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,
$\overline{x}_{ik}$ 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

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

Alternative item difficulty indices that use a transformation based on the inverse of the cumulative distribution function (CDF) of the normal distribution for the $\overline{x}_i$ 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):

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

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):

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

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),

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

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),

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

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 kth 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, $\bar{x}_{ik}$, 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, $\bar{x}_{ik} = 0.5$. Results of these steps for a subsample of the items were presented and analyzed in terms of plotted $\bar{x}_{ik}$ 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 $\Delta_{ik} = 0 - (1)\Phi^{-1}(\bar{x}_{ik})$, where the i subscript references the item and the k subscript references the age group responding to the item.

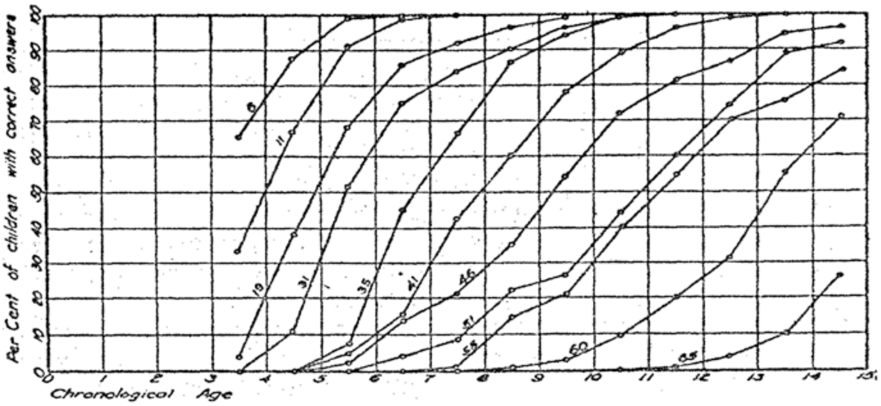


FIG. 5.

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, $\overline{x}_{i7}$ and $\overline{x}_{i8}$, 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,

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

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 $\overline{x}_{ik}$ 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 $\overline{x}_{i3.5}$ 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 $\overline{x}_{i3.5}$ 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

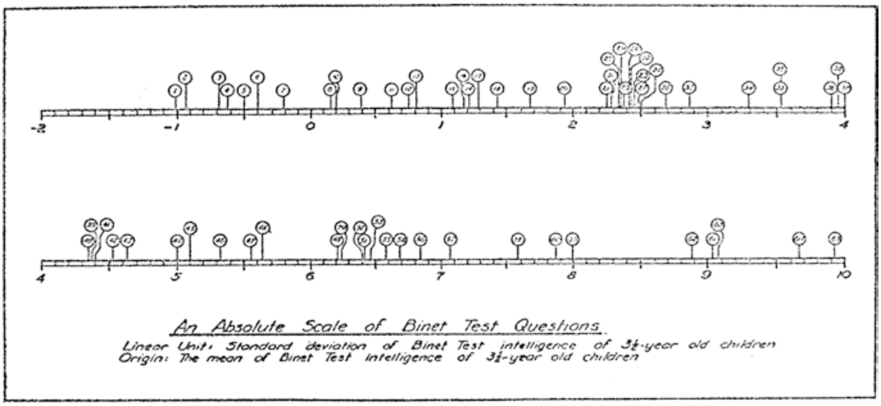


Fig. 6.

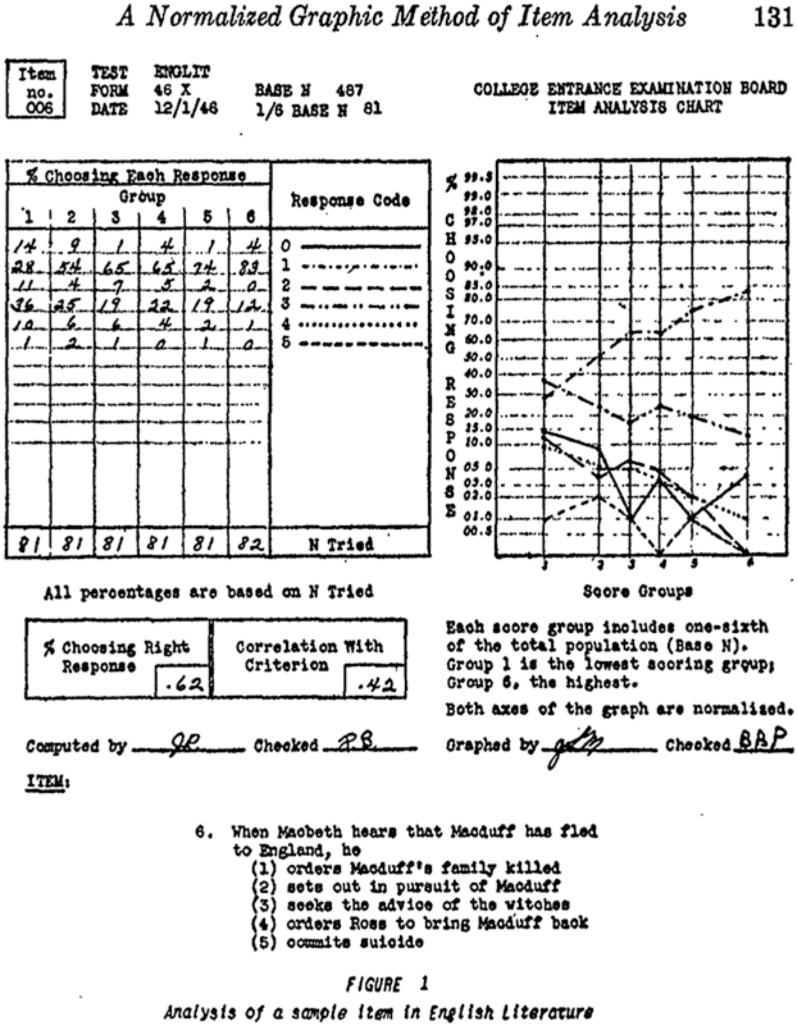
Fig. 2.2 Thurstone's (1925) Figure 6, which represents Binet–Simon test items' average difficulty on an absolute scale

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).


% CHOOSING RESPONSE
99.5
99.0
98.0
97.0
95.0
90.0
85.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
15.0
10.0
05.0
03.0
02.0
01.0
00.5
1
2
3
4
5
6
Score Groups

A Normalized Graphic Method of Item Analysis 131

Item no. 006	TEST	ENGLIT		
	FORM	46 X	BASE N	487
	DATE	12/1/46	1/6 BASE N	81

COLLEGE ENTRANCE EXAMINATION BOARD
ITEM ANALYSIS CHART

% Choosing Each Response						
Group						Response Code
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	Correlation With Criterion
.62	.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:

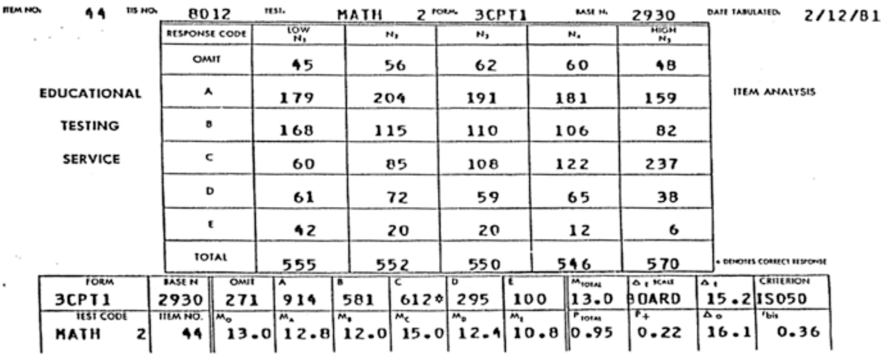
6. When Macbeth hears that Macduff has fled to England, he
   (1) orders Macduff's family killed
   (2) sets out in pursuit of Macduff
   (3) seeks the advice of the witches
   (4) orders Ross to bring Macduff back
   (5) commits suicide

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

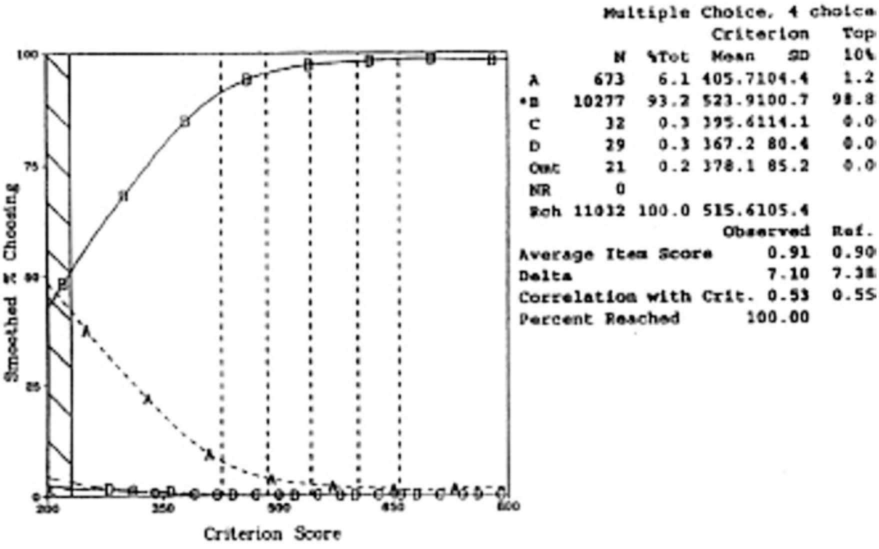
Exhibit 1. Standard ETS Item Analysis Information Strip.

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).

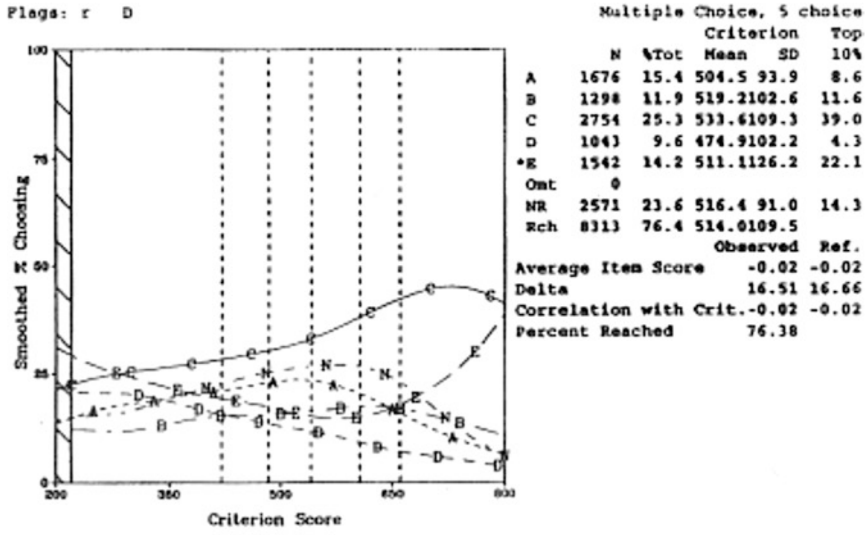
The results of the easy item shown in Fig. 2.5 are distinguished from those of the more difficult items in Figs. 2.6 and 2.7 in that the percentages of examinees choosing the correct option in Fig. 2.5 is 50% or greater for all examinees, and the percentages monotonically increase with the total test 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.**

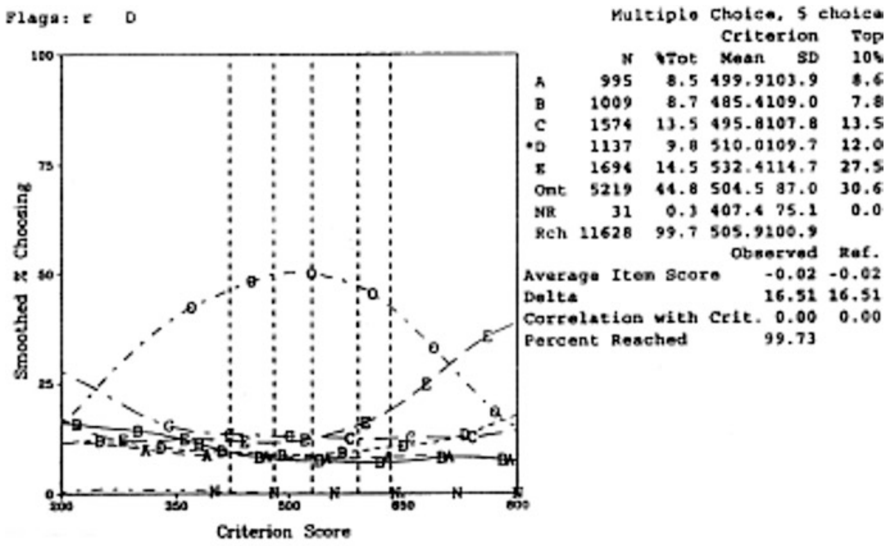
Fig. 2.5 Livingston and Dorans's (2004) Figure 1, which demonstrates classical item analysis results currently used at ETS, for a relatively easy item



```
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