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, $\hat{r}el\left(X\right)$. The conditional standard error estimate based on Keats' correction,

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

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 coeffi-

cient 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.

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

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 ($\boldsymbol{\lambda\Psi\lambda}'$, where Ψ 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, $\mathbf{v}$,

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

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 $\mathbf{v}$ 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 $\mathbf{v}$ are not exactly error variables, but instead are the sum of the error factors and specific factors of the q variables. That is, the $\mathbf{v}$ factors are understood to reflect unreliability (error factors) as well as actual measurement differences (specific factors). The assumption that the $\mathbf{v}$ 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 $\mathbf{v}$.

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 equiv-

alent (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,

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

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 inac-

curate 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*,

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

where $\hat{\beta}_{T_X}$ 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*,

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

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,

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

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

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

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

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,

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

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,

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

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/10.1177/001316445401400401

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