with a reading passage. They called this DIF for a set of items a *testlet DIF*. This methodology paralleled the IRT-based likelihood procedures mentioned by Thissen et al. (1993).

Zwick (1989, 1990) demonstrated that the null definition of DIF for the MH procedure (and hence STAND and other procedures employing observed scores as matching variables) and the null hypothesis based on IRT are different because the latter compares item response curves, which in essence condition on unobserved ability. She also demonstrated that the item being studied for DIF should be included in the matching variable if MH is being used to identify IRT DIF.

7.2.1.6 SIBTEST

Shealy and Stout (1993) introduced a general model-based approach to assessing DIF and other forms of differential functioning. They cited the STAND approach as a progenitor. From a theoretical perspective, SIBTEST is elegant. It sets DIF within a general multidimensional model of item and test performance. Unlike most IRT approaches, which posit a specific form for the item response model (e.g., a 2PL model), SIBTEST does not specify a particular functional form. In this sense it is a nonparametric IRT model, in principle, in which the null definition of STAND involving regressions onto observed scores is replaced by one involving regression onto true scores,

$$\varepsilon_f\left(Y|T_x\right)=\varepsilon_r\left(Y|T_x\right),$$

where T_x represents a true score for$X$. As such, SIBTEST employs a measurement invariance definition of null DIF, while STAND employs a prediction invariance definition (Meredith and Millsap 1992).

Chang et al. (1995, 1996) extended SIBTEST to handle polytomous items. Two simulation studies compared the modified SIBTEST procedure with the generalized Mantel (1963) and SMD or STAND procedures. The first study compared the procedures under conditions in which the generalized Mantel and SMD procedures had been shown to perform well (Zwick et al. 1993a. Results of Study 1 suggested that SIBTEST performed reasonably well, but that the generalized Mantel and SMD procedures performed slightly better. The second study used data simulated under conditions in which observed-score DIF methods for dichotomous items had not performed well (i.e., a short nonrepresentative matching test). The results of Study 2 indicated that, under these conditions, the modified SIBTEST procedure provided better control of impact-induced Type I error inflation with respect to detecting DIF (as defined by SIBTEST) than the other procedures.

Zwick et al. (1997b) evaluated statistical procedures for assessing DIF in polytomous items. Three descriptive statistics – the SMD (Dorans and Schmitt 1991) and two procedures based on SIBTEST (Shealy and Stout 1993) were considered, along with five inferential procedures: two based on SMD, two based on SIBTEST, and one based on the Mantel (1963) method. The DIF procedures were evaluated through applications to simulated data, as well as to empirical data from ETS tests. The simulation included conditions in which the two groups of examinees had the same ability distribution and conditions in which the group means differed by one standard deviation. When the two groups had the same distribution, the descriptive index that performed best was the SMD. When the two groups had different distributions, a modified form of the SIBTEST DIF effect-size measure tended to perform best. The five inferential procedures performed almost indistinguishably when the two groups had identical distributions. When the two groups had different distributions and the studied item was highly discriminating, the SIBTEST procedures showed much better Type I error control than did the SMD and Mantel methods, particularly with short tests. The power ranking of the five procedures was inconsistent; it depended on the direction of DIF and other factors. The definition of DIF employed was the IRT definition, measurement invariance, not the observed score definition, prediction invariance.

Dorans (2011) summarized differences between SIBTEST and its progenitor, STAND. STAND uses observed scores to assess whether the item-test regressions are the same across focal and reference groups. On its surface, the SIBTEST DIF method appears to be more aligned with measurement models. This method assumes that examinee group differences influence DIF or test form difficulty differences more than can be observed in unreliable test scores. SIBTEST adjusts the observed data toward what is suggested to be appropriate by the measurement model. The degree to which this adjustment occurs depends on the extent that these data are unreliable. To compensate for unreliable data on the individual, SIBTEST regresses observed performance on the test to what would be expected for the focal or reference group on the basis of the ample data that show that race and gender are related to item performance. SIBTEST treats true score estimation as a prediction problem, introducing bias to reduce mean squared error. In essence, the SIBTEST method uses subgroup-specific true score estimates as a surrogate for the true score that is defined in the classical test theory model. If SIBTEST regressed all test takers to the same mean it would not differ from STAND.

7.2.2 Matching Variable Issues

Dorans and Holland (1993) laid out an informal research agenda with respect to observed score DIF. The matching variable was one area that merited investigation. Inclusion of the studied item in the matching variable and refinement or purification of the criterion were mentioned. Dimensionality and DIF was, and remains, an important factor; DIF procedures presume that all items measure the same construct in the same way across all groups.

Donoghue and Allen (1993) examined two strategies for forming the matching variable for the MH DIF procedure; "thin" matching on total test score was compared to forms of "thick" matching, pooling levels of the matching variable. Data were generated using a 3PL IRT model with a common guessing parameter. Number of subjects and test length were manipulated, as were the difficulty, discrimination, and presence/absence of DIF in the studied item. For short tests (five or ten items), thin matching yielded very poor results, with a tendency to falsely identify items as possessing DIF against the reference group. The best methods of thick matching yielded outcome measure values closer to the expected value for non-DIF items and a larger value than thin matching when the studied item possessed DIF. Intermediate-length tests yielded similar results for thin matching and the best methods of thick matching.

The issue of whether or not to include the studied item in the matching variable was investigated by many researchers from the late 1980s to early 1990s. Holland and Thayer (1988) demonstrated mathematically that when the data were consistent with the Rasch model, it was necessary to include the studied item in a purified rights-scored matching criterion in order to avoid biased estimates of DIF (of the measurement invariance type) for that studied item. Inclusion of the studied item removes the dependence of the item response on group differences in ability distributions. Zwick (1990) and Lewis (1993) developed this idea further to illustrate the applicability of this finding to more general item response models. Both authors proved mathematically that the benefit in bias correction associated with including the studied item in the matching criterion held true for the binomial model, and they claimed that the advantage of including the studied item in the matching criterion would not be evident for any IRT model more complex than the Rasch model.

Donoghue et al. (1993) evaluated the effect of including/excluding the studied item under the 3PL IRT model. In their simulation, they fixed the discrimination parameters for all items in a simulated test in each studied condition and fixed the guessing parameter for all conditions, but varied the difficulty (*b*) parameters for different items for each studied condition. Although the 3PL model was used to simulate data, only the *b*-parameter was allowed to vary. On the basis of their study, they recommended including the studied item in the matching variable when the MH procedure is used for DIF detection. They also recommended that short tests not be used for matching variables.

Zwick et al. (1993a) extended the scope of their DIF research to performance tasks. In their study, multiple-choice (MC) items and performance tasks were simulated using the 3PL model and the partial-credit model, respectively. The MC items were simulated to be free of DIF and were used as the matching criterion. The performance tasks were simulated to be the studied items with or without DIF. They found that the item should be included in the matching criterion.

Zwick (1990) analytically examined item inclusion for models more complex than the Rasch. Her findings apply to monotone IRFs with local independence for the case where the IRFs on the matching items were assumed identical for the two groups. If the studied item is excluded from the matching variable, the MH null hypothesis will not hold in general even if the two groups had the same IRF for the studied item. It is assured to hold only if the groups have the same ability distribution. If the ability distributions are ordered, the MH will show DIF favoring the higher group (generalization of Holland and Thayer's [1988] Rasch model findings). Even if the studied item is included, the MH null hypothesis will not hold in general. It is assured to hold only if the groups have the same ability distribution or if the Rasch model holds. Except in these special situations, the MH can produce a conclusion of DIF favoring either the focal or reference group.

Tan et al. (2010) studied the impact of including/excluding the studied item in the matching variable on bias in DIF estimates under conditions where the assumptions of the Rasch model were violated. Their simulation study varied different magnitudes of DIF and different group ability distributions, generating data from a 2PL IRT model and a multidimensional IRT model. Results from the study showed that including the studied item leads to less biased DIF estimates and more appropriate Type I error rate, especially when group ability distributions are different. Systematic biased estimates in favor of the high ability group were consistently found across all simulated conditions when the studied item was excluded from the matching criterion.

Zwick and Ercikan (1989) used bivariate matching to examine DIF on the NAEP history assessment, conditioning on number-right score and historical period studied. Contrary to expectation, the additional conditioning did not lead to a reduction in the number of DIF items.

Pomplun et al. (1992) evaluated the use of bivariate matching to study DIF with formula-scored tests, where item inclusion cannot be implemented in a straightforward fashion. Using SAT Verbal data with large and small samples, both male-female and black-white group comparisons were investigated. MH D-DIF values and DIF category classifications based on bivariate matching on rights score and nonresponse were compared with MH D-DIF values and categories based on rights-scored and formula-scored matching criteria. When samples were large, MH D-DIF values based on the bivariate matching criterion were ordered very similarly to MH D-DIF values based on the other criteria. However, with small samples the MH D-DIF values based on the bivariate matching criterion displayed only moderate correlations with MH D-DIF values from the other criteria.

7.2.3 Study Group Definition

Another area mentioned by Dorans and Holland (1993) was the definition of the focal and reference groups. Research has continued in this area as well.

Allen and Wainer (1989) noted that the accuracy of procedures that are used to compare the performance of different groups of examinees on test items obviously depends upon the correct classification of members in each examinee group. They argued that because the number of nonrespondents to questions of ethnicity is often of the same order of magnitude as the number of identified members of most minority groups, it is important to understand the effect of nonresponse on DIF results. They examined the effect of nonresponse to questions of ethnic identity on the measurement of DIF for SAT Verbal items using the MH procedure. They demonstrated that efforts to obtain more complete ethnic identifications from the examinees would lead to more accurate DIF analyses.

DIF analyses are performed on target populations. One of the requirements for inclusion in the analysis sample is that the test taker has sufficient skill in the language of the test. Sinharay (2009b) examined how an increase in the proportion of examinees who report that English is not their first language would affect DIF results if they were included in the DIF analysis sample of a large-scale assessment. The results varied by group. In some combinations of focal/reference groups, the magnitude of DIF was not appreciably affected by whether DIF was performed on examinees whose first language was not English. In other groups, first language status mattered. The results varied by type of test as well. In addition, the magnitude of DIF for some items was substantially affected by whether the DIF was performed on examinees whose first language was not English.

Dorans and Holland (1993) pointed out that in traditional one-way DIF analysis, deleting items due to DIF can have unintended consequences on the focal group. DIF analysis performed on gender and on ethnicity/race alone ignores the potential interactions between the two main effects. Additionally, Dorans and Holland suggested applying a “melting-pot” DIF method wherein the total group would function as the reference group and each gender-by-ethnic subgroup would serve sequentially as a focal group. Zhang et al. (2005) proposed a variation on the melting-pot approach called DIF dissection. They adapted the STAND methodology so that the reference group was defined to be the total group, while each of the subgroups independently acted as a focal group. They argued that using a combination of all groups as the reference group and each combination of gender and ethnicity as a focal group produces more accurate, though potentially less stable, findings than using a simple majority group approach. As they hypothesized, the deletion of a sizable DIF item had its greatest effect on the mean score of the focal group that had the most negative DIF according to the DIF dissection method. In addition, the study also found that the DIF values obtained by the DIF procedure reliably predicted changes in scaled scores after item deletion.

7.2.4 *Sample Size and Power Issues*

From its inaugural use as an operational procedure, DIF has had to grapple with sample size considerations (Zieky 1993). The conflict between performing as many DIF analyses as possible and limiting the analysis to those cases where there is sufficient power to detect DIF remains as salient as ever.

Lyu et al. (1995) developed a smoothed version of STAND, which merged kernel smoothing with the traditional STAND DIF approach, to examine DIF for student produced response (SPR) items on the SAT I Math at both the item and testlet levels. Results from the smoothed item-level DIF analysis showed that regular multiple-choice items have more variability in DIF values than SPRs.

Bayesian methods are often resorted to when small sample sizes limit the potential power of a statistical procedure. Bayesian statistical methods can incorporate, in the form of a prior distribution, existing information on the inference problem at hand, leading to improved estimation, especially for small samples for which the posterior distribution is sensitive to the choice of prior distribution. Zwick et al. (1997a, 1999) developed an empirical Bayes (EB) enhancement to MH DIF analysis in which they assumed that the MH statistics were normally distributed and that the prior distribution of underlying DIF parameters was also normal. They used the posterior distribution of DIF parameters to make inferences about the item's true DIF status and the posterior predictive distribution to predict the item's future observed status. DIF status was expressed in terms of the probabilities associated with each of the five DIF levels defined by the ETS classification system (Zieky 1993). The EB method yielded more stable DIF estimates than did conventional methods, especially in small samples. The EB approach also conveyed information about DIF stability in a more useful way by representing the state of knowledge about an item's DIF status as probabilistic.

Zwick et al. (2000) investigated a DIF flagging method based on loss functions. The approach built on their earlier research that involved the development of an EB enhancement to MH DIF analysis. The posterior distribution of DIF parameters was estimated and used to obtain the posterior expected loss for the proposed approach and for competing classification rules. Under reasonable assumptions about the relative seriousness of Type I and Type II errors, the loss-function-based DIF detection rule was found to perform better than the commonly used ETS DIF classification system, especially in small samples.

Zwick and Thayer (2002) used a simulation to investigate the applicability to computerized adaptive test data of an EB DIF analysis method developed by (Zwick et al. 1997a, 1999) and showed that the performance of the EB DIF approach to be quite promising, even in extremely small samples. When combined with a loss-function-based decision rule, the EB method is better at detecting DIF than conventional approaches, but it has a higher Type I error rate.

The EB method estimates the prior mean and variance from the current data and uses the same prior information for all the items. For most operational tests, however, a large volume of past data is available, and for any item appearing in a current test, a number of similar items are often found to have appeared in past operational administrations of the test. Conceptually, it should be possible to incorporate that past information into a prior distribution in a Bayesian DIF analysis. Sinharay (2009a) developed a full Bayesian (FB) DIF estimation method that used this type of past information. The FB Bayesian DIF analysis method was shown to be an improvement over existing methods in a simulation study.

Zwick et al. (2000) proposed a Bayesian updating (BU) method that may avert the shrinkage associated with the EB and FB approaches. Zwick et al. (2012) implemented the BU approach and compared it to the EB and FB approaches in both simulated and empirical data. They maintained that the BU approach was a natural way to accumulate all known DIF information about an item while mitigating the tendency to shrink DIF toward zero that characterized the EB and FB approaches.

Smoothing is another alternative used for dealing with small sample sizes. Yu et al. (2008) applied smoothing techniques to frequency distributions and investigated the impact of smoothed data on MH DIF detection in small samples. Eight sample-size combinations were randomly drawn from a real data set were replicated 80 times to produce stable results. Loglinear smoothing was found to provide slight-to-moderate improvements in MH DIF estimation with small samples.

Puhan, Moses, Yu, and Dorans (Puhan et al. 2007, 2009) examined the extent to which loglinear smoothing could improve the accuracy of SIBTEST DIF estimates in small samples of examinees. Examinee responses from a certification test were used. Separate DIF estimates for seven small-sample-size conditions were obtained using unsmoothed and smoothed score distributions. Results indicated that for most studied items smoothing the raw score distributions reduced random error and bias of the DIF estimates, especially in the small-sample-size conditions.

7.3 Fair Linking of Test Scores

Scores on different forms or editions of a test that are supposed to be used interchangeably should be related to each other in the same way across different subpopulations. Score equity assessment (SEA) uses subpopulation invariance of linking functions across important subpopulations to assess the degree of interchangeability of scores.

Test score equating is a statistical process that produces scores considered comparable enough across test forms to be used interchangeably. Five requirements are often regarded as basic to all test equating (Dorans and Holland 2000). One of the most basic requirements of score equating is that equating functions should be subpopulation invariant (Dorans and Holland 2000; Holland and Dorans 2006). That is, they should not be influenced by the subpopulation of examinees on which they are computed. The same construct and equal reliability requirements are prerequisites for subpopulation invariance. One way to demonstrate that two test forms are not equatable is to show that the equating functions used to link their scores are not invariant across different subpopulations of examinees. Lack of invariance in a linking function indicates that the differential difficulty of the two test forms is not consistent across different groups. The invariance can hold if the relative difficulty changes as a function of score level in the same way across subpopulations. If, however, the relative difficulty of the two test forms interacts with group membership or an interaction among score level, difficulty, and group is present, then invariance does not hold. SEA uses the subpopulation invariance of linking functions across important subgroups (e.g., gender groups and other groups, sample sizes permitting) to assess the degree of score exchangeability.

In an early study, Angoff and Cowell (1985, 1986) examined the invariance of equating scores on alternate forms of the *GRE*® quantitative test for various populations, including gender, race, major, and ability. Angoff and Cowell conducted equatings for each of the populations and compared the resulting conversions to each other and to differences that would be expected given the standard errors of equating. Differences in the equatings were found to be within that expected given sampling error. Angoff and Cowell concluded that population invariance was supported.

Dorans and Holland (2000) included several examples of linkings that are invariant (e.g., SAT Mathematics to SAT Mathematics and SAT Verbal to SAT Verbal, and SAT Mathematics to ACT Mathematics) as well as ones that are not (e.g., verbal to mathematics, and linkings between non-math ACT subscores and SAT Verbal). Equatability indexes are used to quantify the degree to which linkings are subpopulation invariant.

Since 2000, several evaluations of population invariance have been performed. Yang (2004) examined whether the linking functions that relate multiple-choice scores to composite scores based on weighted sums of multiple choice and constructed response scores for selected *Advanced Placement*® (*AP*®) exams remain invariant over subgroups by geographical region. The study focused on two questions: (a) how invariant were cut-scores across regions and (b) whether the small sample size for some regional groups presented particular problems for assessing linking invariance. In addition to using the subpopulation invariance indexes to evaluate linking functions, Yang also evaluated the invariance of the composite score thresholds for determining final AP grades. Dorans (2004) used the population sensitivity of linking functions to assess score equity for two AP exams.

Dorans et al. (2008) used population sensitivity indexes with SAT data to evaluate how consistent linear equating results were across males and females. Von Davier and Wilson (2008) examined the population invariance of IRT equating for an AP exam. Yang and Gao (2008) looked at invariance of linking computer-administered *CLEP*® data across gender groups.

SEA has also been used as a tool to evaluate score interchangeability when a test is revised (Liu and Dorans 2013). Liu et al. (2006) and Liu and Walker (2007) used SEA tools to examine the invariance of linkages across the old and new versions of the SAT using data from a major field trail conducted in 2003. This check was followed by SEA analyses conducted on operational data (see studies cited in Dorans and Liu 2009).

All these examples, as well as others such as Dorans et al. (2003), are illustrations of using SEA to assess the fairness of a test score by examining the degree to which the linkage between scores is invariant across subpopulations. In some of these illustrations, such as one form of SAT Mathematics with another form of SAT Mathematics, the expectation of score interchangeability was very high since alternate forms of this test are designed to be parallel in both content and difficulty. There are cases, however, where invariance was expected but did not hold. Cook et al. (1988), for example, found that the linking function between two biology exams depended on whether the equating was with students in a December administration, where most of the examinees were seniors who had not taken a biology course for some time, versus a June administration, where most of the examinees had just completed a biology course. This case, which has become an exemplar of lack of invariance where invariance would be expected, is discussed in detail by Cook (2007) and Peterson (2007). Invariance cannot be presumed to occur simply because tests are built to the same blueprint. The nature of the population can be critical, especially when diverse subpopulations are involved. For most testing programs, analysis that focuses on the invariance of equating functions should be conducted to confirm the fairness of the assembly process.

7.4 Limitations of Quantitative Fairness Assessment Procedures

First, not all fairness considerations can be reduced to quantitative evaluation. Because this review was limited to quantitative fairness procedures, it was limited in scope. With this important caveat in mind, this section will discuss limitations with the classes of procedures that have been examined.

Fair prediction is difficult to achieve. Differential prediction studies are difficult to complete effectively because there are so many threats to the subpopulation invariance of regression equations. Achieving subpopulation invariance of regressions is difficult because of selection effects, misspecification errors, predictor unreliability, and criterion issues. Any attempt to assess whether a prediction equation is invariant across subpopulations such as males and females must keep these confounding influences in mind.

To complicate validity assessment even more, there are as many external criteria as there are uses of a score. Each use implies a criterion against which the test's effectiveness can be assessed. The process of validation via prediction studies is an unending yet necessary task.

DIF screening is and has been possible to do. But it could be done better. Zwick (2012) reviewed the status of ETS DIF analysis procedures, focusing on three aspects: (a) the nature and stringency of the statistical rules used to flag items, (b) the minimum sample size requirements that are currently in place for DIF analysis, and (c) the efficacy of criterion refinement. Recommendations were made with respect to improved flagging rules, minimum sample size requirements, and procedures for combining data across administrations. Zwick noted that refinement of the matching criterion improves detection rates when DIF is primarily in one direction but can depress detection rates when DIF is balanced.

Most substantive DIF research studies that have tried to explain DIF have used observational data and the generation of post-hoc explanations for why items were flagged for DIF. The chapter by O'Neill and McPeek (1993) in the Holland and Wainer (1993) DIF book is a good example of this approach. As both those authors and Bond (1993) noted, this type of research with observed data is fraught with peril because of the highly selected nature of the data examined, namely items that have been flagged for DIF. In the same section of the DIF book, Schmitt et al. (1993) provided a rare exemplar on how to evaluate DIF hypotheses gleaned from observational data with experimental evaluations of the hypotheses via a carefully designed and executed experimental manipulation of item properties followed by a proper data analysis.

DIF can be criticized for several reasons. An item is an unreliable measure of the construct of interest. Performance on an item is susceptible to many influences that have little to do with the purpose of the item. An item, by itself, can be used to support a variety of speculations about DIF. It is difficult to figure out why DIF occurs. The absence of DIF is not a prerequisite for fair prediction. In addition, DIF analysis tells little about the effects of DIF on reported scores.

SEA focuses on invariance at the reported score level where inferences are made about the examinee. SEA studies based on counterbalanced single-group designs are likely to give the cleanest results about the invariance of score linking functions because it is a data collection design that allows for the computation of correlations between tests across subpopulations.

This chapter focused primarily on studies that focused on methodology and that were conducted by ETS staff members. As a result, many DIF and differential prediction studies that used these methods have been left out and need to be summarized elsewhere. As noted, qualitative and philosophical aspects of fairness have not been considered.

In addition, ETS has been the leader in conducting routine DIF analyses for over a quarter of century. This screening for DIF practice has made it difficult to find items that exhibit the high degree of DIF depicted on the cover of the Winter 2012 issue of *Educational Measurement: Issues and Practices,* an item that Dorans (2012) cited as a vintage example of DIF. Although item scores exhibit less DIF than they did before due diligence made DIF screening an operational practice, a clear need remains for continued research in fairness assessment. This includes improved methods for detecting evidence of unfairness and the use of strong data collection designs that allow researchers to arrive at a clearer understanding of sources of unfairness.

References

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231–251. https://doi.org/10.1111/j.1745-3984.1996.tb00491.x

Allen, N., & Holland, P. H. (1993). A model for missing information about the group membership of examinees in DIF studies. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 241–252). Hillsdale: Erlbaum.

Allen, N. L., & Wainer, H. (1989). *Nonresponse in declared ethnicity and the identification of differentially functioning items* (Research Report No. RR-89-47). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01331.x

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences*. Paper presented at the meeting of the American Psychological Association, Honolulu.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale: Erlbaum.

Angoff, W. H., & Cowell, W. R. (1985). *An examination of the assumption that the equating of parallel forms is population-independent* (Research Report No. RR-85-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00107.x

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327–345. https://doi.org/10.1111/j.1745-3984.1986.tb00253.x

Angoff, W. H., & Ford, S. F. (1971). *Item-race interaction on a test of scholastic aptitude* (Research Bulletin No. RB-71-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1971.tb00812.x

Angoff, W. H., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–105. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Angoff, W. H., & Sharon, A. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34*, 807–816. https://doi.org/10.1177/001316447403400408

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18*, 59–62.

Bond, L. (1993). Comments on the O'Neill & McPeek chapter. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–279). Hillsdale: Erlbaum.

Campbell, J. T. (1964). *Testing of culturally different groups* (Research Bulletin No. RB-64–34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1964.tb00506.x

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test* (Research Bulletin No. RB-64-61). Princeton: Educational Testing Service.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure* (Research Report No. RR-95-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01640.x

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–354. https://doi.org/10.1111/j.1745-3984.1996.tb00496.x

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Cleary, T. A., & Hilton, T. L. (1966). *An investigation of item bias* (Research Bulletin No. RB-66-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1966.tb00355.x

Cleary, T. A., & Hilton, T. J. (1968). An investigation of item bias. *Educational and Psychological Measurement, 5*, 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 5*, 237–255. https://doi.org/10.1177/001316446802800106

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73–88). New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-49771-6_5

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31–45. https://doi.org/10.1111/j.1745-3984.1988.tb00289.x

Donoghue, J. R., & Allen, N. L. (1993). "Thin" versus "thick" in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131–154. https://doi.org/10.2307/1165084

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale: Erlbaum.

Dorans, N. J. (1982). *Technical review of item fairness studies: 1975–1979* (Statistical Report No. SR-82-90). Princeton: Educational Testing Service.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217–233. https://doi.org/10.1207/s15324818ame0203_3

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J. (2011). Holland's advice during the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). New York: Springer-Verlag.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37. https://doi.org/10.1111/j.1745-3992.2012.00250.x

Dorans, N. L., & Cook, L. L. (Eds.). (2016). *NCME application of educational assessment and measurement: Volume 3. Fairness in educational assessment and measurement*. New York: Routledge.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Research Report No. RR-83-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1983.tb00009.x

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the mini-mental state examination: An application of the Mantel-Haenzel and standardization procedures. *Medical Care, 44*(11), S107–S114. https://doi.org/10.1097/01.mlr.0000245182.36914.4a

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (Research Report No. RR-09-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02165.x

Dorans, N. J., & Potenza, M. T. (1994). *Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning* (Research Report No. RR-94-49). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01622.x

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01414.x

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale: Erlbaum.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (Research Report No. RR-88-31). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00287.x

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309–319. https://doi.org/10.1111/j.1745-3984.1992.tb00379.x

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (Research Report No. RR-03-27, pp. 79–118). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01919.x

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81–97. https://doi.org/10.1177/0146621607311580

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Gulliksen, H. O. (1964). Intercultural studies of attitudes. In N. Frederiksen & H. O. Gulliksen (Eds.), *Contributions to mathematical psychology* (pp. 61–108). New York: Holt, Rinehart & Winston.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport: American Council on Education and Praeger.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (Research Report No. RR-86-31). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

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

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

Lewis, C. (1993). Bayesian methods for the analysis of variance. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. II. Statistical issues* (pp. 233–256). Hillsdale: Erlbaum.

Linn, R. L. (1972). *Some implications of the Griggs decision for test makers and users* (Research Memorandum No. RM-72-13). Princeton: Educational Testing Service.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161. https://doi.org/10.3102/00346543043002139

Linn, R. L. (1975). *Test bias and the prediction of grades in law school* (Report No. LSAC-75-01), Newtown: Law School Admissions Council.

Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13*, 53–58. https://doi.org/10.1111/j.1745-3984.1976.tb00181.x

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale: Erlbaum.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1–4. https://doi.org/10.1111/j.1745-3984.1971.tb00898.x

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*(1), 15–22. https://doi.org/10.1111/emip.12001

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-49771-6_7

Liu, J., Cahn, M., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linking of new SAT to old SAT across gender groups. *Journal of Educational Measurement, 43*, 113–129. https://doi.org/10.1111/j.1745-3984.2006.00008.x

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale: Erlbaum.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Lyu, C. F., Dorans, N. J., & Ramsay, J. O. (1995). *Smoothed standardization assessment of testlet level DIF on a math free-response item type* (Research Report No. RR-95-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01672.x

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700. https://doi.org/10.1080/01621459.1963.10500879

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

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (Research Report No. RR-08-43). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02129.x

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289–311. https://doi.org/10.1007/BF02294510

Myers, C. T. (1975). *Test fairness: A comment on fairness in statistical analysis* (Research Bulletin No. RB-75-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01051.x

O'Neill, K. O., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale: Erlbaum.

Pashley, P. J. (1992). *Graphical IRT-based DIF analysis* (Research Report No. RR-92-66). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01497.x

Petersen, N. S., & Novick, M. R. (1976). An evaluating of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29. https://doi.org/10.1111/j.1745-3984.1976.tb00178.x

Peterson, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-49771-6_4

Pomplun, M., Baron, P. A., & McHale, F. J. (1992). *An initial evaluation of the use of bivariate matching in DIF analyses for formula scored tests* (Research Report No. RR-92-63). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01494.x

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23–37. https://doi.org/10.1177/014662169501900104

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2007). *Small-sample DIF estimation using log-linear smoothing: A SIBTEST application* (Research Report No. RR-07-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02052.x

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2009). Using log-linear smoothing to improve small-sample DIF estimation. *Journal of Educational Measurement, 46*, 59–83. https://doi.org/10.1111/j.1745-3984.2009.01069.x

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143–152. https://doi.org/10.1111/j.1745-3984.1979.tb00095.x

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale: Erlbaum.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194. https://doi.org/10.1007/BF02294572

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009a). Using past data to enhance small-sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*, 74–96. https://doi.org/10.3102/1076998607309021

Sinharay, S., Dorans, N. J., & Liang, L. (2009b). *First language of examinees and its relationship to differential item functioning* (Research Report No. RR-09-11). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02162.x

Stricker, L. J. (1982). Identifying test items that perform differently in population subgroups: A partial correlation index. *Applied Psychological Measurement, 6*, 261–273. https://doi.org/10.1177/014662168200600302

Tan, X., Xiang, B., Dorans, N. J., & Qu, Y. (2010). *The value of the studied item in the matching criterion in differential item functioning (DIF) analysis* (Research Report No. RR-10-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02220.x

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale: Erlbaum.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8*, 63–70. https://doi.org/10.1111/j.1745-3984.1971.tb00907.x

Tucker, L. R. (1951). *Academic ability test* (Research Memorandum No. RM-51-17). Princeton: Educational Testing Service.

Turnbull, W. W. (1949). Influence of cultural background on predictive test scores. In *Proceedings of the ETS invitational conference on testing problems* (pp. 29–34). Princeton: Educational Testing Service.

Turnbull, W. W. (1951a). *Socio-economic status and predictive test scores* (Research Memorandum No. RM-51-09). Princeton: Educational Testing Service.

Turnbull, W. W. (1951b). Socio-economic status and predictive test scores. *Canadian Journal of Psychology, 5*, 145–149. https://doi.org/10.1037/h0083546

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11–26. https://doi.org/10.1177/0146621607311560

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197–219. https://doi.org/10.1111/j.1745-3984.1991.tb00354.x

Yang, W.-L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33–41. https://doi.org/10.1111/j.1745-3984.2004.tb01157.x

Yang, W.-L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination Program examination. *Applied Psychological Measurement, 32*, 45–61. https://doi.org/10.1177/0146621607311577

Yu, L., Moses, T., Puhan, G., & Dorans, N. J. (2008). *DIF detection with small samples: Applying smoothing techniques to frequency distributions in the Mantel-Haenszel procedure* (Research Report No. RR-08-44. Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02130.x

Zhang, Y., Dorans, N. J., & Mathews-Lopez, J. (2005). *Using DIF dissection method to assess effects of item deletion* (Research Report No. 2005-10). New York: The College Board. https://doi.org/10.1002/j.2333-8504.2005.tb02000.x

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale: Erlbaum.

Zieky, M. (2011). The origins of procedures for using differential item functioning statistics at Educational Testing Service. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 115–127). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4419-9389-2_7

Zwick, R. (1989). *When do item response function and Mantel-Haenszel definitions of differential item functioning coincide* (Research Report No. RR-89-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00146.x

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185–197. https://doi.org/10.2307/1165031

Zwick, R. (1994). *The effect of the probability of correct response on the variability of measures of differential item functioning* (Research Report No. RR-94-44). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01617.x

Zwick, R. (2012*). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55–66. https://doi.org/10.1111/j.1745-3984.1989.tb00318.x

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*, 187–201. https://doi.org/10.3102/10769986021003187

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*, 57–76. https://doi.org/10.1177/0146621602026001004

Zwick, R., Donoghue, J. R., & Grima, A. (1993a). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251. https://doi.org/10.1111/j.1745-3984.1993.tb00425.x

Zwick, R., Thayer, D. T., & Wingersky, M. (1993b). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests* (Research Report No. RR-93-11). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01522.x

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121–140. https://doi.org/10.1177/014662169401800203

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341–363. https://doi.org/10.1177/014662169401800203

Zwick, R., Thayer D. T., & Lewis, C. (1997a). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (Research Report No. RR-97-21). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01742.x

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997b). Descriptive and inferential procedures for assessing DIF in polytomous items. *Applied Measurement in Education, 10*, 321–344. https://doi.org/10.1207/s15324818ame1004_2

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*, 225–247. https://doi.org/10.3102/10769986025002225

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel–Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics, 37*, 601–629. https://doi.org/10.3102/1076998611431085