Listening. Learning. Leading.®

ETS R&D Scientific and Policy Contributions Series
ETS SPC-13-03

ETS Psychometric Contributions: Focus on Test Scores

Tim Moses

September 2013

ETS R&D Scientific and Policy Contributions Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS R&D Scientific and Policy Contributions series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS R&D Scientific and Policy Contributions series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of ETS.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Psychometric Contributions: Focus on Test Scores

Tim Moses

ETS, Princeton, New Jersey

ETS Research Report No. RR-13-15

September 2013

Action Editor: James Carlson

Reviewers: Randy Bennett and Neil Dorans

Abstract

The purpose of this report is to review ETS psychometric contributions that focus on test scores. Two major sections review contributions based on assessing test scores' measurement characteristics and other contributions about using test scores as predictors in correlational and regression relationships. An additional section reviews additional contributions that integrate test scores' measurement aspects with correlational and regression uses.

Key words: test scores, measurement, correlation, regression

Foreword

Since its founding in 1947, ETS has conducted a significant and wide-ranging research program that has focused on, among other things, psychometric and statistical methodology; educational evaluation; performance assessment and scoring; large-scale assessment and evaluation; cognitive, developmental, personality, and social psychology; and education policy. This broad-based research program has helped build the science and practice of educational measurement, as well as inform policy debates.

In 2010, we began to synthesize these scientific and policy contributions, with the intention to release a series of reports sequentially over the course of the next few years. These reports constitute the *ETS R&D Scientific and Policy Contributions Series*.

In the sixth report in the series, Tim Moses examines the contributions to the psychometric study of test scores made by Educational Testing in the past 60 years since the founding of the organization. These contributions can be roughly organized into three areas. The first area, the measurement properties of tests and developments of classical test theory, covers ETS contributions that provide culminating and progressively more rigorous formalizations of classical test theory. These contributions and others by ETS researchers in this area have led to improvements in the psychometric (measurement) quality of tests. The second area is the use of tests as predictors in correlational and regression relationships. ETS contributions in this area have helped to define appropriate and valid uses for tests. The third area is the integration and application of measurement theories and correlational and regression analyses to address test-score issues, which is relevant to operational work on testing programs.

This report complements two other upcoming reports on psychometrics in the ETS R&D Scientific and Policy Contributions series, one of which by Neil Dorans will focus on fairness and another by James Carlson that will focus on items.

Moses is a senior psychometrician in the Research & Development division (R&D) at Educational Testing Service (ETS) in Princeton, NJ. He joined the organization in 2003 as an associate measurement statistician. The author of more than 40 research publications, Moses research interests include loglinear smoothing models, measurement and prediction error, and observed-score equating.

Future reports in the *ETS R&D Scientific and Policy Contributions Series* will focus on other major areas of research and education policy in which ETS has played a role.

Ida Lawrence
Senior Vice-President
Research & Development Division
ETS

Table of Contents

Page

This report provides 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 two sections: Test Scores as Measurements, and 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 the section titled 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.

Test Scores as Measurements

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 classical test theory, such as conceptions of test score averages and errors, were borrowed from 19th century astronomers and were probably even informed by Galileo’s work in the 17th 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'*, $E_{X'p}$,

$$X'_p = T_{Xp} + E_{X'p}. \qquad (1)$$

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

$$\mu(X'_p) = T_{Xp} \text{ and } \mu(E_{X'p}) = 0, \qquad (2)$$

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

$$\sigma^2_{Xp} = \sigma^2_{E(Xp)} \,|\, T_{Xp}. \qquad (3)$$

The true score variance, error variance, and observed score variance of an examinee group reflect a summative relationship,

$$\sigma^2_X = \sigma^2_{T(X)} + \sigma^2_{E(X)}, \qquad (4)$$

where the covariance of the true scores and errors, $\sigma_{T(X),E(X)}$, 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 and to summarizing the measurement precision (i.e., reliability) of a test score as a ratio of those variances,

$$\text{rel}(X) = \frac{\sigma^2_{T(X)}}{\sigma^2_X} = 1 - \frac{\sigma^2_{E(X)}}{\sigma^2_X}. \qquad (5)$$

Reliability is often assumed to refer to the 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.

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 the Test Scores as Measurements section 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 **Harold Gulliksen's**[1] (1950) *Theory of Mental Tests*, **Frederic Lord** and **Melvin 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 & 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.

ETS Contributions About $\sigma_{E(X)} | T_{Xp}$

The finding that $\sigma_{E(X)}$ (i.e., the standard error of measurement) may not be constant 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 $\sigma^2_{E(X)} | T_{Xp}$ 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 & 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 the subsection in this report titled Alternative Classical Test Theory Models), 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, **William G. 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, 1955, 1956b). 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,

$$T_{Xp}\left(1-\frac{T_{Xp}}{n}\right). \tag{6}$$

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,

$$\sqrt{\frac{X_p\left(n-X_p\right)}{n-1}}. \tag{7}$$

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 (1955, 1956b) 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 **John A. 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, $\sigma^2_{Xp}[1-\mathrm{rel}_{21}(X)]$ (described in the subsection titled Reliability Estimation in this report), Keats proposed a correction that utilizes the Kuder-Richardson Formula 21 reliability, $\mathrm{rel}_{21}(X)$, and any other reliability estimate of interest, $\widehat{\mathrm{rel}}(X)$. The conditional standard error estimate based on Keats' correction,

$$\sqrt{\frac{X_p\left(n-X_p\right)\left[1-\widehat{\mathrm{rel}}(X)\right]}{(n-1)\left[1-\mathrm{rel}_{21}(X)\right]}}, \tag{8}$$

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

Intervals for True Score Inference

One application of interest of standard errors of measurement in the ETS Contributions About $\sigma_{E(X)} \mid T_{Xp}$ subsection 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, 1965b) and the four-parameter beta model (Keats & Lord, 1962). The beta-binomial model builds on the binomial error model described in the ETS Contributions About $\sigma_{E(X)} \mid T_{Xp}$ section. 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, $0 \le a < b \le 1$. Estimation for two-parameter and four-parameter beta-binomial models can be accomplished by the method of moments (Hanson, 1991; Keats & Lord, 1962; Lord & Novick, 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

& Stocking, 1976), but also providing regression-based estimates of true scores (Lord & Novick, 1968), and providing estimates of consistency and accuracy when examinees are classified at specific scores on a test (Livingston & Lewis, 1995).

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

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 & 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 & Novick, p. 50). Finally, Haertel credits **Karl G. Jöreskog** (1971) 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, **William H. Angoff** (1953) and **Walter Kristof** (1971a) were clearly aware of this model when developing their reliability estimates summarized below.

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 the Foundational Developments for Use of Test Scores as Measurements, Pre-ETS subsection) 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 from the Alternative Classical Test Theory Models subsection 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),

$$\frac{2\rho_{X1,X2}}{1+\rho_{X1,X2}}, \tag{9}$$

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, . . . , n parts assumed to be parallel,

$$\frac{n}{n-1}\left(\frac{\sigma_X^2-\sum_i \sigma_{X,i}^2}{\sigma_X^2}\right)=\frac{n}{n-1}\left(1-\frac{\sum_i \sigma_{X,i}^2}{\sigma_X^2}\right). \tag{10}$$

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 & Richardson, 1937) in terms of the proportion of correct responses on the ith part, $\mu(X_i)$,

$$\frac{n}{n-1}\left(1-\frac{\sum_i \mu(X_i)\left[1-\mu(X_i)\right]}{\sigma_X^2}\right). \tag{11}$$

The Kuder-Richardson Formula 21 (rel$_{21}$ (*X*)) from Equation 8 in the Overview of ETS Contributions subsection) can be obtained as a simplification of Equation 11, by replacing each $\mu(X_i)$ for the dichotomously scored items with the mean score on all the items, $\mu(X)$, resulting in

$$\frac{n}{n-1}\left(1-\frac{\mu(X)\left[n-\mu(X)\right]}{n\sigma_X^2}\right). \tag{12}$$

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, 1959a), 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, 1956c). Lord (1955) 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 Equation 5 to produce the Kuder-Richardson Formula 21 reliability estimate (i.e., the squared values in Equation 7 can be averaged over examinees to estimate $\sigma_{E(X)}^2$. Other ETS researchers have pointed out that if the part tests are not essentially tau equivalent, then coefficient alpha is a lower bound to the internal consistency reliability (Novick & 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 (1963; 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 the Overview of ETS Contributions section 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, Gleser, Nanda, & Rajaratnam, 1972). A description of the discussion of relative and absolute error variance and of applications of Lord's (1955, 1956b) binomial error model results (see the Overview of ETS Contributions subsection) 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 Angoff-Feldt reliability coefficient,

$$\frac{4\sigma(X_1, X_2)}{\sigma_X^2 - \frac{\left[\sigma_{X,1}^2 - \sigma_{X,2}^2\right]^2}{\sigma_X^2}}. \tag{13}$$

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

$$\text{rel}(X) = \frac{\rho_{X,X1}\sigma_X - \sigma_{X1}}{\rho_{X,X1}\left(\sigma_X - \rho_{X,X1}\sigma_{X1}\right)} \text{ and } \text{rel}(X_1) = \frac{\rho_{X,X1}\left(\rho_{X,X1}\sigma_X - \sigma_{X1}\right)}{\sigma_X - \rho_{X,X1}\sigma_{X1}}, \tag{14}$$

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

$$\text{rel}(X) = \frac{\rho_{X,Y}(\sigma_X + \rho_{X,Y}\sigma_Y)}{\sigma_Y + \rho_{X,Y}\sigma_X} \text{ and } \text{rel}(Y) = \frac{\rho_{X,Y}(\sigma_Y + \rho_{X,Y}\sigma_X)}{\sigma_X + \rho_{X,Y}\sigma_Y}. \tag{15}$$

The assumptions and approach of Angoff and Feldt were used by Horst (1951) 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 the Alternative Classical Test Theory Models subsection) 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, 1963). By formulating test length as a parameter in a model for a population covariance matrix of two or more tests, Kristof (1971a) 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.

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

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\lambda\Theta} + \boldsymbol{\nu}\,, \qquad (16)$$

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, $\mathbf{v}$,

$$\boldsymbol{\Sigma} = \boldsymbol{\lambda\Psi\lambda}' + \boldsymbol{D}^2\,. \qquad (17)$$

The overall goal of factor analyses described in Equations 16 and 17 is to meaningfully explain the relationships among multiple test forms and other variables with a small number of common factors (i.e., $k < q$ is often intended to mean, "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 & MacCallum, 2007; Harman, 1967; Lord & Novick, 1968). Factor analysis models are multivariate variations of the classical test theory model described in the Test Scores as Measurements section in this report. In this sense, factor analysis informs a "psychometric school" of inquiry, which views a ". . . battery of tests as a selection from a large domain of tests that could be developed for the same psychological phenomenon and focused on the factors in this domain" (Jöreskog, 2007, p. 47). Similar to the classical test theory assumptions, the means of **v** are assumed to be zero, and the variables' covariance matrix, $\boldsymbol{D}^2$, is diagonal, meaning that the unique factors are assumed to be uncorrelated. Somewhat different from the classical test theory model, the unique factors in **v** are not exactly error variables, but instead are the sum of the error factors and specific factors of the q variables. That is, the **v** factors are understood to reflect unreliability (error factors) as well as actual measurement differences (specific factors). The assumption that the **v** factors are uncorrelated implies that the observed covariances between the observed variables are attributable to common factors and loadings, $\boldsymbol{\lambda\Theta}$. The common factors are also somewhat different from the true scores of the variables because the factor-analysis model implies that the true scores reflect common factors as well as specific factors in **v**.

Many developments in factor analysis are attempts to formulate its inherently subjective nature into mathematical, statistical, and computational solutions. ETS researchers have contributed several solutions pertaining to these interests, which are reviewed in **Harry H. 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 **Micheal W. Browne** (1967) and Jöreskog (Jöreskog, 1965, 1966, 1969; Jöreskog & Lawley, 1967), 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 **Charles Pinzka**, **David R. Saunders**, and **Robert I. Jennrich** include the development of different rotation methods that either allow the common factors to be correlated (oblique) or force the factors to be orthogonal (Browne, 1966, 1972a, 1972b; Green, 1951b; Pinzka & 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, Koopman, & Linn, 1969). Methods for estimating statistical standard errors of estimated factor loadings have been derived (Jennrich, 1973; Jennrich & 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 & 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, 1967; Jöreskog & Lawley, 1967). 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, 1970) and evaluating test scores with respect to psychometric models (Jöreskog, 1969). 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, 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 & van Thillo, 1972).

Applications to Psychometric Test Building 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 (described in Gulliksen, 1950; Mislevy, 1993; Traub, 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, 1956b) and also in empirical verifications (Lord, 1958b). 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, 1957b) 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 & Lord, 1955).

An important limitation of many reliability estimation methods is that they pertain to overall or average score precision. **Samuel Livingston** and **Charles 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 the Alternative Classical Test Theory Models subsection). 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 (1968) 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, 1956a) and as F-ratio tests (Kristof, 1971b). Another F test was developed to assess if two tests differ only with respect to measurement errors, units, and origins of measurement (Lord, 1971). A likelihood ratio test was derived for comparing two or more coefficient alpha estimates obtained from dividing two tests each into two part tests with equivalent error variances using a single sample of examinees (Kristof, 1964). Different maximum likelihood and chi-square procedures have been developed for assessing whether tests have equivalent overall standard errors of measurement, assuming these tests are parallel (Green, 1950), or that they are essentially tau equivalent (Kristof, 1962). 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, 1969).

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. The subsection titled Foundational Developments for Use of Test Scores as Predictors, Pre-ETS reviews the basic correlational and regression developments established prior to ETS. The subsection titled ETS Contributions to the Methodology of Correlations and Regressions and Their Application to the Study of Test Scores as Predictors reviews ETS psychometric contributions involving correlation and regression analyses.

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 19th 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*,

$$Y = \mu(Y \mid X) + \varepsilon = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \varepsilon \ . \tag{18}$$

The prediction error, ε, in Equation 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, (the Test Scores as Measurements section in this report). The linear regression function in Equation 18 is based on least-squares estimation methods because using these methods results in the smallest possible value of $\sigma_\varepsilon^2 = \sigma_Y^2\left[1-\rho_{X,Y}^2\right]$. The multivariate version of Equation 18 is based on predicting the conditional mean of *Y* from a combination of a set of *q* observable predictor variables,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \widehat{\boldsymbol{Y}} + \boldsymbol{\varepsilon}\,, \tag{19}$$

Where $\boldsymbol{Y}$ is an *N*-by-1 column vector of the *N Y* values in the data, $\widehat{\boldsymbol{Y}} = \mathbf{X}\boldsymbol{\beta}$ is an *N*-by-1 column vector of predicted values ($\widehat{Y}$),***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 *X*s 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 *X*s in Equations 18 and 19 can be computed given the $\boldsymbol{\beta}$ parameters (or estimated using estimated parameters, $\widehat{\boldsymbol{\beta}}$) as,

$$\rho_{\widehat{Y},Y}^2 = \frac{\sum_{i=1}^{N}\left(\mathbf{X}_i\boldsymbol{\beta}\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{X}_i\boldsymbol{\beta}\right)^2}{\mathbf{Y}^{\mathbf{t}}\mathbf{Y} - \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{Y}_i\right)^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\,. \tag{20}$$

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).

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.

Relationships of tests in a population's subsamples with partially missing data. Some contributions by ETS scientists, such as Guilliksen, Lord, **Donald Rubin**, **Dorothy Thayer**, **Paul Horst**, and **Tim 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, 1954a, 1954b). 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

$$\mu_{X,S} = \mu_{X,R} - \rho_{X,A,R}\frac{\sigma_{X,R}}{\sigma_{A,R}}(\mu_{A,R} - \mu_{A,S}), \tag{21}$$

and

$$\sigma_{X,S}^{2}=\sigma_{X,R}^{2}-\left[\rho_{X,A,R}\frac{\sigma_{X,R}}{\sigma_{A,R}}\right]^{2}\left[\sigma_{A,R}^{2}-\sigma_{A,S}^{2}\right]. \qquad (22)$$

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 & 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, 1950b). 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 are often regarded as optimal methods for addressing missing 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, 1954a, 1954b; Rubin & 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, Laird, & Rubin, 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, Deng, & Zhang, 2010). 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).

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 the Relationships of Tests in a Population's Subsamples With Partially Missing Data subsection). The use of correlation and regression concepts 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 **Robert Linn**, **Charles Werts**, **Nancy Wright**, **Neil Dorans**, **Paul Holland**, **Paul R. Rosenbaum**, and **Edward F. O'Connor**.

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

Concerns have been raised about the adequacy of Gulliksen's (1950) corrections for the statistics of self-selected groups. In particular, the corrections may be inaccurate if the assumed regression model is incorrect (i.e., is actually nonlinear or if the error variance, $\sigma^2(\varepsilon)$, is not constant), or if the corrections are based on a purported selection variable that is not the actual variable used to select the groups (Linn, 1967; Lord & 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 & 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 & Werts, 1971a; Wright & 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 ($\rho_{X,Y,R+S}\frac{\sigma_{Y,R+S}}{\sigma_{X,R+S}}$) 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,

$$\mu_{Y,R}-\mu_{Y,S}-\rho_{X,Y,R+S}\frac{\sigma_{Y,R+S}}{\sigma_{X,R+S}}\left(\mu_{X,R}-\mu_{X,S}\right). \tag{23}$$

The application of the covariance analyses of Equation 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, $\mu_{Y,R}-\mu_{Y,S}-\left(\mu_{X,R}-\mu_{X,S}\right)$, (Lord, 1965a). 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, 1968). 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 (1971b) 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, 1978a), a combination of Mahalanobis distances and regression (Rubin, 1978b), and propensity-score matching methods (Rosenbaum & Rubin, 1984, 1985).

Detecting group differences in test and criterion regressions. Some ETS scientists such as **Douglas Schultz**, **Samuel Wilks**, **T. Anne Cleary**, **Norman Frederiksen**, and **S. Donald Melville** have developed and applied statistical methods for comparing the regression functions of group. 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 & 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 & Wilks, 1950). Significance procedures have also been described to consider how the correlation from the estimated regression model in Equation 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),

$$\begin{bmatrix} Y_1 \\ Y_1 \\ . \\ . \\ Y_N \end{bmatrix} = \begin{bmatrix} 1_1 & X_1 & G_1 & X_1G_1 \\ 1_2 & X_2 & G_2 & X_2G_2 \\ . & & & \\ . & & & \\ 1_N & X_N & G_N & X_NG_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_X \\ \beta_G \\ \beta_{XG} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_1 \\ . \\ . \\ e_N \end{bmatrix}. \tag{24}$$

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, 1963) and iterative, exploratory procedures for allowing the regression weights of individuals to emerge in ways that maximize prediction accuracy (Cleary, 1965).

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 & Melville, 1953). 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 is biased for that subgroup (Cleary, 1966). The statistical techniques in this section, such as moderated multiple regression (Saunders, 1953b) for assessing differential predictability and Cleary's test bias,[2] help to define appropriate and valid uses for tests.

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, **Paul Horst**, **Bert F. Green, Jr**., and **Lynnette 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 mathematically 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, 1951a). 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 answer-only 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, 1962b).

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 each test in ways that maximize the multiple correlation of the battery with an external criterion (Equation 20). These developments have origins in Horst (1950a), but have been extended to a more general and sophisticated solution by Woodbury and Novick (1967). 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 & Novick, 1969). 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 & Thayer, 1969).

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 the Test Scores as Measurements section and the correlational and regression developments in the Test Scores as Predictors section. As previously stated, describing measurement and correlational concepts as if they are completely independent is an oversimplification. Some of the reliability estimates in the Test Scores as Measurements section explicitly incorporate test correlations. In the Test Scores as Predictors section, 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 & Novick, 1969; Novick & Thayer, 1969; Woodbury & Novick, 1967). The results by Novick and his colleagues are consistent with other results that have shown that observed-score regressions such as Equation 18 can serve as optimal predictors of the true scores of a criterion (Holland & 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,

$$\frac{\rho_{X,Y}}{\sqrt{\text{rel}(X)\text{rel}(Y)}}\,. \tag{25}$$

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

$$\widehat{T}_{Xp} = \text{rel}(X)X_p - \left[1-\text{rel}(X)\right]\mu(X_P)\,. \tag{26}$$

Equations 25 and 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 & Werts, 1971b) and differential prediction studies that evaluate *X's* bias (Linn & Werts, 1971a). Lord (1960) developed an approach for addressing unreliable *X* scores in covariance analyses. In Lord's formulations, the standard covariance analysis model described in Equation 23 is altered to produce an estimate of the covariance results that might be obtained based on a perfectly reliable *X*,

$$\mu_{Y,R} - \mu_{Y,S} - \widehat{\beta}_{T(X)}\left(\mu_{X,R} - \mu_{X,S}\right), \tag{27}$$

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

$$\widehat{\beta}_{T(X)} = \frac{N_R\sigma_{X,Y,R} + N_S\sigma_{X,Y,S}}{N_R\text{rel}_R(X)\sigma^2_{X,R} + N_S\text{rel}_S(X)\sigma^2_{X,S}}\left[1-\frac{k(k-w)}{(N_R+N_S)w^2}\right], \tag{28}$$

where $k = \frac{N_R \sigma^2_{X,R} + N_S \sigma^2_{X,S}}{N_R + N_S}$, $v = \frac{N_R \sigma_{X,Y,R} + N_S \sigma_{X,Y,S}}{N_R + N_S}$,

$w = \frac{N_R \mathrm{rel}_R(X) \sigma^2_{X,R} + N_S \mathrm{rel}_S(X) \sigma^2_{X,S}}{N_R + N_S}$, and the bracketed term in Equation 28 is a correction for sampling bias. Large sample procedures are used to obtain a sample estimate of the slope in Equation 28 and produce a statistical significance procedure for evaluating Equation 27.

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

$$\widehat{T}_Y - \widehat{T}_X = \mu(Y) + \widehat{\beta}_{Y|X}\left[Y - \mu(Y)\right] - \mu(X) - \widehat{\beta}_{X|Y}\left[X - \mu(X)\right], \qquad (29)$$

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

$$\widehat{\beta}_{Y|X} = \frac{\mathrm{rel}(Y) - \rho^2_{X,Y} - \left[1 - \mathrm{rel}(X)\right]\rho_{X,Y}\sigma_X / \sigma_Y}{1 - \rho^2_{X,Y}}, \qquad (30)$$

$$\widehat{\beta}_{X|Y} = \frac{\mathrm{rel}(X) - \rho^2_{X,Y} - \left[1 - \mathrm{rel}(Y)\right]\rho_{X,Y}\sigma_Y / \sigma_X}{1 - \rho^2_{X,Y}}. \qquad (31)$$

Lord also showed that the reliability of the observed change can be estimated from

$$\mathrm{rel}(Y - X) = \frac{\mathrm{rel}(Y)\sigma^2_Y + \mathrm{rel}(X)\sigma^2_X - 2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma^2_Y + \sigma^2_X - 2\rho_{X,Y}\sigma_X\sigma_Y}. \qquad (32)$$

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 Equation 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 Equation 26 of the subscore's true score from its observed score,

$$\mathrm{rel}_{X}\sigma_{X}^{2}\left[1-\mathrm{rel}_{X}\right], \tag{33}$$

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

$$\mathrm{rel}_{X}\sigma_{X}^{2}\left[1-\rho_{T(X),Y}^{2}\right]. \tag{34}$$

The prediction error variance for the regression of the true scores of *X* on both *X* and *Y* is obtained in extensions of Equations 33 and 34,

$$\mathrm{rel}_{X}\sigma_{X}^{2}\left[1-\mathrm{rel}_{X}\right]\left[1-\rho_{Y,T(X).X}^{2}\right], \tag{35}$$

where $\rho_{Y,T(X).X}$ is the partial correlation of the true score of *Y* and *X* given the observed score of *X*. Estimates of the correlations in Equations 34 and 35 are obtained somewhat like the disattenuated correlation in Equation 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 Equations 33, 34, and 35 produce an indication for when the observed subscore has value for reporting (i.e., when Equation 33 is less than Equations 34 and 35, such as when the subscore has high reliability and a moderate correlation with the total test score). Comparisons of Equations 33 to 35 can also suggest when the total test score is a more accurate reflection of the true subscore (i.e., when Equation 34 is less than Equation 33, such as when the subscore has low reliability and/or a high correlation with the total test score). Haberman's (2008) applications to applied testing data 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 (Equation 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, 1957a). 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 & Lord, 1962; Lord, 1965b). 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 & Lewis,1995), smoothing observed test score distributions (Hanson & Brennan, 1990; Kolen & Brennan, 2004), producing interval estimates for true scores (Lord & Novick, 1968), predicting test norms (Lord, 1961), and predicting the bivariate distribution of two tests assumed to be parallel (Lord & Novick, 1968).

Discussion

The purpose of this report 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 & Brennan, 2004), and employ adaptations of Cleary's (1966) test bias[2] approach to evaluate the invariance of test equating functions (Dorans & 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 report.

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 **Frances 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, 1957), and through mathematical functions derived to describe the influence of items with given difficulty levels on the moments of test-score distributions (Lord, 1959b; Lord & 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 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 & Wingersky, 1984) and showing how classical test theory results can be directly obtained from some IRT models (Holland & 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 contributions to DIF and to IRT are just two of several additional areas of psychometrics summarized in other reports in the ETS R&D Scientific and Policy Contributions series (Carlson, 2013; Dorans, in press).

.

References[3]

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*, 1–14.

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. (Reprinted as W. H. Angoff, *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service, 1984).

- Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14–20.

Bridgeman, B., Pollack, J. & Burton, N. (2008). *Predicting grades in different types of college courses* (Research Report No. RR-08-06). Princeton, NJ: Educational Testing Service.

- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Browne, M. W. (1966). *On oblique procrustes rotation* (Research Bulletin No. RB-66-08). Princeton, NJ: Educational Testing Service.

Browne, M. W. (1967). *Fitting the factor analysis model* (Research Bulletin No. RB-67-02). Princeton, NJ: Educational Testing Service.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33,* 267–334.

Browne, M. W. (1972a). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25,* 207–212.

Browne, M. W. (1972b). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25,* 115–120.

Carlson, J. (2013). *Psychometrics: Items* (ETS R&D Scientific and Policy Contributions Series). Manuscript in preparation.

Cleary, T. A. (1965). *An individual differences model for multiple regression* (Research Bulletin No. RB-65-03). Princeton, NJ: Educational Testing Service.

Cleary, T. A. (1966). *Test bias: Validity of the Scholastic Aptitude Test for negro and white students in integrated colleges* (Research Bulletin No. RB-66-31). Princeton, NJ: Educational Testing Service.

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

• 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, NY: Wiley.

• Cudeck, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions.* Mahwah, NJ: 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. (in press). *ETS contributions to the quantitative assessment of item, test, and score fairness* (ETS R&D Scientific and Policy Contributions Series). Princeton, NJ: Educational Testing Service.

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, NJ: Erlbaum.

• Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370.

• Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika, 40,* 557–561.

• 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.

Frederiksen, N., & Melville, S. D. (1953). *Differential predictability in the use of test scores* (Research Bulletin No. RB-53-17). Princeton, NJ: Educational Testing Service.

Green, B. F. (1950). *A test of the equality of standard errors of measurement* (Research Bulletin No. RB-50-25). Princeton, NJ: Educational Testing Service.

Green, B. F. (1951a). *A note on item selection for maximum validity* (Research Bulletin No. RB-51-37). Princeton, NJ: Educational Testing Service.

Green, B. F. (1951b). *The orthogonal approximation of an oblique structure in factor analysis* (Research Bulletin No. RB-51-18). Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33,* 204–229.

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

Harman, H. H. (1967). *Modern factor analysis* (3rd ed.). Chicago, IL: 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, NY: Springer.

Holland, P. W. (2008, March). *The first 4 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.

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, NJ: Erlbaum.

Horst, P. (1936). Item selection by means of a maximizing function. *Psychometrika, 1,* 229–244.

Horst, P. (1950a). *Optimal test length for maximum battery validity* (Research Bulletin No. RB-50-36). Princeton, NJ: Educational Testing Service.

Horst, P. (1950b). *The relationship between the validity of a single test and its contribution to the predictive efficiency of a test battery* (Research Bulletin No. RB-50-32). Princeton, NJ: Educational Testing Service.

Horst, P. (1951). Estimating total test reliability from parts of unequal length. *Educational and Psychological Measurement, 11,* 368–371.

Jackson, P. H., & Novick, M. R. (1969). *Maximizing the validity of a unit-weight composite as a function of relative component lengths with a fixed total testing time* (Research Bulletin No. RB-69-14). Princeton, NJ: Educational Testing Service.

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika, 38,* 593–604.

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.

Jöreskog, K. G. (1965). *Image factor analysis* (Research Bulletin No RB-65-05). Princeton, NJ: Educational Testing Service.

Jöreskog, K. G. (1966). *Some contributions to maximum likelihood factor analysis* (Research Bulletin . No. RB-66-41). Princeton, NJ: Educational Testing Service.

Jöreskog, K. G. (1967). *A general approach to confirmatory maximum likelihood factor analysis* (Research Bulletin No. RB-67-48). Princeton, NJ: Educational Testing Service.

Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika, 34,* 51–75.

Jöreskog, K. G. (1970). *Simultaneous factor analysis in several populations* (Research Bulletin No. RB-70-61). Princeton, NJ: Educational Testing Service.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

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, NJ: Erlbaum.

Jöreskog, K. G., & Lawley, D. N. (1967). *New methods in maximum likelihood factor analysis* (Research Bulletin No. RB-67-49). Princeton, NJ: Educational Testing Service.

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, NJ: Educational Testing Service.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22,* 29–41.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27,* 59–72.

- Kelley, T. L. (1923). *Statistical methods.* New York, NY: 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, NY: Springer-Verlag.

Kristof, W. (1962). *Statistical inferences about the error variance* (Research Bulletin No. RB-62-21). Princeton, NJ: Educational Testing Service.

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

Kristof, W. (1964). Testing differences between reliability coefficients. *British Journal of Statistical Psychology, 17,* 105–111.

Kristof, W. (1968). *Estimation of true score and error variance for tests under various equivalence assumptions* (Research Bulletin No. RB-68-57). Princeton, NJ: Educational Testing Service.

Kristof, W. (1970). On the sampling theory of reliability estimation. *Journal of Mathematical Psychology*, *7,* 371–377.

Kristof, W. (1971a). On the theory of a set of tests which differ only in length. *Psychometrika, 36,* 207–225.

Kristof, W. (1971b). *Testing a linear relation between true scores of two measures* (Research Bulletin No. RB-71-63). Princeton, NJ: Educational Testing Service.

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

• Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Linn, R. L. (1967). *Range restriction problems in the validation of a guidance test battery* (Research Bulletin No. RB-67-08). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika, 33,* 33–71.

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, NJ: Erlbaum.

Linn, R. L., & Werts, C. E. (1971a). Considerations for studies of test bias. *Journal of Educational Measurement*, *8,* 1–4.

Linn, R. L., & Werts, C. E. (1971b). *Errors of inference due to errors of measurement* (Research Bulletin No. RB-71-07). Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

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, NJ: Educational Testing Service.

Lord, F. M. (1950b). *Properties of test scores expressed as functions of the item parameters* (Research Bulletin No. RB-50-56). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1954a). *Equating test scores: The maximum likelihood solution for a common equating problem* (Research Bulletin No. RB-54-01). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1954b). *Estimation of parameters from incomplete data* (Research Bulletin No. RB-54-18). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1955). *Estimating test reliability* (Research Bulletin No. RB-55-07). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1956a). *A significance test for the hypothesis that two variables measure the same trait except for errors of measurement* (Research Bulletin No. RB-56-09). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1956b). *Do tests of the same length have the same standard error of measurement?* (Research Bulletin No. RB-56-07). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1956c). Sampling error due to choice of split in split-half reliability coefficients. *Journal of Experimental Education, 24,* 245–249.

Lord, F. M. (1957a). *Inferring the shape of the frequency distribution of true scores and of errors of measurement* (Research Bulletin No. RB-57-09). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1957b). *Some relations between Guttman's principal components of scale analysis and other psychometric theory* (Research Bulletin No. RB-57-07). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1958a). *An empirical study of the normality and independence of errors of measurement in test scores* (Research Bulletin No. RB-58-14). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1958b). *Tests of the same length do have the same standard error of measurement* (Research Bulletin No. RB-58-07). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1959a). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement, 19,* 233–239.

Lord, F. M. (1959b). *Use of true-score theory to predict moments of univariate and bivariate observed score distributions* (Research Bulletin No. RB-59-06). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307–321.

Lord, F. M. (1961). *Estimating norms by item sampling* (Research Bulletin No. RB-61-02). Princeton, NJ: Educational Testing Service.

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

Lord, F. M. (1962b). *Formula scoring and validity* (Research Bulletin No. RB-62-12). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965a). *A paradox in the interpretation of group comparisons* (Research Bulletin No. RB-65-08). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965b). A strong true score theory with applications. *Psychometrika, 30,* 239–270.

Lord, F. M. (1968). *Statistical adjustments when comparing preexisting groups* (Research Bulletin No. RB-68-67). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1971). *Testing if two measuring procedures measure the same psychological dimension* (Research Bulletin No. RB-71-36). Princeton, NJ: Educational Testing Service.

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

Lord, F. M., & Stocking, M. (1976). An interval estimate for making statistical inferences about true score. *Psychometrika, 41,* 79–87.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement, 8,* 453–461.

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, NJ: Erlbaum.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14,* 189–229.

Moses, T. (2012). Relationships of measurement error and prediction error in observed-score regression. *Journal of Educational Measurement, 49,* 380–398.

Moses, T., Deng, W., & Zhang, Y. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (Research Report No. RR-10-23). Princeton, NJ: Educational Testing Service.

Myers, C. T. (1975). *Test fairness: A comment on fairness in statistical analysis* (Research Bulletin No. RB-75-12). Princeton, NJ: Educational Testing Service.

Novick, M. R. (1965). *The axioms and principal results of classical test theory* (Research Report No. RR-65-2). Princeton, NJ: Educational Testing Service.

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, NJ: Erlbaum.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1–13.

Novick, M. R., & Thayer, D. T. (1969). *Some applications of procedures for allocating testing time* (Research Bulletin No. RB-69-01). Princeton, NJ: Educational Testing Service.

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, NJ: Educational Testing Service.

- 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, England: 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, NJ: Educational Testing Service.

Plumlee, L. B. (1954). Predicted and observed effect of chance on multiple-choice test validity. *Psychometrika, 19*, 65–70.

Potthoff, R. F. (1963). *On the Johnson-Neyman technique and some extensions thereof* (Research Bulletin No. RB-63-04). Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33–8.

Rubin, D. B. (1978a). *Bias reduction using Mahalanobis metric matching* (Research Bulletin No. RB-78-17). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1978b). *Using multivariate matched sampling and regression adjustment to control bias in observational studies* (Research Bulletin No. RB-78-16). Princeton, NJ: Educational Testing Service.

Rubin, D. B., & Thayer, D. (1978). Relating tests given to different samples. *Psychometrika, 43,* 1–10.

Saunders, D. R. (1953a). *An analytic method for rotation to orthogonal simple structure* (Research Bulletin No. RB-53-10). Princeton, NJ: Educational Testing Service.

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, NJ: Educational Testing Service.

Schultz, D. G., & Wilks, S. S. (1950). *A method for adjusting for lack of equivalence in groups* (Research Bulletin No. RB-50-59). Princeton, NJ: Educational Testing Service.

- Spearman, C. (1904a). General intelligence objectively determined and measured. *American Journal of Psychology, 15,* 201–293.
- Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

• Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Swineford, F. (1957). *Some relations between test scores and item statistic*s (Research Bulletin No. RB-57-02). Princeton, NJ: Educational Testing Service.

• Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8–14.

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

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.

Tucker, L. R., & Finkbeiner, C.T. (1981). *Transformation of factors by artificial personal probability functions* (Research Report No. RR-81-58). Princeton, NJ: Educational Testing Service.

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.

Woodbury, M. A., & Lord, F. M. (1955). *The most reliable composite score* (Research Bulletin No. RB-55-02). Princeton, NJ: Educational Testing Service.

Woodbury, M. A., & Novick, M. R. (1967). *Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time* (Research Bulletin No. RB-67-03). Princeton, NJ: Educational Testing Service.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton, NJ: Educational Testing Service.

Notes

[1] Boldface and full spelling of an individual's name indicates an ETS staff member.

[2] Although the summary of Cleary's (1966) work in this report 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, Pollack, & Burton, 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.

[3] The bullet symbol in the reference list (•) indicates work that was not performed at ETS.

Reports in the ETS R&D Scientific and Policy Contributions Series

Reports in the ETS R&D Scientific and Policy Contributions Series document the contributions made by the research program at Educational Testing Service since the founding of the organization in 1947.

Evaluating Educational Programs
by Samuel Ball (2011)
ETS R&D Scientific and Policy Contributions Series No. SPC-11-01
This inaugural report in the series reprints a paper that documented the vigorous program of evaluation research conducted at ETS in the 1960s and 1970, which helped lay the foundation for this fledgling field.

Modeling Change in Large-Scale Longitudinal Studies of Educational Growth: Four Decades of Contributions to the Assessment of Educational Growth
by Donald A. Rock (2012)
ETS R&D Scientific and Policy Contributions Series No. SPC-12-01
Rock reviews ETS's contribution to several large-scale longitudinal assessments over the years, ranging from the National Longitudinal Study of the High School Class of 1972 (NLS-72) to the Early Childhood Longitudinal Studies (ECLS).

Understanding the Impact of Special Preparation for Admissions Tests
by Donald E. Powers (2012)
ETS R&D Scientific and Policy Contributions Series No. SPC-12-02
Special preparation for tests has been a sometimes contentious subject, with disagreement over the effectiveness of preparation, equality of access, and impact on validity. Powers reviews the role ETS has taken over the years in test preparation and in addressing the associated issues.

ETS Research on Cognitive, Personality, and Social Psychology: I
by Lawrence J. Stricker (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-01
Stricker addresses research that ETS has conducted since the organization's inception in cognitive, personality, and social psychology.

Contributions of a Nonprofit Educational Measurement Organization to Education Policy Research
by Richard J. Coley, Margaret E. Goertz, and Gita Z. Wilder (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-02
The authors encapsulate the extensive work conducted at ETS in education policy research, including providing information on educational opportunity and educational outcomes, contributing to the discussion of important education issues, and promoting equal educational opportunity for all.