

Susan J. Maller and Lai-Kwan Pei

In the 1970s, concerns regarding potential bias of intelligence tests led to several court cases (e.g., *Diana v. the California State Board of Education* 1970; *Larry P. v. Wilson Riles* 1979), and studies of item bias, with conflicting findings (cf., Cotter and Berk 1981; Ilai and Willerman 1989; Jastak and Jastak 1964; Koh et al. 1984; Ross-Reynolds and Reschly 1983; Sandoval 1979; Sandoval et al. 1977; Turner and Willerman 1977). Bryk (1980) found methodological flaws in the above-mentioned mean score difference score definition and related item bias studies, noting that the current psychometric methodologies (e.g., latent trait theory) had not even been mentioned by Jensen (1980). However, studies using such methods continue to be promoted as evidence of bias (e.g., Braden 1999; Frisby 1998).

*Bias* refers to systematic error in the estimations of a construct across subgroups (e.g., males vs. females, minority vs. majority). All forms of bias eventually lead to a question of construct validity due to the potential influence of unintended constructs. The presence of bias ultimately suggests that scores have different

meanings for different subgroups. Bias can be investigated empirically at the item or test score levels. The various methods to investigate bias relate to the source of bias or differential validity (content, construct, and criterion-related).

*Fairness* is a more inclusive term and refers specifically to the (a) absence of bias, (b) equitable treatment of examinees during the testing process, (c) equitable test score interpretations for the intended uses, and (d) equitable opportunities to learn the content of the test (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] 2014). Clearly, there is no such thing as a “nonbiased test” or a test that “is fair” or “is valid” for all subgroups under all conditions. Furthermore, test developers can go to extensive lengths to create instruments that lack evidence of bias against subgroups; however, test consumers ultimately are responsible for selecting, administering, and interpreting the results of tests with evidence of validity for the purpose in which tests are used.

Various professional entities have developed guidelines related to fairness in testing. For example, the Standards for Educational and Psychological Testing (AERA, APA, & NCME 2014) devotes an entire chapter to “Fairness in Testing”. The section “fairness as lack of measurement bias” states that mean score differences are insufficient evidence of bias. When mean score differences are found for subgroups, construct irrelevant variance, or construct underrepresentation should be investigated as an explanation. Construct irrelevant

---

S.J. Maller (✉)

Department of Educational Studies, Purdue  
University, West Lafayette, IN 47907, USA  
e-mail: maller@purdue.edu

L.-K. Pei

Houston Independent School District, 4400 West  
18th Street, Houston, TX 77092-8501, USA  
e-mail: lpei@houstonisd.org

variance may occur as a function of test development, administration, and scoring procedures. Four guidelines are provided to help test developers and users to minimize construct irrelevant variance and ensure the validity of the test and test score interpretation.

Code of Professional Responsibilities in Educational Measurement (National Council on Measurement in Education [NCME] 1995) states that those who develop assessments are responsible for making their products “as free as possible from bias due to characteristics irrelevant to the construct being measured, such as gender, ethnicity, race, socioeconomic status, disability, religion, age, or national origin” (Sect. 1.2a).

Code of Fair Testing in Education (Joint Committee on Testing Practices [JCTP] 2004) Section A states that test developers should obtain and provide evidence on the performance of test takers of diverse subgroups, and evaluate the evidence to ensure that differences in performance are related to the skills being assessed, while test users should evaluate the available evidence on the performance of test takers of diverse subgroups, and determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed.

Test publishers routinely enlist the assistance of experts in the test content domain to conduct sensitivity reviews or evaluate the items for content unfairness, including offensive language, insensitivity, or other content that may have unintended influences on the performances of members of various subgroups. Panel reviews of the item contents in several achievement and scholastic aptitude tests have tied differential item performance to differences in opportunities to learn or differences in socialization. For example, items favoring females have been linked to specific topics involving humanities, esthetics, human relationships, whereas items that favoring males have been linked to contents about science, sports, mechanics (Lawrence and Curley 1989; Lawrence et al. 1988; Scheuneman and Gerritz 1990; Wild and McPeck 1986). Unfortunately, panel reviews of the item content

bias have neither yielded consistent nor accurate results (Engelhard et al. 1990; Plake 1980; Sandoval and Miille 1980).

To study whether a test is biased toward specific groups, the psychometric properties of the test can be investigated for invariance (equality) across groups. The type of the invariance investigation depends on the suspected nature of bias and can include a variety of methods to (a) detect differential item functioning (DIF), and (b) examine measurement invariance. Item bias/DIF detection examines the characteristics of the test and item itself to check whether the test/items are measuring irrelevant construct. Measurement invariance refers to whether the scale is measuring the same construct at different occasions or across different groups.

---

### Item Bias and Differential Item Functioning (DIF)

Although the terms item bias and differential item functioning (DIF) are often used interchangeably, the term DIF was suggested (Holland and Thayer 1988) as a somewhat neutral term to refer to differences in the statistical properties of an item between groups of examinees of equal ability. These groups are often referred to as the reference (e.g., majority) and focal (e.g., minority) groups. DIF detection methods “condition on” or control for ability, meaning that examinees are necessarily matched on ability; thus, only examinees of equal ability (e.g., overall test score) in the reference and focal groups are compared. The item that is being tested for DIF is referred to as the study item. There are two types of DIF: uniform DIF and non-uniform DIF. Uniform DIF, also called unidirectional DIF, occurs when an item favors one group over another across all ability levels. Alternatively, non-uniform DIF, also called crossing DIF, occurs when an item discriminates across the ability levels differently for the groups. Items that exhibit DIF threaten the validity of a test and may have serious consequences for

groups as well as individuals, because the probabilities of correct responses are determined not only by the trait that the test claims to measure, but also by factors specific to group membership, such as ethnicity or gender. Thus, it is critical to identify DIF items in a test.

Numerous methods have been proposed for detecting DIF. The methods can be classified into two groups, depending on whether the method is based on item response theory (IRT). In non-IRT DIF detection methods, the observed total score is usually used to indicate the examinee's ability level. Non-IRT methods are better than IRT methods when the sample size is small, because they do not require item parameter estimation. However, without item parameter estimation, it is more difficult to figure out the source of DIF when an item is flagged as a DIF item. Non-IRT detection methods include: the (a) Mantel–Haenszel procedure (Holland and Thayer 1988; Mantel and Haenszel 1959); and (b) logistic regression modeling (Zumbo 1999); and (c) SIBTEST (Shealy and Stout 1993).

In IRT DIF detection methods, the probability of a correct response to an item is assumed to follow an IRT model. The examinee ability levels and item parameters are estimated based on item responses. DIF detection is then performed by comparing the estimated models for the reference and focal groups. IRT methods require larger sample size and more computational load for parameter estimation. However, with the known item parameters, the test developers can learn more about the source of the DIF and revise the item/test. IRT DIF detection methods include: (a) Lord's chi-square test (Kim et al. 1995; Lord 1980); (b) area method (Raju 1988, 1990); and (c) IRT likelihood ratio test (Thissen et al. 1988, 1993).

In the following sections, the more popular DIF detection methods, including Mantel–Haenszel procedure, logistic regression modeling, SIBTEST, and IRT likelihood ratio test, are described. Details of the other methods, as well as some older methods not mentioned above, can be found in the overviews given by Camilli and Shepard (1994), Clauser and Mazor (1998), Holland and Wainer (1993), Millsap and Everson

(1993), Osterlind and Everson (2009), and Penfield and Camilli (2007).

## Ability Matching

When detecting DIF, only examinees of equal ability in the reference and focal groups are compared. Thus, ability matching is very important in DIF detection. For example, if examinees with different ability levels are matched by mistake, then a non-DIF item could be flagged incorrectly as a DIF item. If external criterion is not available to match the reference and focal groups, then the matching has to be performed with the item responses of the study items. Because the inclusion of DIF items in the matching step would likely result in incorrect matches, a purification step is usually used to remove DIF items that might contaminate the matching criterion. The remaining DIF-free items, also known as *anchor items*, can then be used in ability matching. The purification is usually performed with an initial Mantel–Haenszel procedure, in which the observed total score from all items is used to match ability levels.

## Mantel–Haenszel Procedure (MH)

Mantel and Haenszel (1959) introduced a procedure to study matched groups. Holland (1985) and later Holland and Thayer (1988) adapted the procedure for detection of DIF. The MH procedure compares the odds of a correct response of the reference and focal groups on a dichotomous item. For each ability level, a contingency table is constructed for the study item, resulting in  $J \times 2 \times 2$  tables, where  $J$  is the number of ability levels. Each cell of the table indicates the frequency of correct/incorrect responses to the item for reference/focal group. An example of contingency table is shown in Table 2.1.

Under the null hypothesis of no DIF, the two groups have the same odds of getting a correct response in all ability levels, i.e.,  $A_j/B_j = C_j/D_j$  for all ability levels  $j$ . The chi-square statistic for the null hypothesis is

**Table 2.1** Example of a  $2 \times 2$  contingency table for an item at ability level  $j$ 

	1 (correct response)	0 (incorrect response)	Total
Reference group	$A_j$	$B_j$	$N_{Rj}$
Focal group	$C_j$	$D_j$	$N_{Fj}$
Total	$N_{1j}$	$N_{0j}$	$N_j$

$$MH - \chi^2 = \frac{\left[ \left| \sum_j (A_j - E(A_j)) \right| - 0.5 \right]^2}{\sum_j \text{var}(A_j)}, \quad (2.1)$$

where

$$E(A_j) = \frac{N_{Rj}N_{1j}}{N_j} \quad \text{and} \\ \text{var}(A_j) = \frac{N_{Rj}N_{Fj}N_{1j}N_{0j}}{(N_j)^2(N_j - 1)}.$$

The statistic follows chi-square distribution with one degree of freedom. The  $-0.5$  term in the statistic is a continuity correction, which is supposed to improve accuracy of Type I error (Holland and Thayer 1988). Note that when detecting DIF on a study item, examinees are matched on the purified subtest and the study item. Although it may be counter-intuitive to include the study item in the matching, exclusion of the studied item would change the calculation of the test statistic (Holland and Thayer 1988) (Table 2.2).

An additional statistic can be used with the MH procedure to facilitate interpretation of DIF by taking the natural logarithm of the chi-square statistic. Zieky (1993) suggested multiplying  $\ln(MH - \chi^2)$  by  $-2.35$ , denoted as  $\delta$ , resulting in a statistic that centers at zero and ranges from  $-4$  to  $+4$ . Educational testing service (ETS) developed a scheme for classifying a dichotomous item into one of three categories of DIF: A (negligible), B (slight to moderate), and C (moderate to severe). The classification guidelines are as follows (Dorans and Holland 1993; Zieky 1993):

Level	$\delta$
A	$ \delta  < 1.0$
B	$1.0 <  \delta  < 1.5$
C	$ \delta  > 1.5$

Samples of 100 examinees are adequate for the MH procedure (Hills 1989), and as small as 50 examinees in the reference group and 10 examinees in the focal group have been suggested for MH DIF screening (Kromrey and Parshall 1991). The MH procedure is not designed to detect non-uniform DIF. If non-uniform DIF is present, the term  $A_j - E(A_j)$  in Eq. (2.1) is positive for some ability levels and negative for the others. The statistic will then be small because of cancellation, giving a false conclusion of no DIF. To detect non-uniform DIF, the modified version proposed by Mazor et al. (1994) can be used. The MH procedure has been extended to detect DIF for polytomous items, and to detect DIF for multiple groups simultaneously (Penfield 2001; Zwick et al. 1993). The MH procedure can easily and quickly be run in statistical analysis packages such as SAS, SPSS, Stata, R, and Systat.

### Logistic Regression DIF Detection Method (LR DIF)

Unlike the MH procedure, LR DIF can be used to test non-uniform DIF directly (Rogers and Swaminathan 1993; Swaminathan and Rogers 1990). In LR DIF, the probability of a correct response to an item follows a logistic regression model:

$$P(x = 1 | \theta, g) = \frac{\exp(\beta_0 + \beta_1\theta + \beta_2g + \beta_3(\theta g))}{1 + \exp(\beta_0 + \beta_1\theta + \beta_2g + \beta_3(\theta g))}, \quad (2.2)$$

where  $x$  is the item response,  $\theta$  is the ability level, and  $g$  is the group membership, which is usually coded as 1 for reference group and 0 for focal

**Table 2.2** Summary of bias studies for several nonverbal intelligence tests

Test	Item bias/DIF	Factor invariance	Differential prediction	Other
CTONI-2	Logistic regression was used to detect DIF across gender, race, and ethnicity. No DIF item was found	No examination was conducted	<p>Sensitivity and specificity indices between CTONI-2 and TONI-4, PTONI were larger than 0.70, which meet acceptability Level I when predicting TONI-4 and PTONI</p> <p>Correlation coefficients were used to be the predictive validity for achievement test</p> <p>Differential prediction across subgroups was not investigated</p>	The reliability coefficients are consistently large for a general population and subgroups: gender, racial, and exceptionality categories
Leiter-3	Correlations between separately calibrated difficulty parameters of 1-PL IRT model were used to detect DIF across race and ethnicity. Two DIF items were found	No examination was conducted	<p>Correlation between Leiter-3 and the California Achievement Test was used as the predictive validity coefficient. It was not significant for samples of 50 Caucasians/non-Hispanic and 20 African Americans</p> <p>Differential prediction across subgroups was not investigated</p>	The lack of mean score differences was used as evidence of fairness, although a few differences were found between the following: Caucasians/Hispanics, Normative sample members/Navajos, and males/females
UNIT2	Logistic regression was used to detect DIF across gender, race, and ethnicity. No DIF item was found according to Jodoin and Gierl's (2001) criteria	CFA for various subgroups, including males, females, African Americans, Hispanics, all showed good fit to the data, although hypotheses regarding invariance were not tested using simultaneous, multisample CFA	<p>The correlations between the UNIT2 and seven criterion measures was used as the predictive validity</p> <p>The correlations between the UNIT2 composites and the criterion tests range from moderate to nearly perfect</p> <p>However, the predictive validity across subgroups were not examined</p>	<p>Mean score comparisons of several subgroups indicated some differences between groups</p> <p>Internal consistency reliability coefficients were consistently high across subgroups (males, females, African Americans, Hispanics)</p>

(continued)

**Table 2.2** (continued)

Test	Item bias/DIF	Factor invariance	Differential prediction	Other
TONI-4	Logistic regression was used to detect DIF across gender, race, and ethnicity. One DIF item was found	No examination was conducted	Predictive validity was measured by the correlations between TONI-4 and two selected criterion measures, TONI-3 and CTONI-2  Differential prediction across subgroups was not investigated	Internal consistency reliability coefficients were uniformly high across subgroups (gender, race/ethnicity, intellectual ability, English language proficiency)
Wechsler Nonverbal Scale of Ability (WNV)	No DIF detection was conducted	No examination was conducted	Correlations between corresponding subtests on the WNV and other selected cognitive tests were moderate  Differential prediction across subgroups was not investigated	The authors indicated that the majority of the subtests reliability coefficients across special groups were similar to, or higher than, those of the U.S. normative sample, and claimed that the WNV was an equally reliable tool for this population

group. Equation (2.2) can be written as an additive function by taking the log odd ratio:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1\theta + \beta_2g + \beta_3(\theta g), \quad (2.3)$$

where the parameters  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  represent the intercept, slopes for ability, membership and the interaction term (ability  $\times$  group), respectively. The item exhibit non-uniform DIF if  $\beta_3 \neq 0$ , and exhibit uniform DIF if  $\beta_2 \neq 0$  and  $\beta_3 = 0$ . If  $\beta_2 = \beta_3 = 0$ , then the item does not exhibit DIF.

To test whether the item exhibits non-uniform DIF, two different models are fitted to the data, yielding two likelihood ratio chi-squares. The *compact* model only has the first three terms of Eq. (2.3), while the *augmented* model has all the terms. Because chi-squares are additive, the explanatory power of the interaction term can be tested by subtracting the likelihood ratio of the less restrictive (augmented) model from the more restrictive (compact) model, yielding a difference chi-square with one degree of freedom. If the difference is significant, the interaction term is necessary, and the item is concluded to exhibit non-uniform DIF. Otherwise, the item is tested for uniform DIF, in which a compact model including only the first two terms ( $\beta_0$ ,  $\beta_1$ ), and an augmented model including the first three terms ( $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ), are fitted to the data. If the difference chi-square between the compact and augmented models is significant, the item is concluded to exhibit uniform DIF, and the direction of uniform DIF is indicated by the sign of  $\beta_2$ . Uniform DIF favors the reference group when  $\beta_2 > 0$ , and favors the focal group when  $\beta_2 < 0$ . Zumbo and Thomas (1997) proposed to use the difference between Nagelkerke's  $R^2$  (1991) of two logistic models, denoted  $\Delta R^2$ , to be the effect size of DIF. They provided the following interpretation of DIF: (a) negligible DIF if  $\Delta R^2 \leq 0.13$ ; (b) moderate DIF if  $0.13 < \Delta R^2 \leq 0.26$ ; and (c) large DIF if  $\Delta R^2 > 0.26$ . LR DIF is widely available in statistical software likes SAS, SPSS, Stata, R, and Systat. The method has been extended to

polytomous items and to multiple groups (Agresti 1996; Magis et al. 2011).

## SIBTEST Procedure

Simultaneous item bias test (SIBTEST) was developed by Shealy and Stout (1993) to detect and estimate DIF. The procedure was later extended to polytomous items (Chang et al. 2005). SIBTEST tests the following hypothesis:

$$H_0 : \beta_{\text{uni}} = 0 \text{ versus } H_1 : \beta_{\text{uni}} \neq 0,$$

where  $\beta_{\text{uni}}$  is the parameter specifying the magnitude of unidirectional DIF.  $\beta_{\text{uni}}$  is defined as:

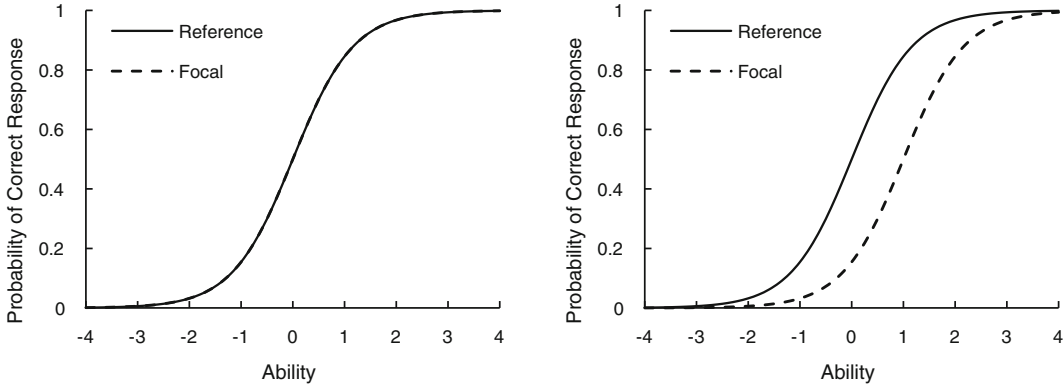
$$\beta_{\text{uni}} = \int_0^1 d(\theta) f_F(\theta) d\theta, \quad (2.4)$$

where  $d(\theta) = P(x = 1|\theta, F) - P(x = 1|\theta, R)$  is the difference in probability of correct response at ability  $\theta$ , and  $f_F(\theta)$  is the density function of ability in the focal group. When the reference and focal groups have the same ability distribution, the observed total score is an unbiased estimator of ability, and  $\beta_{\text{uni}}$  can be estimated by

$$\hat{\beta}_{\text{uni}} = \sum_{j=0}^J p_j (\bar{Y}_{Fj} - \bar{Y}_{Rj}), \quad (2.5)$$

where  $\bar{Y}_{gj}$  is the average score on the study item for the group  $g$  examinees with observed score  $j$ ,  $p_j$  is the proportion of examinees with observed score  $j$ , and  $J$  is the total number of items (Bolt and Stout 1996). In practice, the two groups usually have different ability distributions, and the observed total score is a biased estimator of ability. To adjust the estimation bias, Shealy and Stout (1993) introduced regression correction step into SIBTEST procedure to correct for mean difference in the ability distribution of the reference and focal groups. The regression correction step was later improved by Jiang and Stout (1998) for better control of Type I error inflation. Replacing the observed item score  $\bar{Y}_{gj}$  in





**Fig. 2.1** Item characteristic curves for non-DIF (*left*) and DIF (*right*) items

Eq. (2.5) with the adjusted item score  $\bar{Y}_{gj}^*$  yields an unbiased estimator for the DIF size regardless of the difference in the ability distribution of two groups. Positive values of  $\hat{\beta}_{\text{uni}}$  indicate DIF favoring the reference group and negative values indicate DIF favoring the focal group. The test statistic for the null hypothesis of no DIF is then given by

$$B_{\text{uni}} = \frac{\hat{\beta}_{\text{uni}}}{\hat{\sigma}(\hat{\beta}_{\text{uni}})},$$

where  $\hat{\sigma}(\hat{\beta}_{\text{uni}})$  is the standard error of the estimator  $\hat{\beta}_{\text{uni}}$ .  $B_{\text{uni}}$  follows a normal distribution with mean 0 and standard deviation 1 under the null hypothesis (Shealy and Stout 1993).

SIBTEST was designed to detect uniform DIF. If non-uniform DIF is present, the term  $d(\theta)$  in Eq. (2.4) changes sign at a certain ability. The magnitude of  $\beta_{\text{uni}}$  will then be small because of cancelation, giving a false conclusion of no DIF. To address this problem, crossing simultaneous item bias test (CSIBTEST) was developed by Li and Stout (1996) to detect crossing DIF. Unfortunately, the distribution of the test statistic in CSIBTEST cannot be derived easily, and a randomization test has to be used to determine statistical significance. SIBTEST is the computer program for this DIF detection method (Li and Stout 1994).

## IRT-Based DIF Detection

One problem of non-IRT DIF detection methods is the use of the observed score as an indicator of ability level, which may not be reliable. For example, both theoretical studies and simulation studies showed that when the item responses are generated by complex IRT models, the MH procedure can falsely indicate DIF when no bias is present (Meredith and Millsap 1992; Millsap and Meredith 1992; Uttaro 1992; Zwick 1990). In IRT models, ability is conceptualized as a latent trait. The probability of a correct response to an item for a given ability is given by the *item characteristic curve* (ICC). Figure 2.1 shows the ICCs for non-DIF and DIF items. The first set of ICCs, which are the same for the reference and focal groups, shows that the item does not exhibit DIF. The second set of ICCs is for an item that exhibit uniform DIF, because the reference group always has a higher probability of correct response than the focal group. If an item exhibits non-uniform DIF, then the ICCs will cross each other.

IRT models commonly used to investigate DIF in intelligence tests include the one-, two-, and three-parameter models, as well as Samejima's (1969) graded response model. In the two-parameter logistic (2-PL) model, the probability of correct response is for an examinee with ability  $\theta$  is:



$$P(x = 1|\theta) = \frac{1}{1 + e^{-a(\theta-b)}},$$

where  $x$  is the item response,  $a$  is the item discrimination parameter (proportional to the slope of the ICC), and  $b$  is the item difficulty parameter (at which the examinee has a 50% probability of correct response). The one-parameter logistic (1-PL; also known as Rasch) model differs from the 2-PL model in that the discrimination parameter is held constant across items. This is a very stringent assumption that rarely can be met in practice. However, examination of fit statistics can indicate whether the assumption is met. Regardless, if sufficient sample sizes are available, the 2-PL model is generally preferable to test the invariance of item discriminations across groups. A three-parameter logistic (3-PL) model is recommended for multiple choice items, because the model includes a guessing parameter  $c$ . The parameter ranges from 0 to 1, but is typically  $<0.3$ . The 3-PL model is defined as:

$$P(x = 1|\theta) = c + \frac{1 - c}{1 + e^{-a(\theta-b)}}.$$

When items are scored using necessarily ordered categories, they can be fitted with Samejima's graded response model (Samejima 1969). For example, for an item scored 0, 1, or 2, the graded response model provides two item difficulty estimates (based on the probability of scoring 1 or the probability of scoring 2). The graded response model is as follows:

$$P_k^*(\theta) = P(x \geq k|\theta) = \frac{1}{1 + e^{-a(\theta-b_k)}},$$

where  $P_k^*(\theta)$  is the probability of an examinee with ability  $\theta$  reaching category  $k$  or higher, and  $b_k$  is the difficulty parameter in reaching category  $k$ . For an examinee with ability  $\theta$ ,  $P_0(\theta) = 1 - P_1^*(\theta)$  is the probability of scoring 0,  $P_1(\theta) = P_2^*(\theta) - P_1^*(\theta)$  is the probability of scoring 1, and  $P_2(\theta) = P_2^*(\theta)$  is the probability of scoring 2.

## IRT Likelihood Ratio Test (IRT-LR)

IRT-based likelihood ratio test for DIF is designed to determine whether the ICC of the study item differs for the reference and focal groups. The method used the likelihood ratio test statistic to test the null hypothesis that the item parameters of the study item do not differ between groups. In this method, two models are fitted for the anchor items and the study item. In the *free* model, all parameters for the anchor items are constrained to be equal across groups, whereas the parameters for the study item are not. The *constrained* model poses an additional equality constraint on one of the parameters for the study item, such as lower asymptote parameter, discrimination parameter, or difficulty parameter. The likelihood goodness-of-fit statistic,  $G^2$ , is then used to test the hypothesis that the parameter estimate is invariant across groups:

$$G^2 = 2 \sum_{g \in \{R, F\}} \sum_{\mathbf{x}} n_g(\mathbf{x}) \cdot \ln \left( \frac{P_{\text{free}}(\mathbf{x}|g)}{P_{\text{constrained}}(\mathbf{x}|g)} \right),$$

where  $g$  is the group (reference or focal);  $\mathbf{x}$  is a response pattern;  $n_g(\mathbf{x})$  is the count for pattern  $\mathbf{x}$  in group  $g$ ;  $P_{\text{free}}(\mathbf{x}|g)$  and  $P_{\text{constrained}}(\mathbf{x}|g)$  are the probabilities of pattern  $\mathbf{x}$  under the free and constrained models, respectively. The statistic follows the chi-square distribution approximately with degree of freedom equal to the difference in the number of free parameters in the two models. IRT-LR test can be carried out with IRTLRDIF (Thissen 2001).

## Test Bias

Evidence of *Test bias* is reflected in test/subtest scores if there is differential validity as a function of group membership. Investigations of test bias usually include studies of (a) unequal psychometric properties, (b) unequal factor structures, or (c) differential prediction of performance between groups. Traditionally, test developers

and consumers believed that special subgroup norms may be useful for comparing individuals to a more representative peer group. For example, special norms were developed for Wechsler Intelligence Scale for Children—Revised Performance Scale for deaf children (Anderson and Sisco 1977). However, subgroup norms may be a superficial solution to a larger problem concerning content and construct validity. If test items have different meanings for examinees belonging to different subgroups, then subgroup norms result in comparing members to other members on some trait not claimed to be measured by the test (Maller 1996).

Differences in reliability coefficients also may indicate bias. Reliability coefficients provide an indication of how consistently a construct, such as intelligence, is measured across groups. Statistical tests are used to assess differences in the reliability coefficients (Feldt and Brennan 1989). Differences found in the internal consistency coefficients between groups may indicate bias. However, differences in the test–retest and alternate forms coefficients may also be a result of the time between testings (test–retest) or nonequivalent forms (alternate forms) and not a result of bias.

## Factor Invariance

*Construct equivalence* suggests that test constructs are conceptualized and measured similarly across groups (Shelley-Sireci and Sireci 1998; Sireci et al. 1998). Factor analytic methods are used to examine the internal structure of a test and to investigate whether a construct is equally indicated for groups. Exploratory (EFA) and confirmatory (CFA) factor analyses are used to examine the similarity of the factor structures. In EFA, the *coefficient of congruence*, a type of correlation, is used to determine the similarity of the factor loadings for groups. Values above 0.90 indicate factor invariance, meaning factors are equivalently indicated across groups and provides evidence against test bias (Cattell 1978).

Reynolds (1982) stated “bias exists in regard to construct validity when a test is shown to

measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with different degrees of accuracy” (p. 194). Furthermore, Reynolds added that multisample CFA based on the techniques of Jöreskog (1971) is a more promising and sophisticated method in detecting such construct bias than the method of exploratory factor analysis, which examines factorial similarity using the coefficient of congruence.

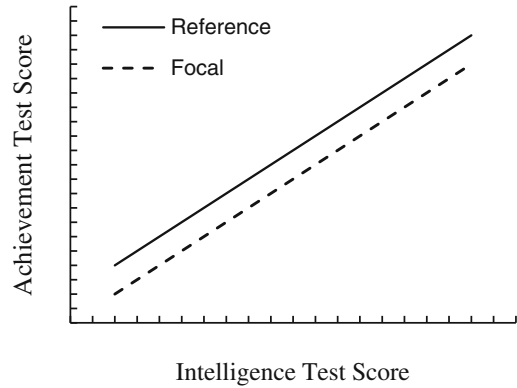
Multisample CFA has been used to test the invariance of factor structures (Alwin and Jackson 1981; Bollen 1989; Jöreskog and Sörbom 1989; Jöreskog 1971; McGaw and Jöreskog 1971). Following the procedures recommended by Bollen (1989) and Jöreskog and Sörbom (1989), the general form (hereafter referred to as the Model<sub>baseline</sub>) of the theoretical model is tested for invariance across samples which equal the sum of the chi-squares for the individual group analyses, and to obtain fit statistics of the model across groups. To assess the fit of the model, the following fit indices can be used: GFI, TLI, CFI, and RMSEA. The GFI is interpreted as the proportion of the observed variances and covariances that can be accounted for by the model. The TLI is recommended by Tucker and Lewis (1973), with values greater than or equal to 0.90 indicating reasonable fit. The CFI is recommended by Bentler (1990, 1992) and Rigdon (1996) to indicate the difference in fit of the null and target models relative to the fit of the null model, with values greater than or equal to 0.90 indicating reasonable fit. The RMSEA is recommended by Browne and Cudeck (1993) and Rigdon (1996) to indicate the fit of the empirical and modeled variance-covariance matrices, with values less than 0.05 indicating excellent fit and values less than 0.08 indicating reasonable fit (Rigdon 1996). In addition, the Satorra–Bentler scaled chi-square (Satorra and Bentler 1988) also might be examined, because it has been reported to be reliable for various distributional conditions and sample sizes (Hu et al. 1992).

If the general form does not fit across groups, test constructs are measured differently across the groups and a more exploratory approach might be

taken to reveal a model that fits the data. These approaches may include exploratory factor analytic studies or model fitting approaches in CFA. If, however, the general form of the model adequately fits across groups, progressively more restrictive models are then tested for invariance. Three progressively more restricted models may be tested by adding one additional constrained matrix of: (a) factor loadings or path coefficients, describing the relationships between the latent and observed variables and are interpreted like regression coefficients, (b) error variances, and (c) factor variances and covariances. The chi-squares for each of the restricted models,  $Model_{nested}$ , are compared to the chi-square for the  $Model_{baseline}$ , using a difference chi-square test, which involves subtracting the  $Model_{baseline}$  chi-square from the chi-square obtained for the restricted model, with degrees of freedom equal to the degrees of freedom for the  $Model_{nested}$  minus the degrees of freedom for the  $Model_{baseline}$ .

Factor loading invariance is the most critical concern regarding construct validity, because factor loadings indicate the relationship between the observable item response and factor (construct). If the matrix of factor loadings is not invariant, at least one element of the matrix lacks invariance, individual elements of the matrix subsequently should be individually tested for invariance to isolate the source(s) of invariance (Maller and Ferron 1997; Maller et al. 1998). The restricted model is the  $Model_{nested}$  with one equality constraint of the studied parameter. The chi-square difference is obtained by comparing the restricted and  $Model_{baseline}$  chi-squares with one degree of freedom. A lack of factor loading invariance suggests that factor loadings should not be constrained to be invariant when testing the invariance of error variances and factor variances and covariances. In fact, a lack of factor loading invariance is sufficient to lead to the conclusion of differential validity.

If the factor loadings are invariant, the matrix of error variances should be tested for invariance. If the matrix is not invariant, individual elements subsequently can be tested for invariance, as described above. A lack of error variance invariance suggests that the measurement of the



**Fig. 2.2** Regression lines for reference and focal groups where intelligence scores under-predict achievement test scores for the focal group

variables (subtests) is differentially affected by extraneous sources of variance.

If factor covariances are found to lack invariance, differential variability in the factors may be the source of invariance, resulting in smaller or greater redundancy in the constructs claimed to be measured by the factors. In other words, the “separate” factors may be measuring overlapping abilities for one of the groups.

If factor variances and covariances are invariant, it makes sense to do a follow-up test of the invariance of means structures to investigate whether the latent means differ across groups. A lack of invariance suggests that, although the measurement of test constructs do not differ, the groups differ in terms of ability.

## Prediction Bias

The examination of differential predictive validity is especially important when tests are used for placement and selection decisions. Differential prediction has been used as an indication of test bias (Cleary 1968). Predictive validity coefficients that significantly differ between groups indicate that the test has different relationships with the criterion across the groups. Another type of differential prediction refers to a systematic under or overestimation of a criterion for a given group (Cleary 1968; Scheuneman and Oakland

1998). Specifically, differential prediction occurs when examinees belonging to different subgroups, but with comparable ability based on some predictor test score, tend to obtain different scores on some criterion test. To investigate differential prediction, regression lines for criterion (e.g., intelligence) and predictor (e.g., achievement) test scores are compared for reference and focal groups.

Figure 2.2 depicts an example of regression lines with different intercepts. The criterion is underpredicted for the focal group through achievement test. Suppose the achievement test in Fig. 2.2 is required for admission to a gifted education program. Members of the focal group actually will obtain lower scores on the achievement test than would be expected based on their intelligence scores, when using the regression line for the reference group. Focal group members who would be successful on the criterion may be denied acceptance into the gifted program based on their achievement test scores. A test that does not exhibit differential predictive validity still may be biased based on other definitions of bias. Furthermore, predictor and criterion tests may be spuriously correlated due to systematic factors, including construct bias. That is, factors specific to group membership that similarly affect scores on both tests may actually inflate predictive validity coefficients. Consistent with Messick's (1989) concerns, this method is not recommended in the absence of other bias investigations related to construct validity.

---

## Current Status and Recommendations

The best practices in detecting bias in nonverbal tests are really no different from the best practices for detecting bias in other psychoeducational tests. Until recently, there were few published studies of invariance at the item or test levels in intelligence tests using state-of-the-art methods, though these methods have been used for quite some time to study bias in various scholastic aptitude tests (e.g., Dorans and Kulick 1983;

Green et al. 1989; Holland and Thayer 1988; Linn et al. 1981; Scheuneman 1987). Recently, nonverbal and verbal intelligence test manuals and independent researchers have begun to report investigations of DIF and factor invariance. However, some popular test manuals do not include DIF investigation, such as Wechsler Nonverbal Scale of Ability technical manual (WNV; Wechsler and Naglieri 2006).

The comprehensive test of nonverbal intelligence—second edition manual includes a report of DIF analysis for three dichotomous groups (male vs. female, African American vs. non-African American, Hispanic vs. non-Hispanic) (CTONI-2; Hammill et al. 2009). Using the entire normative sample as subjects, the LR DIF approach was applied to all items contained in each of the CTONI-2 subtests. Of the 150 items, at least 24 were found to be statistically significant at the 0.001 level, but had negligible effect sizes according to Jodoin and Gierl's (2001) criteria ( $\Delta R^2 < 0.035$ ).

The Leiter International Performance Scale-3 manual includes a report of DIF analysis for two dichotomous groups (Caucasian vs. African American, Anglo vs. Hispanic) (Leiter-3; Roid and Miller 2013). For each item, the difficulty parameters for the 1-PL IRT model were derived separately for each ethnic/racial sample. The correlations between difficulty parameters were then used to indicate the uniformity of indices across groups. Out of the 152 items tested, 2 items were found to departed slightly from the linear trend in the scatter plots. However, this method suffers from at least two flaws. First, no mention was made regarding whether item difficulty estimates were placed on a similar scale. Second, like traditional methods, this method used a summary statistic, ignoring the functioning of specific items.

The Universal Nonverbal Intelligence Test—Second Edition manual includes a report of DIF analysis for three dichotomous groups (male vs. female, African American vs. non-African American, Hispanic vs. non-Hispanic) (UNIT2; Bracken and McCallum 2016). The LR DIF approach was applied to all items contained in

each of the UNIT2 subtests. Of the 241 items, 25 were found to be statistically significant at the 0.001 level, but had negligible effect sizes according to Jodoin and Gierl's (2001) criteria. The manual also reports a multigroup invariance study across gender, race, and ethnic groups. The TLI, CFI, and RMSEA fit indices were reported for four different models, with TLI and CFI values greater than 0.90, and RMSEAs of less than 0.12.

The Test of Nonverbal Intelligence–Fourth Edition manual includes a report of DIF analysis for three dichotomous groups (male vs. female, African American vs. non-African American, Hispanic vs. non-Hispanic) (TONI-4; Brown et al. 2010). The LR DIF approach was applied to all items contained in TONI-4. Of the 120 items, at least 5 were found to be statistically significant at the 0.001 level, with 1 item found to have moderate effect size according to Jodoin and Gierl's (2001) criteria.

The Wechsler Intelligence Scale for Children–Fifth Edition technical manual states that MH DIF analysis and IRT-LR approach were used to examine DIF across race (WISC-V; Wechsler 2014). However, no details were provided on specific items in terms of results. A study of invariance across age groups with a five-factor higher order models was reported in the technical manual. However, Canivez and Watkins (in press) was not able to replicate the five-factor baseline structural model in WISC-V, which was used for invariance study in the technical model. Therefore, the conclusion of the invariance study in the technical manual may be questionable. Besides, to capture the bias of the test, the invariance study should be conducted across gender, race groups instead of age group to ensure the test is free of bias against any one minority group.

A test may contain considerable DIF, yet focal and reference groups may have similar score distributions due to cancelation DIF, which occurs when some items favor the reference group and other favor the focal group. Scores may be based in part on different items systematically scored as correct. Although some might believe that DIF cancelation results

in a fairer test, the presence of even a one point systematic raw score difference on individual subtests due to DIF may result in systematic age-based standard score differences at the subtest level and may have cumulative effects at the scale score level for individuals. Furthermore, when ceiling rules are used and numerous adjacent items exhibit DIF against one group, individual examinees may reach a ceiling for reasons related to both group membership and intelligence. It is very likely that different items systematically scored as correct comprise the scores of examinees from different groups with the same test scores.

The scores from tests that lack item or test invariance cannot be assumed to have the same meaning across groups. Differential prediction studies are not recommended in the absence of DIF and factor invariance investigations, because tests may be correlated due to construct irrelevant factors. Thus, bias studies should begin with DIF studies, move to factor invariance studies, and conclude with differential prediction studies. The results of bias studies are crucial to the interpretation of test scores. A lack of item and test score invariance can be a function of possible differential opportunities to learn or other differences in socialization. Unfortunately, results of state-of-the-art item and test structure invariance investigations traditionally have not been reported for individually administered intelligence tests. Thus, conclusions regarding intellectual similarities or differences may be unfounded, and the interpretation of test scores influenced by unintended constructs may have serious consequences for individuals and groups. Although such investigations are labor intensive and expensive, and it is impossible to compare psychometric properties for all possible groups, test developers are encouraged to conduct more invariance investigations for nonverbal and other psychoeducational tests used for high-stakes educational decisions.

Even if a test developer makes a thorough attempt to create a test that lacks evidence of bias against a variety of subgroups, the test cannot be assumed to be fair for all subgroups under all conditions. Ultimately, practitioners must take



responsibility for understanding the psychometric properties and potential unintended consequences, as discussed by Messick (1989), of using tests without the necessary validity evidence. Specifically, practitioners should question whether (a) the test should be used for a given purpose, based on the empirical validity evidence, and (b) score interpretation reflects intended test constructs. That is, adverse outcomes for examinees should not be a result of construct irrelevant variance. Messick (1989) points out that, given the social consequences of test use and value implications of test score interpretation, testing practices should be based on both scientific evidence and ethical consideration.

## References

- Agresti, A. (1996). Logit models with random effects and quasi-symmetric loglinear models. In Proceedings of the 11th International Workshop on Statistical Modeling (pp. 3–12).
- Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249–279). Beverly Hills: Sage.
- Anderson, R. J., & Sisco, F. H. (1977). *Standardization of the WISC-R performance scale for deaf children* (Office of Demographic Studies Publication Series T, No. 1). Washington, DC: Gallaudet College.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P. M. (1992). On the fit of models to covariances and methodology in the Bulletin. *Psychological Bulletin*, 112, 400–404.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bolt, D., & Stout, W. (1996). Differential item functioning. Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, 23, 67–95.
- Bracken, B. A., & McCallum, R. S. (2016). *Examiner's manual: Universal nonverbal intelligence test-second edition (UNIT2)*. Austin, TX: PRO-ED.
- Braden, J. P. (1999). Straight talk about assessment and diversity: What do we know? *School Psychology Review*, 14, 343–355.
- Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Test of nonverbal intelligence (TONI-4)*. Austin, TX: PRO-ED.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.
- Bryk, A. (1980). Review of Bias in mental testing. *Journal of Educational Measurement*, 17, 369–374.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Canivez, G. L., & Watkins, M. W. (in press). Review of the Wechsler intelligence scale for children-fifth edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Authors), *Intelligent testing with the WISC-V* (pp. xx–xx). Hoboken, NJ: Wiley.
- Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.
- Chang, H. H., Mazzeo, J., & Roussos, L. (2005). Detecting DIF for polytomously scored items: An adaption of the SIBTEST procedure. *Journal of Educational Measurement*, 33, 333–353.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31–44.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124.
- Cotter, D. E., & Berk, R. A. (1981, April). *Item bias in the WISC-R using Black, White, and Hispanic learning disabled children*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.
- Diana v. the California State Board of Education. Case No. C-70-37 RFP. (N.D. Cal., 1970).
- 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., & 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. *ETS Research Report Series*, 1983 (pp. i–14).
- Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Psychological Measurement*, 3, 347–360.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education & Macmillan.
- Frisby, C. L. (1998). Poverty and socioeconomic status. In J. L. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 241–270). Washington, DC: American Psychological Association.

- Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, 26, 147–160.
- Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (2009). *Comprehensive test of nonverbal intelligence–2 (CTONI–2)*. Austin, TX: PRO-ED.
- Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice*, 8, 5–11.
- Holland, P. W. (1985). On the study of differential item performance without IRT. In Proceedings of the 27th Annual Conference of the Military Testing Association (Vol. 1, pp. 282–287). San Diego, CA.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112, 351–362.
- Ila, D., & Willerman, L. (1989). Sex differences in WAIS-R item performance. *Intelligence*, 13, 225–234.
- Jastak, J. E., & Jastak, S. R. (1964). Short forms of the WAIS and WISC vocabulary subtests. *Journal of Clinical Psychology*, 20, 167–199.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational Statistics*, 23, 291–322.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329–349.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 57, 409–426.
- Jöreskog, K. G., & Sörbom, D. (1989). *LISREL7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.
- Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261–276.
- Koh, T., Abbatiello, A., & McLoughlin, C. S. (1984). Cultural bias in WISC subtest items: A response to Judge Grady's suggestions in relation to the PASE case. *School Psychology Review*, 13, 89–94.
- Kromrey, J. D., & Parshall, C. G. (1991, November). *Screening items for bias: An empirical comparison of the performance of three indices in small samples of examinees*. Paper presented at the annual meeting of the Florida Educational Research Association, Clearwater, FL.
- Larry P. v. Wilson Riles, Superintendent of Public Instruction for the State of California. Case No. C-71-2270 (N.D. Cal., 1979).
- Lawrence, I. M., & Curley, W. E. (1989, March). *Differential item functioning of SAT-Verbal reading subscore items for males and females: Follow-up study*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988, April). *Differential item functioning of SAT-Verbal reading subscore items for male and female examinees*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Li, H.-H., & Stout, W. (1994). *SIBTEST: A FORTRAN-V program for computing the simultaneous item bias DIF statistics [Computer software]*. Urbana-Champaign, IL: University of Illinois, Department of Statistics.
- Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, 61, 647–677.
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159–173.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, 11, 365–386.
- Maller, S. J. (1996). WISC-III Verbal item invariance across samples of deaf and hearing children of similar measured ability. *Journal of Psychoeducational Assessment*, 14, 152–165.
- Maller, S. J., & Ferron, J. (1997). WISC-III factor invariance across deaf and standardization samples. *Educational and Psychological Measurement*, 7, 987–994.
- Maller, S. J., Konold, T. R., & Glutting, J. J. (1998). WISC-III Factor invariance across samples of children displaying appropriate and inappropriate test-taking behavior. *Educational and Psychological Measurement*, 58, 467–475.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from the retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, 54, 284–291.
- Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57, 289–311.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.



- McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematical and Statistical Psychology*, 24, 154–168.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297–334.
- Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, 16, 389–402.
- Nagelkerke, N. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.
- National Council on Measurement in Education. (1995). Code of professional responsibilities in educational measurement. *Educational Measurement: Issues and Practice*, 14, 17–24.
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.
- Penfield, R. D. (2001). Assessing differential item functioning across multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, 14, 235–259.
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Amsterdam: Elsevier.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, 40, 397–404.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207.
- Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–208). New York: Wiley.
- Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, 3, 369–379.
- Roid, G. H., & Miller, L. J. (2013). *Leiter international performance scale-3<sup>rd</sup> edition (Leiter-3) manual*. Wood Dale, IL: Stoelting.
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and the Mantel-Haenszel procedures for detecting differential item functioning. *Applied Measurement in Education*, 17, 105–116.
- Ross-Reynolds, J., & Reschly, D. J. (1983). An investigation of item bias on the WISC-R with four sociocultural groups. *Journal of Consulting and Clinical Psychology*, 51, 144–146.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph Series No. 17). Richmond, VA: Psychometric Society.
- Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 47, 919–927.
- Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology*, 48, 249–253.
- Sandoval, J., Zimmerman, I. L., & Woo-Sam, J. M. (1977). Cultural differences on the WISC-R verbal items. *Journal of School Psychology*, 21, 49–55.
- Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association* (pp. 303–313).
- Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement*, 24, 97–118.
- Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27, 109–131.
- Scheuneman, J. D., & Oakland, T. (1998). High stakes testing in education. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenirer (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 77–103). Washington, DC: American Psychological Association.
- Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159–194.
- Shelley-Sireci, & Sireci, S. G. (1998, August). *Controlling for uncontrolled variables in cross-cultural research*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.
- Sireci, S. G., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Thissen, D. (2001). Psychometric engineering as art. *Psychometrika*, 66, 473–486.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 149–169). Hillsdale, NJ: Erlbaum.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response theory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–8.

- Turner, R. G., & Willerman, L. (1977). Sex differences in WAIS item performance. *Journal of Clinical Psychology*, 33, 795–798.
- Uttaro, T. (1992). *Factors influencing the Mantel–Haenszel procedure in the detection of differential item functioning*. Unpublished doctoral dissertation, Graduate Center, City University of New York.
- Wechsler, D. (2014). *Wechsler intelligence scale for children-fifth edition technical and interpretive manual*. San Antonio, TX: NCS Person.
- Wechsler, D., & Naglieri, J. A. (2006). *Wechsler nonverbal scale of ability technical and interpretive manual*. San Antonio, TX: Pearson.
- Wild, C. L., & McPeck, W. M. (1986, August). *Performance of the Mantel–Haenszel statistic in identifying differentially functioning items*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Erlbaum.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia, Prince George, B.C.
- 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., Donoghue, J. R., & Grima, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement*, 30, 233–251.

Handbook of Nonverbal Assessment

McCallum, R.S. (Ed.)

2017, XII, 320 p. 15 illus., 8 illus. in color., Hardcover

ISBN: 978-3-319-50602-9