Listening. Learning. Leading.®

ETS R&D Scientific and Policy Contributions Series
ETS SPC-13-04

ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness

Neil J. Dorans

December 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 Contributions to the Quantitative Assessment of Item, Test, and Score Fairness

Neil J. Dorans

Educational Testing Service, Princeton, New Jersey

RR-13-27

Action Editor: James Carlson

Reviewers: Randy Bennett and Rebecca Zwick

Abstract

Quantitative fairness procedures have been developed and modified by ETS staff over the past several decades. ETS has been a leader in fairness assessment, and its efforts are reviewed in this report. The first section deals with differential prediction and differential validity procedures that examine whether test scores predict a criterion, such as performance in college, across different subgroups in a similar manner. The bulk of this report focuses on item level fairness, or differential item functioning, which is addressed in the various subsections of the second section. In the third section, I consider research pertaining to whether tests built to the same set of specifications produce scores that are related in the same way across different gender and ethnic groups. Limitations with the approaches reviewed here are discussed in the final section.

Key words: fairness, differential prediction, differential item functioning, score equity assessment, ETS, quantitative methods

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 seventh report in the series, Neil Dorans looks at quantitative fairness assessment procedures developed and modified by ETS staff, which have helped to make ETS a leader in fairness assessment. Almost since the inception of the organization in 1947, ETS has been concerned with the issues of fairness. In the late 1940s and early 1950s, William Turnbull, who later became the second president of ETS, was an early advocate of fairness, recommending the comparison of prediction equations as a method for assessing test fairness. In the 1980s, interest in fairness in the assessment community shifted from scores to items, as evidenced in widespread studies of differential item functioning (DIF). ETS, under the direction of Gregory Anrig, the third ETS president, established the industry standard for fairness assessment at the item level, and ETS has been in the vanguard in conducting DIF analyses as a standard psychometric check of test quality for over a quarter of a century.

Dorans is a distinguished presidential appointee at ETS. A major focus of his operational and research efforts during his career at ETS has been on the quantitative evaluation of fairness at the item and score levels. Since the early 1980s, he has been involved with most score equatings of the *SAT*® test. Equating is an essential to the process of producing fair scores. The architect for the recentered SAT scales, he has also performed score linking studies relating the SAT I to the ACT and the Prueba de Aptitud Academica (*PAA*™). Dorans co-edited a book on score linking and scale aligning, and co-authored a book on computer adaptive testing. He has written book chapters on DIF, context effects, item response theory, and score linking. He was awarded the ETS Measurement Statistician Award for, among other things, his work on recentering the SAT and the role he played in mentoring staff on score linking and fairness

issues. He received the National Council on Measurement in Education's Career Contributions Award in recognition of his substantial and creative theoretical and technical developments and his innovative ideas that have significantly affected measurement practices.

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

ETS was founded in 1947 as a not-for-profit organization (Bennett, 2005). Fairness concerns have been an issue at ETS almost since its inception. William Turnbull (1949, 1951a, 1951b), who in 1970 became the second president of ETS, addressed the Canadian Psychological Association on socioeconomic status and predictive test scores. He made a cogent argument for rejecting the notion that differences in subgroup performance on a test means that a test score is biased. He also advocated the comparison of prediction equations as a means of assessing test fairness. His article was followed by a number of articles by ETS staff on the issue of differential prediction. By the 1980s, under the direction of its third president, Gregory Anrig, ETS established the industry standard for fairness assessment at the item level (Holland & Wainer, 1993; Zieky, 2011). This century, fairness analyses have begun to focus on relationships between tests that purport to measure the same thing in the same way across different subgroups (Dorans & Liu, 2009; Liu & Dorans, 2013).

In this report, I review quantitative fairness procedures that have been developed and modified by ETS staff over the past decades. While some reference is made to events external to ETS, the focus is on ETS, which has been a leader in fairness assessment. In the first section, Fair Prediction of a Criterion, I consider differential prediction and differential validity, procedures that examine whether test scores predict a criterion, such as performance in college, across different subgroups in a similar manner. The bulk of this review is in the second section, Differential Item Functioning (DIF), which focuses on item-level fairness, or DIF. Then in the third section, Fair Linking of Test Scores, I consider research pertaining to whether tests built to the same set of specifications produce scores that are related in the same way across different gender and ethnic groups. In the final section, Limitations of Quantitative Fairness Assessment Procedures, limitations of these procedures are mentioned.

Fair Prediction of a Criterion

Turnbull (1951a) concluded his early ETS treatment of fairness with the following statement: “Fairness, like its amoral brother, validity, resides not in tests or test scores but in the relation to its uses” (p. 4–5).

While several ETS authors had addressed the relative lower performance of minority groups on tests of cognitive ability and its relationship to grades (e.g., Campbell, 1964), Cleary (1968) conducted one of the first differential prediction studies. That study has been widely cited and critiqued. A few years after the Cleary article, the field was replete with differential validity

studies, which focus on comparing correlation coefficients, and differential prediction studies, which focus on comparing regression functions, in large part because of interest engendered by the Supreme Court decision *Griggs v. Duke Power Co.* in 1971. This decision included the terms *business necessity* and *adverse impact*, both of which affected employment testing. Adverse impact is a substantially different rate of selection in hiring, promotion, transfer, training, or other employment-related decisions for any race, sex, or ethnic group. Business necessity can be used by an employer as a justification for using a selection mechanism that appears to be neutral with respect to sex, race, national origin, or religious group even though it excludes members of one sex, race, national origin, or religious group at a substantially higher rate than members of other groups. The employer must prove that the selection requirement having the adverse impact is job related and consistent with business necessity. In other words, in addition to avoiding the use of race/ethnic/gender explicitly as part of the selection process, the selection instrument had to have demonstrated validity for its use. Ideally, this validity would be the same for all subpopulations.

Linn (1972) considered the implications of the Griggs decision for test makers and users. A main implication was that there would be a need for empirical demonstrations that test scores predict criterion performance, such as how well one does on the job. (In an educational context, test scores may be used with other information to predict the criterion of average course grade). Reliability alone would not be an adequate justification for use of test scores. Linn also noted that for fair prediction to hold, the prediction model must include all the appropriate variables in the model. Otherwise misspecification of the model can give the appearance of statistical bias. The prediction model should include all the predictors needed to predict Y and the functional form used to combine the predictors should be the correct one. The reliabilities of the predictors also were noted to play a role. These limitations with differential validity and differential predictions studies were cogently summarized in four pages by Linn and Werts (1971). One of the quandaries faced by researchers that was not noted in this 1971 study is that some of the variables that contribute to prediction are variables over which a test taker has little control, such as gender, race, parent's level of education and income, and even zip code. Use of variables such as zip code to predict grades in an attempt to eliminate differential prediction would be unfair.

Linn (1975) later noted that differential prediction analyses should be preferred to differential validity studies because differences in predictor or criterion variability can produce

differential validity even when the prediction model is fair. Differential prediction analyses examine whether the same prediction models hold across different groups. Fair prediction or selection requires invariance of prediction equations across groups,

$$R(Y \mid \mathbf{X}, G = 1) = R(Y \mid \mathbf{X}, G = 2) = ... = R(Y \mid \mathbf{X}, G = g),$$

where *R* is the symbol for the function used to predict *Y*, the criterion score, from **X**, the predictor. *G* is a variable indicating subgroup membership.

Petersen and Novick (1976) compared several models for assessing fair selection, including the regression model (Cleary, 1968), the constant ratio model (Thorndike, 1971), the conditional probability model (Cole, 1973), and the constant probability model (Linn, 1973) in the lead article in a special issue of the *Journal of Educational Measurement* dedicated to the topic of fair selection.[1] They demonstrated that the regression, or Cleary, model, which is a differential prediction model, was a preferred model from a logical perspective in that it was consistent with its converse (i.e., fair selection of applicants was consistent with fair rejection of applicants). In essence, the Cleary model examines whether the regression of the criterion onto the predictor space is invariant across subpopulations.

Linn (1976) in his discussion of the Petersen and Novick (1976) analyses noted that the quest to achieve fair prediction is hampered by the fact that the criterion in many studies may itself be unfairly measured. Even when the correct equation is correctly specified in the full population and the criterion is measured well, invariance may not hold in subpopulations because of selection effects. Linn (1983) described how predictive bias may be an artifact of selection procedures. Linn used a simple case to illustrate his point. He posited that a single predictor *X* and linear model were needed to predict *Y* in the full population *P*. To paraphrase his argument, assume that a very large sample is drawn from *P* based on a selection variable *U* that might depend on *X* in a linear way. Errors in the prediction of *Y* from *X* and *U* from *X* are thus also linearly related because of their mutual dependence on *X*. Linn showed that the sample regression for the selected sample, *R* (*Y*|*X*, *G*) equals the regression in the full unselected population if the correlation between *X* and *U* is zero, or if errors in prediction of *Y* from *X* and *U* from *X* are uncorrelated. In other words, the slope of the relationship for predicting *U* from X must be zero or *Y* and *U* must be linearly independent given *X*.

Myers (1975) criticized the regression model because regression effects can produce differences in intercepts when two groups differ on *X* and *Y* and the predictor is unreliable, a point noted by Linn and Werts (1971). Myers argued for a linking or scaling model for assessing fairness. He noted that his approach made sense when *X* and *Y* were measures of the same construct, but admitted that scaling test scores to grades or vice versa had issues. This brief report by Myers can be viewed as a remote harbinger of work on the population invariance of score linking functions done by Dorans and Holland (2000), Dorans (2004), Dorans and Liu (2009), and Liu and Dorans (2013).

As can be inferred from the studies above, in particular Linn and Werts (1971) and Linn (1975, 1983), there are many ways in which a differential prediction study can go awry, and even more ways that differential validity studies can be problematic.

Differential Item Functioning (DIF)

During the 1980s, the focus in the profession shifted to DIF studies. Although interest in item bias studies began in the 1960s (Angoff, 1993), it was not until the 1980s that interest in fair assessment at the item level became widespread. During the 1980s, the measurement profession engaged in the development of item level models for a wide array of purposes. DIF procedures developed as part of that shift in attention from the score to the item.

Moving the focus of attention to prediction of item scores, which is what DIF is about, represented a major change from focusing primarily on fairness in a domain, where so many factors could spoil the validity effort, to a domain where analyses could be conducted in a relatively simple, less confounded way. While factors such as multidimensionality can complicate a DIF analysis, as described by Shealy and Stout[2] (1993), they are negligible compared to the many influences that can undermine a differential prediction study, as described in Linn and Werts (1971). In a DIF analysis, the item is evaluated against something designed to measure a particular construct and something that the test producer controls, namely a test score.

Around 100 ETS research bulletins, memoranda, or reports have been produced on the topics of item fairness, DIF, or item bias. The vast majority of these studies were published in the late 1980s and early 1990s. The major emphases of these reports can be sorted into categories and are treated in subsections of this section: Differential Item Functioning Methods, Matching Variable Issues, Study Group Definitions, and Sample Size and Power Issues. The DIF methods section begins with some definitions followed by a review of procedures that were suggested

before the term DIF was introduced. Most of the section then describes the following procedures: Mantel-Haenszel (MH), standardization (STAND), item response theory (IRT), and SIBTEST.

Differential Item Functioning (DIF) Methods

Two reviews of DIF methods were conducted by ETS staff: Dorans and Potenza (1994), which was shortened and published as Potenza and Dorans (1995), and Mapuranga, Dorans, and Middleton (2008), which then superseded Potenza and Dorans. In the last of these studies, the criteria for classifying DIF methods were a) definition of null DIF, b) definition of the studied item score, c) definition of the matching variable, and d) the variable used to define groups.

Null DIF is the absence of DIF. One definition of null DIF, observed score null DIF, is that all individuals with the same score on a test should have the same proportions answering the item correctly regardless of whether they are from the reference or focal group. The latent variable definition of null DIF can be used to compare the performance of focal and reference subgroups that are matched with respect to a latent variable. An observed difference in average item scores between two groups that may differ in their distributions of score on the matching variable is referred to as *impact*. With impact, we compare groups that may or may not be comparable with respect to the construct being measured by the item; using DIF, we compare groups that are comparable with respect to an estimate of their standing on that construct.

The s*tudied item score* refers to the scoring rule used for the items being studied for DIF. Studied items can either be scored as correct/incorrect (i.e., binary) or scored using more than two response categories (i.e., polytomous).

The *matching variable* is a variable used in the process of comparing the reference and focal groups (e.g., total test score or subscore) so that comparable groups are formed. In other words, matching is a way of establishing score equivalence between groups that are of interest in DIF analyses. The matching variable can either be an observed score or an unobserved latent variable consistent with a specific model for item performance, and can be either a univariate or multivariate variable.

In most DIF analyses, a single focal group is compared to a single reference group where the subgroup-classification variable (gender, race, geographic location, etc.) is referred to as the *grouping variable*. This approach ignores potential interactions between types of subgroups, (e.g., male/female and ethnic/racial). Although it might be better to analyze all grouping variables for DIF simultaneously (for statistical and computational efficiency), most DIF

methods compare only two groups at a time. While convention is often the reason for examining two groups at a time, small sample size sometimes makes it a necessity.

The remainder of this section describes briefly the methods that have been developed to assess what has become known as DIF. After reviewing some early work, I turn to the two methods that are still employed operationally here at ETS: the MH method and the STAND method. After briefly discussing IRT methods, I mention the SIBTEST method. Methods that do not fit into any of these categories are addressed in what seems to be the most relevant subsection.

Early developments: The years before differential item functioning (DIF) was defined at ETS. While most of the focus in the 1960s and 1970s was on the differential prediction issue, several researchers turned their attention to item-level fairness issues. Angoff (1993) discussed several, but not all of these efforts. Cardall and Coffman (1964) and Cleary and Hilton (1966, 1968) defined *item bias*, the phrase that was commonly used before DIF was introduced, as an item-by-subgroup interaction. Analysis of variance was used by both studies of DIF. Identifying individual problem items was not the goal of either study.

Angoff and Sharon (1974) also employed an analysis of variance (ANOVA) method, but by then the transformed item difficulty (TID) or delta-plot method had been adopted for item bias research. Angoff (1972) introduced this approach, which was rooted in Thurstone's absolute scaling model. This method had been employed by Tucker (1951) in a study of academic ability on vocabulary items and by Gulliksen (1964) in a cross-national study of occupation prestige. This method uses an inverse normal transformation to convert item proportion-correct values for two groups to normal deviates that are expect to form an ellipse. Items that deviate from the ellipse exhibit the item difficulty by group interaction that is indicative of what was called item bias. Angoff and Ford (1971, 1973) are the standard references for this approach.

The delta-plot method is akin to the Rasch (1960) model approach to assessing DIF. If items differ in their discriminatory power and the groups under study differ in terms of proficiency, then items will exhibit item-by-group interactions even when there are no differences in item functioning. This point was noted by several scholars including Lord (1977) and affirmed by Angoff (1993). As a consequence, the delta-plot method is rarely used for DIF assessment, except in cases where small samples are involved.

Two procedures may be viewed as precursors of the eventual move to condition directly on total score that was adopted by the STAND (Dorans & Kulick, 1983) and MH (Holland & Thayer, 1988) DIF approaches. Stricker (1982) recommended a procedure that looks for DIF by examining the partial correlation between group membership and item score with the effect of total test score removed. Scheuneman (1979) proposed a test statistic that was like a chi-square. This method was shown by Baker[3] (1981) and others to be affected inappropriately by sample size and to possess no known sampling distribution.

The late 1980s and the early 1990s were the halcyon days of DIF research and development at ETS and in the profession. Fairness was of paramount concern, and practical DIF procedures were developed and implemented (Dorans & Holland, 1993; Zieky, 1993). In October 1989, ETS and the Air Force Human Resources Laboratory sponsored a DIF conference that was held at ETS in October 1989. The papers presented at that conference, along with a few additions, were collected in the volume edited by Holland and Wainer (1993), known informally as the DIF book. It contains some of the major work conducted in this early DIF era, including several chapters about MH and STAND. The chapter by Dorans and Holland (1993) is the source of much of the material in the next two sections, which describe the MH and STAND procedures in some detail because they have been used operationally at ETS since that time. Dorans (1989) is another source that compares and contrasts these two DIF methods.

Mantel-Haenszel (MH): Original implementation at ETS. In their seminal paper, Mantel and Haenszel[4] (1959) introduced a new procedure for the study of matched groups. Holland and Thayer (1986, 1988) adapted the procedure for use in assessing DIF. This adaptation, the MH method, is used at ETS as the primary DIF detection device. The basic data used by the MH method are in the form of M 2-by-2 contingency tables or one large three dimensional 2-by-2-by-M table, where M is the number of levels of the matching variable.

Under rights-scoring for the items in which responses are coded as either correct or incorrect (including omissions), proportions of rights and wrongs on each item in the target population can be arranged into a contingency table for each item being studied. There are two levels for group: the focal group (f) that is the focus of analysis, and the reference group (r) that serves as a basis for comparison for the focal group. There are also two levels for item response: right (R) or wrong (W), and there are M score levels on the matching variable, (e.g., total score). Finally, the item being analyzed is referred to as the studied item. The 2 (groups)-by-2 (item

scores)-by-M (score levels) contingency table (see Table 1) for each item can be viewed in *2-by-2 slices.*

Table1

2-by-2-by-M Contingency Table for an Item, Viewed in a 2-by-2 Slice

	Item score		
Group	Right	Wrong	Total
Focal group (f)	R_{fm}	W_{fm}	N_{fm}
Reference group (r)	R_{rm}	W_{rm}	N_{rm}
Total group (t)	R_{tm}	W_{tm}	N_{tm}

The null DIF hypothesis for the MH method can be expressed as

$$H_0: [R_{rm}/W_{rm}] = [R_{fm}/W_{fm}] \qquad m = 1,...,M.$$

In other words, the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and the reference group portions of the population, and this equality holds across all M levels of the matching variable.

In their original work, Mantel and Haenszel (1959) developed a chi-square test of the null hypothesis against a particular alternative hypothesis known as the constant odds ratio hypothesis,

$$H_a: [R_{rm}/W_{rm}] = \alpha\,[R_{fm}/W_{fm}] \qquad m = 1,...,M \text{ and } \alpha \neq 1.$$

Note that when $\alpha = 1$, the alternative hypothesis reduces to the null DIF hypothesis. The parameter α is called the *common odds ratio* in the M 2-by-2 tables because under H_a, the value of α is the odds ratio that is the same for all m.

Holland and Thayer (1988) reported that the MH approach is the test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds-ratio hypothesis.

Mantel and Haenszel (1959) also provided an estimate of the constant odds -ratio that ranges from 0 to ∞, for which a value of 1 can be taken to indicate null DIF. This odds-ratio metric is not particularly meaningful to test developers who are used to working with numbers

on an item difficulty scale. In general, odds are converted to $\log_e$ odds because the latter is symmetric around zero and easier to interpret.

At ETS, test developers use item difficulty estimates in the *delta metric*, which has a mean of 13 and a standard deviation of 4. Large values of delta correspond to difficult items, while easy items have small values of delta. Holland and Thayer (1985) converted the estimate of the common odds ratio, α_{MH} ,into a difference in deltas via:

$$\text{MH D-DIF} = -2.35 \ln[\alpha_{MH}].$$

Note that positive values of MH D-DIF favor the focal group, while negative values favor the reference group. A standard error for MH D-DIF was provided in Dorans and Holland (1993)**.**

Subsequent developments with the Mantel-Haenszel (MH) approach. Subsequent to the operational implementation of the MH approach to DIF detection by ETS in the late 1980s (Zieky, 1993, 2011), there was a substantial amount of DIF research conducted by ETS staff through the early 1990s. Some of this research was presented in Holland and Wainer (1993); other presentations appeared in journal articles and ETS Research Reports. This section contains a partial sampling of research conducted primarily on the MH approach.

Donoghue, Holland, and Thayer (1993) varied six factors in an IRT-based simulation of DIF in an effort to better understand the properties of the MH and STAND (to be described in the next section) effect sizes and their standard errors. The six factors varied were level of the IRT discrimination parameter, the number of DIF items in the matching variable, the amount of DIF on the studied item, the difficulty of the studied item, whether the studied item was included in the matching variable, and the number of items in the matching variable. Donoghue et al. found that both the MH and STAND methods had problems detecting IRT DIF in items with nonzero lower asymptotes. Their two major findings were the need to have enough items in the matching variable to ensure reliable matching for either method and the need to include the studied item in the matching variable in MH analysis. This study thus provided support for the analytical argument for inclusion of the studied item that had been made by Holland and Thayer (1986). As will be seen later, Zwick, Donaghue, and Grima (1993), Zwick (1990), Lewis (1993), and Tan, Xiang, Dorans, and Qu (2010) also addressed the question of inclusion of the studied item.

Longford, Holland, and Thayer (1993) demonstrated how to use a random-effect or variance-component model to aggregate DIF results for groups of items. In particular they showed how to combine DIF estimates from several administrations to obtain variance components for administration differences for DIF within an item. In their examples, they demonstrated how to use their models to improve estimations within an administration, and how to combine evidence across items in randomized DIF studies. Subsequently, ETS researchers have employed Bayesian methods with the goal of pooling data across administrations to yield more stable DIF estimates within an administration. These approaches are discussed in the section on sample size and power issues.

Allen and Holland (1993) used a missing data framework to address the missing data problem in DIF analyses where "no response" to the self-reported group identification question is large, a common problem in applied settings. They showed how MH and STAND statistics can be affected by different assumptions about nonresponses.

Zwick and her colleagues examined DIF in the context of computer adaptive testing (CAT) in which tests are tailored to the individual test taker on the basis of his or her response to previous items. Zwick, Thayer, and Wingersky (1993) described in great detail a simulation study in which they examined the performance of MH and STAND procedures that had been modified for use with data collected adaptively. The modification to the DIF procedures involved replacing the standard number-right matching variable with a matching variable based on IRT, which was obtained by converting a maximum likelihood estimate of ability to an expected number-right true score on all items in the reference pool. Examinees whose expected true scores fell in the same one-unit intervals were considered to be matched. They found that DIF statistics computed in this way for CAT were similar to those obtained with the traditional matching variable of performance on the total test. In addition they found that pretest DIF statistics were generally well behaved, but the MH DIF statistics tended to have larger standard errors for the pretest items than for the CAT items.

Zwick, Thayer, and Wingersky (1994) addressed the effect of using alternative matching methods for pretest items. Using a more elegant matching procedure did not lead to a reduction of the MH standard errors and produced DIF measures that were nearly identical to those from the earlier study. Further investigation showed that the MH standard errors tended to be larger when items were administered to examinees with a wide ability range, whereas the opposite was

true of the standard errors of the STAND DIF statistic. As reported in Zwick (1994), there may be a theoretical explanation for this phenomenon.

CAT can be thought of as a very complex form of item sampling. The sampling procedure used by the National Assessment of Educational Progress (NAEP) is another form of complex sampling. Allen and Donoghue (1995) used a simulation study to examine the effect of complex sampling of items on the measurement of DIF using the MH DIF procedure. Data were generated using a three-parameter logistic (3PL) IRT model according to the balanced incomplete block design. The length of each block of items and the number of DIF items in the matching variable were varied, as was the difficulty, discrimination, and presence of DIF in the studied item. Block, booklet, pooled booklet, and other approaches to matching on more than the block, were compared to a complete data analysis using the transformed log-odds on the delta scale. The pooled booklet approach was recommended for use when items are selected for examinees according to a balanced incomplete block (BIB) data collection design.

Zwick, Donoghue, and Grima (1993) noted that some forms of performance assessment may in fact be more likely to tap construct-irrelevant factors than multiple-choice items are. The assessment of DIF can be used to investigate the effect on subpopulations of the introduction of performance tasks. Two extensions of the MH procedure were explored: the test of conditional association proposed by Mantel (1963) and the generalized statistic proposed by Mantel and Haenszel (1959). Simulation results showed that, for both inferential procedures, the studied item should be included in the matching variable, as in the dichotomous case. Descriptive statistics that index the magnitude of DIF, including that proposed by Dorans and Schmitt (1991; described below) were also investigated.

Standardization (STAND). Dorans (1982) reviewed item bias studies that had been conducted on data from the *SAT*® exam in the late 1970s, and concluded that these studies were flawed because either DIF was confounded with lack of model fit or it was contaminated by impact as a result of *fat matching*, the practice of grouping scores into broad categories of roughly comparable ability. A new method was needed. Dorans and Kulick (1983, 1986) developed the STAND approach after consultation with Holland. The formulas in the following section can be found in these articles and in Dorans and Holland (1993) and Dorans and Kulick (2006).

Standardization's (STAND's) definition of differential item functioning (DIF). An item exhibits DIF when the expected performance on an item differs for matched examinees from different groups. Expected performance can be estimated by nonparametric item-test regressions. Differences in empirical item-test regressions are indicative of DIF.

The first step in the STAND analysis is to use all available data in the target population of interest to estimate nonparametric item-test regressions in the reference group and in the focal group. Let $\boldsymbol{\varepsilon}_f(Y \mid X)$ define the empirical item-test regression for the focal group *f*, and let $\boldsymbol{\varepsilon}_r(Y \mid X)$ define the empirical item-test regression for the reference group *r*, where Y is the item-score variable and X is the matching variable. For STAND, the definition of null-DIF conditions on an observed score is $\boldsymbol{\varepsilon}_f(Y \mid X) = \boldsymbol{\varepsilon}_r(Y \mid X)$. Plots of difference in empirical item-test regressions, focal minus reference, provide visual descriptions of DIF in fine detail for binary as well as polytomously scored items. For illustrations of nonparametric item-test regressions and differences for an actual SAT item that exhibits considerable DIF, see Dorans and Kulick (1986).

Standardization's (STAND's) primary differential item functioning (DIF) index. While plots described DIF directly, there was a need for some numerical index that targets suspect items for close scrutiny while allowing acceptable items to pass swiftly through the screening process. For each score level, the focal group supplies specific weights that are used for each individual D_m before accumulating the weighted differences across score levels to arrive at a summary item-discrepancy index, $STD-EISDIF$, which is defined as:

$$STD-EISDIF = \frac{\sum_{m=1}^{M} N_{fm} * E_f(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}} - \frac{\sum_{m=1}^{M} N_{fm} * E_r(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}}$$

where $N_{fm} \; / \sum_{m=1}^{M} N_{fm}$ is the weighting factor at score level X_m supplied by the focal group to weight differences in expected item performance observed in the focal group $E_f(Y \mid X)$ and expected item performance observed in the reference group $E_r(Y \mid X)$.

In contrast to impact, in which each group has its relative frequency serve as a weight at each score level, STAND uses a standard or common weight on both $\varepsilon_f(Y \mid X = m)$ and $\varepsilon_r(Y \mid X = m)$, namely $N_{fm} / \sum_{m=1}^{M} N_{fm}$. The use of the same weight on both $\varepsilon_f(Y \mid X = m)$ and $\varepsilon_r(Y \mid X = m)$ is the essence of the STAND approach. Use of N_{fm} means that $EISDIF$ equals the difference between the observed performance of the focal group on the item and the predicted performance of selected reference group members who are matched in ability to the focal group members. This difference can be derived very simply; see Dorans and Holland (1993).

Extensions to standardization (STAND). The generalization of the STAND methodology to all response options including omission and not reached is straightforward and is known as standardized distractor analysis (Dorans & Schmitt, 1993; Dorans, Schmitt, & Bleistein, 1988, 1992). It is as simple as replacing the keyed response with the option of interest in all calculations. For example, a standardized response-rate analysis on Option A would entail computing the proportions choosing A in both the focal and reference groups. The next step is to compute differences between these proportions at each score level. Then these individual score-level differences are summarized across score levels by applying some standardized weighting function to these differences to obtain $STD-DIF(A)$, the standardized difference in response rates to Option A. In a similar fashion one can compute standardized differences in response rates for Options B, C, D, and E, and for nonresponses as well. This procedure is used routinely at ETS.

Application of the STAND methodology to counts of examinees at each level of the matching variable who did not reach the item results in a standardized not-reached difference. For items at the end of a separately timed section of a test, these standardized differences provide measurement of the differential speededness of a test. *Differential speededness* refers to the existence of differential response rates between focal group members and matched reference group members to items appearing at the end of a section. Schmitt, Holland, and Dorans (1993) reported that excluding examinees who do not reach an item from the calculation of the DIF statistic for that item partially compensates for the effects of item location on the DIF estimate.

Dorans and Schmitt **(**1991) proposed an extended version of STAND for ordered polytomous data. This extension has been used operationally with NAEP data since the early 1990s. This approach, called standardized mean difference (*SMD*) by Zwick, Donoghue, and Grima (1993), provides an average DIF value for describing DIF on items with ordered categories. At each matching score level, there exist distributions of ordered item scores, I, for both the focal group (e.g., females) and the reference group (e.g., males). The expected item scores for each group at each matching score level can be computed by using the frequencies to obtain a weighted average of the score levels. The difference between these expected items scores for the focal and reference groups, $STD-EISDIF$ **,** is the DIF statistic. Zwick and Thayer (1996) provide standard errors for *SMD* **(**or $STD-EISDIF$ **).**

Item response theory (IRT). DIF procedures differ with respect to whether the matching variable is explicitly an observed score (Dorans & Holland, 1993) or implicitly a latent variable (Thissen, Steinberg, & Wainer, 1993). Observed score DIF and DIF procedures based on latent variables do not measure the same thing, and both are not likely to measure what they strive to measure, which is DIF with respect to the construct that the item purports to measure. The observed score procedures condition on an observed score, typically the score reported to a test taker, which contains measurement error and clearly differs from a pure measure of the construct of interest, especially for test scores of inadequate reliability. The latent variable approaches in essence condition on an unobservable that the test is purportedly measuring. As such they employ what Meredith and Millsap[5] (1992) would call a measurement invariance definition of null DIF, while methods like MH and STAND employ a prediction invariance definition, which may be viewed as inferior to measurement invariance from a theoretical perspective. On the other hand, procedures that purport to assess measurement invariance employ a set of assumptions; in essence they are assessing measurement invariance under the constraint that the model they assume to be true is in fact true.

The observed score methods deal with the fact that an unobservable is unknowable by replacing the null hypothesis of measurement invariance (i.e., the items measure the construct of interest in the same way in the focal and reference groups with a prediction invariance assumption and use the data directly to assess whether expected item score is a function of observed total score in the same way across groups). The latent variable approaches retain the measurement invariance hypothesis and use the data to estimate and compare functional forms of

the measurement model relating item score to a latent variable in the focal and reference groups. The assumptions embodied in these functional forms may or may not be correct, however, and model misfit might be misconstrued as a violation of measurement invariance, as noted by Dorans and Kulick (1983). For example applying the Rasch model to data fit by the two-parameter logistic (2PL) model would flag items with lower IRT slopes as having DIF favoring the lower scoring group, while items with higher slopes would favor the higher scoring group.

Lord (1977, 1980) described early efforts to assess DIF from a latent trait variable perspective. Lord recommended a statistical significance test on the joint difference between the IRT difficulty and discrimination parameters between the two groups under consideration. Thissen, Steinberg, and Wainer (1993) discussed Lord's procedure and described the properties of four other procedures that used IRT. All these methods used statistical significance testing. They also demonstrated how the IRT methods can be used to assess differential distractor functioning. Thissen et al. remains a very informative introduction and review of IRT methods circa 1990.

Pashley (1992) suggested a method for producing simultaneous confidence bands for the difference between item response curves. After these bands have been plotted, the size and regions of DIF can be easily identified. Wainer (1993) provided an IRT-based effect size of amount of DIF that is based on the STAND weighting system that allows one to weight difference in the item response functions (IRF) in a manner that is proportional to the density of the ability distribution.

Zwick et al. (1994) and Zwick, Thayer, and Wingersky (1995) applied the Rasch model to data simulated according to the 3PL model. They found that the DIF statistics based on the Rasch model were highly correlated with the DIF values associated with the generated data, but that they tended to be smaller in magnitude. Hence the Rasch model did not detect DIF as well, which was attributed to degradation in the accuracy of matching. Expected true scores from the Rasch-based computer-adaptive test tended to be biased downward, particularly for lower-ability examinees. If the Rasch model had been used to generate the data, different results would probably have been obtained.

Wainer, Sireci, and Thissen (1991) developed a procedure for examining DIF in collections of related items, such as those associated with a reading passage. They called this

DIF for a set of items a *testlet DIF*. This methodology paralleled the IRT-based likelihood procedures mentioned by Thissen et al. (1993).

Zwick (1989, 1990) demonstrated that the null definition of DIF for the MH procedure (and hence STAND and other procedures employing observed scores as matching variables) and the null hypothesis based on IRT are different because the latter compares item response curves, which in essence condition on unobserved ability. She also demonstrated that the item being studied for DIF should be included in the matching variable if MH is being used to identify IRT DIF.

SIBTEST. Shealy and Stout (1993) introduced a general model-based approach to assessing DIF and other forms of differential functioning. They cited the STAND approach as a progenitor. From a theoretical perspective, SIBTEST is elegant. It sets DIF within a general multidimensional model of item and test performance. Unlike most IRT approaches, which posit a specific form for the item response model (e.g., a 2PL model), SIBTEST does not specify a particular functional form. In this sense it is a nonparametric IRT model, in principle, in which the null definition of STAND is replaced by

$$\boldsymbol{\varepsilon}_f(Y \mid T_x) = \boldsymbol{\varepsilon}_r(Y \mid T_x),$$

where T_x represents a true score for X. As such, SIBTEST employs a measurement invariance definition of null DIF, while STAND employs a prediction invariance definition (Meredith & Millsap, 1992).

Chang, Mazzeo, and Roussos (1995, 1996) extended SIBTEST to handle polytomous items. Two simulation studies compared the modified SIBTEST procedure with the generalized Mantel (1963) and SMD or STAND procedures. The first study compared the procedures under conditions in which the generalized Mantel and SMD procedures had been shown to perform well (Zwick, Donoghue, & Grima, 1993). Results of Study 1 suggested that SIBTEST performed reasonably well, but that the generalized Mantel and SMD procedures performed slightly better. The second study used data simulated under conditions in which observed-score DIF methods for dichotomous items had not performed well (i.e., a short nonrepresentative matching test). The results of Study 2 indicated that, under these conditions, the modified SIBTEST procedure

provided better control of impact-induced Type I error inflation with respect to detecting DIF (as defined by SIBTEST) than the other procedures.

Zwick, Thayer, and Mazzeo (1997) evaluated statistical procedures for assessing DIF in polytomous items. Three descriptive statistics—the SMD (Dorans & Schmitt, 1991) and two procedures based on SIBTEST (Shealy & Stout, 1993) were considered, along with five inferential procedures: two based on SMD, two based on SIBTEST, and one based on the Mantel (1963) method. The DIF procedures were evaluated through applications to simulated data, as well as to empirical data from ETS tests. The simulation included conditions in which the two groups of examinees had the same ability distribution and conditions in which the group means differed by one standard deviation. When the two groups had the same distribution, the descriptive index that performed best was the SMD. When the two groups had different distributions, a modified form of the SIBTEST DIF effect-size measure tended to perform best. The five inferential procedures performed almost indistinguishably when the two groups had identical distributions. When the two groups had different distributions and the studied item was highly discriminating, the SIBTEST procedures showed much better Type I error control than did the SMD and Mantel methods, particularly with short tests. The power ranking of the five procedures was inconsistent; it depended on the direction of DIF and other factors. The definition of DIF employed was the IRT definition, measurement invariance, not the observed score definition, prediction invariance.

Dorans (2011) summarized differences between SIBTEST and its progenitor, STAND. STAND uses observed scores to assess whether the item-test regressions are the same across focal and reference groups. The SIBTEST DIF method appears to be more aligned with measurement models. This method assumes that examinee group differences influence DIF or test form difficulty differences more than can be observed in unreliable test scores. SIBTEST adjusts the observed data toward what is suggested to be appropriate by the measurement model. The degree to which this adjustment occurs depends on the extent that these data are unreliable. To compensate for unreliable data on the individual, SIBTEST regresses observed performance on the test to what would be expected for the focal or reference group on the basis of the ample data that show that race and gender are related to item performance. SIBTEST treats true score estimation as a prediction problem, introducing bias to reduce mean squared error. In essence, the SIBTEST method uses subgroup-specific true score estimates as a surrogate for the true score

that is defined in the classical test theory model. If SIBTEST regressed all test takers to the same mean it would not differ from STAND.

Matching Variable Issues

Dorans and Holland (1993) laid out an informal research agenda with respect to observed score DIF. The matching variable was one area that merited investigation. Inclusion of the studied item in the matching variable and refinement or purification of the criterion were mentioned. Dimensionality and DIF was, and remains, an important factor; DIF procedures presume that all items measure the same construct in the same way across all groups.

Donoghue and Allen (1992, 1993) examined two strategies for forming the matching variable for the MH DIF procedure; "thin" matching on total test score was compared to forms of "thick" matching, pooling levels of the matching variable. Data were generated using a 3PL IRT model with a common guessing parameter. Number of subjects and test length were manipulated, as were the difficulty, discrimination, and presence/absence of DIF in the studied item. For short tests (5 or 10 items), thin matching yielded very poor results, with a tendency to falsely identify items as possessing DIF against the reference group. The best methods of thick matching yielded outcome measure values closer to the expected value for non-DIF items and a larger value than thin matching when the studied item possessed DIF. Intermediate-length tests yielded similar results for thin matching and the best methods of thick matching.

The issue of whether or not to include the studied item in the matching variable was investigated by many researchers from the late 1980s to early 1990s. Holland and Thayer (1988) demonstrated mathematically that when the data were consistent with the Rasch model, it was necessary to include the studied item in a purified rights-scored matching criterion in order to avoid biased estimates of DIF (of the measurement invariance type) for that studied item. Inclusion of the studied item removes the dependence of the item response on group differences in ability distributions. Zwick (1990) and Lewis (1993) developed this idea further to illustrate the applicability of this finding to more general item response models. Both authors proved mathematically that the benefit in bias correction associated with including the studied item in the matching criterion held true for the binomial model, and they claimed that the advantage of including the studied item in the matching criterion would not be evident for any IRT model more complex than the Rasch model.

Donoghue et al. (1993) evaluated the effect of including/excluding the studied item under the 3PL IRT model. In their simulation, they fixed the discrimination parameters for all items in a simulated test in each studied condition and fixed the guessing parameter for all conditions, but varied the difficulty (*b*) parameters for different items for each studied condition. Although the 3PL model was used to simulate data, only the *b*-parameter was allowed to vary. On the basis of their study, they recommended including the studied item in the matching variable when the MH procedure is used for DIF detection. They also recommended that short tests not be used for matching variables.

Zwick, Donoghue, and Grima (1993) extended the scope of their DIF research to performance tasks. In their study, multiple-choice (MC) items and performance tasks were simulated using the 3PL model and the partial-credit model, respectively. The MC items were simulated to be free of DIF and were used as the matching criterion. The performance tasks were simulated to be the studied items with or without DIF. They found that the item should be included in the matching criterion.

Zwick (1990) analytically examined item inclusion for models more complex than the Rasch. Her findings apply to monotone IRFs with local independence for the case where the IRFs on the matching items were assumed identical for the two groups. If the studied item is excluded from the matching variable, the MH null hypothesis will not hold in general even if the two groups had the same IRF for the studied item. It is assured to hold only if the groups have the same ability distribution. If the ability distributions are ordered, the MH will show DIF favoring the higher group (generalization of Holland & Thayer's [1988] Rasch model findings). Even if the studied item is included, the MH null hypothesis will not hold in general. It is assured to hold only if the groups have the same ability distribution or if the Rasch model holds. Except in these special situations, the MH can produce a conclusion of DIF favoring either the focal or reference group.

Tan et al. (2010) studied the impact of including/excluding the studied item in the matching variable on bias in DIF estimates under conditions where the assumptions of the Rasch model were violated. Their simulation study varied different magnitudes of DIF and different group ability distributions, generating data from a 2PL IRT model and a multidimensional IRT model. Results from the study showed that including the studied item leads to less biased DIF estimates and more appropriate Type I error rate, especially when group ability distributions are

different. Systematic biased estimates in favor of the high ability group were consistently found across all simulated conditions when the studied item was excluded from the matching criterion.

Zwick and Ercikan (1989) used bivariate matching to examine DIF on the NAEP history assessment, conditioning on number-right score and historical period studied. Contrary to expectation, the additional conditioning did not lead to a reduction in the number of DIF items.

Pomplun, Baron, and McHale (1992) evaluated the use of bivariate matching to study DIF with formula-scored tests, where item inclusion cannot be implemented in a straightforward fashion. Using SAT Verbal data with large and small samples, both male-female and black-white group comparisons were investigated. MH D-DIF values and DIF category classifications based on bivariate matching on rights score and nonresponse were compared with MH D-DIF values and categories based on rights-scored and formula-scored matching criteria. When samples were large, MH D-DIF values based on the bivariate matching criterion were ordered very similarly to MH D-DIF values based on the other criteria. However, with small samples the MH D-DIF values based on the bivariate matching criterion displayed only moderate correlations with MH D-DIF values from the other criteria.

Study Group Definition

Another area mentioned by Dorans and Holland (1993) was the definition of the focal and reference groups. Research has continued in this area as well.

Allen and Wainer (1989) noted that the accuracy of procedures that are used to compare the performance of different groups of examinees on test items obviously depends upon the correct classification of members in each examinee group. They argued that because the number of nonrespondents to questions of ethnicity is often of the same order of magnitude as the number of identified members of most minority groups, it is important to understand the effect of nonresponse on DIF results. They examined the effect of nonresponse to questions of ethnic identity on the measurement of DIF for SAT verbal items using the MH procedure. They demonstrated that efforts to obtain more complete ethnic identifications from the examinees would lead to more accurate DIF analyses.

DIF analyses are performed on target populations. One of the requirements for inclusion in the analysis sample is that the test taker has sufficient skill in the language of the test. Sinharay, Dorans, and Liang (2009) examined how an increase in the proportion of examinees who report that English is not their first language would affect DIF results if they were included

in the DIF analysis sample of a large-scale assessment. The results varied by group. In some combinations of focal/reference groups, the magnitude of DIF was not appreciably affected by whether DIF was performed on examinees whose first language was not English. In other groups, first language status mattered. The results varied by type of test as well. In addition, the magnitude of DIF for some items was substantially affected by whether the DIF was performed on examinees whose first language was not English.

Dorans and Holland (1993) pointed out that in traditional one-way DIF analysis, deleting items due to DIF can have unintended consequences on the focal group. DIF analysis performed on gender and on ethnicity/race alone ignores the potential interactions between the two main effects. Additionally, Dorans and Holland suggested applying a "melting-pot" DIF method wherein the total group would function as the reference group and each gender-by-ethnic subgroup would serve sequentially as a focal group. Zhang, Dorans, and Matthews-Lopez (2005) proposed a variation on the melting-pot approach called DIF dissection. They adapted the STAND methodology so that the reference group was defined to be the total group, while each of the subgroups independently acted as a focal group. They argued that using a combination of all groups as the reference group and each combination of gender and ethnicity as a focal group produces more accurate, though potentially less stable, findings than using a simple majority group approach. As they hypothesized, the deletion of a sizable DIF item had its greatest effect on the mean score of the focal group that had the most negative DIF according to the DIF dissection method. In addition, the study also found that the DIF values obtained by the DIF procedure reliably predicted changes in scaled scores after item deletion.

Sample Size and Power Issues

From its inaugural use as an operational procedure, DIF has had to grapple with sample size considerations (Zieky 1993). The conflict between performing as much DIF as possible and limiting the analysis to those cases where there is sufficient power to detect DIF remains as salient as ever.

Lyu, Dorans, and Ramsay (1995) developed a smoothed version of STAND, which merged kernel smoothing with the traditional STAND DIF approach, to examine DIF for student produced response (SPR) items on the SAT I Math at both the item and testlet levels. Results from the smoothed item-level DIF analysis showed that regular multiple-choice items have more variability in DIF values than SPRs.

Bayesian methods are often resorted to when small sample sizes limit the potential power of a statistical procedure. Bayesian statistical methods can incorporate, in the form of a prior distribution, existing information on the inference problem at hand, leading to improved estimation, especially for small samples for which the posterior distribution is sensitive to the choice of prior distribution. Zwick, Thayer, and Lewis (1997, 1999) developed an empirical Bayes (EB) enhancement to MH DIF analysis in which they assumed that the MH statistics were normally distributed and that the prior distribution of underlying DIF parameters was also normal. They used the posterior distribution of DIF parameters to make inferences about the item's true DIF status and the posterior predictive distribution to predict the item's future observed status. DIF status was expressed in terms of the probabilities associated with each of the five DIF levels defined by the ETS classification system (Zieky, 1993). The EB method yielded more stable DIF estimates than did conventional methods, especially in small samples. The EB approach also conveyed information about DIF stability in a more useful way by representing the state of knowledge about an item's DIF status as probabilistic.

Zwick, Thayer, and Lewis (2000) investigated a DIF flagging method based on loss functions. The approach built on their earlier research that involved the development of an EB enhancement to MH DIF analysis. The posterior distribution of DIF parameters was estimated and used to obtain the posterior expected loss for the proposed approach and for competing classification rules. Under reasonable assumptions about the relative seriousness of Type I and Type II errors, the loss-function-based DIF detection rule was found to perform better than the commonly used ETS DIF classification system, especially in small samples.

Zwick and Thayer (2002) used a simulation to investigate the applicability to computerized adaptive test data of an EB DIF analysis method developed by Zwick, Thayer, and Lewis (1997; Zwick et al., 1999) and showed that the performance of the EB DIF approach to be quite promising, even in extremely small samples. When combined with a loss-function-based decision rule, the EB method is better at detecting DIF than conventional approaches, but it has a higher Type I error rate.

The EB method estimates the prior mean and variance from the current data and uses the same prior information for all the items. For most operational tests, however, a large volume of past data is available, and for any item appearing in a current test, a number of similar items are often found to have appeared in past operational administrations of the test. Conceptually, it

should be possible to incorporate that past information into a prior distribution in a Bayesian DIF analysis. Sinharay, Dorans, Grant, and Blew (2009) developed a full Bayesian (FB) DIF estimation method that used this type of past information. The FB Bayesian DIF analysis method was shown to be an improvement over existing methods in a simulation study.

Zwick et al. (2000) proposed a Bayesian updating (BU) method that may avert the shrinkage associated with the EB and FB approaches. Zwick, Ye, and Isham (2012) implemented the BU approach and compared it to the EB and FB approaches in both simulated and empirical data. They maintained that the BU approach was a natural way to accumulate all known DIF information about an item while mitigating the tendency to shrink DIF toward zero that characterized the EB and FB approaches.

Smoothing is another alternative used for dealing with small sample sizes. Yu, Moses, Puhan, and Dorans (2008) applied smoothing techniques to frequency distributions and investigated the impact of smoothed data on MH DIF detection in small samples. Eight sample-size combinations were randomly drawn from a real data set were replicated 80 times to produce stable results. Loglinear smoothing was found to provide slight-to-moderate improvements in MH DIF estimation with small samples.

Puhan, Moses, Yu, and Dorans (2007, 2009) examined the extent to which loglinear smoothing could improve the accuracy of SIBTEST DIF estimates in small samples of examinees. Examinee responses from a certification test were used. Separate DIF estimates for seven small-sample-size conditions were obtained using unsmoothed and smoothed score distributions. Results indicated that for most studied items smoothing the raw score distributions reduced random error and bias of the DIF estimates, especially in the small-sample-size conditions.

Fair Linking of Test Scores

Scores on different forms or editions of a test that are supposed to be used interchangeably should be related to each other in the same way across different subpopulations. Score equity assessment (SEA) uses subpopulation invariance of linking functions across important subpopulations to assess the degree of interchangeability of scores.

Test score equating is a statistical process that produces scores considered comparable enough across test forms to be used interchangeably. Five requirements are often regarded as basic to all test equating (Dorans & Holland, 2000). One of the most basic requirements of score

equating is that equating functions should be subpopulation invariant (Dorans & Holland, 2000; Holland & Dorans, 2006). That is, they should not be strongly influenced by the subpopulation of examinees on which they are computed. The same construct and equal reliability requirements are prerequisites for subpopulation invariance. One way to demonstrate that two test forms are not equatable is to show that the equating functions used to link their scores are not invariant across different subpopulations of examinees. Lack of invariance in a linking function indicates that the differential difficulty of the two test forms is not consistent across different groups. The invariance can hold if the relative difficulty changes as a function of score level in the same way across subpopulations. If, however, the relative difficulty of the two test forms interacts with group membership or an interaction among score level, difficulty, and group is present, then invariance does not hold. SEA uses the subpopulation invariance of linking functions across important subgroups (e.g., gender groups and other groups, sample sizes permitting) to assess the degree of score exchangeability.

In an early study, Angoff and Cowell (1985, 1986) examined the invariance of equating scores on alternate forms of the *GRE*® quantitative test for various populations, including gender, race, major, and ability. Angoff and Cowell conducted equatings for each of the populations and compared the resulting conversions to each other and to differences that would be expected given the standard errors of equating. Differences in the equatings were found to be within that expected given sampling error. Angoff and Cowell concluded that population invariance was supported.

Dorans and Holland (2000) included several examples of linkings that are invariant (e.g., SAT Mathematics to SAT Mathematics and SAT Verbal to SAT Verbal, and SAT Mathematics to ACT Mathematics) as well as ones that are not (e.g., verbal to mathematics, and linkings between non-math ACT subscores and SAT Verbal). Equatability indexes are used to quantify the degree to which linkings are subpopulation invariant.

Since 2000, several evaluations of population invariance have been performed. Yang (2004) examined whether the linking functions that relate multiple-choice scores to composite scores based on weighted sums of multiple choice and constructed response scores for selected *Advanced Placement*® (*AP*®) exams remain invariant over subgroups by geographical region. The study focused on two questions: (a) how invariant were cut-scores across regions and (b) whether the small sample size for some regional groups presented particular problems for

assessing linking invariance. In addition to using the subpopulation invariance indexes to evaluate linking functions, Yang also evaluated the invariance of the composite score thresholds for determining final AP grades. Dorans (2004) used the population sensitivity of linking functions to assess score equity for two AP exams.

Dorans, Liu, and Hammond (2008) used population sensitivity indexes with SAT data to evaluate how consistent linear equating results were across males and females. Von Davier and Wilson (2008) examined the population invariance of IRT equating for an AP exam. Yang and Gao (2008) looked at invariance of linking computer-administered *CLEP*® data across gender groups.

SEA has also been used as a tool to evaluate score interchangeability when a test is revised (Liu & Dorans, 2013). Liu, Cahn, and Dorans (2006) and Liu and Walker (2007) used SEA tools to examine the invariance of linkages across the old and new versions of the SAT using data from a major field trail conducted in 2003. This check was followed by SEA analyses conducted on operational data (see studies cited in Dorans & Liu, 2009).

All these examples, as well as others such as Dorans, Holland, Thayer, and Tateneni (2003), are illustrations of using SEA to assess the fairness of a test score by examining the degree to which the linkage between scores is invariant across subpopulations. In some of these illustrations, such as one form of SAT Mathematics with another form of SAT Mathematics, the expectation of score interchangeability was very high since alternate forms of this test are designed to be parallel in both content and difficulty. There are cases, however, where invariance was expected but did not hold. Cook, Eignor, and Taft (1988), for example, found that the linking function between two biology exams depended on whether the equating was with students in a December administration, where most of the examinees were seniors who had not taken a biology course for some time, versus a June administration, where most of the examinees had just completed a biology course. This case, which has become an exemplar of lack of invariance where invariance would be expected, is discussed in detail by Cook (2007) and Petersen (2007). Invariance cannot be presumed to occur simply because tests are built to the same blueprint. The nature of the population can be critical, especially when diverse subpopulations are involved. For most testing programs, analysis that focuses on the invariance of equating functions should be conducted to confirm the fairness of the assembly process.

Limitations of Quantitative Fairness Assessment Procedures

First, not all fairness considerations can be reduced to quantitative evaluation. Because this review was limited to quantitative fairness procedures, it was limited in scope. With this important caveat in mind, this section will discuss limitations with the classes of procedures that have been examined.

Fair prediction is difficult to achieve. Differential prediction studies are difficult to complete effectively because there are so many threats to the subpopulation invariance of regression equations. Achieving subpopulation invariance of regressions is difficult because of selection effects, misspecification errors, predictor unreliability, and criterion issues. Any attempt to assess whether a prediction equation is invariant across subpopulations such as males and females must keep these confounding influences in mind.

To complicate validity assessment even more, there are as many external criteria as there are uses of a score. Each use implies a criterion against which the test's effectiveness can be assessed. The process of validation via prediction studies is an unending yet necessary task.

DIF screening is and has been possible to do. But it could be done better. Zwick (2012) reviewed the status of ETS DIF analysis procedures, focusing on three aspects: (a) the nature and stringency of the statistical rules used to flag items, (b) the minimum sample size requirements that are currently in place for DIF analysis, and (c) the efficacy of criterion refinement. Recommendations were made with respect to improved flagging rules, minimum sample size requirements, and procedures for combining data across administrations. Zwick noted that refinement of the matching criterion improves detection rates when DIF is primarily in one direction but can depress detection rates when DIF is balanced.

Most substantive DIF research studies that have tried to explain DIF have used observational data and the generation of post-hoc explanations for why items were flagged for DIF. The chapter by O'Neill and McPeek (1993) in the Holland and Wainer (1993) DIF book is a good example of this approach. As both those authors and Bond[6] (1993) noted, this type of research with observed data is fraught with peril because of the highly selected nature of the data examined, namely items that have been flagged for DIF. In the same section of the DIF book, Schmitt et al. (1993) provided a rare exemplar on how to evaluate DIF hypotheses gleaned from observational data with experimental evaluations of the hypotheses via a carefully designed and executed experimental manipulation of item properties followed by a proper data analysis.

DIF can be criticized for several reasons. An item is an unreliable measure of the construct of interest. Performance on an item is susceptible to many influences that have little to do with the purpose of the item. An item, by itself, can be used to support a variety of speculations about DIF. It is difficult to figure out why DIF occurs. The absence of DIF is not a prerequisite for fair prediction. In addition, DIF analysis tells little about the effects of DIF on reported scores.

SEA focuses on invariance at the reported score level where inferences are made about the examinee. SEA studies based on counterbalanced single-group designs are likely to give the cleanest results about the invariance of score linking functions because it is a data collection design that allows for the computation of correlations between tests across subpopulations.

This report focused primarily on studies that focused on methodology and that were conducted by ETS staff members. As a result, many DIF and differential prediction studies that used these methods have been left out and need to be summarized elsewhere. As noted, qualitative and philosophical aspects of fairness have not been considered.

In addition ETS has been the leader in conducting routine DIF analyses for over a quarter of century. This screening for DIF practice has made it difficult to find items that exhibit the high degree of DIF depicted on the cover of the Winter 2012 issue of *Educational Measurement: Issues and Practices,* an item that Dorans (2012) cited as a vintage example of DIF. Although item scores exhibit less DIF than they did before due diligence made DIF screening an operational practice, a clear need remains for continued research in fairness assessment. This includes improved methods for detecting evidence of unfairness and the use of strong data collection designs that allow researchers to arrive at a clearer understanding of sources of unfairness.

References

Allen, N. L., & Donoghue, J. R. (1995). *Application of the Mantel-Haenszel procedure to complex samples of items* (Research Report No. RR-95-04). Princeton, NJ: Educational Testing Service.

Allen, N., & Holland, P. H. (1993). A model for missing information about the group membership of examinees in DIF studies. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 241–252). Hillsdale, NJ: Erlbaum.

Allen, N. L., & Wainer, H. (1989). *Nonresponse in declared ethnicity and the identification of differentially functioning items* (Research Report No. RR-89-47). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the meeting of the American Psychological Association, Honolulu, HI. (ERIC Document Reproduction Service No. ED 069686).

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Erlbaum.

Angoff, W. H., & Cowell, W. R. (1985). *An examination of the assumption that the equating of parallel forms is population-independent* (Research Report No. RR-85-22). Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327–345.

Angoff, W. H., & Ford, S. F. (1971). *Item-race interaction on a test of scholastic aptitude* (Research Bulletin No. RB-71-59). Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–106.

Angoff, W. H., & Sharon, A. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34,* 807–816.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18,* 59–62.

Bennett, R. E. (2005). *What does it mean to be a nonprofit educational measurement organization in the 21st century?* Retrieved from http://wwwets.org/Media/Research/pdf/Nonprofit.pdf

Bond, L. (1993). Comments on the O'Neill & McPeek chapter. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–279). Hillsdale, NJ: Erlbaum.

Campbell, J. T. (1964). *Testing of culturally different groups* (Research Bulletin No. RB-64–34). Princeton, NJ: Educational Testing Service.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test* (Research Bulletin No. RB-64-61). Princeton, NJ: Educational Testing Service.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure* (Research Report No. RR-95-05). Princeton, NJ: Educational Testing Service.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–354.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cleary, T. A., & Hilton, T. L. (1966). *An investigation of item bias* (Research Bulletin No. RB-66-17). Princeton, NJ: Educational Testing Service.

Cleary, T. A., & Hilton, T. J. (1968). An investigation of item bias. *Educational and Psychological Measurement, 5,* 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 5,* 237–255.

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73–88). New York, NY: Springer Science & Business Media.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25,* 31–45.

Donoghue, J. R., & Allen, N. L. (1992). *"Thin" versus "thick" in the Mantel-Haenszel procedure for detecting DIF* (Research Report No. RR-92-76). Princeton, NJ: Educational Testing Service.

Donoghue, J. R., & Allen, N. L. (1993). “Thin” versus “thick” in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18,* 131–154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). *A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Erlbaum.

Dorans, N. J. (1982). *Technical review of item fairness studies: 1975–1979* (Statistical Report No. SR-82-90). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Dorans, N. J. (2004). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41,* 43–68.

Dorans, N. J. (2011). Holland’s advice during the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). New York, NY: Springer-Verlag.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37.

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

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

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (Research Report No. RR-03-27, pp. 79-118). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of*

the standardization approach (Research Report No. RR-83-09). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel-Haenzel and standardization procedures. *Medical Care, 44*(S3), S107–S114.

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., & Potenza, M. T. (1994). *Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning* (Research Report No. RR-94-49). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (Research Report No. RR-88-31). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29,* 309–319.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Gulliksen, H. O. (1964). Intercultural studies of attitudes. In N. Frederiksen & H. O. Gulliksen (Eds.), *Contributions to mathematical psychology* (pp. 61–108). New York, NY: Holt, Rinehart & Winston.

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

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (Research Report No. RR-86-31). Princeton, NJ: Educational Testing Service.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Lewis, C. (1993). Bayesian methods for the analysis of variance. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. II. Statistical issues* (pp. 233–256). Hillsdale, NJ: Erlbaum.

Linn, R. L. (1972). *Some implications of the Griggs decision for test makers and users* (Research Memorandum No. RM-72-13). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161.

Linn, R. L. (1975). *Test bias and the prediction of grades in law school* (Report No. LSAC-75-01). Newtown, PA: Law School Admissions Council.

Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13,* 53–58.

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. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Liu, J., Cahn, M., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linking of New SAT to Old SAT across gender groups. *Journal of Educational Measurement, 43,* 113–129.

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*(1), 15–22.

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York, NY: Springer Science & Business Media.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, The Netherlands: Swets and Zeitlinger.

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

Lyu, C. F., Dorans, N. J., & Ramsay, J. O. (1995). *Smoothed standardization assessment of testlet level DIF on a math free-response item type* (Research Report No. RR-95-38). Princeton, NJ: Educational Testing Service.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association,* 58, 690–700.

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

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008, March). *A review of recent developments in differential item functioning*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

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

O'Neill, K. O., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Erlbaum.

Pashley, P. J. (1992). *Graphical IRT-based DIF analysis* (Research Report No. RR-92-66). Princeton, NJ: Educational Testing Service.

Peterson, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York, NY: Springer-Verlag.

Petersen, N. S., & Novick, M. R. (1976). An evaluating of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Pomplun, M., Baron, P. A., & McHale, F. J. (1992). *An initial evaluation of the use of bivariate matching in DIF analyses for formula scored tests* (Research Report No. RR-92-63). Princeton, NJ: Educational Testing Service.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23–37.

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2007). *Small-sample DIF estimation using log-linear smoothing: A SIBTEST application* (Research Report No. RR-07-10). Princeton, NJ: Educational Testing Service.

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2009). Small-sample DIF estimation using log-linear smoothing: A SIBTEST Application. *Journal of Educational Measurement, 46,* 59–83.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Denmarks Paedagogiske Institut.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315), Hillsdale, NJ: Erlbaum.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias. *Psychometrika, 58,* 159–194.

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small-sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34,* 74–96.

Sinharay, S., Dorans, N. J., & Liang, L. (2009). *First language of examinees and its relationship to differential item functioning* (Research Report No. RR-09-11). Princeton, NJ: Educational Testing Service.

Stricker, L. J. (1982). Identifying test items that perform differently in population subgroups: A partial correlation index. *Applied Psychological Measurement, 6,* 261–273.

Tan, X., Xiang, B., Dorans, N. J., & Qu, Y. (2010). *The value of the studied item in the matching criterion in differential item functioning (DIF) analysis* (Research Report No. RR-10-13). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 63–70.

Tucker, L. R. (1951). *Academic ability test* (Research Memorandum No. RM-51-17). Princeton, NJ: Educational Testing Service.

Turnbull, W. W. (1949). Influence of cultural background on predictive test scores. In *Proceedings of the ETS invitational conference on testing problems* (pp. 29–34). Princeton, NJ: Educational Testing Service.

Turnbull, W. W. (1951a). *Socio-economic status and predictive test scores* (Research Memorandum No. RM-51-09). Princeton, NJ: Educational Testing Service.

Turnbull, W. W. (1951b). Socio-economic status and predictive test scores. *Canadian Journal of Psychology, 5,* 145–149.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32,* 11–26.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197–219.

Yang, W.-L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41,* 33–41.

Yang, W., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based CLEP examination. *Applied Psychological Measurement, 32,* 45–61.

Yu, L., Moses, T., Puhan, G., & Dorans, N. J. (2008). *DIF detection with small samples: Applying smoothing techniques to frequency distributions in the Mantel-Haenszel procedure* (Research Report No. RR-08-44). Princeton, NJ: Educational Testing Service.

Zhang, Y., Dorans, N. J., & Mathews-Lopez, J. (2005). *Using DIF dissection method to assess effects of item deletion* (Research Report No. 2005-10). New York, NY: The College Board.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zieky, M. (2011). The origins of procedures for using differential item functioning statistics at Educational Testing Service. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 115–127). New York, NY: Springer-Verlag.

Zwick, R. (1989). *When do item response function and Mantel-Haenszel definitions of differential item functioning coincide* (Research Report No. RR-89-32). Princeton, NJ: Educational Testing Service.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185–197.

Zwick, R. (1994). *The effect of the probability of correct response on the variability of measures of differential item functioning* (Research Report No. RR-94-44). Princeton, NJ: Educational Testing Service.

Zwick, R. (2012*). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment (with K. Ercikan). *Journal of Educational Measurement, 26,* 55–66.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21,* 187–201.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement, 26,* 57–76.

Zwick, R., Thayer D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (Research Report No. RR-97-21). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36, 1–28*.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25,* 225–247.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing DIF in polytomous items. *Applied Measurement in Education, 10,* 321–344.

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests* (Research Report No. RR-93-11). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121–140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32,* 341–363.

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel–Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics, 37*(5), 601–629.

Notes

[1] Thorndike was never an ETS staff member.

[2] Neither Shealy nor Stout was an ETS staff member.

[3] Baker was never an ETS staff member.

[4] Neither Mantel nor Haenszel was an ETS staff member.

[5] Neither Meredith nor Millsap was an ETS staff member.

[6] Bond was never an ETS staff member.

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

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

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

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.

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

ETS Psychometric Contributions: Focus on Test Scores
by Tim Moses (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-03