7

Item Response Theory: Tests of Invariance

This chapter assumes that a common baseline item response theory (IRT) model has been found to fit adequately in all groups of interest defined by **V**. This baseline model now becomes the starting point for further analyses that will investigate bias. This chapter describes a variety of approaches that can be taken in this investigation. We wish to not only determine if bias is present but also provide some estimate of the size of the bias. Bias "effect size" estimates are needed because simple hypothesis-testing methods are influenced by the available sample sizes. In large samples, bias that is trivial in absolute terms will still lead to rejection of the null hypothesis of no bias. Conversely, small samples may offer insufficient power to detect meaningful bias.

Given a common baseline model across groups, the bias investigation will examine group differences in the item parameter values in this model. If no group differences in parameter values are found, there is no evidence of bias. If some items are found to have parameter values that vary over groups, attention will then focus on describing these parameter differences, along with the implications of these differences for the size of the bias. The bias effect size need not be a simple function of the difference in parameter values (Linn, Levine, Hastings, & Wardrop, 1981).

Depending on the method used to study group differences in parameter values, it may be necessary to explicitly rescale the item parameter estimates in each group to place the estimates on a common metric. Recall that in any IRT model, some parameters must be constrained for model identification. In multiple-group applications, these constraints are usually placed on one or more item parameters. Even if there is no measurement bias with respect to these groups, we expect the item parameter estimates to vary across groups because of sampling error. Assuming that the IRT model fits, we should be able to find a simple (often linear) transformation that will put the item parameter estimates on a common scale (Lord, 1980). The process of finding and applying such transformations is known as parameter linkage. Linkage is needed when direct comparisons are to be made between item parameter estimates from different groups. When item bias is present, however, there may exist no simple linking transformation that will place all parameter estimates on a common scale. In this case, the best strategy may be to set aside the biased items through preliminary screening and then base the linkage on the items that remain. Strategies for addressing this problem are discussed in this chapter.

This chapter is organized as follows. Commonly used distinctions among different forms of bias in IRT-based bias investigations are described first. Likelihood-ratio (LR) methods for evaluating bias are presented next. These methods build upon the Pearson and LR chi-square methods of Chapter 6. In the following section, we turn to direct comparisons of item parameters across groups using Wald test statistics. Direct comparisons require some attention to the issues of parameter linkage, and so linkage methods are also discussed. We then describe different approaches to the estimation of bias effect size. Area indices that measure group differences in item response functions are useful for dichotomous items. Effect size measures for ordered-categorical items are also discussed. Next, the DFIT method of evaluating bias is described. An illustration of the use of IRT in item bias detection is given at the end of this chapter.

Forms of Bias

In IRT-based bias studies with dichotomous item data, it has become common to distinguish between *uniform* DIF or bias and *nonuniform* bias. As discussed by Hanson (1998), the precise meaning of uniform bias varies to some degree among authors. Mellenbergh (1989) originally defined uniform bias as holding when the relationship between group membership and the item score is the same at all levels of the variable used to match examinees across groups. In IRT applications, the matching variable is the latent variable score W. Since Mellenbergh, other authors have used a less restrictive definition of uniform bias. This definition simply requires the group difference in item response functions to have the same sign over the range of W (Shealy & Stout, 1993a; Uttaro & Millsap, 1994). For the reference and focal groups, for example, the definition is that there is no reversal in sign for the difference $P_{jr}(W) - P_{jf}(W)$ throughout the range of W. Shealy and Stout (1993a) denote this condition as *unidirectional* bias. Unidirectional bias need not be uniform in the sense of Mellenbergh because under unidirectional bias, the statistical relationship between the item score and group membership may vary over the range of the matching variable as long as there is no reversal in sign. A third type of "uniform" bias distinguished by Hanson is *parallel* bias in which the item response function for one group is simply shifted in location relative to the other group. In other words, for the reference and focal groups, we must have $P_{jr}(W) = P_{jf}(W + \varepsilon_j)$, with ε_j being a constant that may be unique to each item. If both item response functions are Rasch models, for example, any bias that is present is parallel bias because the functions may only differ in location. Parallel bias need not be uniform in the sense of Mellenbergh. For example, when both groups fit the 3PL model, the resulting bias may not be uniform yet may be parallel.

Given these distinctions, the definition of nonuniform bias is unclear unless one is careful to stipulate the definition of "uniform" bias. To avoid confusion, in this book, we will largely rely on the distinction between unidirectional bias as defined above and *bidirectional* bias that exists whenever the difference in item response functions reverses in sign over the range of the latent variable (Shealy & Stout, 1993a). Hanson (1998) notes that both uniform and parallel bias are forms of unidirectional bias. We will refer to these special cases as the need arises, but we will generally avoid the use of the phrase "nonuniform bias."

In polytomous item data, the category response functions are not ordinarily classified in terms of directionality or uniformity because they need not be monotonic. Instead, the directionality distinction can be applied to the item true-score functions across groups. From Chapter 6, it will be recalled that the true-score function for the *j*th item is the conditional expected value $E(X_j|W)$ given in Equation 6.11. Ordinarily, this function will be an increasing function of *W*. Once the category response functions have been estimated in each group, the true-score function can be estimated as well. With regard to these true-score functions, we will say that unidirectional bias exists if the difference between the true-score functions for two groups never reverses its sign throughout the range of *W*. In reference and focal groups, for example, directional bias means that the difference $E_r(X_j|W) - E_f(X_j|W)$ does not reverse its sign over the range of *W*. If sign reversals do occur, the bias that is present will be denoted as bidirectional bias. These distinctions will allow us to draw parallels between the dichotomous and polytomous cases.

Likelihood-Ratio Tests

Given either conditional maximum likelihood (CML) or marginal maximum likelihood (MML) estimation for the item parameters in each group, LR tests offer a general approach to the investigation of item bias (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). In this approach, likelihood functions are evaluated under two models. The first model typically includes only the parameter constraints needed to identify the model. No attempt is made in this model to impose any invariance constraints on the item parameters, apart from constraints needed for identification. These identification constraints are discussed below.

This first model is essentially identical to the baseline model discussed in Chapter 6, except that the placement of identification constraints must be coordinated across groups. Let this first model be denoted M_1. The second model will add constraints to M_1 to achieve some degree of invariance in the item parameters across groups. One option is to constrain all item parameters to invariance in this second model. Alternatively, one can constrain a subset of these parameters, depending on the nature of the baseline model. For example, in ordered-categorical models, one may wish to constrain only a given type of parameter (e.g., category thresholds). Let the second model with invariance constraints be denoted M_0. Clearly, M_0 is nested within M_1. Suppose that one is able to obtain sample likelihood function values, L_0 and L_1, for the two models simultaneously across all groups. These likelihood values may be obtained via either CML or MML methods. Then under fairly general conditions, the test statistic

$$\chi^2_{\mathrm{LR}} = -2\ln\left(\frac{L_0}{L_1}\right) \tag{7.1}$$

will have a chi-square distribution in large samples if M_0 fits in the population (Amemiya, 1985; Rao, 1973). The degrees of freedom for this test statistic are equal to the number of constraints needed to obtain M_0 from M_1. This test statistic provides an omnibus test statistic that can be used to jointly test an entire set of invariance constraints.

This LR test procedure is quite general. The choice of baseline model is flexible, as both dichotomous and ordered-categorical response models are potential candidates. More than two examinee groups may be compared at once. When more than two groups are studied, the M_0 model may incorporate constraints on parameters in only a subset of groups, permitting one or more other groups to differ from the constrained groups. Furthermore, the LR test statistic in Equation 7.1 may be used even when the C^p contingency table formed by the response patterns to the p items is sparse, as discussed in Chapter 6.

The LR test approach also has some weaknesses. Two evaluations of the likelihood are needed to obtain the test statistic in Equation 7.1, one under each of M_0 and M_1. For the restricted model M_0, the method of estimation must incorporate the invariance constraints. These constraints require a method of estimation that will use data simultaneously from all groups. Software for performing such analyses with general IRT models is limited at present. Furthermore, some consideration must be given to the placement of identification constraints in M_1, so that legitimate evaluations of invariance in M_0 will be possible. The typical approach to this problem is to designate one or more items as anchor items in M_1 (Thissen et al., 1993; Wang & Yeh, 2003; V. S. L. Williams, 1997; Woods, 2008). Finally, if the test statistic in Equation 7.1 leads to rejection of the model M_0, most investigators will wish to determine which item parameters differ across groups. Many individual post hoc tests are possible, and the best strategy for sequencing these tests is not obvious.

The next two sections provide a more detailed description of the use of the LR test procedure to detect bias in dichotomous and polytomous items.

Dichotomous Items

For the Rasch model, Andersen (1973) proposed an LR test that is easily adapted for the investigation of item bias. In this approach, M_1 represents a model that permits separate item parameter values among examinee groups defined by **V**. The likelihood L_1 is evaluated as a CML and can be calculated as the sum of the separate sample conditional likelihood values in K independent groups: $L_1 = \sum_{k=1}^{K} L_{1k}$. The likelihood L_0 is evaluated in the combined sample after pooling examinees across groups. Unlike other LR applications, Andersen's test does not require simultaneous multiple-group estimation under M_0. The test is limited to the global M_0 in which all items have invariant item parameters in all groups. Less constrained versions of M_0 cannot be tested using Andersen's test. Once L_0 and L_1 are calculated, the test statistic is given in Equation 7.1. With K groups and p items, the degrees of freedom will be $(K - 1)(p - 1)$. No explicit transformations to link parameters are needed in this test. Within each group under M_1, one constraint on the location parameters is needed to identify the model. One choice for this constraint is to identify a single item that will be an anchor item for all groups, fixing the location parameter for this item to a chosen value that is the same in all groups. The anchor item should have no measurement bias. If bias is present in this item, its use as an anchor will distort the estimates of the $p - 1$ other location parameters, as discussed in Chapter 6. Within the M_0 model, the location parameter for the same anchor item should be fixed to the same value used in M_1.

While the Andersen LR test is easily implemented, the test is limited to global invariance constraints in M_0. A more diverse array of constraints and M_0 models can be tested using MML under simultaneous multiple-group estimation. In this approach, L_1 represents the marginal likelihood for **X** after specification of the item response function and the prior distribution for W in the kth group, $g_k(W)$. The item response function need not be the Rasch function. Multiparameter models can be handled in this approach. Ordinarily, $g_k(W)$ will be taken as normal with an unspecified mean μ_{wk} and standard deviation σ_{wk}. In single-group applications, the mean and standard deviation for W are commonly taken as 0 and 1, respectively. In the multiple-group case, the standardized metric for W should not be uniformly imposed because there is generally no reason to suppose that the groups are identical in their prior distributions for W. A better approach is to require $\mu_{wk} = 0$ and $\sigma_{wk} = 1$ within one chosen group, and let μ_{wk} and σ_{wk} vary in the remaining groups. If this approach is adopted, some additional constraints are needed for the item parameters. A subset of items (perhaps only one item) will be designated as an anchor subset. These items are assumed to be free of bias. The item parameters for these items are constrained to be invariant over all groups. These invariance constraints, in addition to the fixed values of μ_{wk} and σ_{wk} for some choice of k, are sufficient to identify the remaining item parameters in L_1. All items are permitted to vary in their item parameter values in M_1, with the exception of the anchor items.

Under M_0, the item parameters for some further set of items are constrained to invariance. This additional set of items may include all items that are not anchor items or, alternatively, may include only a subset of the remaining items. In either case, no further constraints are introduced for the prior distribution parameters μ_{wk} and σ_{wk}. The likelihood L_0 is calculated under the invariance constraints on the item parameters of the additional set of items. If the item response functions are multiparameter models (e.g., 2PL), M_0 may incorporate invariance constraints on only some of the item parameters, rather than the entire set. For example, M_0 may first restrict the discrimination parameters in the 2PL to invariance, leaving the location parameters to vary over groups. Many possibilities exist for choices of which items and/or parameters to constrain in M_0.

Once M_0 and M_1 are specified, the sample likelihood values L_0 and L_1 can be calculated, and the LR test statistic in Equation 7.1 is found. The degrees of freedom for this test statistic correspond to the number of constraints needed to create M_0 from M_1. Note that unlike the Andersen test statistic described above, this test statistic requires some assumptions about the prior distributions $g_k(W)$. These assumptions are part of both M_0 and M_1 and are not directly tested by the LR statistic. In large samples, the LR statistic has a chi-square distribution under the null hypothesis that M_0 holds.

The MML LR approach is highly flexible, but if the test statistic leads to rejection of M_0, the test does not indicate which of the constraints in M_0 is responsible for the lack of fit. An obvious approach to obtaining additional information about individual constraints is to perform a set of additional LR tests, each with a single degree of freedom, by successively relaxing each invariance constraint in turn. This approach is unwieldy when the number of items is large. A more efficient approach would be to generate Lagrange multiplier (LM) statistics for each constrained item parameter (Glas, 1999). These statistics evaluate the expected drop in the LR chi-square in Equation 7.1 if the constraint under study is relaxed. Under M_0, the LM statistic just described is equivalent to the single degree of freedom chi-square from the LR test in large samples (Buse, 1982). The LM statistic is more easily calculated, however, because it does not require two likelihood evaluations and is based only on M_0. Unfortunately, it appears that no current IRT software package will produce the LM statistics for multiple-group MML estimation.

Polytomous Items

Under polytomous Rasch-family models, it is possible to develop an LR test procedure using CML estimation that parallels Andersen's (1973) procedure for dichotomous items. As in the dichotomous case, M_1 represents separate, group-specific models that introduce constraints needed for identification but that include no additional constraints. The choice of identification constraints will depend on which polytomous model is adopted for the items. For example, in the rating-scale model, one will need to impose one constraint on the location parameters in each group. An anchor item could be selected for this purpose, fixing the location parameter to a known value (e.g., zero). The anchor item should be selected to be an item that is thought to be free of bias, with the same item serving as an anchor in each group. The likelihood L_1 is then a sum of group-specific conditional likelihoods $\sum_{k=1}^{K} L_{1\mathrm{k}}$. The model M_0 is based on the pooled sample that combines data from all K groups. As in M_1, an anchor item is chosen to carry the identification constraints. The conditional likelihood L_1 is based on the pooled sample. Once both L_0 and L_1 are available, the test statistic in Equation 7.1 can be calculated. Under the null hypothesis that no items are biased, this test statistic is distributed as a chi-square variate in large samples. The degrees of freedom are equal to $(K - 1)(h)$, where h is the number of free parameters in a single group under the chosen polytomous model.

The above test procedure has the same drawback found in the dichotomous LR procedure: Only global tests of "no bias" that apply to the entire set of items are possible. LR tests that focus on subsets of items require multiple-group simultaneous estimation procedures, and these procedures are available under MML estimation. The use of marginal likelihood methods also permits the extension of the LR test to models outside of the Rasch family. In this approach, M_1 represents a multiple-group model with identification constraints in each group, but without the invariance constraints that are the focus of interest. The prior distributions for W are permitted to vary across groups but commonly take the same form with varying parameter values (e.g., normal with varying means and variances). As in the dichotomous case, one approach to defining M_1 is to fix the mean and standard deviation of W in one group to known values (e.g., $\mu_{\mathrm{wk}} = 0$, $\sigma_{\mathrm{wk}} = 1$) and also to designate one or more items as anchor items whose parameter values will be invariant across groups. The combination of these two types of constraints should be sufficient to identify all remaining parameters in M_1. The anchor items must be carefully chosen to ensure that these items show no bias in relation to the groups under study.

The restricted model M_0 adds further invariance constraints to model M_1. Item parameters for some or all of the items outside of the anchor subset will be constrained to invariance in M_0. Using the likelihood values corresponding to M_0 and M_1, the test statistic in Equation 7.1 is calculated and is used to test whether the restricted model M_0 provides an adequate fit, given the fit of M_1. This test statistic has a chi-square distribution in large samples under M_0 with degrees of freedom equal to the number of independent constraints needed to create M_0 from M_1.

While M_1 may be initially structured to include invariance constraints only on the anchor items, other versions of M_1 could be created that include invariance constraints on items that are not anchor items. In the extreme case, M_1 may include invariance constraints on $p - 1$ items, permitting only a single item to vary in its parameters across groups. The restricted model M_0 would then add invariance constraints on this item to M_1. The resulting chi-square test examines whether the studied item displays bias, given invariance in the remaining $p - 1$ items. In theory, these tests on individual items could be conducted for each non-anchor item in turn. The presence of bias in more than one item, however, implies that many of the M_1 models in this sequence of tests are themselves incorrect. The typical response to this problem is to iteratively purify the test by dropping items that are initially flagged as biased, followed by further rounds of tests for the items that remain. This process is continued until no further items are found to be biased. An obvious difficulty with this iterative approach is that many hypothesis tests are required if the number of items is moderate or large, inflating the Type I error rate for the set of tests. One approach to reducing the inflation problem is to apply more stringent alpha levels to each individual test. Ideally, one could base decisions to flag an item on some measure of the size of the bias, in addition to the hypothesis test. The LR test statistic itself does not provide such an effect size measure. We will return to this problem below.

Evaluative Studies

The MULTILOG (Thissen, 1991) program gives information for calculation of the LR chi-square statistic once the M_0 and M_1 models have been run and works with all of the common binary models as well as the graded response and nominal polytomous model. The IRTLRDIF program (Thissen, 2001) makes the process of LR chi-square testing more convenient, however, because it will sequentially test all items automatically using a designated anchor set of items. CFA software can also be used for purposes of LR testing for any IRT model that is equivalent to a common factor model.

While the large-sample statistical theory underlying the LR test procedure is well understood, less is known about the small-sample behavior of the LR chi-square statistic in bias applications. Simulation studies have examined a number of potential influences on the Type I error and power performance of the LR chi-square statistic (Ankenmann, Witt, & Dunbar, 1999; Cohen, Kim, & Wollack, 1996; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Finch, 2005; Kim & Cohen, 1998; Wang & Yeh, 2003; Woods, 2008, 2009). These influences include sample size, size of the bias, direction of bias, type of model, test length, choice of anchor, and the presence of biased items in the anchor. Non-normality in the distribution of the latent variable is another possible influence, although fewer studies have examined this issue (Woods, 2006, 2007, 2008; Woods & Thissen, 2006).

Several large simulation studies have appeared that reported results on the Type I error performance of the LR chi-square statistic. Cohen et al. (1996) studied the test statistic in the dichotomous item case under the 2PL and 3PL models. Sample sizes studied were $N = 250$ and 1,000, with each sample size being applied to both the reference and the focal group. Test length was fixed at $p = 50$ items. Prior distributions for W in the two groups were both standard normal distributions, and MML estimation was used. The results showed that the LR test statistic adhered fairly closely to the nominal Type I error rate under the 2PL and 3PL models. Some inflation in the error rate was noted under the 3PL model, but the inflation was not large. Finch (2005) also examined the Type I error performance under either 2PL or 3PL models, with reference group sample sizes of 500 and focal group sample sizes of either 100 or 500. Either $p = 20$ or 50 items were used. Finch also manipulated whether biased items were used as anchor items, with either no biased items or 15% biased items. When no biased items were present in the anchor, the LR chi-square adhered to the nominal error rate. When biased items were present, the error rate was inflated. Kim and Cohen (1998) reported Type I error results for the LR chi-square, this time under the graded response model with five-category items. Sample sizes for the reference and focal groups were either 300 or 1,000, with a test length of $p = 30$ items. Normal prior distributions for W were used, either with identical means across groups or with different means. The results showed that the LR test procedure adhered closely to the nominal Type I error rate even in the smaller sample size condition. Wang and Yeh (2003) reported Type I error results under either a 2PL, 3PL, or graded response model with five-category items. Sample sizes were 1,000 per group, bias direction was either one sided or mixed, and the proportion of biased items was either 0%, 12%, 20%, or 32% of the test. Type of anchor was also manipulated, with either all items, 1 item, 4 items, or 10 items. The LR statistic adhered to the nominal error rate well except when the "all item" anchor method was used with higher proportions of the test items being biased and with one-sided bias. Ankenmann et al. (1999) examined Type I error performance in graded response model items with binary 3PL items in the anchor. No bias contamination in the anchor was introduced. Test length was $p = 26$ items with sample sizes of either 500 per group, 2,000 per group, or reference group with 2,000 and focal group with 500. Bias direction was either one or two sided. The LR statistic showed good adherence to the nominal rate throughout. Woods (2008) studied Type I error in the context of manipulating normality of the latent variable distributions. The latent distributions were normal in both the reference and focal groups, normal in one group but skewed in the other, or skewed in both groups. The reference group N was 1,500, with $N = 500$ in the focal group. Test length was $p = 24$ items. No biased items were present in the anchor set. All data were based on the 2PL model. It was found that the Type I error rate was higher than the nominal rate when only one group had a skewed latent variable distribution and the groups differed in latent means.

The general implications of the available simulation studies are that the LR chi-square statistic adheres well to the nominal error rate when the anchor items contain no biased items, or when the average bias across the biased anchor items is near zero (e.g., mixed directions for biased items). If the anchor contains a substantial proportion of biased items and those items tend to be one sided, the Type I error rate can be large. Also, non-normality of the latent distribution can inflate the error rate when the groups have different distributions and different means. Some effort is needed to remove biased items from the anchor through preliminary tests or through careful anchor choice. Alternatives to assuming normal prior distributions, such as those proposed by Woods (2006, 2007), are worth considering as well.

One question of interest about the anchor items, apart from the presence of biased items in the anchor, is the appropriate choice of length for the anchor: how many items are needed? Simulations have shown that even with only a single unbiased anchor item, the LR chi-square statistic adheres reasonably well to the nominal Type I error rate if other assumptions are met (Stark, Chernyshenko, & Drasgow, 2006; Wang, 2004; Wang & Yeh, 2003). In low sample size situations (e.g., combined N of 800 across groups), Woods (2009) found that the use of single anchor items produced low power to detect bias. Woods recommended that the anchor be chosen to be 10%–20% of the total number of items being studied and pointed to the trade-off between expanding the size of the anchor to improve validity of the anchor and running an increasing risk of including biased items in the anchor.

A few studies have examined the power of the LR chi-square statistic (Ankenmann et al., 1999; Finch, 2005; Wang & Yeh, 2003; Woods, 2008). Some indication exists that the power of the test is low at smaller sample sizes. Ankenmann et al. found lower power when $N = 500$ in both the reference and focal groups. Power is generally higher when longer anchor sets are used and when the bias is one sided rather than mixed (Wang & Yeh; Woods, 2009). Woods (2008) found fairly good power even with nonnormal latent variable distributions, but the Type I error inflation under some conditions makes the power results difficult to judge in some cases. Further studies of the power of the LR chi-square statistic are needed, but the statistic seems to work well when the anchor is free of biased items, sample sizes are substantial, the latent distribution is normal, and the bias in the studied item is not too small in size.

Wald Statistics

Given that item bias within IRT is defined by group differences in item parameters, the most direct approach to investigating bias would seem to be to compare the values of the item parameter estimates across groups. Any such comparison must consider parameter estimate differences in relation to sampling error. Group differences that are larger than expected given the sample sizes and the IRT model under study are taken as evidence of bias. Generally speaking, statistical tests for group differences in item parameters in IRT are not exact in small samples. These tests instead rely on large-sample approximations provided by Wald (1943) and can be collectively denoted as Wald test statistics (Thissen et al., 1993).

To illustrate the principles underlying Wald statistics in a general context, suppose that scores on the measured variables $\mathbf{X}$ have a likelihood function $L(\mathbf{\Gamma})$, where $\mathbf{\Gamma}$ is an $s \times 1$ vector of parameters of interest. Let $\hat{\mathbf{\Gamma}}$ be the maximum likelihood estimator of Γ, and assume that $\hat{\mathbf{\Gamma}}$ is distributed as multivariate normal in large samples. We wish to test a null hypothesis that places $t < s$ restrictions on $\mathbf{\Gamma}$ of the form $c(\mathbf{\Gamma}) = \mathbf{0}$, where $c()$ is a differentiable function. Then the Wald test statistic d^2 is

$$d^2 = -c(\hat{\mathbf{\Gamma}})'[\hat{\mathbf{\Delta}}\hat{\mathbf{H}}^{-1}\hat{\mathbf{\Delta}}']^{-1}c(\hat{\mathbf{\Gamma}}), \tag{7.2}$$

where $\mathbf{\Delta}$ is an $s \times 1$ vector whose ith element is the first partial derivative of $c(\mathbf{\Gamma})$,

$$\Delta_{\mathrm{j}} = \frac{\partial c(\mathbf{\Gamma})}{\partial \Gamma_{\mathrm{i}}}, \tag{7.3}$$

evaluated at Γ_i, and **H** is the $s \times s$ Hessian matrix for log$L(\mathbf{\Gamma})$, whose *ij*th element is

$$H_{ij} = \frac{\partial^2 \log L(\mathbf{\Gamma})}{\partial \Gamma_i \partial \Gamma_j} \tag{7.4}$$

evaluated at the *i*th and *j*th elements of $\mathbf{\Gamma}$. The matrix $\widehat{\mathbf{H}}^{-1}$ in Equation 7.2 is the estimated covariance matrix for $\widehat{\mathbf{\Gamma}}$ given the likelihood function $L(\mathbf{\Gamma})$. In large samples, d^2 will have a chi-square distribution under the null hypothesis with t degrees of freedom (Rao, 1973; Wald, 1943). Amemiya (1985) notes that d^2 will have the stated large-sample distribution under the null hypothesis even if the likelihood $L(\mathbf{\Gamma})$ is unspecified, as long as $\widehat{\mathbf{\Gamma}}$ is multivariate normal in large samples and $\widehat{\mathbf{H}}^{-1}$ in Equation 7.2 is replaced by any consistent estimator of the covariance matrix for $\widehat{\mathbf{\Gamma}}$. This fact is useful, as direct estimates of $\mathbf{H}^{-1}$ are sometimes difficult to obtain.

The above principles provide the basis for a variety of tests for group differences in item parameters, depending on the IRT model of interest and the number of item parameters to be investigated. An example will illustrate the application. Suppose that $L(\mathbf{\Gamma})$ represents a conditional likelihood under the Rasch model for two independent groups of examinees, with possibly different item location parameters for some of the p items in each group. Let $\mathbf{\Gamma} = (\mathbf{b}_r, \mathbf{b}_f)$ be a $2p \times 1$ vector of item location parameters for the reference and focal groups, respectively. For the *j*th item, we wish to test the restriction $b_{jr} = b_{jf}$ or $b_{jr} - b_{jf} = 0$. This null hypothesis is represented as a simple linear restriction $c'\mathbf{\Gamma} = 0$, where c is a $2p \times 1$ vector with zeros in all positions except the *j*th and $(p + j)$th, which are 1 and –1, respectively. Note also that since the two groups are independent, the $2p \times 2p$ Hessian matrix **H** has a block diagonal structure

$$\mathbf{H} = \begin{vmatrix} \mathbf{H}_r & 0 \\ 0 & \mathbf{H}_f \end{vmatrix}, \tag{7.5}$$

where $\mathbf{H}_r$ and $\mathbf{H}_f$ are each $p \times p$ matrices. The diagonal elements of $\mathbf{H}_r^{-1}$ and $\mathbf{H}_f^{-1}$ are the variances for the elements of $\hat{\mathbf{b}}_r$ and $\hat{\mathbf{b}}_f$, the CML estimates of $\mathbf{b}_r$ and $\mathbf{b}_f$. Then from Equation 7.2, the Wald statistic for testing the null hypothesis is

$$d^2 = (\hat{b}_{jr} - \hat{b}_{jf})[\mathbf{c}' \, \widehat{\mathbf{H}}^{-1} \, \mathbf{c}]^{-1}(\hat{b}_{jr} - \hat{b}_{jf}) = \frac{(\hat{b}_{jr} - \hat{b}_{jf})^2}{\mathrm{Var}(\hat{b}_{jr}) + \mathrm{Var}(\hat{b}_{jf})}, \tag{7.6}$$

where $\mathrm{Var}(\hat{b}_{jr})$ and $\mathrm{Var}(\hat{b}_{jf})$ are the estimated variances of the location parameter estimates from the reference and focal groups, respectively.

In this case, d^2 has $df = 1$. Alternatively, we can regard $d = \sqrt{d^2}$ as a standard normal variate and refer d to the standard normal table. The above test is easily expanded to simultaneously consider restrictions on parameters for more than one item or to test hypotheses for more than two examinee groups.

The great advantage of the Wald statistic in comparison to the LR procedures considered earlier is that there is no need to evaluate the likelihood under restrictions. Only the unrestricted likelihood L_1 is needed to obtain the estimates $\hat{\mathbf{\Gamma}}$. It is necessary to estimate the covariance matrix for $\hat{\mathbf{\Gamma}}$, however, a requirement that is simple in principle yet often difficult in practice, depending on the model, sample sizes, and number of items. In multiparameter models, composite hypotheses that evaluate invariance jointly for different parameters (e.g., location and discrimination parameters) must consider both the variances and covariances among these parameters within each examinee group. The required estimates for this covariance structure may be inaccurate in small samples. The small-sample behavior of d^2 is not well understood in general, although simulations have been conducted for particular cases as discussed below.

A further difficulty facing applications of the Wald statistic in bias investigations is the need to achieve linkage across groups in the scales used for the item parameters, as noted earlier. Assuming that no bias exists, we expect the item parameter estimates to differ across groups both as a function of sampling error and as a function of any arbitrary scaling differences induced by group-specific standardizations. As an example of such a standardization, item parameters are commonly estimated in single-group applications under the restriction that $\mu_{\text{wk}} = 0$ and $\sigma_{\text{wk}} = 1$ in MML estimation. If this scaling is adopted in multiple-group applications, however, the resulting item parameter estimates will be artifactually different across groups whenever the latent variable distributions $g_{\text{k}}(W)$ have varying means and/or variances across groups. If these item parameter estimates are then compared across groups using the Wald statistic, spurious findings of bias may result. The traditional approach to this problem has been to place identification constraints on the item parameters, rather than on the parameters governing $g_{\text{k}}(W)$. For example, we could require that the average location parameter value be zero in each group in the Rasch model. It will still be true that sampling error, in combination with these constraints on the item parameters, will create group differences in the item parameter estimates even when no bias is present. If the average location parameter value is constrained to zero, for example, sampling error in one item's location parameter estimate will affect the parameter estimates for all other items. To surmount this difficulty, the location parameter estimates in each group are transformed to a common metric across groups prior to any direct comparisons. Various approaches to creating such "linkage" transformations exist.

The presence of bias in some items complicates the development of any linkage transformations. If the entire item set, including the biased items, is used to generate the linkage, no simple transformation may be found that will yield a common metric. The ideal approach in this case would be to weed out the clearly biased items using a preliminary screening and then base the linkage on the subset of unbiased (or less biased) items. The preliminary screening must identify the biased items using a method that does not require direct comparisons of parameter values. Once a linkage transformation is created and applied to the items that remain, these items are again evaluated for bias, now using direct comparisons via the Wald statistic. Some methods for developing linkage transformations in bias investigations are discussed below.

The next section describes the use of Wald statistics to evaluate bias in dichotomous items. The earliest applications of Wald statistics to investigate bias were developed for such items. Following this section, we discuss the use of Wald statistics in polytomous items. Applications to polytomous items are complex because the models include multiple item parameters. After the dichotomous and polytomous item sections, studies that have evaluated the Wald statistic approach in simulated data are described. Linkage transformations are discussed in the last section.

Dichotomous Items

The earliest application of the Wald statistic to the problem of item bias may be the statistic developed by Wright, Mead, and Draba (1976) for the Rasch model. This statistic is simply the square root of the Wald statistic, with the sign of the location parameter difference included:

$$d_{\mathrm{j}} = \frac{\hat{b}_{\mathrm{jr}} - \hat{b}_{\mathrm{jf}}}{\sqrt{\mathrm{Var}(\hat{b}_{\mathrm{jr}}) + \mathrm{Var}(\hat{b}_{\mathrm{jf}})}}. \tag{7.7}$$

In large samples, the above statistic is distributed as a standard normal variate under the null hypothesis of invariance in the location parameter for the jth item. Separate values of d_{j} are calculated for each item, $j = 1, \ldots, p$. Items whose d_{j} values exceed a chosen threshold for significance are flagged as biased. Within the Rasch model, the sampling variance of b_{j} under CML estimation depends to some extent on the choice of identification constraints.

A general approach to obtaining large-sample standard errors or sampling variances $\mathrm{Var}(\hat{b}_{\mathrm{j}})$ in the Wald statistic in Equation 7.7 is to use the appropriate diagonal elements of the inverse of the information matrix associated with the location parameter estimates b_j (Andersen, 1980). Under the Rasch model, the information matrix $\mathbf{I}_k(\mathbf{b})$ for the *k*th group is the $p \times p$ matrix whose element in the *j*th row and *m*th column is

$$-E\left[\frac{\partial^2 L(\mathbf{b})}{\partial b_j \partial b_m}\right], \tag{7.8}$$

where $L(\mathbf{b})$ is the conditional log likelihood function used to obtain estimates of $\mathbf{b}$. An estimate of the information matrix is obtained using estimates $\hat{\mathbf{b}}$. Assuming that the resulting estimate of the information matrix is positive definite, the information matrix is inverted and the appropriate diagonal element of the inverse is used in Equation 7.7. Separate estimates of the sampling variances are obtained for each group using that group's estimate of the information matrix.

Wright et al. (1976) developed the statistic in Equation 7.7 specifically for the two-group Rasch case. Other IRT models require different expressions for the sampling variances of $\hat{b}_j$ and would also incorporate Wald statistics for group differences in other parameters. Lord (1980) developed Wald statistics for use with the 2PL and 3PL models. In the 3PL case, Lord recommended that the test statistic be used only for hypotheses involving the location and discrimination parameters. The pseudo-guessing parameter c_j would be fixed to a common value across groups based on pooled data. The suggested sequence of steps in conducting Wald tests in the 3PL case began with the pooled sample, ignoring group membership. Item parameters are estimated in this pooled sample, with the scaling chosen to force the average location parameter value to zero and the standard deviation of the location parameters to one. The pseudo-guessing parameters are then fixed to their estimated values for this pooled analysis, and they retain those values in all subsequent steps. Next, the groups are separated and group-specific estimates of the location and discrimination parameters are obtained. The scaling used earlier for the location parameters is again adopted, this time imposed separately within each group. The resulting location and discrimination parameter estimates are then used in Wald tests to evaluate bias. For tests involving only location parameters, the statistic in Equation 7.7 is used. For tests on discrimination parameters, this statistic is simply modified as

$$d_j = \frac{\hat{a}_{jr} - \hat{a}_{jf}}{\sqrt{\mathrm{Var}(\hat{a}_{jr}) + \mathrm{Var}(\hat{a}_{jf})}}. \tag{7.9}$$

In large samples, d_j in Equation 7.9 is distributed as standard normal under the null hypothesis of invariance.

Lord (1980) also considered simultaneous tests for invariance in both location and discrimination parameters. This test requires estimates of the 2 × 2 covariance matrix for ($\hat{a}_j, \hat{b}_j$) within each group. Let $\boldsymbol{\Sigma}_{jr}$ and $\boldsymbol{\Sigma}_{jf}$ be these covariance matrices within the reference and focal groups, respectively. Also let

$$\hat{\boldsymbol{\delta}}_j = [\hat{a}_{jr} - \hat{a}_{jf}, \hat{b}_{jr} - \hat{b}_{jf}]. \tag{7.10}$$

Then the test statistic is

$$d_j^2 = \hat{\boldsymbol{\delta}}_j' \, [\hat{\boldsymbol{\Sigma}}_{jr} + \hat{\boldsymbol{\Sigma}}_{jf}]^{-1} \hat{\boldsymbol{\delta}}_j. \tag{7.11}$$

This test statistic is distributed as a chi-square variate with $df = 2$ under the null hypothesis of invariance in both location and discrimination parameters for the *j*th item.

Polytomous Items

Wald statistics generalize easily to the case of polytomous items, apart from the difficulties created by the greater number of parameters found in models for such items. Each item will include multiple parameters. To fully evaluate bias in a given item, group differences in all of these parameters must be investigated. To illustrate, suppose that the partial credit model is found to fit the data in all groups and that we wish to compare parameters for the *j*th item across groups. Within the *k*th group, the vector of parameters for the *j*th item is $\boldsymbol{\Gamma}_{jk} = (\tau_{j1}, \tau_{j2} \ldots, \tau_{jC-1})$, a $(C - 1) \times 1$ vector of category location parameters. The estimates for these parameters will be correlated in general, regardless of the method of estimation. In the partial credit model, we could use CML estimates in the vector $\hat{\boldsymbol{\Gamma}}_{jk}$. Let $L(\boldsymbol{\Gamma})$ be the conditional log likelihood function to be maximized in obtaining $\hat{\boldsymbol{\Gamma}}_{jk}$. Also let the entire vector of parameters for *p* items in the *k*th group be $\boldsymbol{\Gamma}_k = (\boldsymbol{\Gamma}_{1k}, \boldsymbol{\Gamma}_{2k}, \ldots, \boldsymbol{\Gamma}_{pk})$. This vector is $p(C - 1) \times 1$. We can define the $p(C - 1) \times p(C - 1)$ information matrix $\mathbf{I}_k(\boldsymbol{\Gamma}_k)$ for the item parameters across *p* items in the *k*th group as having elements

$$-E\left[\frac{\partial^2 L(\boldsymbol{\Gamma})}{\partial\gamma_n \partial\gamma_m}\right], \tag{7.12}$$

in the *n*th row, *m*th column, with γ_n and γ_m being the *n*th and *m*th elements of $\boldsymbol{\Gamma}_k$. An estimate of the information matrix $\mathbf{I}_k(\boldsymbol{\Gamma}_k)$ is obtained by substituting estimates $\hat{\boldsymbol{\Gamma}}_k$ into the expressions in Equation 7.12. If the resulting matrix is positive definite, the matrix can be inverted to obtain an estimate of the large-sample covariance matrix for $\hat{\boldsymbol{\Gamma}}_k$.

In the two-group case with $\boldsymbol{\Gamma}_r$ and $\boldsymbol{\Gamma}_f$ being the item parameter vectors for the reference and focal groups, respectively, suppose that we wish to compare $\boldsymbol{\Gamma}_{jr}$ and $\boldsymbol{\Gamma}_{jf}$ for the *j*th item. Let the difference in the parameter vectors for the two groups be

$$\boldsymbol{\delta}_j = \boldsymbol{\Gamma}_{jr} - \boldsymbol{\Gamma}_{jf} = (\tau_{j1r} - \tau_{j1f}, \tau_{j2r} - \tau_{j2f}, \ldots, \tau_{jC-1r} - \tau_{jC-1f}), \tag{7.13}$$

where τ_{jcr}, τ_{jcf} represent the *c*th category location parameters for the reference and focal groups, respectively. Then $\hat{\boldsymbol{\delta}}_j$ will represent the sample estimate of the difference vector in Equation 7.13 after the estimates $\hat{\boldsymbol{\Gamma}}_{jr}$ and $\hat{\boldsymbol{\Gamma}}_{jf}$ are substituted. Finally, let

$$\hat{\boldsymbol{\Sigma}}_{jr} = [\mathbf{I}_r(\hat{\boldsymbol{\Gamma}}_{jr})]^{-1}, \quad \hat{\boldsymbol{\Sigma}}_{jf} = [\mathbf{I}_f(\hat{\boldsymbol{\Gamma}}_{jf})]^{-1} \tag{7.14}$$

be the sample estimates of the large-sample covariance matrices for the item parameter estimates in each group. We can then calculate the Wald statistic for the *j*th item as

$$d_j^2 = \hat{\boldsymbol{\delta}}_j' \, [\hat{\boldsymbol{\Sigma}}_{jr} + \hat{\boldsymbol{\Sigma}}_{jf}]^{-1} \, \hat{\boldsymbol{\delta}}_j. \tag{7.15}$$

This statistic has $df = C - 1$, corresponding to the $C - 1$ restrictions imposed on $\boldsymbol{\Gamma}_{jr}$ and $\boldsymbol{\Gamma}_{jf}$. In large samples, d_j^2 will have a chi-square distribution under the null hypothesis that $\boldsymbol{\delta}_j = 0$, which corresponds to no bias in the *j*th item.

The above Wald statistic could be modified in several ways, depending on the needs of a particular investigation. For example, one could investigate group differences in the item parameters for an entire block of items, considered simultaneously. This extension would require that the large-sample covariance matrix be expanded to include elements for all parameters from the block of items under consideration. Alternatively, one could modify the statistic to compare parameters across more than two groups simultaneously. In a three group problem, for example, one could expand $\boldsymbol{\delta}_j$ to include pairwise contrasts across both groups 1 and 2, and groups 2 and 3. The dimension of this difference vector is then $2(C - 1) \times 1$. Analogously, the large-sample covariance matrix is expanded to a block diagonal matrix of dimension $2(C - 1) \times 2(C - 1)$, with each block being the sum of the appropriate pair of covariance matrices. In either of these extensions, the main difficulty would be that rejection of the null hypothesis would require further investigation to locate which set of item parameters are responsible for the rejection.

Evaluative Studies

A number of studies have evaluated the performance of Wald statistics in bias applications, in terms of Type I error behavior, Type II error behavior, or both (Cohen & Kim, 1993; Cohen, Kim, & Baker, 1993; Donoghue & Isham, 1998; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987). More evaluative studies of the Wald statistic have been performed in the dichotomous case than in the polytomous case. Kim et al. investigated the Type I error behavior of Lord's chi-square statistic in data simulated to fit either the 2PL or 3PL models. Two sample size conditions were created (N = 250 or 1,000), with test length fixed at p = 50. Prior distributions for W were standard normal in both the focal and reference groups. The results showed that under the 3PL, the effective Type I error rate exceeded the nominal rate, even in the larger sample size condition. Conversely, under the 2PL model, the error rate was lower than the nominal rate, resulting in a conservative test. This result contradicts an earlier finding by McLaughlin and Drasgow, who found inflated Type I error rates under the 2PL. McLaughlin and Drasgow used joint maximum likelihood estimation in their study, while Kim et al. used MML. The error rate inflation for the 3PL found by Kim et al. was lessened if a 3PL model with a common pseudo-guessing parameter was used in the 3PL data, even though the data were generated under the unrestricted 3PL model. Kim et al. suggested that the distortions in the Type I error rates found in their study were probably due to inaccuracy in the estimation of the covariance matrix for the item parameter estimates, a problem noted by Thissen et al. (1993).

Cohen and Kim (1993) studied both the Type I and Type II error behavior of Lord's chi-square test under the 2PL model. Two sample size conditions (N = 100 or 500) were factorially combined with two test length conditions (p = 20 or 60). Bias was simulated for either no items, 10% of the items, or 20% of the items. The size of the bias, when present, was varied across items. Prior normal distributions for W were either invariant over groups or differed in means. The results were that in the null bias condition, the observed Type I error rate was held to the nominal rate, although the rate was higher when the prior distributions differed across groups. In the bias conditions, the false-negative rate (Type II error rate) was unacceptably high when N = 100 but was considerably better at N = 500, especially when no group differences existed in the prior distributions. The false-negative rate was higher when the percentage of biased items was higher. In general, the results supported the use of Lord's test under the 2PL when the sample size was at least 500.

Donoghue and Isham (1998) reported simulation results on the Type I error and power performance of Lord's test in the detection of item parameter drift. Item parameter drift refers to changes in the parameter values for an item across repeated measurement occasions. Item parameter drift differs from the bias phenomenon by being a within-group change, rather than a group difference. Formally, however, the use of Lord's test to detect drift is similar to its use in bias applications. Donoghue and Isham generated data to fit a 3PL model with a common value for the pseudo-guessing parameter across items. Sample size was varied (N = 300, 600, or 1,000), as was the test length (p = 30 or 60) and the number of items showing drift (0, 3, 6, or 12). Drift was manifested in changes in the difficulty parameters, changes in the discrimination parameters, or both. The results were that Lord's test performed well overall as long as no attempt was made to study group differences in the pseudo-guessing parameters. Lord's test adhered closely to the Type I error rate overall in the no-drift condition. When drift was present, Lord's test had fairly good power when $N > 300$.

In the polytomous item case, Cohen et al. (1993) studied Lord's test in data simulated to fit the graded response model. The sample size was fixed at N = 1,000, and the test length was fixed at p = 40. Each item was scored in five categories. Bias was simulated in six items by varying the discrimination parameters, the location parameters, or both. Prior distributions for W were standard normal in each group. The results showed that Lord's test detected bias in five of the six items simulated to be biased and (falsely) in one item that was simulated as unbiased. While limited in scope, these results suggest that Lord's chi-square test is feasible for polytomous items.

From this brief review, it is clear that Lord's test can be used effectively with the 2PL and 3PL models if the sample sizes are 300 or more, and if no group differences in pseudo-guessing parameters are sought under the 3PL. Attempts to evaluate bias in all three parameters of the 3PL do not seem effective, at least using current MML estimation methods. Further power studies that would focus specifically on the amount of bias that can reliably be detected for various sample sizes would be useful. Also, studies of the accuracy of estimates for parameter sampling covariance matrices are needed, as this accuracy affects the performance of the Wald statistic. It is also clear that much more work is needed for the polytomous item case.

Parameter Linkage

In any direct comparison of item parameter estimates across groups, preliminary adjustments to the parameter estimates are usually needed to remove the effects of any group-specific scaling adopted during the estimation process. These adjustments are especially important when the parameter estimation is done separately within each group defined by **V**. When estimation is done separately in this manner, constraints are placed on the item parameters within each group to achieve identification, as discussed in Chapter 6. For example, in the Rasch model for dichotomous items, we might require the location parameter estimates to have a mean of zero within each examinee group defined by **V**. Even if there is no bias present in any item in relation to **V**, sampling error will result in some group differences in the location parameter estimates. The group-specific identification constraints will combine with sampling error to widen group differences in parameter estimates for some items and to narrow group differences for other items. The net effect of these influences is that it becomes more difficult to accurately compare parameter estimates across groups for purposes of bias detection.

In the absence of any bias in the items, it should be possible to find a transformation to the item parameters that will place the estimates on a common scale across groups, removing the influence of any group-specific identification constraints. Lord (1980) illustrates the use of linear transformations for dichotomous item data under logistic models in two examinee groups. In this case, the objective is to find constants *A* and *B* to apply to the item parameter estimates in group one that will place these estimates on the scale used for group two. This strategy is based on the idea that the item parameters in the two groups are related as

$$b_{j2} = Ab_{j1} + B, \quad a_{j2} = \frac{a_{j1}}{A} \tag{7.16}$$

for location parameters (b_{j1}, b_{j2}) and discrimination parameters (a_{j1}, a_{j2}), $j = 1, \ldots, p$. Hence, knowledge of *A* and *B* will permit the transformation of the parameters in group one to the scale for group two (or the reverse). Note that *A* and *B* are not regression parameters but are scaling constants used to relate sets of item parameter values. In the 3PL model, the pseudo-guessing parameter c_j need not be transformed, as this parameter is not directly affected by the identification constraints. Methods for generating the needed constraints are discussed below. The problem of linkage transformations in polytomous items has received less attention. Methods are available for the graded-response model (Baker, 1992a; Cohen & Kim, 1998; Kim & Cohen, 1995). In polytomous Rasch models such as the rating scale or partial credit models, only the additive constant *B* is needed to link the metrics for the category location parameters.

In bias investigations, the development of the linkage transformation may be complicated by the presence of bias in some items. If bias is present, there will exist no linear transformation that will place the item parameter estimates on a common scale. If bias is ignored and a transformation is developed based on the entire item set, distortions may be produced in the transformed estimates for the unbiased items. The ideal strategy in this case is to base the generation of the linkage transformation on only the unbiased items. This strategy requires a preliminary screening for bias prior to development of the linkage transformation. Studies of linkage in the dichotomous case have generally found the optimal approach to be an iterative sequence in which a preliminary screening is followed by linkage based on the remaining items, followed in turn by another screening, and so forth (Candell & Drasgow, 1988; Kim & Cohen, 1992; Lautenschlager & Park, 1988; Lord, 1980; Marco, 1977; Park, 1988; Park & Lautenschlager, 1990; Segall, 1983). Presumably, this iterative strategy would be effective in the polytomous case as well.

We next describe some general methods for generating linkage transformations and then discuss their use in bias investigations.

Creating Linkage Functions

In the dichotomous item case, the linkage transformation is linear, and this linear transformation requires the constants A and B described above. Two broad methods exist for estimating these constants. The first method bases the constants on the means and standard deviations of the distributions of item location parameter estimates from the two groups under study (Linn et al., 1981; Loyd & Hoover, 1980; Marco, 1977; Vale, 1986; Warm, 1978). These methods will be denoted *moment methods* here. The second group of methods derive constants that minimize group differences in item or test response functions (Divgi, 1985; Haebara, 1980; Stocking & Lord, 1983). These methods will be denoted *response function methods* here.

The simplest moment methods take advantage of the fact that if the item location and discrimination parameters are related as in Equation 7.16, it should be true that

$$\bar{b}_2 = A\bar{b}_1 + B, \quad s_{\mathrm{b}2} = \frac{S_{\mathrm{b}1}}{A}, \tag{7.17}$$

where

$\bar{b}_{\mathrm{k}}$ is the average location parameter estimate in the kth group

S_{bk} is the standard deviation of the location parameter estimates in the kth group (Marco, 1977)

Then we can choose

$$A = \frac{s_{\mathrm{b}1}}{s_{\mathrm{b}2}}, \quad B = \bar{b}_2 - A\bar{b}_1, \tag{7.18}$$

as the required constants. This approach would work perfectly if there is no sampling error in the item parameter estimates. Unfortunately, large sampling errors can strongly influence the means and standard deviations in Equation 7.18. In response to this problem, a number of modifications have been suggested that make the required means and standard deviations more robust. Linn et al. (1981) proposed that the item location parameter estimates be weighted to reflect their standard errors in the process of deriving *A* and *B*. Stocking and Lord (1983) applied a different weighting procedure. These weighting procedures are most useful when the available sample sizes for one or both groups are likely to lead to substantial estimation errors. No single method has been shown to be uniformly superior over variations in sample sizes, test lengths, and parameter values (Baker & Al-Karni, 1991; Candell & Drasgow, 1988; Kim & Cohen, 1992; Stocking & Lord).

Response function methods do not lead to simple formulas for *A* and *B* but instead find *A* and *B* values that will minimize some measure of distance between the item response functions in the two groups. The solution for *A* and *B* generally requires iterative numerical procedures. Divgi (1985) proposed that *A* and *B* be selected to minimize the sum, across items, of the quadratic forms

$$\boldsymbol{\delta}_{\mathrm{j}}^{*\prime}(\boldsymbol{\Sigma}_{\mathrm{j}1}+\boldsymbol{\Sigma}_{\mathrm{j}2})^{-1}\boldsymbol{\delta}_{\mathrm{j}}^{*}, \tag{7.19}$$

where

$$\boldsymbol{\delta}_{\mathrm{j}}^{*\prime} = (a_{\mathrm{j}1}-a_{\mathrm{j}2}^{*},\ b_{\mathrm{j}1}-b_{\mathrm{j}2}^{*}) \tag{7.20}$$

with

$$b_{\mathrm{j}2}^{*} = Ab_{\mathrm{j}1}+B, \quad a_{\mathrm{j}2}^{*} = \frac{a_{\mathrm{j}1}}{A}. \tag{7.21}$$

The matrix $\boldsymbol{\Sigma}_{\mathrm{j}1}$ is the 2 × 2 covariance matrix for $(a_{\mathrm{j}1}, b_{\mathrm{j}1})$, and $\boldsymbol{\Sigma}_{\mathrm{j}2}^{*}$ is the analogous matrix for $(a_{\mathrm{j}2}^{*}, b_{\mathrm{j}2}^{*})$. These matrices are estimated as described in the earlier discussion of Wald statistics. The quadratic form in Equation 7.19 resembles the Wald statistic that would be used to test for parameter invariance under the 2PL model. As shown by Divgi (1985), minimization of the sum of the quadratic forms in Equation 7.19 across items leads to a simple expression for *B* in terms of *A*. The solution for *A* is found iteratively.

Stocking and Lord (1983) presented a response function method that selects *A* and *B* to minimize the difference between two estimates of the examinee's test true score. The first estimate uses the item parameter estimates from one examinee group. The second estimate uses the item parameter estimates from the second examinee group after linear transformation as in Equation 7.21. The objective is to find A and B to minimize the average squared difference between the two true-score estimates across examinees. Unlike the method of Divgi (1985), Lord and Stocking's method requires estimates of the latent trait values for the examinees, so that the true-score estimates can be calculated. This requirement makes Lord and Stocking's method more difficult to use, even though the minimization problem itself is not especially difficult.

Most of the research on linkage has addressed the dichotomous item case, but some work is available for the case of polytomous items. The available methods are simple extensions of the linkage methods used with dichotomous items. Baker (1992a) applied Stocking and Lord's response function method to the linkage problem for the graded response model. The extension modified the calculation of the true-score estimates to include the chosen scaling for the response categories. Baker (1993) provided a software program for implementing the method. Kim and Cohen (1995) modified Divgi's method for use with the graded response model. Cohen and Kim (1998) described the use of several moment methods with the graded response model. They evaluated three moment methods, along with the extensions of Divgi's method and Lord and Stocking's method, in simulated data from the graded response model. All of the methods performed well under MML estimation, suggesting that the simpler moment methods should be useful in the graded response model.

A general review of parameter linkage methods in IRT, apart from bias investigations, can be found in Kolen and Brennan (2010). Most of the linking methods described here can be implemented within the free R program plink (Weeks, 2010).

Bias Investigations

The presence of items that are biased in relation to the groups under study complicates the linkage problem. If the linkage transformation is based on the entire item set, the biased items may distort the transformation, leading to subsequent errors when bias is evaluated using the transformed parameters. McCauley and Mendoza (1985) showed that when biased items were included in the set that generates the linkage transformation, the bias investigation of the item set led to some items being falsely labeled as biased. In response to this problem, the ideal strategy would base the linkage transformation on only the unbiased items. This strategy is difficult to implement because the biased items cannot be identified a priori. Instead, an iterative scheme is introduced that alternates item screening for bias with generation of the linkage transformation based on the remaining items.

Lord (1980; Marco, 1977) presented one iterative scheme to be used with the 3PL model. The steps in this approach were as follows:

1. Combine data from all groups and estimate the item parameters. Use identification constraints that force the distribution of item location parameters to have a mean of zero and a standard deviation of one.
2. Fixing the pseudo-guessing parameters at the values estimated in step (1), separate the groups and estimate the location and discrimination parameters within each group, using the standardization from (1) separately within each group. Evaluate all items for bias.
3. Remove any biased items from the pool.
4. Combine all groups and estimate the latent trait values *W* using the remaining items.
5. Using the estimates of *W* from (4), estimate the item parameters for all items, including those dropped earlier, separately within each group.
6. Evaluate each item for bias using the item parameter estimates found in (5).

Note that only the location and discrimination parameters enter into the evaluations of bias in steps (2) and (6). The pseudo-guessing parameters are given the same values across groups in this approach. The bias evaluations in steps (2) and (6) could use Wald statistics or could be based on other approaches.

Lord's procedure conducts one initial screen for biased items, followed by a second estimation phase, and a final bias screen. Park and Lautenschlager (1990; Park, 1988) modified Lord's procedure by including additional iterations. The steps in this modified procedure are as follows:

1. Combine data from all groups and estimate latent trait values *W*.
2. Separate the groups and estimate all item parameters within each group using the latent trait estimates from step (1). Evaluate all items for bias.
3. Remove any biased items.
4. Working with the remaining items, combine data from all groups and reestimate the latent trait values *W*.
5. Separate the groups and estimate item parameters for all items within each group using the latent trait estimates from step (4). Evaluate the items for bias.
6. Repeat steps (3) through (5) until the same items are identified as biased on successive iterations.

The above method is not limited to the 3PL model and can potentially be used with any model. Both Park and Lautenschlager's method and Lord's method have the drawback that repeated estimation of latent trait values is needed. Given that latent trait estimates are generally not needed for bias investigations, the need for these estimates here is a significant inefficiency.

To avoid the need for repeated latent trait estimation, Segall (1983) suggested the following alternative:

1. Estimate the item parameters within each group separately.
2. Generate a linking function using the estimates from step (1), using one of the linking methods reviewed earlier.
3. After linkage, evaluate all items for bias. Remove any biased items.
4. Using the remaining items, generate new linking functions. Use these new functions to link the parameters for all items, including those removed earlier for bias.
5. Evaluate all items for bias and remove any biased items.
6. Repeat steps (4) and (5) until the same items are flagged as biased on successive iterations.

Segall's method has been used successfully in both real applications and simulations (Candell & Drasgow, 1988; Drasgow, 1987; Park & Lautenschlager, 1990). The available evidence suggests that the use of multiple iterations to establish the parameter linkage in bias investigations leads to better results in comparison to single linkage methods (Candell & Drasgow; Kim & Cohen, 1992; Lautenschlager & Park, 1988; Park & Lautenschlager). The use of multiple iterations may be especially advantageous when the samples are small (Kim & Cohen).

Effect Size Measures

A finding of statistically significant bias is simply the statement that the amount of bias is larger than we would expect to find by chance, given the sample size, number of items, and the model. The statistical significance does not tell us whether the bias found is of a magnitude that is meaningful in practical terms. To determine whether the bias is meaningful, it may be useful to create a measure of the "effect size" for the bias. Some approaches to developing such measures are described here. As will become clear, no fully general measure of effect size is yet available for IRT-based bias investigations. Measures of effect size that are useful with specific models and item formats are available, however. For example, measures for logistic models in dichotomous item data have received substantial attention. Measures that are useful in the polytomous case are less studied. We will begin with the dichotomous case, followed by the polytomous case.

Dichotomous Items

Assuming that estimates of the group-specific item response functions for the biased items are available, it is tempting to evaluate the size of the bias directly by calculating group differences in item parameters. In the case of the Rasch model, the group difference in the item location parameter is directly related to the log of the odds ratio. Letting $P_{\mathrm{r}}(W)$ and $P_{\mathrm{f}}(W)$ be the item response functions for the reference and focal groups, we have

$$\ln\left(\frac{P_{\mathrm{r}}(W)/[1-P_{\mathrm{r}}(W)]}{P_{\mathrm{f}}(W)/[1-P_{\mathrm{f}}(W)]}\right)=b_{\mathrm{f}}-b_{\mathrm{r}}. \tag{7.22}$$

In this case, the relative advantage of the reference group in the odds of passing the item can be written as a simple function of the group difference in location parameters, and it does not depend on W. When more than two groups are involved, each group can be compared to the one group that serves as a reference group, or each pair of groups can be compared. Given the familiarity of the log odds as a measure of effect size in other contexts, the relationship in Equation 7.22 is highly useful.

In multiple parameter models in which group differences exist in additional parameters, the use of the log odds to characterize the bias effect size is less successful. Under these models, the bias effect size will ordinarily depend on W. The bias will be larger within a certain range on the latent variable scale and will be smaller outside of that range. The direction of the bias may even reverse itself along the latent variable scale. The general approach to be taken under these multiple parameter models is to summarize the group difference in item response functions over the latent trait scale. The resulting measures are generically denoted as "area measures" of bias. The label denotes the idea that these measures gauge the magnitude of the bias by the area between the item response functions for a pair of groups.

In creating an area measure, several choices must be made regarding (a) the range on W to include, (b) whether a discrete approximation is to be used, and (c) whether absolute or signed measures of area will be used. For (a), the simplest choice is to use the full range of W. To do so, however, one must include regions of the latent variable scale in which very few examinees are found. Arguably, an area measure should focus on regions in which most examinees appear. For 3PL models, use of the full range for the latent variable leads to undefined (infinite) area measures whenever the pseudo-guessing parameter estimates differ between groups. The use of a bounded area measure that restricts W (e.g., $-3.0 < W < 3.0$) will eliminate these problems. Under (b), early area measures used discrete approximations to the true area (Ironson & Subkoviak, 1979; Linn et al., 1981; Rudner, Getson, & Knight, 1980). If the area between the item response functions is to be calculated without any differential weighting, discrete approximations are unnecessary because exact formulas are available for the Rasch, 2PL, and 3PL models (Kim & Cohen, 1991; Raju, 1988, 1990). The decision about (c) is only relevant when bidirectional bias is present, resulting in a reversal of the two groups in their rank-ordering on the probability of passing the item. Bidirectional bias can only arise in either the 2PL, normal ogive, or 3PL models. An unsigned area measure will gauge the area between the item response functions as an absolute quantity. A signed area measure will calculate the area between the item response functions as a sum of positive and negative areas, the sign reflecting which group is highest in the probability of passing in any given region of the latent variable scale. For example, a signed area index could have a value near zero even though the absolute area between the IRFs is substantial.

The general continuous area measure can be described for the jth item as

$$A_{\mathrm{j}} = \int_S f[P_{\mathrm{jk}}(W) - P_{\mathrm{jm}}(W)]\,dW \tag{7.23}$$

for item response functions from the kth and mth groups, $P_{\mathrm{jk}}(W)$ and $P_{\mathrm{jm}}(W)$, respectively, with $f[\,]$ being a chosen function and S being a chosen range of integration. An unbounded measure will use $S = [-\infty, +\infty]$, while a bounded measure will pick $S = [W_{\mathrm{L}}, W_{\mathrm{U}}]$ for lower bound W_{L} and upper bound W_{U}. An unsigned measure might select

$$f = \left|P_{\mathrm{jk}}(W) - P_{\mathrm{jm}}(W)\right|, \tag{7.24}$$

the absolute value of the difference in item response functions. A signed measure could use the simple difference

$$f = P_{\mathrm{jk}}(W) - P_{\mathrm{jm}}(W). \tag{7.25}$$

For the unbounded case in which $S = [-\infty, +\infty]$, Table 7.1 gives expressions for both signed and unsigned area measures under the Rasch, 2PL, and 3PL

TABLE 7.1

Signed and Unsigned Unbounded Area Measures

Rasch model, *j*th item, comparing groups *k* and *m*

Signed area: $b_{jk} - b_{jm}$

Unsigned area: $|\, b_{jk} - b_{jm} \,|$

Two-parameter logistic model, *j*th item, comparing groups *k* and *m*

Signed area: $b_{jk} - b_{jm}$

Unsigned area, $a_{jk} = a_{jm}$: $|\, b_{jk} - b_{jm} \,|$

Unsigned area, $a_{jk} \neq a_{jm}$:

$$UA_{2pl} = \left| \frac{2(a_{jk} - a_{jm})}{Da_{jk}a_{jm}} \ln\left[1 + \exp\left(\frac{Da_{jk}a_{jm}(b_{jk} - b_{jm})}{a_{jk} - a_{jm}}\right)\right] - (b_{jk} - b_{jm}) \right|$$

Three-parameter logistic model, *j*th item, comparing groups *k* and *m* (common *c* value)

Signed area: $(1 - c)(b_{jk} - b_{jm})$

Unsigned area, $a_{jk} = a_{jm}$: $(1 - c)\ |\, b_{jk} - b_{jm} \,|$

Unsigned area, $a_{jk} \neq a_{jm}$: $(1 - c)UA_{2pl}$

models (Raju, 1988). The 3PL case in which the pseudo-guessing parameters vary over groups is omitted because the area measure is infinite in this case. In all three models, invariance in the discrimination parameters leads to area measures that are simple functions of the location parameter differences. In the 2PL and 3PL models, the signed area measures are zero when the discrimination parameters differ across groups. Given invariance in the discrimination parameters, we would expect the signed and unsigned measures to yield identical results for the size of the bias, apart from sign. In practice, the area measures in Table 7.1 are calculated by replacing the item parameters with their sample estimates. Raju (1990) provides large-sample standard errors for the resulting indices. These standard errors can be used to construct test statistics for testing the null hypothesis that the area measure is zero.

Expressions for the bounded area measures in the Rasch, 2PL, and 3PL models were given by Kim and Cohen (1991). The expressions assume that fixed boundary points $S = [W_L, W_U]$ are chosen by the investigator. The resulting area measures are more complex than those given in Table 7.1, especially when either the discrimination or pseudo-guessing parameters vary over groups. In these cases, the item response functions may intersect, and the formula for the area measure depends on whether the intersection points lie within the interval S or outside of this interval. Under the 3PL, the intersection points do not themselves have simple formulas when $c_{jk} \neq c_{jm}$ and $a_{jk} \neq a_{jm}$. Kim and Cohen suggest that the required intersection points be found using a Newton–Raphson algorithm. There may be zero, one, or two intersection points in this 3PL case. The same authors provide software for computing area measures (Kim & Cohen, 1992).

The few comparisons of performance between the unbounded and bounded continuous area measures have revealed few differences in the two types of measures (Kim & Cohen, 1991), apart from the 3PL case in which the unbounded measure is infinite. Bounded measures have a clear advantage in this case. On the other hand, the bounded measures employ bounds that are arbitrary to some extent, and their calculation requires greater effort. At present, the advantages of the bounded measures in comparison to the simpler unbounded measures have not been clearly demonstrated.

The choice between signed and absolute area measures depends on the item response functions in the groups being compared. The two types of area measures will be identical when the Rasch model holds, and when either the 2PL or 3PL models hold with invariant discrimination and pseudo-guessing parameters. When these parameters are not invariant, the signed and absolute area measures will differ, and the difference can be substantial. If the bias in the items is bidirectional, the signed measure may be close to zero while the absolute measure is substantial. For this reason, the two area measures can provide quite different pictures of the magnitude of the bias across items. The two types of measures have been compared in both real and simulated data (Cohen, Kim, & Subkoviak, 1991; Ironson & Subkoviak, 1979; Kim & Cohen, 1991; McCauley & Mendoza, 1985; Raju, 1990; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984, 1985; Subkoviak, Mack, Ironson, & Craig, 1984). Neither measure has emerged from this research as uniformly better. The choice between them should depend on whether directional or bidirectional bias is found. If the bias is directional, the two measures should yield the same value. If bidirectional bias is present, the absolute measure provides a more realistic index of the difference in the item response functions.

All of the area measures discussed thus far have omitted any explicit differential weighting of the group differences in the item response functions. This weighting would assign varying weights to the differences as a function of *W*. The weights might reflect the relative frequencies or densities of examinees at a given *W*. Alternatively, the weights might be inversely related to the standard error associated with the estimated item response function at a given *W*. As Wainer (1993) emphasizes, a great advantage in weighting by the distribution of examinees is that bias in regions of low density will be given little weight because few examinees are affected by bias in this region.

Early weighting schemes were based on discrete approximations to various indices in Table 7.1 (Linn et al., 1981; Shepard et al., 1984, 1985). In practice, these investigators found that the weighted and unweighted versions of different area measures yielded similar results. Wainer (1993) proposed four weighted continuous area measures, in each case using weights based on the latent variable distribution in the focal group. He termed these area measures "standardized indices of impact." Two of the indices are unsigned measures, and two of them are signed. Letting N_f be the number of focal group examinees in the sample, the four indices are

$$T_{1j} = \int_{-\infty}^{+\infty} [P_{jf}(W) - P_{jr}(W)] g_f(W)\, dW \tag{7.26}$$

$$T_{3j} = \int_{-\infty}^{+\infty} [P_{jf}(W) - P_{jr}(W)]^2 g_f(W)\, dW \tag{7.27}$$

with $T_{2j} = N_f T_{1j}$ and $T_{4j} = N_f T_{3j}$. The inclusion of focal group sample sizes in T_{2j} and T_{4j} is intended to gauge "total impact" by multiplying the "average" impacts T_{1j} and T_{3j} by the number of focal group examinees. Indices T_{3j} and T_{4j} are unsigned measures. All four indices may be used with any dichotomous item IRT model.

Wainer's indices have clear advantages as measures of bias effect size. All of the indices are bounded yet do not require the use of arbitrary boundary values for W. Furthermore, it makes sense to downplay the bias in regions of W in which few examinees are found. At least three difficulties must be overcome before the measures will be widely used. First, it is unclear at present which values of the indices should be considered "large." Second, the relative performance of these weighted indices in comparison to the unweighted continuous indices is unknown. It may be true that in many applications, the weighted and unweighted yield similar rank-orderings of the items in terms of bias. Finally, no simple formulas for $\{T_{1j}, T_{2j}, T_{3j}, T_{4j}\}$ exist that are analogous to the continuous unweighted area measures. Wainer (1993) suggests several computational methods that use existing IRT software. These methods take advantage of the quadrature already performed in MML software such as BILOG-MG (Zimowsk, Muraki, Mislevy, & Bock, 2003). To date, none of the indices $\{T_{1j}, T_{2j}, T_{3j}, T_{4j}\}$ have been used extensively in bias research.

In addition to the direct area measures, some researchers have developed confidence region methods for either the item response functions or the group difference in item response functions (Hauck, 1983; Linn et al., 1981; Lord & Pashley, 1988; Pashley, 1992; Thissen & Wainer, 1990). Instead of summarizing the bias effect size with a single measure, these methods express the effect size as a function of W, graphically displaying the amount of bias as it varies over the latent variable scale. In addition, the confidence region methods incorporate statistical uncertainty into the display by creating confidence bounds on the item response functions or their differences, at any desired level of confidence. Pashley presents a method for creating confidence bands for the difference between item response functions under the 3PL model. The method relies on large-sample approximations to the standard errors of the item parameter estimates and also ignores variability associated with the estimation of the latent variable scores. Preliminary evidence using the method in real data showed consistency with the results of Mantel–Haenszel tests in the same data. Thissen and Wainer present methods for creating confidence "envelopes" for individual item response functions and suggest that in bias applications, the degree of overlap in the envelopes for each group's function can be used to informally assess the extent of the bias. None of these confidence region methods have yet been widely used in bias applications.

Polytomous Items

In the polytomous item case, area measures of bias are most easily defined in relation to the item true-score functions within each group: $E_{\mathrm{k}}(X_{\mathrm{j}}|W)$ for $k = 1,\ldots,K$. These true-score functions are monotonically increasing in W under very general conditions (Molenaar, 1997). For commonly used polytomous IRT models such as the partial credit, rating scale, graded response, and generalized partial credit models, bias in the item true-score functions can exist if and only if the category response functions also exhibit bias (Chang & Mazzeo, 1994). Assuming that one of these models is an adequate baseline model for the data, we can evaluate the size of any bias by comparing the item true-score functions across groups.

As in the dichotomous case, we can distinguish two forms of bias in the true-score functions. Let $E_{\mathrm{r}}(X_{\mathrm{j}}|W)$ and $E_{\mathrm{f}}(X_{\mathrm{j}}|W)$ be the true-score functions for the jth item in the reference and focal groups, respectively. Directional bias will be said to exist for item j when

$$D_{\mathrm{Ej}}(W) = E_{\mathrm{r}}(X_{\mathrm{j}}\,|\,W) - E_{\mathrm{f}}(X_{\mathrm{j}}\,|\,W) \tag{7.28}$$

is nonzero for some W and does not reverse its sign throughout the range of W. Bidirectional bias is present when $D_{\mathrm{Ej}}(W)$ is nonzero for some W and reverses its sign across the range of W. The form of bias that is present is an important consideration in choosing an area measure. An absolute area measure for the jth item is

$$A_{\mathrm{Eaj}} = \int_{-\infty}^{+\infty} |\,D_{\mathrm{Ej}}(W)\,|\,dW. \tag{7.29}$$

This area measure will be positive if either directional or bidirectional bias is present in the jth item. A signed area measure is

$$A_{\mathrm{Asj}} = \int_{-\infty}^{+\infty} D_{\mathrm{Ej}}(W)\,dW. \tag{7.30}$$

This area measure need not be positive and may be close to zero when bidirectional bias is present. Both of these area measures are discussed by Cohen et al. (1993). Expressions for the large-sample standard errors for both area measures are given in that paper, but the expression for the standard error of A_{Eaj} requires distributional assumptions that are probably unrealistic in most applications. Cohen et al. note that A_{Esj} may be small when bidirectional bias is present. They suggest that other measures be used in such cases.

The evaluation of either area measure in Equations 7.29 and 7.30 requires that the true-score function be estimated in each group. For the *k*th group,

$$E_{\text{k}}(X_{\text{j}} \mid W) = \sum_{m=0}^{c-1} mP_{\text{k}}(X_{\text{j}} = m \mid W), \tag{7.31}$$

with $P_{\text{k}}(X_{\text{j}} = m|W)$ being the category response function for the *m*th category of the *j*th item in the *k*th group. Then for the reference and focal groups, we have

$$D_{\text{Ej}}(W) = \sum_{m=0}^{c-1} m[P_{\text{r}}(X_{\text{j}} = m \mid W) - P_{\text{f}}(X_{\text{j}} = m \mid W)]. \tag{7.32}$$

The value of $D_{\text{Ej}}(W)$ is easily calculated at any given value of *W*, given estimates for the category response functions in the reference and focal groups. It is then possible to graph $D_{\text{Ej}}(W)$ as a function of *W* for display purposes. The area measures A_{Eaj} and A_{Esj} are then calculable in theory, but in practice the required integral may be difficult to solve in closed form. Quadrature approximations will be necessary in such cases. Analytical formulas for A_{Esj} are available for the graded response model, as are formulas for a bounded version of A_{Eaj} (Cohen et al., 1993).

Summary

As the foregoing makes clear, a bias effect size measure can be defined for nearly any IRT model. It is fair to ask, however, whether any of the measures so defined are really useful in a practical sense. What is a "large" effect size in any of these cases? How should one judge whether an effect is large or small? Answers to these and other similar questions are not yet clear in general. In fact, it may not make sense to formulate a general rule for what constitutes a large effect size across all situations in which such effect sizes might be considered. Items are embedded in test forms, and the same effect size might be considered large or small depending on the length of a test. In short tests, a single item will have more influence on the resulting test score, and a given effect size will have more impact. In long tests such as those used in many educational settings, a single item ordinarily has less relative influence on the test score. The purpose of the test should also be considered. High-stakes tests used for selection in educational or employment settings may require stricter standards for exclusion of biased items, even when the bias effect size seems small. Remarkably, while much effort has gone into research on the development of IRT methods for detecting biased items, questions about the practical impact of item bias have not received much systematic attention within IRT (for an interesting exception, see Stark, Chernyshenko, & Drasgow, 2004). The next section describes a general method for item bias detection that explicitly considers the impact of item bias on the test score.

The DFIT Approach

Raju, Van der Linden, and Fleer (1995) proposed an IRT-based framework for evaluating bias at both the item and test levels for tests consisting of dichotomous items. This approach to bias evaluation is known by the acronym DFIT, which stands for *differential functioning of items and tests*. The DFIT approach provides effect size measures for bias at the level of the test item but also at the whole test level. The latter is an especially useful feature of the method, as the area measures described earlier have no simple extensions to the test level. The DFIT approach has been generalized to include polytomous items (C. P. Flowers, Oshima, & Raju, 1999) and dichotomous items modeled using multidimensional latent variables (Oshima, Raju, & Flowers, 1997). We first focus on the dichotomous, unidimensional application. The polytomous application is discussed below.

The DFIT approach begins with the assumption that a baseline IRT model has been found to fit in both the reference and focal groups, with separate group-specific estimates of all item parameters and latent variable scores W being available. It is also assumed that the metrics of the item parameter estimates have been linked using one of the methods discussed earlier. The DFIT method is not model specific. Any IRT model for dichotomous items may be used as a baseline model, providing that the model fits the data in both groups. The method is also flexible with regard to the estimation method chosen for the item parameters and latent variable scores. Raju et al. (1995) used MML estimation for the item parameters, with Bayesian maximum a posteriori estimation for the latent variable scores.

The DFIT method begins by defining a measure of differential test functioning (DTF) that applies to the p items taken as a whole. Assuming that the test score is calculated as an unweighted sum of the item scores, the conditional expected value for the test score T_i in the kth group for the ith person is

$$E_k(T_i \mid W_i) = \sum_{j=1}^{p} P_{jk}(W_i), \tag{7.33}$$

where $k = r, f$ for the reference and focal groups, $j = 1, \ldots, p$. Assuming that the item response functions in the reference and focal groups are known and that a score on W for the ith person is known, it is possible to evaluate the conditional expected value in Equation 7.33 for the ith person under *both* the reference and focal group models. If no bias exists in any of the p items, the expected test score is the same for the ith person regardless of which model is used because the models are identical across groups. On the other hand, if any of the p items are biased, the item response functions for these items differ across groups, and so the expected test score for the ith person may have different values depending on the group. To capture this difference, define

$$D_{i|W} = E_f(T_i \mid W_i) - E_r(T_i \mid W_i) \tag{7.34}$$

as the difference in the conditional expected test scores for the ith person under the two item response functions. We can then define an index of DTF as

$$DTF = E_f(D_{i|W}^2) = \int D_{i|W}^2 g_f(W)\, dW, \tag{7.35}$$

where $g_f(W)$ is the density function for W in the focal group. In practice, DTF is defined in relation to focal group members only, so that only persons who are members of the focal group receive values on $D_{i|W}$.

Several features of the DTF measure in Equation 7.35 should be noted. First, weighting by $g_f(W)$ will give more weight to values of $D_{i|W}$ in regions of the latent variable scale in which most focal group members are found. Second, the presence of item-level bias may, or may not, lead to meaningful bias at the test level as measured by DTF. As argued by Raju et al. (1995), item level biases that operate in different directions, with some items favoring the reference group and others favoring the focal group, may cancel out at the test level and yield low values for DTF. This compensatory feature of the DTF measure is useful when the test-level score is of inherent interest, as when an existing test is evaluated for bias.

In practice, the DTF measure in Equation 7.35 is estimated by replacing $P_{jk}(W_i)$ with $\hat{P}_{jk}(\hat{W}_i)$ in Equation 7.33, using estimated item parameters and latent variable scores. The integral in Equation 7.35 is avoided by realizing that

$$DTF = \sigma_D^2 + \mu_D^2. \tag{7.36}$$

Here $\mu_D = \mu_{Tf} - \mu_{Tr}$, and

$$\mu_{Tf} = \int E_f(T_i \mid W_i) g_f(W)\, dW, \tag{7.37}$$

$$\mu_{Tr} = \int E_r(T_i \mid W_i) g_f(W)\, dW. \tag{7.38}$$

The parameter σ_D^2 is the variance of $D_{i|W}$ in the focal group. Note that both μ_{Tr} and μ_{Tf} are defined with respect to the focal group also. We can estimate μ_D by estimating $D_{i|W}$ for every focal group member and then using the sample mean of $\hat{D}_{i|W}$ as $\hat{\mu}_D$. Similarly, the variance σ_D^2 is estimated using the sample variance of $\hat{D}_{i|W}$ within the focal group. Finally, DTF is estimated by substituting $\hat{\mu}_D$ and $\hat{\sigma}_D^2$ in Equation 7.36.

The DTF estimate indicates the magnitude of the test-level bias. Raju et al. (1995) suggest several alternative test statistics that might be used to test the null hypothesis that DTF is zero in the population from which the focal group sample is drawn. The proposed test statistics are preliminary and do not formally incorporate adjustments needed to acknowledge the degrees of freedom lost due to estimation of the item parameters and latent variable scores. The tests also assume a normal distribution for $D_{i|W}$ in the focal group, an assumption that will be difficult to verify in practice. In spite of these potential problems, simulation evidence for the proper Type I error and power behavior of the test statistics is promising (Raju et al.).

The DFIT approach extends to the item level in two ways, either by considering the bias for a given item in relation to the bias shown by other items in the test (compensatory DIF, or CDIF) or by considering bias in each item as if that item is the only item in the test that manifests bias (non-compensatory DIF, or NCDIF). The rationale for CDIF begins by noting that DTF can be expressed as

$$DTF = E_f\left[\left(\sum_{j=1}^{p} d_{ij}\right)^2\right]. \tag{7.39}$$

Here $d_{\text{ij}} = P_{\text{jf}}(W_{\text{i}}) - P_{\text{jr}}(W_{\text{i}})$. But the expression in Equation 7.39 can be rewritten as

$$DTF = \sum_{j=1}^{p} [\text{Cov}(d_{\text{ij}}, D_{\text{i|W}}) + \mu_{\text{dj}}\mu_{\text{D}}]. \tag{7.40}$$

Here $\text{Cov}(d_{\text{ij}}, D_{\text{i|W}})$ is the covariance between d_{ij} and $D_{\text{i|W}}$ in the focal group, and μ_{dj} is the expected value of d_{ij} in the focal group. Equation 7.40 suggests that DTF can be additively decomposed into a sum of DIF terms across p items. CDIF is then defined for the jth item as the term for that item in Equation 7.40

$$CDIF_{\text{j}} = \text{Cov}(d_{\text{ij}}, D_{\text{i|W}}) + \mu_{\text{dj}}\,\mu_{\text{D}}. \tag{7.41}$$

CDIF is unlike the standard definition of DIF in that it considers DIF for the jth item as it relates to the DIF in the other $p - 1$ items. The stronger the association between DIF on the jth item and DIF in the other items, the greater will be the CDIF for the jth item. Also, nonzero values for CDIF across items need not combine to yield nonzero DTF because the item-level DIF may go in different directions, with mutual cancellation at the test level.

NCDIF is defined as the special case of CDIF in which the other $p - 1$ items have no DIF, leading to

$$NCDIF_{\text{j}} = \sigma_{\text{dj}}^2 + \mu_{\text{df}}, \tag{7.42}$$

where σ_{dj}^2 is the variance of d_{ij} in the focal group. We can rewrite $NCDIF_{\text{j}}$ as

$$E_{\text{f}}[(P_{\text{jf}}(W_{\text{i}}) - P_{\text{jr}}(W_{\text{i}}))^2] = \int (P_{\text{jf}}(W_{\text{i}}) - P_{\text{jr}}(W_{\text{i}}))^2 g_{\text{f}}(W)\,dW. \tag{7.43}$$

NCDIF is closely related to the unsigned weighted area measure suggested by Wainer (1993). Raju et al. (1995) described the relationship of $NCDIF_{\text{j}}$ to other standard indices of bias, such as Lord's chi-square index. Unlike $CDIF_{\text{j}}$, $NCDIF_{\text{j}}$ values for the p items do not sum to the DTF value for the set of items. The unsigned nature of $NCDIF_{\text{j}}$ removes any cancellation in this sum that would occur when the directions of the biases differ across items.

$NCDIF_{\text{j}}$ and $CDIF_{\text{j}}$ are estimated by first estimating $P_{\text{jf}}(W_{\text{i}})$ and $P_{\text{jr}}(W_{\text{i}})$ for every member of the focal group once parameter estimates and latent variable scores are available. Then $\hat{d}_{\text{ij}}$ and $\hat{D}_{\text{i|W}}$ can be calculated, and

$\text{Cov}(d_{ij}, D_{i|W})$ is estimated using these values. Similarly, μ_{dj} and μ_D are estimated as averages in the focal group. Finally, $NCDIF_j$ and $CDIF_j$ are estimated by substituting all of these quantities in Equations 7.41 and 7.42. Raju et al. (1995) did not develop any significance test procedures for $CDIF_j$. A test statistic is available for $NCDIF_j$. A chi-square statistic for the null hypothesis that the $NCDIF_j$ is zero for the jth item is

$$\chi^2_{\text{NCDj}} = \frac{\text{est}(NCDIF_j)}{\hat{\sigma}^2_{dj}/N_f}, \tag{7.44}$$

where $\text{est}(NCDIF_j)$ is the estimated $NCDIF_j$ for the jth item. This test statistic has $df = N_f$. The square root of the chi-square statistic can be regarded as a standard normal deviate and can be referred to the standard normal distribution.

When the bias investigation is focused on an intact test and interest lies in the total test score, Raju et al. (1995) suggest that DTF first be examined. If meaningful levels of DTF are found, Raju et al. suggest that $CDIF_j$ values be examined for the p items, with the item with the largest $CDIF_j$ value being dropped from the test. DTF is then reestimated for the shorter test. This iterative process is continued until the level of DTF falls below an acceptable limit. The $NCDIF_j$ values are useful if the individual items are specifically of interest. For example, when tests are being assembled from a larger item pool and the goal is to purge this larger pool of biased items, the $NCDIF_j$ values should be the focus of interest. The $NCDIF_j$ values will be less dependent than the $CDIF_j$ values on which other items were selected for the particular test form under study.

C. P. Flowers et al. (1999) describe the extension of the DFIT approach to ordered-categorical items. Any of the available polytomous IRT models may be used for the required calculations, assuming that the chosen model provides an adequate fit. It is also assumed that any needed parameter linkage has been achieved prior to the DFIT calculations. The first step begins with the item parameter estimates obtained within the reference and focal groups, along with the latent variable score estimates W_i for all focal group members. Let $P_{jmf}(W)$ and $P_{jmr}(W)$ be the category response functions for the jth item and mth category in the focal and reference groups, respectively. Assume that the item response categories are scored as $m = 0, 1, \ldots, C$ in each group. Then the conditional expected score for the ith focal person on the jth item in the reference group is

$$E_r(X_j \mid W_i) = \sum_{m=0}^{C} mP_{jmr}(W_i). \tag{7.45}$$

Similarly, the conditional expected score for the *i*th focal person in the focal group is

$$E_{\mathrm{f}}(X_{\mathrm{j}} \mid W_{\mathrm{i}}) = \sum_{m=0}^{C} mP_{\mathrm{jmf}}(W_{\mathrm{i}}). \tag{7.46}$$

Then the difference d_{ij} is defined as

$$d_{\mathrm{ij}} = E_{\mathrm{f}}(X_{\mathrm{j}} \mid W_{\mathrm{i}}) - E_{\mathrm{r}}(X_{\mathrm{j}} \mid W_{\mathrm{i}}). \tag{7.47}$$

At the test level, the conditional expected test score T_{j} for the *i*th person in the reference group is

$$E_{\mathrm{r}}(T_{\mathrm{i}} \mid W_{\mathrm{i}}) = \sum_{j=1}^{p} E_{\mathrm{r}}(X_{\mathrm{j}} \mid W_{\mathrm{i}}), \tag{7.48}$$

and for the *i*th person in the focal group, the conditional expected value is

$$E_{\mathrm{f}}(T_{\mathrm{i}} \mid W_{\mathrm{i}}) = \sum_{j=1}^{p} E_{\mathrm{f}}(X_{\mathrm{j}} \mid W_{\mathrm{i}}) \tag{7.49}$$

Given these definitions, the difference $D_{\mathrm{i|W}}$ is

$$D_{\mathrm{i|W}} = E_{\mathrm{f}}(T_{\mathrm{i}} \mid W_{\mathrm{i}}) - E_{\mathrm{r}}(T_{\mathrm{i}} \mid W_{\mathrm{i}}). \tag{7.50}$$

From this point on, the definitions of DTF, $CDIF_{\mathrm{j}}$, and $NCDIF_{\mathrm{j}}$ are identical to the dichotomous case. No other special formulas are needed for the polytomous item case. Test statistics are also identical in the dichotomous and polytomous cases.

Oshima et al. (1997) extend the DFIT framework to items that are fit by a multidimensional latent variable model. The dichotomous item case is illustrated in Oshima et al., although no barrier appears to exist that would prevent the application to polytomous items also. The multidimensional extension will not be described here.

An Example

To illustrate the use of IRT in the detection of measurement bias, we use data on a subtest within the College Basic Academic Subjects Examination (CBASE). CBASE is a set of achievement tests in four subjects: mathematics, English, science, and social studies. The tests are intended for students enrolled in college and designed to assess knowledge and skills of the sort that would be part of most undergraduate general curricula. The entire exam consists of 180 multiple-choice items, along with an additional essay portion for writing assessment. For more information on CBASE, see Osterlind, Robinson, and Nickens (1997), L. Flowers, Osterlind, Pascarella, and Pierson (2001), and Pike (1992).

The analyses to be described here used 11 items from the geometry section of the mathematics portion of CBASE. The mathematics test contains three subtests: general mathematics, algebra, and geometry. We will compare males and females on the geometry items. A total of 5,486 examinees provided data, with 1,034 males and 4,452 females. Table 7.2 gives the proportions passing each of the 11 multiple-choice items by gender. Males show a consistently higher proportion passing, although the gap between males and females varies considerably across items. The geometry subtest was selected for these analyses in preference to the entire mathematics test in hopes that the subtest would more closely adhere to unidimensionality for purposes of IRT analyses.

The first set of analyses examined the dimensionality of the 11 items using CFA in Mplus 5.21 (Muthén & Muthén, 1998–2006). A single-factor model was fit simultaneously within the male and female groups using the factor model for ordered-categorical data (see Chapter 5). In these analyses, no invariance constraints were imposed apart from those needed for identification.

Given that the items are scored as binary, all threshold parameters were initially constrained to invariance. The loading for item *Q*87 was

TABLE 7.2

Proportions Passing the CBASE Geometry Items by Gender

Item Number	Males (*N* = 1,034)	Females (*N* = 4,452)
*Q*83	.766	.618
*Q*84	.601	.488
*Q*85	.813	.742
*Q*86	.618	.496
*Q*87	.860	.809
*Q*88	.653	.583
*Q*89	.590	.387
*Q*90	.544	.375
*Q*91	.808	.728
*Q*92	.837	.784
*Q*93	.779	.642

fixed to one in both groups, with the remaining loadings free. This item was selected because of the small group difference in proportion passing for that item. The factor mean for males was fixed to zero, with the factor mean in the female group being free. Factor variances were also free. In addition, the scaling parameter for item *Q*87 was fixed to invariance at 1.0 (Millsap & Yun-Tien, 2004). This model led to numerical problems due to the constraints on the thresholds for items *Q*88 and *Q*89. To circumvent this problem, the thresholds were permitted to vary across groups for these two items, and the scaling constants for the items were constrained to invariance instead. The new model produced convergence on a solution. The global fit statistics suggest that while the model is wrong, the approximate fit is reasonably good. The chi-square is 346.69 with *df* = 82. The CFI value is .977 and the RMSEA value is .034. Examining the residuals for the sample tetrachoric correlation matrices in each group reveals that for the male group, 9 of 55 residuals exceed .05 in absolute value, and in the female group, 12 of 55 residuals exceed .05 in absolute value. The largest residual in the male group is −.123 for the correlation between *Q*86 and *Q*92. The largest residual in the female group is .135 for the correlation between *Q*92 and *Q*93. The approximate fits afforded by the single-factor model in each group are good enough to proceed to the next stage. The results also suggest that a 2PL model might fit the data in each group.

In the next step, the IRTLRDIF (version 2.0b, Thissen, 2001) program was used to provide LR tests of invariance for each item under the 2PL model. Given no prior research on which of the 11 items might violate invariance over gender, we begin with a procedure that evaluates each item for possible bias, using all other items as the anchor. Items that were flagged for bias using this procedure and anchor definition were dropped from the anchor before the next step. The next step used a designated anchor consisting only of items that had not been flagged as biased in the first round. All items were again evaluated for bias using this designated anchor. This process was repeated until all of the 11 items were divided into two categories: anchor items showing no evidence of bias and items that were flagged for evidence of statistically significant bias. For the 2PL model, an item is flagged for bias if the LR chi-square exceeds the critical value at an alpha of .05 with *df* = 2. Once flagged, separate *df* = 1 tests were performed for invariance in the difficulty and discrimination parameters separately.

Table 7.3 provides the parameter estimates under the 2PL model for each of the 11 items in the male and female groups. These estimates were actually provided by MULTILOG 7.0 (Thissen, 1991), following completion of the IRTLRDIF analyses. At the end of those analyses, items *Q*83, *Q*84, *Q*90, *Q*91, and *Q*93 were designated anchor items. Their parameters were estimated in MULTILOG under invariance constraints across gender. The IRTLRDIF analyses found statistically significant group differences

TABLE 7.3

Item Parameter Estimates for the CBASE Geometry Items by Gender

	Males (N = 1,034)		Females (N = 4,452)	
Item Number	a_j	b_j	a_j	b_j
*Q*83	1.97	−.42	1.97	−.42
*Q*84	1.21	.06	1.21	.06
*Q*85[a]	1.73	−.70	1.73	−.92
*Q*86[a]	1.35	.11	1.35	.01
*Q*87[a]	1.62	−1.03	1.62	−1.27
*Q*88[a]	1.14	−.11	1.14	−.38
*Q*89[a]	1.03	.16	1.03	.54
*Q*90	1.22	.51	1.22	.51
*Q*91	1.21	−1.02	1.21	−1.02
*Q*92[a]	1.51	−.93	1.01	−1.53
*Q*93	1.40	−.59	1.40	−.59

[a] Statistically significant group differences at $p < .05$.

in difficulty parameters for items *Q*85, *Q*86, *Q*87, *Q*88, *Q*89, and *Q*92. Significant group differences in discrimination parameters were found only for item *Q*92, and those differences were marginally significant. The estimates in Table 7.3 reflect the corresponding pattern of group differences. For items in which group differences in difficulty parameters and no group differences in discrimination parameters were found, separate difficulty estimates by group are given, with invariant discrimination parameters. Only item *Q*92 shows group differences on both parameters. Finally, the estimates in Table 7.3 were obtained assuming that the latent variable is distributed normally in both groups, with $\mu_f = 0$, $\sigma_f = 1.0$ for females and $\mu_m = 60$, $\sigma_m = 1.0$, for males.

Among the items showing group differences in Table 7.3, items *Q*88 and *Q*89 had the largest chi-square statistics from IRTLRDIF for the test of invariance in difficulty parameters (9.5 and 14.1, respectively, at $df = 1$). For item *Q*89, the item is considerably harder for females than males of equal status on the latent variable. This result is consistent with the proportions passing the item in Table 7.2. On the other hand, the difficulty estimates for *Q*88 show that it is actually easier for females than we would expect, even though males pass the item at a higher rate overall. In other words, we expect to see an even larger gap in the proportion passing the item if the item is functioning identically across gender. In fact, five of the six items flagged as biased show bias against males, a surprising result. The item with the largest bias, however, is item *Q*89, and it is harder for females than males.