Appendix 1

CAOC Items

Item	Pessimism
1	Most of the programs that are supposed to solve problems around here will not do much good.
3	Attempts to make things better around here will not produce good results.
5	Suggestions on how to solve problems won't produce much real change.
7*	Plans for future improvements will not amount to much.
	Dispositional Attribution
2*	The people responsible for solving problems around here do not try hard enough to solve them.
4	The people responsible for making improvements do not know enough about what they are doing.
6	The people responsible for making changes around here don't care enough about their jobs.
8	The people responsible for making changes around here do not have the skills needed to do their jobs.

* deleted and do not form part of the revised 3 item sub-scales.

JOURNAL OF APPLIED MEASUREMENT, *9*(4), 387-408
Copyright© 2008

Assessment of Differential Item Functioning

Wen-Chung Wang
Hong Kong Institute of Education

This study addresses several important issues in assessment of differential item functioning (DIF). It starts with the definition of DIF, effectiveness of using item fit statistics to detect DIF, and linear modeling of DIF in dichotomous items, polytomous items, facets, and testlet-based items. Because a common metric over groups of test-takers is a prerequisite in DIF assessment, this study reviews three such methods of establishing a common metric: the equal-mean-difficulty method, the all-other-item method, and the constant-item (CI) method. A small simulation demonstrates the superiority of the CI method over the others. As the CI method relies on a correct specification of DIF-free items to serve as anchors, a method of identifying such items is recommended and its effectiveness is illustrated through a simulation. Finally, this study discusses how to assess practical significance of DIF at both item and test levels.

Requests for reprints should be sent to Wen-Chung Wang, Hong Kong Institute of Education, Department of Educational Psychology, Counseling and Learning Needs, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong, e-mail: wcwang@ied.edu.hk.

In recent years, the assessment of differential item functioning (DIF) (Holland and Wainer, 1993) has become a routine practice of item analysis. Many commercial tests have undergone comprehensive DIF analysis before being released, especially high-stakes tests. DIF is also a very popular topic of academic research. A survey of DIF entries on the PsycINFO database yields 427 journal articles up through 2007. This chapter not only provides an introduction to DIF analysis, but it also addresses several complicated issues on DIF analysis that should be more fully investigated. In particular, this chapter makes a distinction between item misfit and DIF. Through simulation studies, this chapter demonstrates that the standard item infit and outfit statistics (Wright and Masters, 1982) are not powerful enough to detect DIF. Hence, more advanced detection methods are needed. This chapter develops a theoretical framework to model DIF in dichotomous items, polytomous items, facet structures, and testlet-based items in which one or multiple grouping variables with more than two categories are simultaneously analyzed.

In DIF analysis, different groups of test-takers need to be placed on the same metric such that their responses to a studied item (the one to be detected for DIF) can be compared. Otherwise, DIF detection is, by definition, impossible. A common metric over groups, or a matching variable as it is used to match test-takers with identical latent trait levels, is a prerequisite of DIF analysis. Three major methods of establishing a common metric over groups have been developed. Two of them, although widely employed in practice and implemented in computer programs such as ConQuest (Wu, Adams, and Wilson, 1998) and Winsteps (Linacre, 2003), are based on assumptions that are too stringent to implement in reality. The third method, although not as widely recognized and practiced, yields appropriate DIF detection as long as one DIF-free item is chosen to function as an anchor such that a common metric can be established. In this chapter, the limitations and advantages of these three methods in establishing a common metric over groups are not only clarified but also demonstrated through a simulation study. Through the analysis of a simulated data set, this chapter demonstrates an iterative procedure to locate a set of DIF-free items needed for the third method. Finally, in addition to appraising the statistical significance of DIF, this chapter explains how to ascertain the practical significance of DIF at item level and test level.

A Definition of DIF

An item is said to exhibit DIF when it functions differently for different groups of test-takers. Specifically, it occurs when test-takers having identical levels on the latent trait that the test was designed to measure but belonging to different groups, have different probabilities of endorsing (or answering correctly) a particular item. For example, an item that is intended to measure mathematical proficiency may contain slang that is unintelligible to people from minority groups, such that a minority test-taker would have a lower probability of answering that item correctly than would a majority test-taker, even they both possesses equal proficiency in mathematics. When a test contains DIF items, strictly speaking, the test no longer measures the same construct for different groups of test-takers and the test scores can no longer be considered comparable over groups. In other words, the test is not invariant. However, a real test can never be perfect and always contains DIF to some degree. In practice, as long as the magnitude of DIF is reasonably small, then the test is practically invariant.

In a typical DIF study, the item responses of two groups of test-takers are examined: a reference group, which is often the majority, and a focus group, which is often the minority. The grouping variable in DIF assessment, although usually bi-categorical, can be multi-categorical (e.g., ethnicity) or even continuous (e.g., age). Moreover, the grouping variable does not need to be a demographic category (e.g., gender, ethnicity, or socio-economic status). It can be any variable of interest. Many tests, for example, have two delivery systems: paper-and-pencil and computer-based testing. It is highly desirable to maintain construct invariance over the delivery systems. In this case, delivery system is the grouping variable. Today, many countries participate in international large-scale educational assessment, such as the Programme for International Student Assessment and the Trends in International Mathematics and Science Study. It is important to ensure that the test is invariant over countries such that international comparison is possible. Here, country is the grouping variable.

It is important to make a distinction between DIF and misfit. Item misfit can have many causes, such as local dependence, multidimensionality, inappropriate discrimination power, guessing, etc. In contrast, an item that exhibits DIF functions differently for different groups of test-takers. Within the context of Rasch measurement (Rasch, 1960), or item response theory (IRT; Lord, 1980), an item is said to exhibit DIF when it meets the model's expectation reasonably well for each group of test-takers, but does not have the same parameters for different groups. In accordance with this definition, one might conduct a Rasch analysis of the whole data set (including all groups of test-takers) and a separate Rasch analysis of the data set for each group. If an item is found to be poor-fitting in the whole data set or within any group of test-takers, it should be removed from subsequent DIF analysis. This procedure ensures that item parameter estimates obtained from each group are meaningful and may be compared for evidence of DIF.

One might speculate whether the standard item infit and outfit statistics can detect DIF since DIF is actually a type of model-data misfit. In the event they can, practitioners might simply complete a single Rasch analysis of the entire data set and use the resulting statistics to identify poor-fitting items. Unfortunately, the results of studies that have attempted to detect DIF via fit statistics are disappointing. Fit statistics are used to assess the overall degree to which an item meets the model's expectation. If an item fits the model's expectation reasonably well within groups, but also has different parameters for different groups (a definition of DIF), then fit statistics will not be powerful enough to detect such model-data misfit (Smith, 1994, 1996; Smith and Suh, 2003). In order to verify this argument, data for a 50-item test is simulated according to the Rasch model where both the reference and focal groups consist of 500 persons. Item 25 is simulated to have DIF: it has a difficulty of 0 logits for the reference group and 1 logit for the focal group. The other 49 items are all simulated as DIF-free. The Rasch model is then fitted to the whole data set (i.e., 1000 persons) using Winsteps and ConQuest. Table 1 summarizes the infit and outfit mean square error statistics for the 50 items. The infit and outfit mean square errors for item 25 are 1.03 and 1.00 given by Winsteps and 1.03 and 1.01 given by ConQuest. These statistics are very similar to those of the DIF-free items and therefore in this case, the infit and outfit statistics are not powerful in detecting DIF.

Linear Modeling of DIF

In addition to statistically testing the difference in item parameter estimates between groups for evidence of DIF, one can develop a model to take into account the effect of grouping variables on item parameters. Such a modeling makes DIF analysis more structural and general, especially in complicated testing situations (Wang, 2000a, 2000b, 2000c). In order to begin, assume there

Table 1

Infit and outfit mean square errors for the simulated 50-item test in which item 25 has DIF

	WINSTEPS		CONQUEST	
	Infit	Outfit	Infit	Outfit
Mean	1.00	1.00	1.00	1.00
Maximum	1.06	1.13	1.06	1.10
Minimum	0.94	0.91	0.93	0.91
Item 25	1.03	1.00	1.03	1.01

are a set of G ($g = 1, \ldots, G$) groups of test-takers of interest for DIF. For a dichotomous item, assume the Rasch model holds for every group of test-takers:

$$\ln\left(p_{ni1}/p_{ni0}\right)_g = \theta_n - B_{ig}, \tag{1}$$

where p_{ni1} is the probability of answering correctly (or endorsing) item i for person n with a latent trait level of θ_n; p_{ni0} is the probability of answering incorrectly; and B_{ig} is the difficulty of item i for group g. The subscript ig for the difficulty parameter suggests that DIF is a form of item-group interaction. DIF detection therefore tests whether the difficulties are all identical over groups:

$$B_{i1} = \ldots = B_{iG}. \tag{2}$$

Statistically, one can form two nested models—a full model in which each group has its own item parameter of that studied item, and a reduced model in which all groups are restricted to the same item parameter—and then apply the likelihood ratio test to test their difference. If the two models are statistically significantly different, then the studied item is said to have DIF (Thissen, Steinberg, and Wainer, 1988; Wang and Chang, 1998; Wang and Yeh, 2003).

In a typical DIF analysis, there is a reference group (usually the majority) and a focal group (usually the minority). In practice, more than two groups are sometimes of interest for DIF, such as multiple ethnic groups. If one employs standard two-group methods, one can compare item parameter estimates over groups pairwise, two groups at a time, or set one group as the reference and compare every other group with this reference for DIF. This procedure, analogous to multiple two-group *t*-tests, is not only cumbersome, but also statistically inefficient. As analysis of variance (ANOVA) is more powerful than multiple two-group *t*-tests, so too is the simultaneous testing of Equation 2 with the likelihood ratio test more powerful than traditional pairwise procedures.

Figure 1 presents the item response functions in which the item difficulties for the reference and focal groups are 0 and 0.5 logits, respectively. The DIF amount for this item, defined as the difference in the item parameters between groups, is 0.5 logits (a positive value indicates the item favors the reference group). The larger the DIF amount, the greater the difference between the two functions. Figure 1 also suggests that it is justifiable to compare the difference in the item parameters between groups for evidence of DIF only when the item meets the model's expectation reasonably well for each group of test-takers. Otherwise, the item parameters for individual groups would be meaningless.

The parameter B_{ig} in Equation 1 can be re-parameterized as a mean plus a deviation from the mean:

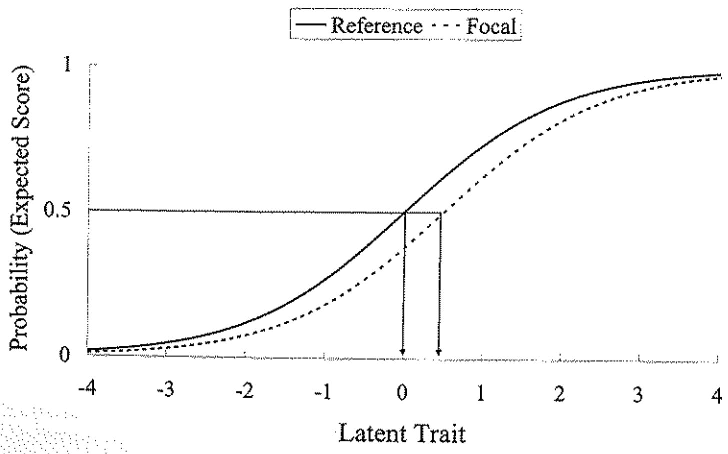

Figure 1. Item response functions in which the reference and focal groups have a difficulty of 0 and 0.5 logits, respectively.

$$B_{ig} = B_i + b_{ig}, \tag{3}$$

$$\sum\nolimits_g b_{ig} = 0, \tag{4}$$

where B$_i$ denotes the grand difficulty of item i, and b_{ig} denotes the effect of group g on item i. Equation 4 indicates that the sum of the deviations from the mean is equal to zero. Testing the null hypothesis of Equation 2 is equivalent to testing the following null hypothesis:

$$b_{i1} = \ldots = b_{iG} = 0. \tag{5}$$

If there are only two groups, then $2b_{ig}$ is the difference in item parameters between groups, which describes the DIF amount of that item. Equation 3 is analogous to the structural model in one-way ANOVA. The major difference between the two is that the former does not incorporate an error term whereas the latter does. In spite of this difference, the two may be considered to be conceptually equivalent.

Equation 3 involves only a single grouping variable. In practice, there may be more than one grouping variable of interest, such as ethnicity, gender, location, country, culture, etc. Although it is possible to combine all grouping variables into one single variable and conduct DIF analysis as described above, important information concerning the contribution of each grouping variable to DIF then becomes invisible. In order to retrieve this information, another kind of modeling is required. Just as one-way ANOVA can be extended to factorial ANOVA when there are multiple independent variables such that main effects and interaction effects on the dependent variable can be partitioned, so too can Equation 3 be expanded.

Suppose there are two grouping variables X and Y indexed as g and h, respectively. Let B_{igh} denote item i's difficulty for a group whose membership is gh. If one adopts the structural formulation in linear models, one has

$$B_{igh} = B_i + x_{ig} + y_{ih} + (xy)_{igh}, \tag{6}$$

$$\begin{aligned}\sum\nolimits_g x_{ig} &= \sum\nolimits_h y_{ih} \\ &= \sum\nolimits_g (xy)_{igh} \\ &= \sum\nolimits_h (xy)_{igh} = 0,\end{aligned} \tag{7}$$

where B_i denotes the grand difficulty of item i, x_{ig} denotes the "main" effect of X_g on the difficulty of item i, y_{ih} denotes the "main" effect of Y_h on the difficulty of item i, and $(xy)_{igh}$ denotes the "interaction" effect of X_g by Y_h on the difficulty of item i. Equation 7 indicates that the sum of the deviations from the mean is equal to zero. The likelihood ratio test can be employed to test the statistical significance of the main and interaction effects. Equation 6 can be directly generalized with three or more grouping variables. This linear modeling of DIF allows one to obverse closely the sources of DIF and determine whether the DIF comes from the main effect of grouping variable X, or Y, or their interaction.

Polytomous Items

The linear modeling of DIF can also be applied to polytomous items. Assume that the partial credit model (PCM; Masters, 1982) holds for each group of test-takers, such that

$$\ln\left(p_{nij}/p_{ni(j-1)}\right)_g = \theta_n - \left(B_{ig} + C_{ijg}\right), \tag{8}$$

where p_{nij} and $p_{ni(j-1)}$ are the probabilities of scoring j and $j-1$ on item i for person n with a latent trait level of θ_n; B_{ig} is the overall item difficulty of item i for group g; and C_{ijg} is the jth threshold difficulty of item i for group g. As in Equation 3, the item parameters in Equation 8 can be re-parameterized as a mean plus a deviation from the mean:

$$B_{ig} = B_i + b_{ig}, \tag{9}$$

$$C_{ijg} = C_{ij} + c_{ijg}, \tag{10}$$

$$\sum\nolimits_g b_{ig} = \sum\nolimits_g c_{ijg} = 0, \tag{11}$$

where B_i is the grand overall difficulty of item i, b_{ig} is the effect of group g on the overall difficulty of item i, C_{ij} is the grand threshold difficulty of step j in item i, and c_{ijg} is the effect of group g on the jth threshold difficulty of item i. Equation 11 indicates that the sum of the deviations from the mean is equal to zero. Note that if the subscript i in C_{ijg} of Equation 8 and the subsequent equations is dropped, then the PCM becomes the rating scale model (RSM; Andrich, 1978). Equation 8 and the subsequent equations can be easily generalized to incorporate multiple grouping variables, such as Equation 6 for dichotomous items. The likelihood ratio test can be employed to test the statistical significance of the main and interaction effects.

Facets

The above equations involve only two facets, person and item. In some testing situations, additional facets may be involved. For example, when constructed-response items are judged by raters, in addition to the two facets of person latent trait and item difficulty, rater severity might also play a role in determining the item response (score) for a specific person on a specific item. In such a case, rater is the third facet. The facets model (Linacre, 1989) is especially suitable for this type of data (Lunz, Wright, and Linacre, 1990; Myford and Wolfe, 2003, 2004).

Assume that the three-facet model holds for every group of test-takers:

$$\ln\left(p_{nijk}/p_{ni(j-1)k}\right)_g = \theta_n - \left(B_{ig} + C_{ijg}\right) - D_k, \tag{12}$$

where p_{nijk} and $p_{ni(j-1)k}$ are the probabilities of scoring j and $j-1$ on item i for person n in group g with a latent trait level of θ_n, when judged by rater k, B_{ig} is the overall item difficulty of item i for group g, C_{ijg} is the jth threshold difficulty of item i for group g, and D_k is the severity of rater k. Equations 9 through 11 can be directly incorporated into Equation 12 for DIF detection. Equation 12 can also be extended to incorporate multiple grouping variables.

The rater severity D_k in Equation 12 is assumed to be independent of items and grouping variables. Sometimes, raters may exhibit different degrees of severity in judging different items or different groups of test-takers and there may be rater-item interaction, rater-group interaction (e.g., some raters may favor those test-takers with the same ethnicity as theirs), or even rater-item-group interaction. Equation 12 can be extended to take these interactions into account by replacing D_k in Equation 12 with D_{ik} (for item-rater interaction), D_{kg} (for rater-group interaction), or D_{ikg} (for item-rater-group interaction). The likelihood ratio test can be applied to test whether these interactions are statistically significant.

Testlet-based Items

Testlet designs have been widely used in educational and psychological tests. A testlet is a bundle of items that share a common stimulus or other common feature, e.g., a reading comprehension passage or a figure. Another name for a testlet is an item bundle. Standard Rasch models have been extended to testlet-based items by adding an additional variable to account for possible interaction between items and persons within a testlet, one variable for each testlet (Wang and Wilson, 2005b). Specifically, the Rasch model for dichotomous items may be extended as

$$\ln\left(p_{ni1}/p_{ni0}\right) = \theta_n - B_i + \gamma_{nd(i)}, \tag{13}$$

where p_{ni1} and p_{ni0} are the probabilities of scoring 1 and 0 in item i for person n, respectively; and $\gamma_{nd(i)}$ represents the interaction effect between item i and person n within testlet d. For polytomous testlet-based items, one may add an additional variable to the PCM (or RSM) to account for possible interaction between items and persons within a testlet:

$$\ln\left(p_{nij}/p_{ni(j-1)}\right) = \theta_n - \left(B_i + C_{ij}\right) + \gamma_{nd(i)}, \tag{14}$$

where B_i is the overall item difficulty of item i; and C_{ij} is the jth threshold difficulty of item i. For model identification and ease of interpretation, we assume:

$$\gamma_{nd(i)} \sim N\left(0, \sigma^2_{\gamma_{d(i)}}\right), \tag{15}$$

and the latent trait θ and the random testlet variables γs are all independent. The variance indicates the amount of interaction effect between items and persons within a testlet. If $\sigma^2_{\gamma_{d(i)}}$ is equal to zero, Equations 13 and 14 then become the standard Rasch model and the PCM, respectively.

In order to detect DIF in testlet-based items, assume the testlet model holds for each group of test-takers:

$$\ln\left(p_{ni1}/p_{ni0}\right)_g = \theta_n - B_{ig} + \gamma_{nd(i)}, \tag{16}$$

$$\ln(p_{nij}/p_{ni(j-1)})_g = \theta_n - (B_{ig} + C_{ijg}) + \gamma_{nd(i)}, \tag{17}$$

where B_{ig} is the difficulty of item i for group g; C_{ijg} is the j^{th} threshold difficulty of item i for group g; and the others are defined as above. As usual, the linear modeling of DIF can be applied to B_{ig} and C_{ijg}, and they can be extended to incorporate multiple grouping variables (Wang and Wilson, 2005a).

The linear modeling of DIF is conceptually equivalent to the linear logistic test model (LLTM, Fischer, 1973) and the facets model (Linacre, 1989), in that the item parameter is modeled by a linear combination of a set of variables. In this study, the variables are grouping variables; in the LLTM, they are item features; and in the facets model, they are facets. For a discussion of the relationship between these models, the reader should see De Boeck and Wilson (2004).

It should be noted that the establishment of a complicated model (e.g., a model with DIF parameters, item-rater-group interactions, or testlet effects) to account for data should be taken as a means, not an end. In general, the more complicated a model that is needed to account for test data, the poorer the quality of measurement, the more difficult the interpretation of the test scores, and the less generalizable the findings will be. In effect, complicated models should be employed to diagnose sources of noise in test data (e.g., DIF) and to identify possible test revisions for better quality measurement, rather than to simply describe the data.

Three Methods of Establishing a Common Metric over Groups

By definition, an item has DIF when test-takers having identical latent trait levels but belonging to different groups have different probabilities of endorsing an item. According to this definition, person measures of the latent trait should be known a priori so that test-takers from different groups can be matched according to their measures, and then their responses to the studied item can be compared for evidence of DIF. In order to obtain person measures, a perfect test (one that does not contain any DIF items) that measures the same latent trait as the studied item must be administered to the test-takers. Note that this "external" test should not contain any DIF items; otherwise, it measures qualitatively different latent traits for different groups of test-takers and thus the establishment of a common metric is not possible. By "perfect," we do not mean an absolute sense; rather, we mean that the DIF in the external test should be negligible in practice. Unfortunately, such an external test is unlikely to exist in practice.

In most practical cases of DIF studies, no external perfect tests are available to establish a common metric. Instead, a common metric must be established through the studied test itself, which is called the "internal" matching variable (Welch and Miller, 1995). If the studied test, serving as a matching internal variable, contains DIF items, then DIF analysis is based on a biased matching variable. If the internal matching variable is free from DIF (meaning that the studied test does not contain DIF items), then DIF analysis is no longer necessary, which is a problem of circularity.

There are three major methods of establishing a common metric in DIF detection through the use of an internal matching variable, namely, the equal-mean-difficulty (EMD), the all-other-item (AOI), and the constant-item (CI) methods (Wang and Yeh, 2003; Wang, 2004). Different methods often lead to different results in DIF detection. Practitioners often rely on computer programs to detect DIF without knowing what method is actually employed in the programs to establish a common metric. The EMD method and the AOI method are more popular than the CI method, especially in the context of Rasch measurement or IRT. The EMD method, for example, is implemented in the computer programs ConQuest and Bilog-MG (Zimowski, Muraki, Mislevy, and Bock, 1996) and the AOI method is implemented in Winsteps, such that users can easily perform

DIF detection with simple commands. In addition to these computer programs, the Mantel-Haenszel (Holland and Thayer, 1988) and logistic regression (Swaminathan and Rogers, 1990) DIF detection techniques can be viewed as examples of the AOI method, in which raw test score is used to establish a common metric over groups. Waller (1998) developed the computer program EZDIF to facilitate the use of the Mantel-Haenszel and logistic regression methods. The popularity of the EMD and AOI methods does not necessarily justify their use.

The EMD Method

Assume a typical testing situation in which a test is administered to a reference group and a focal group and all items need to be examined for DIF. In the EMD method, as its name implies, the mean item difficulty of the test is constrained to be equal across groups. In introducing this constraint, the user assumes that a correct common metric has been established and the difference in the item parameter estimates between groups can be directly compared to detect DIF. The EMD method is simple and easy to implement. A common way of employing the EMD method is to conduct separate Rasch or IRT calibrations, one for each group, and then directly compare the differences in the item parameter estimates between groups. By default, the mean difficulty in a Rasch or IRT calibration is often set at zero for model identification and the EMD method is automatically actualized in separate calibrations. The use of separate calibrations in DIF detection can be dated back to Wright, Mead, and Draba (1976) and Wright and Stone (1979). In order to reduce the burden of running separate calibrations, ConQuest and Bilog-MG implement the EMD method in such a way that users are able to perform DIF analysis with one single command file.

By definition, the assumption of equal mean difficulty between groups holds only when either: (a) the test does not contain any DIF items, or (b) the test contains multiple DIF items in which some favor one group and the other DIF items favor the other group by exactly the same amount such that the mean difficulties for the two groups are identical. However, these two conditions are highly unlikely to occur in reality. If there is only one single DIF item in the test, the mean difficulties of the two groups will never be equal. Moreover, the more DIF items favoring one group (often the reference group), the greater the severity of the false assumption, and thus the worse the EMD will perform. A direct consequence of employing the EMD method to detect DIF in any imperfect test (real tests are always imperfect) is that approximately one half of the items will be detected as favoring the reference group while the others will be detected as favoring the focal group. This is inevitable because the DIF amount in a test is forced to be balanced between groups, once the constraint of equal mean difficulty between groups is imposed (see also Luppescu, 1993).

The AOI Method

In the AOI method, when examining an item for DIF, all other items in the test serve as an anchor (i.e., they are all assumed to be DIF-free) to establish a common metric. Unlike the EMD method where there is only one single common metric for the whole test, there are as many common metrics as the number of studied items (often all items in the test have to be examined for DIF) when the AOI method is adopted. By definition, the metrics cannot be simultaneously correct, unless they are actually the same. Only when the test does not contain any DIF items, will all metrics be identical and pure (not containing any DIF items) so as to yield appropriate DIF detection. Under all other conditions, the AOI method cannot yield correct detection. Within the context of Rasch measurement or IRT, the AOI method can be cumbersome, because each studied item needs a separate calibration. In order to facilitate the use of the AOI method and ease the burden of annoying programming, Winsteps provides user-friendly commands for such kinds of DIF analysis. The AOI method has also been incorporated into IRT-based DIF detection techniques (Bolt, 2002; Cohen, Kim, and Wollack, 1996; Kim and Cohen, 1998; Wang and Yeh, 2003).

The "implicit" assumption of the AOI method—all but the studied item are DIF-free—holds only when: (a) the test does not contain any DIF items, or (b) the studied item is the only DIF item in the test. As an example, imagine a test in which item 1 is the only item that has DIF. When item 1 is the studied item, the assumption of the AOI method holds, such that the resulting DIF detection on item 1 will be correct. However, when another item is the studied item, the assumption that all other items are DIF-free no longer holds because item 1 has DIF. As the number of DIF items in the test increases, so does the severity of the false assumption, and thus the worse the AOI method will perform.

Real tests may contain a high percentage of DIF items. Using the AOI method, the matching variable (the one containing all other items) may be so contaminated by the inclusion of many DIF items that its DIF detection becomes problematic. In order to purify the DIF contamination of the matching variable due to the inclusion of DIF items, scale purification procedures have been strongly advocated and widely employed (e.g., Lord, 1980; Holland and Thayer, 1988). Scale purification involves the following major steps:

1. Use the AOI method to detect DIF.
2. Remove those items found to exhibit DIF in the previous step from the matching variable and examine all items in the test for DIF again, using the new matching variable.
3. Repeat Step 2 until the same set of items are found to have DIF or a maximum number of iterations (say, 20) is reached.

Unfortunately, scale purification procedures can only partially eliminate the DIF contamination in the matching variable when a test contains a high percentage of DIF items. It has been found that when tests consist of more than 20% or 30% DIF items, the AOI method with scale purification procedures, although superior to the AOI method without scale purification procedures, begins to yield an inflated Type I error rate and a decreased power of DIF detection (Candell and Drasgow, 1988; Clauser, Mazor, and Hambleton, 1993; Hidalgo-Montesinos and Gómez-Benito, 2003; Navas-Ara and Gómez-Benito, 2002; Wang and Su, 2004a, 2004b).

The CI Method

The CI method is not as commonly used as the other two methods. In the CI method, a set of items is specified to serve as an anchor set to establish a common metric over groups, and all other items in the test are then examined for DIF. This is called the constant-item method because the same set of items serve as an anchor, no matter what item is studied. As already discussed, the matching variable has to be pure, otherwise, the resulting DIF detection will be misleading. In other words, only DIF-free items should serve as an anchor. In general, the more items chosen to serve as an anchor (assuming they are all DIF-free), the higher the power of DIF detection will be, and a 4-item anchor is generally enough to yield a high power (Thissen et al., 1988; Wang, 2004; Wang and Yeh, 2003). The CI method is also implemented in non-IRT-based approaches, such as the SIBTEST method (Chang, Mazzeo, and Roussos, 1996; Shealy and Stout, 1993). Users of SIBTEST have to specify a set of DIF-free items to serve as an anchor. If the specification is impossible, then SIBTEST switches to the AOI method for DIF detection.

Comparison of the Three Methods: A Simulation Study

In order to demonstrate the advantages and limitations of these three methods in establishing a common metric over groups for DIF detection, a brief simulation study is conducted. In the following section the design, analysis, and hypothesis of the simulation are described. The results are then summarized.

Design

Item responses were simulated according to the Rasch model, in which five items were administered to a reference and a focal group, each with a sample size of 5,000. The five item difficulties were set at −1, −1, 0, 1, and 1 logits for the reference group, and 0, 0, 0, 1, and 1 logits for the focal group. That is, both items 1 and 2 were set as favoring the reference group by 1 logit (i.e., DIF amount = 1; a positive value indicates the reference group is favored), and the other three items were set as DIF-free (DIF amount = 0). With this setting, the mean difficulties for the five items were 0 and 0.4 logits for the reference and focal groups, respectively. The test-takers of the reference and focal groups were sampled from $N(0, 1)$ and $N(-1, 1)$, suggesting that the mean ability difference between the two groups is 1 logit (i.e., impact = –1; a negative value indicates the focal group has a lower mean). The sample sizes of the two groups were set so large that the effect of sampling fluctuation on parameter estimation could be reduced and one single replication would be enough to draw reliable conclusion.

Analysis

After the item responses were simulated, the EMD, AOI, and CI methods were applied by using ConQuest. In the EMD method, the simulated datasets for the reference and focal groups were calibrated separately. Since by default ConQuest sets the mean difficulty at zero for identification, the EMD method was automatically employed in separate calibrations. In the AOI method, the two datasets were combined into one large dataset. Five calibrations were then made on this combined dataset, one calibration for each studied item. In the CI method, five calibrations were made on the combined dataset, one calibration for each item serving as an anchor.

Hypothesis

The mean difficulties of the focal and the reference groups are simulated as 0.4 and 0 logits. Employing the EMD method will force the 0.4-logit difference in the mean difficulty between groups to be zero. In order to maintain a linear relationship between items and persons within groups, the 1-logit DIF amount for items 1 and 2 will be estimated as 0.6 logits, the 0-logit DIF amount for items 3, 4 and 5 as –0.4 logits, and the –1-logit impact as –1.4 logits.

In the AOI method, when detecting item 1 for DIF, items 2, 3, 4 and 5 serve as an anchor. The mean difficulties of the four anchored items for the reference and focal groups are 0.25 logits and 0.50 logits, respectively. When the four items serve as an anchor, the 0.25-logit difference in the mean difficulty between groups will be forced to zero. In order to maintain a linear relationship between items and persons within groups, the 1-logit DIF amount for item 1 will be estimated as 0.75 logits and the –1-logit impact as –1.25 logits. Likewise, when item 2 is examined for DIF, the 1-logit DIF amount for item 2 will be estimated as 0.75 logits, and the –1-logit impact as –1.25 logit. When examining item 3 for DIF, items 1, 2, 4 and 5 serve as an anchor. The mean difficulties of the four anchored items for the reference and focal groups are 0.0 logits and 0.50 logits. The 0.50-logit difference in the mean difficulty between groups will be forced to zero when the four items serve as an anchor. As a result, the 0-logit DIF amount for item 3 will be estimated as –0.50 logits, and the –1-logit impact as –1.50 logits. When detecting item 4 (or item 5), the mean difficulty of the four anchored items for the focal group is higher than that for the reference group by 0.5 logits. When the 0.50-logit difference is forced to zero, the 0-logit DIF amount for item 4 (or item 5) will be estimated as –0.50 logits, and the –1-logit impact as –1.50 logits.

In the CI method, DIF detection will be correct if the anchored item is indeed DIF-free. Specifically, when item 3, 4 or 5 serves as an anchor, both the DIF detection and the person parameter estimation will be appropriate. On the contrary, when item 1 (or item 2) serves as an anchor, its 1-logit DIF amount will be forced to zero. In order to maintain a linear relationship between items and persons within groups, the 1-logit DIF amount of item 2 (or item 1) will be estimated as 0 logits, the 0-logit DIF amount for items 3, 4 and 5 as –1 logit, and the –1-logit impact as –2 logits.

Results

The parameter recovery of the five items and the person measures for the three methods is summarized in Table 2. When the EMD method was employed, items 1 and 2 were estimated as having a DIF amount of 0.60 and 0.62 logits, items 3, 4, and 5 as having a DIF amount of –0.38, –.0.45, and –0.39 logits, and the impact as –1.40 logits. As hypothesized, both the DIF amount and the impact were underestimated by approximately 0.4 logits. The DIF amount for the five items added up to zero. This result (some items favor the focal group and others favor the reference group such that overall neither group is favored) is inevitable when the EMD method is applied to any imperfect test. Using the EMD method, practitioners may overlook the consequences of DIF on test fairness and argue that no actions need to be taken, because overall neither group is favored nor disfavored.

Table 2

Parameter recovery for the three methods when items 1 and 2 in the five-item test have a DIF amount of 1 logit

	Item 1	Item 2	Item 3	Item 4	Item 5	Mean Ability	Variance
True Parameter (R)	–1	–1	0	1	1	0	1
True Parameter (F)	0	0	0	1	1	–1	1
True DIF Amount / Impact	1	1	0	0	0	–1	
EMD Method							
Estimated Parameter (R)	–1.01	–1.02	0.01	1.02	1.00	0.00	0.88
Estimated Parameter (F)	–0.41	–0.40	–0.37	0.57	0.61	–1.40	1.12
Estimated DIF Amount / Impact	0.60	0.62	–0.38	–0.45	–0.39	–1.40	
AOI Method							
Estimated DIF Amount / Impact	0.72	+	+	+	+	–1.26	0.95
Estimated DIF Amount / Impact	+	0.74	+	+	+	–1.26	0.95
Estimated DIF Amount / Impact	+	+	–0.55	+	+	–1.55	0.94
Estimated DIF Amount / Impact	+	+	+	–0.61	+	–1.52	0.94
Estimated DIF Amount / Impact	+	+	+	+	–0.55	–1.51	0.94
CI Method							
Estimated DIF Amount / Impact	+	0.02	–0.99	–1.09	–1.03	–2.00	0.99
Estimated DIF Amount / Impact	–0.02	+	–1.01	–1.11	–1.05	–2.01	0.99
Estimated DIF Amount / Impact	0.99	1.01	+	–0.10	–0.04	–1.00	0.99
Estimated DIF Amount / Impact	1.09	1.11	0.10	+	0.06	–0.91	0.99
Estimated DIF Amount / Impact	1.03	1.05	0.04	–0.06	+	–0.96	0.99

Note: + anchored; R = reference group; F = focal group.

When the AOI method was employed, as hypothesized, the 1-logit DIF amount for items 1 and 2 were estimated as 0.72 and 0.74 logits, the 0-logit DIF amount for items 3, 4, and 5 as –0.55, –0.61, and –0.55 logits, and the –1-logit impact as –1.26, –1.26, –1.55, –1.52, and –1.51 in the five calibrations. When the CI method was employed, the DIF detection and parameter recovery were of the appropriate values as long as the anchored item was correctly chosen (i.e., item 3, 4, or 5). In the case when item 3 served as an anchor, the 1-logit DIF amount of items 1 and 2 was estimated as 0.99 and 1.01 logits, the 0-logit DIF amount of items 4 and 5 as –0.10 and –0.04 logits, and the –1-logit impact as –1.00 logits. In contrast, when item 1 or item 2 was mistakenly selected as an anchor, both the estimates for the DIF amount and impact were misleading. In the case when item 1 served as an anchor, the 1-logit DIF amount of item 2 was estimated as 0.02 logits, the 0-logit DIF amount of items 3, 4 and 5 as –0.99 and –1.09 and –1.03 logits, and the –1-logit impact as –2.00 logits. In summary, all the hypotheses about the three methods were confirmed.

Locating DIF-free Items to Serve as an Anchor

As demonstrated in the previous simulation, the superiority of the CI method over the other two methods is possible when a DIF-free item is correctly selected as an anchor. An immediate question that follows is: how may one locate a set of DIF-free items to function as an anchor in order for the CI method to proceed appropriately? Non-statistical procedures (e.g., content expert review or previous experiences) might provide some suggestions about which items are DIF-free. However, relying on solely non-statistical procedures to locate DIF-free items might be infeasible and potentially controversial. Statistical procedures together with non-statistical procedures would provide more comprehensive information about the location of DIF-free items than either procedure alone.

In this study, we focused on statistical procedures. The logic is to turn DIF analysis from finding items that have DIF to finding items that do not have DIF. Those methods that can accurately distinguish DIF items from DIF-free items (i.e., high power of DIF detection and well-controlled Type I error rate) are good candidates for the location of DIF-free items. As indicated in the literature (Wang, 2004; Wang and Yeh, 2003) and the demonstration above, the CI method is very powerful in DIF detection. It seems reasonable to adopt this method in order to locate a set of items that are the most likely to be DIF-free to serve as an anchor. Such a statistical procedure, called the iterative constant-item (denoted as ICI) procedure, was recently developed (Wang, Shih, and Su, 2007). It consists of the following steps:

1. Set item 1 as an anchor and assess all other items for DIF with the CI method; obtain an estimate of DIF amount for each studied item.
2. Set the next item as an anchor and assess all other items for DIF as in Step 1.
3. Repeat Step 2 until the last item is set as an anchor.
4. Compute the sum of the DIF amount estimates (in absolute value) over iterations for each item, rank the sum, and select the desired number of items (e.g., 4) with the smallest DIF amount estimates to serve as an anchor.

Wang et al. (2007) conducted a series of simulations to ascertain the accuracy of the ICI procedure in locating DIF-free items and found that it almost always yields a perfect rate of accuracy when locating one through four DIF-free items; as DIF amount and sample size are increased, so is the rate of accuracy; even when tests contain as many as 40% DIF items, it yields a very satisfactory rate of accuracy, which is also much higher than the accuracy rate of random selection.

A simulated data set can serve as an example to demonstrate the ICI procedures. A 40-item test was generated according to the Rasch model, in which items 1 through 16 were simulated to favor the reference group by 0.8 logits and the other 24 items were DIF-free. The sample sizes of the reference and the focal groups were both 500. The test-takers of the reference and focal groups were generated from N(0, 1) and N(–1, 1), respectively. The ICI procedure, implemented with ConQuest, was applied to this dataset to locate DIF-free items. The results show that the first 18 located items were all DIF-free items. Such a rate of accuracy was extremely high. If items were randomly selected from a 40-item test with 24 DIF-free items, then the chances of correctly locating one, two, three, four, or five DIF-free items would be .60, .35, .20, .12, and .06, respectively. The chance of correctly locating 18 DIF-free items would be as small as 1.19×10^{-6}. Hence, the ICI procedure yielded a high rate of accuracy in locating DIF-free items.

Computer Programs

All the above linear modeling of DIF (Equations 1 through 17), the three methods of establishing a common metric and the ICI procedure can be implemented with ConQuest. By default, ConQuest adopts the EMD method to establish a common metric and provides several annotated examples for standard DIF analysis. Through manipulation of the design matrix of ConQuest, more complicated DIF modeling (e.g., testlet-based items), the CI method, and the ICI procedure can be carried out. Unfortunately, this implementation requires a certain level of sophistication of ConQuest syntaxes and may not be intelligible to ordinary users. As the importance of the linear modeling of DIF, the CI method, and the ICI procedure becomes more widely recognized, customized computer programs might be available in the near future.

In addition to ConQuest, several other software packages are also available for the analyses as shown in this chapter, including SAS NLMIXED (SAS Institute., 1999; Sheu, Chen, Su, and Wang, 2005; Wolfinger and SAS Institute., n.d.) and STATA GLLAMM (Rabe-Hesketh, Skrondal, and Pickles, 2003, 2004; Skrondal and Rabe-Hesketh, 2004). Based on the author's experience, ConQuest is more efficient and user-friendly than the other two programs. NLMIXED and GLLAMM have an advantage over ConQuest in fitting a wider range of models, in addition to those shown in this chapter. However, since these two programs are not developed specifically for DIF detection or Rasch analysis, users need an even higher level of sophistication of their syntaxes to carry out such analyses.

Practical Significance of DIF at Item and Test Levels

Item Level

No matter how minute an item's amount of DIF, it will be detected as statistically significant as long as the sample size is sufficiently large. Thus, in addition to testing the statistical significance of DIF, it is also necessary to ascertain its practical significance. Within the family of Rasch models, the DIF amount (defined as the difference in item parameters between groups) is actually an effect size measure of DIF. A DIF amount of d logits represents an odds-ratio of 2.72^d. A DIF amount of 0.5 logits, representing an odds-ratio of 1.65, is sometimes treated as a cut-off point for substantial DIF. For polytomous items, cut-off points are more difficult to determine. There are two kinds of parameters in a polytomous items—the overall difficulty and the threshold difficulty. Each kind of parameter can have different values for different groups of test-takers. It is not appropriate to apply the same cut-off point (e.g., 0.5 logits) to these two kinds of parameters. A DIF amount of 0.5 logits, for example, in the overall difficulty is much more influential on the item expected score than on the threshold parameter. To make things even more complicated, polytomous items with different categories may require different cut-off points. A DIF amount of 0.5 logits for a polytomous item with many categories (e.g., 7) may not have the same meaning as it has for a polytomous item with few categories (e.g., 3).

For polytomous items, an inspection of item expected score curves over groups is more helpful in determining the practical significance of DIF than merely an inspection of the magnitudes of DIF amount. For example, Figure 2a shows the item expected score curves on a 3-category item with a DIF amount of 0.5 logits in the overall difficulty, and Figure 2b shows the curves for an item with a DIF amount of 0.5 logits in the first threshold difficulty. Figures 3a and 3b show the expected score curves for a 7-category item with the same types of DIF described in Figures 2a and 2b, respectively. In terms of the difference in the item expected score curves between groups, DIF in the overall difficulty is more substantial than it is in the threshold difficulty. Besides, when DIF occurs in the overall difficulty, one group always has a higher expected score than the other group, throughout the latent trait level; whereas when DIF occurs in the threshold difficulty, one group has a higher expected score than the other group only within a certain range of latent trait level. A comparison of Figures 2 and 3 suggests that a DIF amount of 0.5 logits for a 3-category DIF item is slightly more substantial than it is for a 7-category item. Practitioners should inspect the pattern of the item expected score curves over groups, such as those shown in Figures 2 and 3, to determine whether the DIF is substantial or not.

Test Level

Test scores or person measures are often used to make important decisions about individuals (e.g., admission to colleges, personnel selection and placement, etc.). Hence, in addition to examining the practical significance of DIF at the item level, it is also necessary to ascertain it at the test level. Two major methods are widely used. One is to compare the test expected score curves (also called test characteristic curves) for different groups of test-takers, which is referred to as the assessment of differential test functioning (Raju, van der Linden, and Fleer, 1995). If the test expected score curves for different groups are far apart, then the DIF items are practically significant. On the contrary, if the curves nearly overlap, then the DIF items are not practically significant at the test level, even though the test may contain a high percentage of DIF items that are practically significant at the item level. The other method is to compare person measures obtained from a model in which DIF items are excluded with those obtained from another model in which DIF items are not excluded and are treated as DIF-free. If the person measures obtained from these two models are very different, then the inclusion of DIF items substantially affects person measures.

In order to explain the procedures of ascertaining the practical significance of DIF at test level, the previous data set (the 40-item test with 16 DIF items) is taken as an example. In the previous simulation, the first four items located by the ICI procedure are items 33, 24, 20, and 21. All of them are generated as DIF-free. Only these four items are selected to serve as an anchor because a 4-item anchor is often powerful enough to detect DIF (Thissen et al., 1988; Wang, 2004; Wang and Yeh, 2003), and the more items that are selected to serve as an anchor, the more likely some DIF items will be mistakenly selected. These four items are then used to establish a common metric such that the CI method can proceed to assess DIF for the remaining 36 items. The true values and estimates of the DIF amount, and their standard errors for the 36 items, are listed in the left-hand side of Table 3. All estimates for the DIF amount are very close to their true values. Besides, those estimates for the first 16 items are statistically significant (power = 1.00) whereas those for the remaining 20 items were not (Type I error rate = .00). Hence, the CI method together with a 4-item anchor yields a very accurate DIF

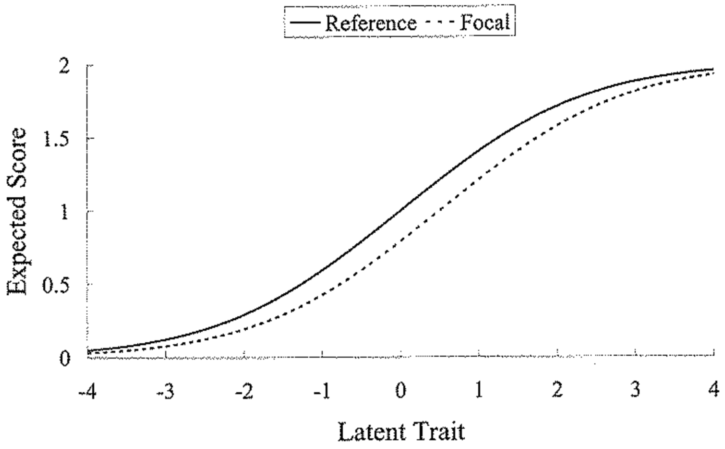

Figure 2a. Item expected score curves of a 3-category item with a DIF amount of 0.5 logits in the overall difficulty.

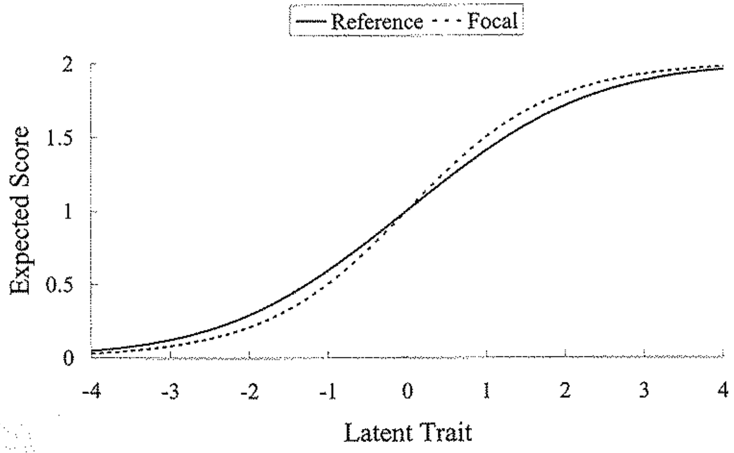

Figure 2b. Item expected score curves of a 3-category item with a DIF amount of 0.5 logits in the first threshold difficulty.

Table 3

True values and estimates of the DIF amount and their standard errors for the CI and EMD methods

		CI			EMD		
Item	True Value	Estimate	*SE*	*Z*	Estimate	*SE*	*Z*
1	0.80	0.95	0.17	5.51*	0.70	0.15	4.52*
2	0.80	0.70	0.16	4.25*	0.45	0.15	3.10*
3	0.80	0.83	0.18	4.59*	0.59	0.16	3.57*
4	0.80	0.73	0.23	3.17*	0.48	0.21	2.25*
5	0.80	0.84	0.23	3.71*	0.59	0.21	2.79*
6	0.80	1.12	0.26	4.36*	0.87	0.24	3.61*
7	0.80	0.61	0.17	3.64*	0.36	0.15	2.43*
8	0.80	0.47	0.17	2.86*	0.23	0.15	1.53
9	0.80	0.83	0.21	3.88*	0.58	0.20	2.93*
10	0.80	0.80	0.24	3.38*	0.55	0.22	2.48*
11	0.80	0.86	0.17	5.19*	0.61	0.15	4.14*
12	0.80	0.79	0.17	4.72*	0.55	0.15	3.64*
13	0.80	0.57	0.17	3.43*	0.32	0.15	2.18*
14	0.80	0.71	0.16	4.33*	0.46	0.15	3.18*
15	0.80	0.50	0.16	3.05*	0.25	0.14	1.72
16	0.80	0.50	0.17	3.04*	0.26	0.15	1.73
17	0.00	0.08	0.16	0.49	−0.17	0.14	−1.20
18	0.00	0.06	0.16	0.39	−0.19	0.14	−1.33
19	0.00	−0.07	0.16	−0.42	−0.32	0.14	−2.19*
20	0.00	0.00$^+$			−0.25	0.14	−1.76
21	0.00	0.00$^+$			−0.25	0.14	−1.77
22	0.00	−0.08	0.16	−0.49	−0.33	0.14	−2.33*
23	0.00	−0.25	0.19	−1.32	−0.50	0.17	−2.87*
24	0.00	0.00$^+$			−0.25	0.14	−1.75
25	0.00	0.14	0.16	0.86	−0.11	0.14	−0.77
26	0.00	−0.05	0.16	−0.30	−0.30	0.14	−2.11*
27	0.00	−0.34	0.20	−1.67	−0.58	0.18	−3.17*
28	0.00	0.14	0.19	0.76	−0.11	0.17	−0.62
29	0.00	−0.11	0.21	−0.53	−0.36	0.20	−1.85
30	0.00	−0.12	0.16	−0.74	−0.37	0.14	−2.61*
31	0.00	−0.09	0.19	−0.47	−0.34	0.17	−1.94
32	0.00	−0.31	0.16	−1.88	−0.56	0.15	−3.81*
33	0.00	0.00$^+$			−0.25	0.16	−1.59
34	0.00	−0.02	0.18	−0.11	−0.27	0.17	−1.61
35	0.00	−0.34	0.20	−1.71	−0.59	0.18	−3.25*
36	0.00	−0.18	0.16	−1.11	−0.43	0.14	−3.04*
37	0.00	−0.03	0.16	−0.16	−0.27	0.14	−1.96
38	0.00	−0.19	0.18	−1.05	−0.43	0.16	−2.71*
39	0.00	−0.10	0.16	−0.63	−0.35	0.14	−2.49*
40	0.00	−0.03	0.16	−0.16	−0.27	0.14	−1.96

* $p < .05$; $^+$ fixed at zero

assessment. In comparison, the estimates for the DIF amount obtained from the EMD method are reported in the right-hand side of Table 3. All the estimates for the DIF amount are far from their true values; 13 of the 16 DIF items are correctly detected as having DIF (power = .82), and 11 of the 24 DIF-free items are incorrectly detected as having DIF (Type I error rate = .46). Besides, the estimates for DIF amount of the 40 items sum to zero, indicating that overall, neither group of test-takers is favored nor disfavored. This result certainly contradicts the simulation design.

With the item parameter estimates obtained from the CI method with a 4-item anchor, the test expected score curves for the two groups are shown in Figure 4a. The reference group always has a higher test expected score than the focal group, which is expected because all the DIF items are generated to favor the reference group. Two test-takers from different groups having the same latent trait levels can have a difference in the test expected scores as large as roughly 2.5. In a scale of 0–40, such a difference is certainly substantial, especially if the test is of great importance (i.e., high-stakes).

In comparison, the item parameter estimates obtained from the EMD method are used to produce the test expected score curves for the two groups. These two curves, shown in Figure 4b, are visually indistinguishable. Facing such a figure, one may draw the conclusion that the DIF in the test is not substantial at all, even though one realizes from Table 3 that the EMD method has detected many DIF items. The phenomenon that the test expected score curves produced by the EMD method nearly overlap is often a methodological artifact rather than a reality.

To ascertain how the DIF items affected person measures, a Rasch model is fitted to the 40-item test in which all items are treated as DIF-free, and another Rasch model to the 24-item test in which the detected 16 DIF items (by the CI method) were excluded and the 24 items are treated as DIF-free. If these two models yield very similar person measures, then the inclusion of DIF items is not practically significant. Figure 5 presents the relationship between the person measures derived from these two models. A high degree of variation is found, suggesting the inclusion of DIF items can substantially alter person measures. In practice, test scores or person measures are often ranked in order to make comparative decisions (e.g., admission to college). If the rank orders of person measures obtained

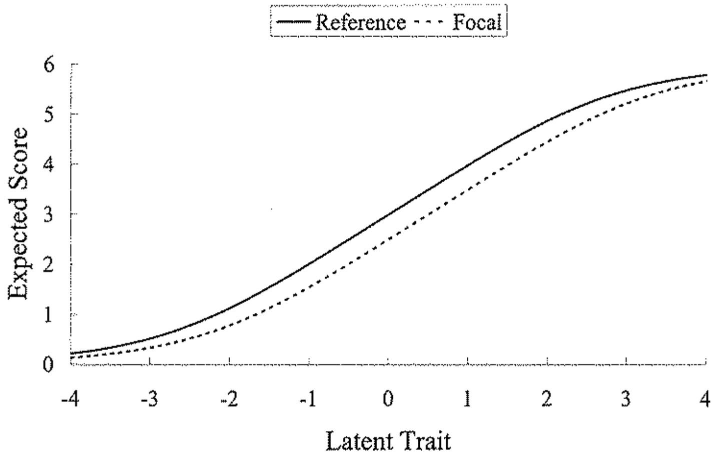

Figure 3a. Item expected score curves of a 7-category item with a DIF amount of 0.5 logits in the overall difficulty.

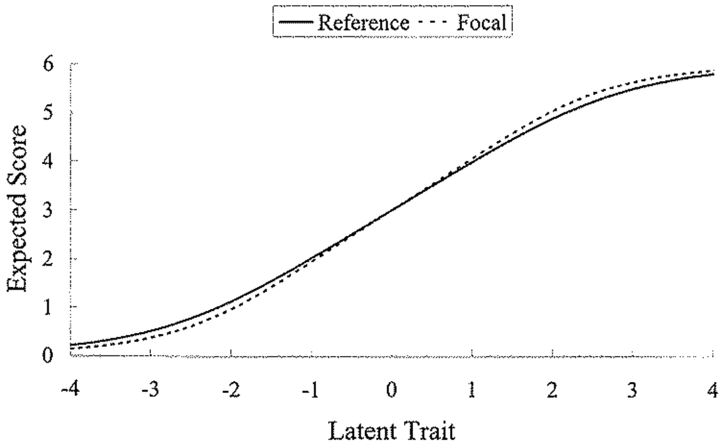

Figure 3b. Item expected score curves of a 7-category item with a DIF amount of 0.5 logits in the first threshold difficulty.

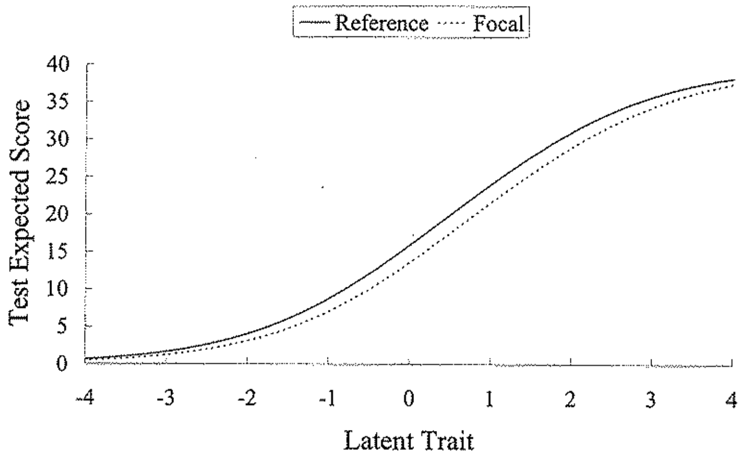

Figure 4a. Test expected score curves for two groups where the detected 16 DIF items are not excluded using the CI method with a 4-item anchor.

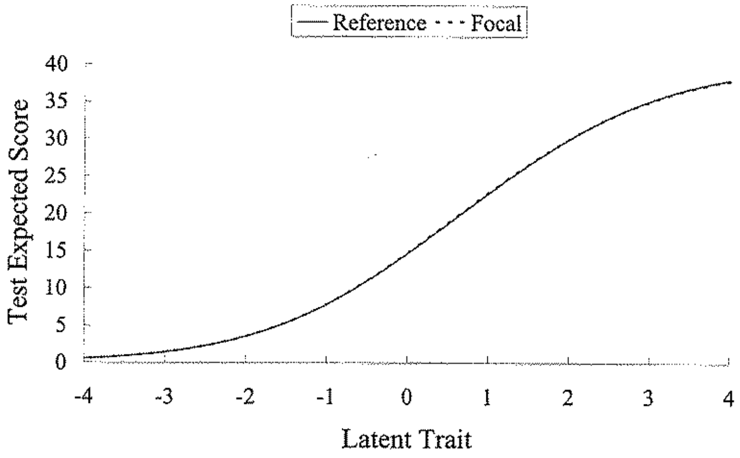

Figure 4b. Test expected score curves for two groups where the detected 16 DIF items are not excluded using the CI method with the EMD method.

from these two models change substantially, then the DIF items are practically significant. Figure 6 shows the changes in rank orders (in absolute value) of person measures, ranging from 0 to 365 with a median of 48. A total of 217 persons have a rank order change larger than 100. For a sample size of 1000 persons, such a rank order change is certainly substantial.

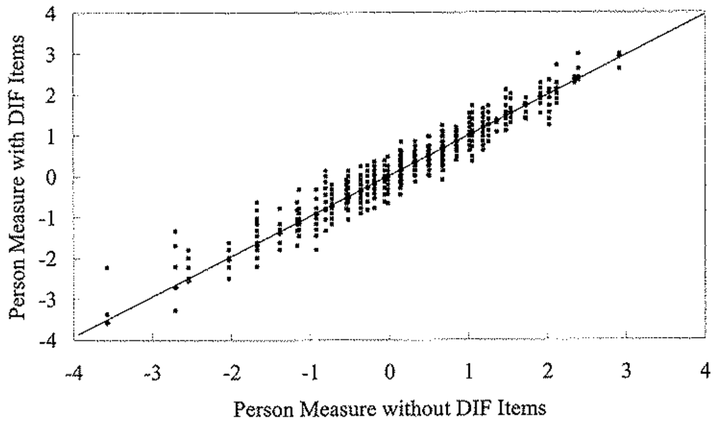

Figure 5. Relationship between person measures before and after the detected 16 DIF items are excluded.

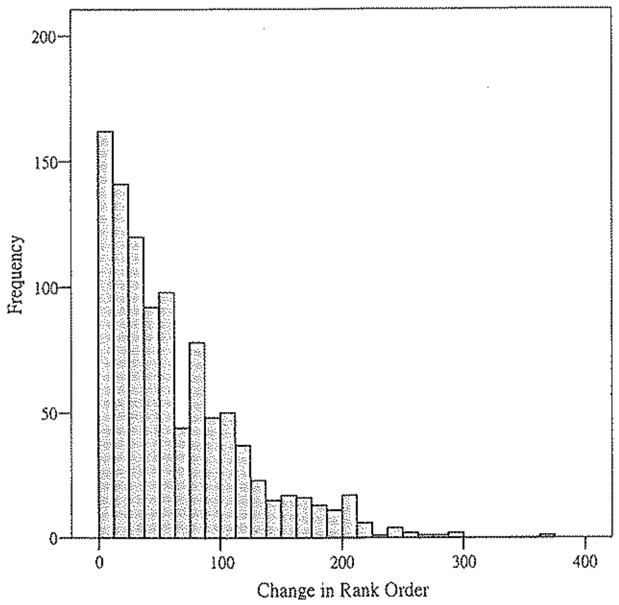

Figure 6. Changes in rank orders (in absolute value) of person measures before and after the 16 detected DIF items are excluded.

Table 4

Change in the classification of test-takers before and after the DIF items are excluded

		DIF Items Included				
	Grade	A	B	C	D	Total
DIF Items Excluded	A	129	31	0	0	160
	B	11	279	41	0	331
	C	0	48	280	40	368
	D	0	0	17	124	141
	Total	140	358	338	164	1000

Test scores or person measures are sometimes used to classify test-takers into a few categories (e.g., criterion-reference tests). For example, students are given grades (e.g., A, B, C, and D) based on their test performances, and individuals are classified as seriously disturbed, marginally disturbed, or normal based on their performances in a clinical psychological assessment. It is important to ascertain the way in which DIF items affect the classification of test-takers. If the classification alters substantially before and after the DIF items are excluded, then they are practically significant. To illustrate, suppose for theoretical or practical reasons, test-takers are classified into four categories: A test-taker receives an A if the measure is above 1 logit, a B if the measure is between 0 logits and 1 logit, a C if the measure is between –1 logit and 0 logits, and a D if the measure is below –1 logit. Table 4 shows the change in the classification before and after the 16 detected DIF items are excluded. Altogether, 812 out of 1000 test-takers receive the same grades and 188 (18.8%) test-takers receive different grades, yielding a kappa coefficient of .74. These represent the practical significance of these DIF items.

Summary

DIF detection is not only a popular research topic but also a routine component of item analysis. In the context of Rasch measurement, DIF detection appears extraordinarily simple and direct. All one needs to do is to compare the difference in the item parameters between groups. This seemingly simple task, however, can involve many complex issues. The overall item infit and outfit statistics are not powerful enough for DIF detection, even though DIF is a sort of item misfit. When a grouping variable has more than two categories or there is more than one grouping variable, traditional two-group DIF detection methods are cumbersome and inefficient. If one adopts the procedure of factorial ANOVA to linearly decompose DIF as main effects and interaction effects of grouping variables, one can better diagnose sources of DIF. This linear modeling makes DIF detection more general and gives it more structure. It can be applied to dichotomous items, polytomous items, facet structures, and testlet-based items.

The establishment of a common metric over groups is a prerequisite of DIF analysis. Only if all test-takers are placed on the same metric can the difference in the item parameters between groups be compared for DIF detection. Different methods of establishing a common metric over groups often lead to quite different results of DIF detection. Practitioners usually rely on computer programs to conduct DIF detection without knowing which method is actually incorporated in the program they employ. The EMD method yields appropriate DIF detection only when the test does not contain any DIF items, or it contains multiple DIF items in which some of them favor one group and others favor the other group at exactly the same amount such that overall the DIF amount is cancelled out between groups. The AOI method performs appropriately only when the test does not contain any DIF items or the studied item is the only item that has DIF. The CI method performs very well as long as one or more DIF-free items are chosen as an anchor.

The ICI procedure performs very well in locating a set of DIF-free items. The basic idea is to identify items that are the most likely to be DIF-free. The ICI procedure is in accordance with Rasch measurement and is flexible for complicated item formats (e.g., polytomous items, facet structures or testlet-based items). In practice, one may adopt the ICI procedure to locate a set of items and then request content experts to verify whether they might have substantial DIF. In short, DIF detection should consist of two steps: (a) adopt a procedure (e.g., the ICI procedure) to locate a set of DIF-free items to serve as an anchor (content expert approval is recommended), and then (b) apply another procedure (e.g., the CI method) to detect all other items in the test for evidence of DIF. This is referred to as the DIF-free-then-DIF strategy (Wang et al., 2007).

The linear modeling of DIF, the CI method, and the ICI procedure can be implemented with computer programs such as ConQuest, NLMIXED, and GLLAMM, although certain level of sophistication of their syntaxes is required. As these advanced issues in DIF detection become widely recognized, commercial computer programs might be available in the near future.

In addition to the detection of the statistical significance of DIF, it is also necessary to ascertain its practical significance, both at the item level and the test level. At the item level, the DIF amount (i.e., the difference in the item parameter estimates between groups) is an effect size measure of DIF. A one logit DIF amount constitutes an odds-ratio of 2.72. If an item has a DIF amount greater than a cut-off point (e.g., 0.5 logits), then it may be removed from the test. Practitioners should be very cautious when using cut-off points on polytomous items, because the overall difficulty and the threshold difficulty may require different cut-off points. It is often helpful to inspect item expected score curves over groups on the studied item to ascertain whether the DIF is practically significant. At the test level, an inspection of the test expected score curves over groups provides a direct assessment of the practical significance of a set of DIF items on test scores. It is also of great value to compare the change in person measures and classification of test-takers before and after DIF items are excluded.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.

Candell, G. L., and Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Chang, H. H., Mazzeo, J., and Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Clauser, B., Mazor, K., and Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279.

Cohen, A. S., Kim, S.-H., and Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15-26.

De Boeck, P., and Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Hidalgo-Montesinos, M. D., and Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinominal logistic regression. *European Journal of Psychological Assessment, 19*, 1-11.

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

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

Kim, S.-H., and Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2003). A user's guide to WINSTEPS: Rasch-model computer program [Computer program and manual]. Chicago: Winsteps.com.

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

Lunz, M. E., Wright, B. D., and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Luppescu, S. (1993). DIF detection examined. *Rasch Measurement Transactions, 7*(2), 285-286.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Myford, C. M., and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., and Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Navas-Ara, M. J., and Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2003). Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal, 3*, 385-410.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*, 167-190.

Raju, N. S., van der Linden, W. J., and Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

SAS Institute. (1999). The NLMIXED procedure [Computer program]. Cary, NC: Author.

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

Sheu, C.-F., Chen, C.-T., Su, Y.-H., and Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods, 37*, 202-218.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC Press.

Smith, R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54*, 42-55.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56*, 403-418.

Smith, R. M., and Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement, 4*, 153-163.

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum .

Waller, N. G. (1998). EZDIF: A program for the analysis of uniform and nonuniform differential item functioning. *Applied Psychological Measurement, 22*, 391.

Wang, W.-C. (2000a). Factorial modeling of differential distractor functioning in multiple-choice items. *Journal of Applied Measurement, 1*, 238-256.

Wang, W.-C. (2000b). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement. 1*, 63-82.

Wang, W.-C. (2000c). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research*, *5*(1), 56-76.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wang, W.-C., and Chang, C. (1998). Rasch likelihood ratio test of differential item functioning. *Chinese Journal of Psychology, 40*, 15-32.

Wang, W.-C., Shih, C.-L., and Su, Y.-H. (2007). *Establishing a common metric over groups for detection of differential item functioning.* Paper submitted for publication.

Wang, W.-C., and Su, Y.-H. (2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*, 113-144.

Wang, W.-C., and Su, Y.-H. (2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

Wang, W.-C., and Wilson, M. R. (2005a). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*, 549-576.

Wang, W.-C., and Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Wang, W.-C., and Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Welch, C., and Miller, T. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32*, 163-178.

Wolfinger, R. D., and SAS Institute. (n.d.). *Fitting nonlinear mixed models with the new NLMIXED procedure.* Retrieved June 16, 2007, from http://support.sas.com/rnd/app/papers/nlmixedsugi.pdf

Wright, B. D., and Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., Mead, R. J., and Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum 22). Chicago: MESA Statistical Laboratory, Department of Education, University of Chicago.

Wu, M., Adams, R. J., and Wilson, M. R. (1998). ACER ConQuest: Generalized item response modeling software [Computer program and manual]. Camberwell, VIC, Australia: Australian Council for Educational Research.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (1996). Bilog-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer program and manual]. Chicago: Scientific Software International.

Journal of Applied Measurement

Volume 9

Author and Title Index