An NCME Instructional Module on

Raju's Differential Functioning of Items and Tests (DFIT)

T. C. Oshima, *Georgia State University*, and
S. B. Morris, *Illinois Institute of Technology*

Nambury S. Raju (1937–2005) developed two model-based indices for differential item functioning (DIF) during his prolific career in psychometrics. Both methods, Raju's area measures (Raju, 1988) and Raju's DFIT (Raju, van der Linden, & Fleer, 1995), are based on quantifying the gap between item characteristic functions (ICFs). This approach provides an intuitive and flexible methodology for assessing DIF. The purpose of this tutorial is to explain DFIT and show how this methodology can be utilized in a variety of DIF applications.

Keywords: differential item functioning (DIF), differential test functioning (DTF), measurement equivalence, item response theory (IRT)

On October 27, 2005, Nambury Raju unexpectedly passed away while working on refining DFIT. Raju was very excited about the arrival of the new significance test for DFIT and he was a few weeks shy of completing his most recent DFIT program. His sudden passing, however, does not by any means indicate the end of DFIT which he worked on during the past two decades. In fact, I believe that he left us enough ground work to move DFIT from the theoretical concept to a common practice in the field of differential item functioning (DIF). The purpose of this paper is to didactically explain how DFIT has been developed and describe how DFIT is simply related to the fundamental principle of item response theory (IRT). Our hope is that some of the readers find DFIT to be amazingly simple and consider adding DFIT in his/her tool box of psychometrics.

Dr. Nambury S. Raju (1937–2005)

DFIT is one of the many indices proposed in the past three decades to investigate DIF. DIF analyses are important in the field of educational and psychological measurement as they address measurement equivalence across subgroups of examinees. Common and popular indices include the

T. C. Oshima is a Professor at Georgia State University. Her research specialties are item response theory, differential item functioning, and multidimensionality. She can be reached at Educational Policy Studies, P.O. Box 3977, Atlanta, GA 30302-3977; Oshima@gsu.edu. Scott B. Morris is an Associate Professor at Illinois Institute of Technology. His research specialties are differential item functioning, adverse impact analysis, and meta-analysis. He can be reached at: Institute of Psychology, Illinois Institute of Technology, 3105 S. Dearborn, Chicago, IL 60616; scott.morris@iit.edu.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Information regarding the development of new ITEMS modules should be addressed to Dr. Mark Gierl, Canada Research Chair in Educational Measurement and Director, Center for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, University of Alberta, Edmonton, Alberta, Canada T6G 2G5.

Mantel–Hanszel technique (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993). More recently, various new methods have been introduced such as DIF effect variance estimators (Camilli & Penfield, 1997; Penfield & Algina, 2006), an empirical Bayes approach to Mantel–Haenszel DIF (Zwick, Thayer, & Lewis, 1999), and hierarchical generalized linear model DIF (Kamata, 2001; Williams & Beretvas, 2006). These DIF analyses do not utilize item parameters estimates from the parametric IRT calibration. Examples of another camp of DIF methods which makes use of the estimated item parameters from an IRT calibration include Lord's χ^2 (Cohen, Kim, & Baker, 1993; Lord, 1980), the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), area measures (Cohen et al., 1993; Kim & Cohen, 1991; Raju, 1988), Muraki's methods for polytomous items (Muraki, 1999) and the methods based on the DFIT framework (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995).

Two indices developed by Raju and his colleagues (area measures and DFIT) are two of the parametric IRT-based indices listed above. Given the wide use of IRT calibration of items in large-scale testing with the help of computer programs such as BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2002) and PARSCALE (Muraki & Bock, 1996) to name a few, it is only natural to make use of the estimated item parameters from the IRT calibration to conduct a DIF study. If the scores are reported using those IRT estimates (that is, if the ability estimates, or thetas, are reported), then, why not report DIF also using those estimates? Assuming one is familiar with IRT, both indices developed by Raju are easy to conceptualize, since DIF is defined as some type of difference (whether area-based or DFIT-based) between the two item characteristic functions (ICFs).

Raju's DFIT has several characteristics that make it a powerful and flexible approach to assessing measurement equivalence: (1) It can be used for dichotomous and polytomous scoring schemes; (2) it can handle both unidimensional and multidimensional IRT models; (3) it provides not only DIF but also differential test functioning (DTF); (4) it provides two types of DIF, compensatory DIF (CDIF) and noncompensatory DIF (NCDIF); and (5) it has been extended to a variety of applications such as differential bundle functioning (DBF) and conditional DIF. These capabilities were realized over the past fifteen years. Figure 1 shows the history of the development of DFIT.

Given those promising properties of DFIT, Raju spent the last part of his career refining DFIT. One of his last accomplishments was the development of the new significance test for polytomous DFIT (Raju, Oshima, Fortmann, Nering, & Kim, 2006).

Item Response Theory Models

Item response theory methods are used to model the functional relationship between item responses and an individual's standing on an underlying latent trait, typically denoted by θ. When analyzing test scores, θ represents a person's ability in a particular domain, but IRT models can also be applied to other types of constructs (e.g., personality traits, attitudes). A variety of IRT models have been developed to address different types of item response formats.

For dichotomously scored items, IRT models the probability of an individual s answering item i correctly as a function of ability (θ). Several dichotomous IRT models have been developed. Here, we will focus on the three-parameter logistic model,

$$P_i(\theta_s) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_s - b_i)}}{1 + e^{Da_i(\theta_s - b_i)}}. \quad (1)$$

The item characteristics are represented by the a_i, b_i, and c_i parameters. The location parameter, b_i reflects the difficulty of the item. The discrimination parameter, a_i, relates to the steepness of the curve. The c_i parameter reflects the lower asymptote, or the probability that a person with extremely low ability would get the item correct. D is a scaling constant typically set at 1.702.

It is useful to depict the model graphically, by plotting probability (Y-axis) against θ (X-axis). The curve plotted is known as item characteristic curve (ICC) or ICF. Two such curves are depicted in Figure 2.

Polytomous IRT models have been developed to analyze items with more than two response categories, such as data from attitude questionnaires. Here we will focus on Samejima's (1969) graded response model (GRM), although the DFIT framework can also be adapted for other polytomous models as well. The GRM is designed for ordered response categories.

Polytomous IRT models require the estimation of multiple ICFs representing the different response categories. For an item with m response categories, there will be m–1 boundary response functions (BRF). A BRF represents the probability of person s responding above response category k on item i,

$$P^*_{ik}(\theta_s) = \frac{e^{Da_i(\theta_s - b_{ik})}}{1 + e^{Da_i(\theta_s - b_{ik})}}, \quad (2)$$

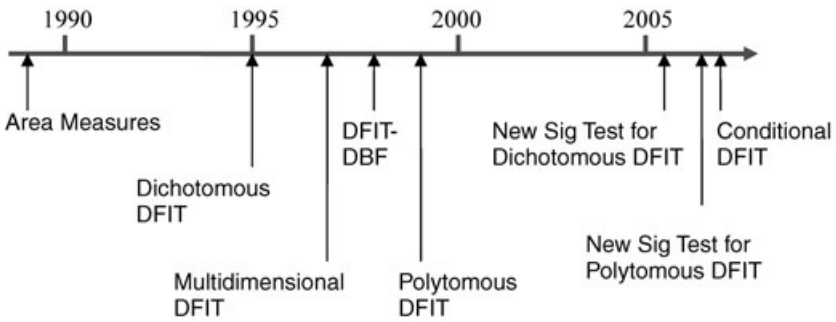

FIGURE 1. History of DFIT.

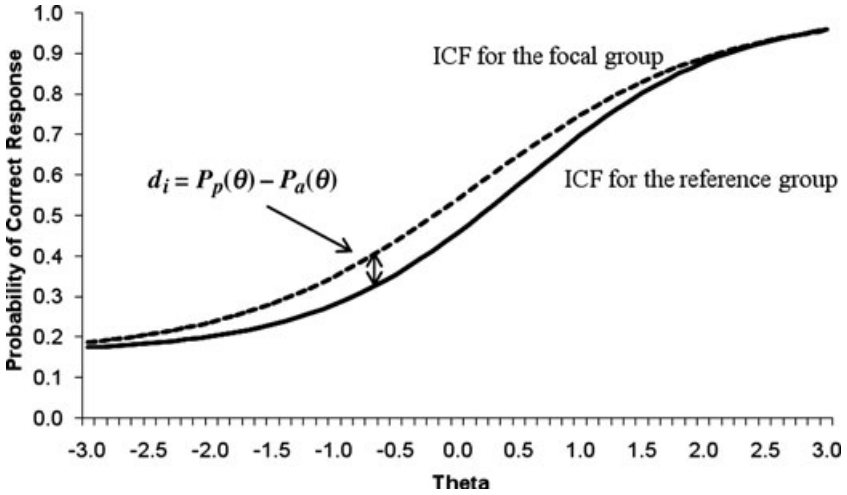

FIGURE 2. IRT and DFIT.

where b_{ik} is a location parameter that designates the boundary between response categories k and $k + 1$, and a_i is the item discrimination parameter.

The probability of responding in a particular response category can be computed from the difference between adjacent BRFs. This function is referred to as the category response function (CRF):

$$P_{ik}(\theta_s) = P^*_{i(k-1)}(\theta_s) - P^*_{ik}(\theta_s). \quad (3)$$

Because the first and last response categories lack an adjacent boundary, Samejima (1969) defined $P^*_{i0}(\theta_s) = 1$, and $P^*_{im}(\theta_s) = 0$. There will be as many CRFs for an item as there are response categories.

It is also useful to generate a single function relating responses on an item to ability. The expected score of individual s on item i, $\mathrm{ES}_{si}(\theta_s)$, can be defined as a weighted average of the category values, where the weights reflect the probability of the individual selecting each category (i.e., the CRFs):

$$\mathrm{ES}_i(\theta_s) = \sum_{k=1}^{m} P_{ik}(\theta_s) X_{ik}, \quad (4)$$

where X_{ik} is the value assigned to category k on item i. For a dichotomously scored item, the expected score function is equal to the ICF.

The total score on a test can be defined as the sum of the scores on the individual items. This total test score can also be modeled as a function of ability and the resulting curve is called the test characteristic function (TCF). The TCF is defined as sum of expected score functions across n items

$$T(\theta_s) = \sum_{i=1}^{n} \mathrm{ES}_i(\theta_s). \quad (5)$$

DFIT statistics rely on item parameter estimates from an IRT model such as those described above. As such, DFIT analysis will be useful only when these item parameters are accurately estimated. Before conducting DFIT analysis, it is strongly recommended that researchers check the assumptions of the particular IRT model they are using (e.g., unidimensionality, model fit), and proceed with DFIT analysis only if these assumptions are met. Methods for evaluating IRT model assumptions are described in Hambleton, Swaminathan, and Rogers (1991).

In addition, accurate parameter estimation typically requires large samples of examinees, and DFIT is not recommended for small samples. For dichotomous IRT models, the required sample size per group will depend on the number of parameters estimated: $N > 200$ for the one-parameter model, $N > 500$ for the two-parameter model, and $N > 1000$ for the three-parameter model (Crocker & Algina, 1986). For polytomous IRT models, $N > 500$ is recommended (Reise & Yu, 1990).

Differential Functioning of Items and Tests (DFIT)

DFIT begins with separate item parameter calibrations for two groups. The resulting ICFs are then compared to determine whether DIF exists. One of the major advantages of IRT over the classical test theory is that the ICFs are invariant over subgroups of examinees (Hambleton et al., 1991). This fundamental property makes IRT an excellent choice for the analysis of differential functioning of items across groups. Simply put, if there is no DIF, the ICF from the focal group (traditionally, it is the minority group) should be the same as ICF from the reference group (traditionally, it is the nonminority group) when they are put on a common scale. If the ICFs are not the same, the item responses do not carry the same meaning for individuals from different groups, and the use of test scores to make comparisons across groups may be inappropriate.

Even though IRT parameters are invariant over subgroups, the scaling of the IRT model is arbitrary, and parameter estimates from separate calibrations will not necessarily be on the same scale. Therefore, before comparing ICFs the two calibrations must be put on a common scale using a process called linking or equating. A variety of linking methods have been developed (Kolen & Brennan, 2004), and there are various linking software programs one can use, such as EQUATE (Baker, 1993) for dichotomous and polytomous models, or IPLINK (Lee & Oshima, 1996) for dichotomous and multidimensional models.

Linking requires the identification of a set of anchor items that are free of DIF, and it is generally not possible to identify these items prior to the DIF analysis. To get around this problem, two-stage linking is recommended (Candell

& Drasgow, 1988). Initially, the scales are linked using all items as anchor items. Then an initial DFIT analysis identifies items with large DIF. A second-stage linking is conducted using the remaining non-DIF items. The linking coefficients obtained from the second-stage linking are then used for calculating the final DFIT indices.

Even after linking, the two ICFs from the focal group and the reference group will never be identical, even when there is no DIF. One has to, of course, allow for sampling error. However, the gap can be larger than what would be expected due to sampling. Figure 2 depicts the gap between two curves. Where there is a gap, the probability of answering an item correctly at a given θ is not the same for the focal group and the reference group. For example, in Figure 2, the probability of answering an item correctly is higher for the focal group than the reference group for most of the θ range. Therefore, this particular item favors the focal group.

Can this gap be an index of DIF? Certainly, and many indices have been proposed. The most obvious would be to measure the area between the two ICFs (area measures by Raju, 1988). The larger the area is, the larger the DIF is. The standard error associated with the area was also developed (Raju, 1990) so that a significance test can be conducted. A limitation of the area measures is that differences between ICFs at all levels of θ contribute equally to the measure of DIF. However, a difference between ICFs will be of greater importance if it occurs in a θ range where there are many examinees. Raju must have realized the shortcomings of the area measures, as he soon developed another measure, DFIT (Raju, van der Linden, & Fleer, 1995). In the DFIT framework, the squared difference is integrated after being multiplied by the density function of ability in the focal group, and therefore represents the typical magnitude of the gap in the actual ability range of interest.

Noncompensatory DIF

The first DIF index in the DFIT framework is noncompensatory DIF (NCDIF), which is defined as the average squared distance between the ICFs for the focal and reference groups. For a dichotomous IRT model, the gap is defined as the difference in the probability of a correct response,

$$d_i(\theta_s) = P_{iF}(\theta_s) - P_{iR}(\theta_s). \quad (6)$$

For a polytomous IRT model, the gap is defined as the difference in expected scores for the focal and reference groups,

$$d_i(\theta_s) = \mathrm{ES}_{iF}(\theta_s) - \mathrm{ES}_{iR}(\theta_s). \quad (7)$$

In either case, NCDIF is defined as the expected value of the squared distance,

$$\mathrm{NCDIF}_i = E_F\left[d_i(\theta_s)^2\right], \quad (8)$$

where E_F denotes the expectation taken over the θ distribution from the focal group.

Taking the square of the d is important so that differences in opposite directions will not cancel each other out. This allows NCDIF to capture both uniform and nonuniform DIF. When the ICFs differ only on the b parameters, DIF is called uniform, and direction of the gap will be the same across the ability distribution (as in Figure 2). Nonuniform DIF occurs when the a parameters differ across groups. In this case, the focal group curve will be higher at some ability levels, but lower at others. When DIF is nonuniform, differences in both directions will contribute to NCDIF.

Differential Test Functioning

One of the advantages of the DFIT framework is the ability to assess differential functioning not only at the item level, but also at the test level. DTF is similar to NCDIF, except that the two curves being compared are the TCFs,

$$D(\theta_s) = T_F(\theta_s) - T_R(\theta_s). \quad (9)$$

Two such curves are depicted in Figure 3. DTF is defined as the expected value of the squared difference between focal and reference groups, where the expectation is taken across the θ distribution from the focal group,

$$\mathrm{DTF} = E_F[D(\theta_s)^2]. \quad (10)$$

Despite the similarity in how NCDIF and DTF are defined, the relationship between the two statistics is not straightforward. Like most item-level DIF indices, NCDIF assumes that all items other than the studied item are DIF free. DTF, on

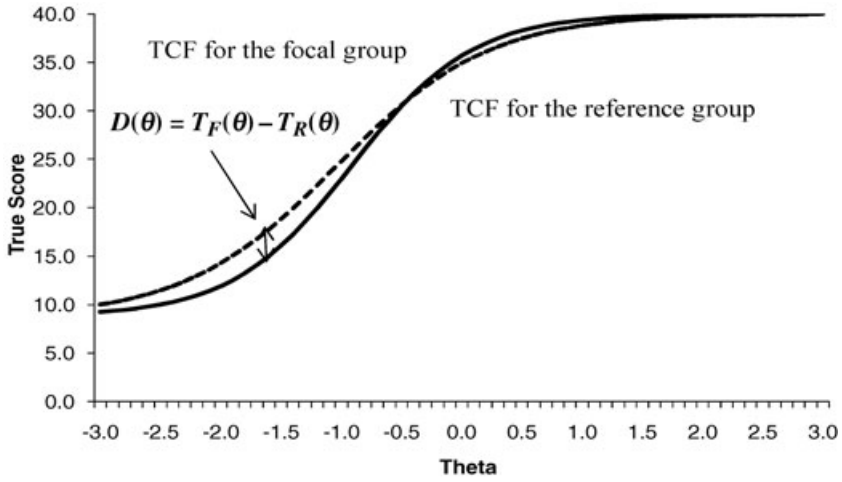

FIGURE 3. A graphic presentation of DTF.

the other hand, depends not only on the level of DIF on each item, but also the pattern of DIF across items. Thus, removing an item with large NCDIF will not necessarily result in a large decrease in DTF. A third DFIT index, CDIF, better reflects the item's contribution to DTF.

Compensatory DIF

By taking the item covariances into account, Raju et al. (1995) were able to develop another index called CDIF (compensatory DIF) which relates item and test level differential functioning in an amazingly simple relationship. CDIF is defined as

$$\text{CDIF}_i = E_F(d_i D) = \text{Cov}(d_i, D) + \mu_{d_i}\mu_D, \quad (11)$$

where Cov stands for covariance. The CDIF index is additive such that:

$$\text{DTF} = \sum_{i=1}^{n} \text{CDIF}_i. \quad (12)$$

The CDIF index is unique among DIF indices due to its additive nature. A researcher can investigate the net effect of removing certain items on DTF. Theoretically speaking, it is possible for an item to have NCDIF but not CDIF. For example, if one item favors the focal group (thus showing NCDIF) and another item favors the reference group by the same amount (thus again showing NCDIF), the DIF on the two items will cancel out. Neither item will show CDIF, and there would be no DTF.

The concept of compensatory DIF deserves further investigation as it helps test developers focus on differential functioning at the test level and not at the item level. It is not always easy or practical to generate totally gender-free or ethnicity-free (or any subgroup-free) items. With the help of the CDIF index, one can try to develop a test with the least amount of DTF.

Extensions to Other Models

In the DFIT framework, the extensions of NCDIF, CDIF, and DTF to more complex IRT models are straightforward. We have seen that DFIT can be generalized from the dichotomous case to the polytomous case by replacing the item probability, $P(\theta)$ with the expected score, $\text{ES}(\theta)$. This was demonstrated using the GRM. Other polytomous models will differ only in the details of computing the expected score.

For multidimensional IRT models, the only difference is reflected in how one calculates the ICF. The ICF will be calculated based on multiple thetas in the multidimensional models. Once $T(\theta)$ (see Equation (5)) is calculated accordingly, the definitions of DFIT indices stay the same (Oshima, Raju, & Flowers, 1997).

Differential Bundle Functioning

Besides the item-level and the test-level differential functioning, recently more attention has been paid at the item-bundle level (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). DBF can help researchers identify possible cause of DIF as well as examining differential functioning for a cluster of items such as those from reading paragraphs. Not surprisingly, due to the additive nature of the DFIT framework, defining DBF is straightforward. Bundle NCDIF is basically DTF for the bundle. Bundle CDIF is simply the sum of item CDIF. The sum of bundle CDIF, then, is DTF. More detailed explanations can be found in Oshima, Raju, Flowers, and Slinde (1998).

Conditional DFIT

Just as DBF focuses on a cluster of items, it is also possible to examine DFIT for clusters of examinees. DFIT statistics are computed by taking an expected value across examinee ability levels. Conditional DFIT (Oshima, Raju, & Domaleski, 2006) is defined by taking the expectation across a subset of examinees. By doing so, one can examine DIF/DTF/DBF for a certain group of people. For example, if a test has a cutoff score, it is of particular importance to examine DIF/DTF/DBF around the cutoff score. One can also examine DIF/DTF/DBF for a subgroup of interest. For example, one can assess gender DIF separately for each ethnic group.

Significance Tests in the DFIT Framework

Raju et al. (1995) originally developed significance tests for DFIT based on the χ^2 statistic. Tests were developed for both NCDIF and DTF. The significance of CDIF is not tested directly. Instead, items with large CDIF are removed one by one until DTF reaches nonsignificance. Those removed CDIF items are then considered significant.

These χ^2 tests, however, turned out to be overly sensitive in large samples and tended to falsely identify non-DIF items as having significant DIF. Based on simulation studies, Raju then recommended a predetermined cutoff score of NCDIF $>$ .006 for the dichotomous items, or NCDIF $>$ $.006(k-1)^2$ for polytomous items. This one-size-fits-all approach, of course, was too simplistic. Additional simulations found that the appropriate cutoff depended on factors such as the sample size and the particular IRT model used (Bolt, 2002; Chamblee, 1998). Therefore, the cutoff scores developed in prior research may not generalize to all situations. Further, typical practitioners may not have the expertise or time to conduct their own simulations. Until recently, the lack of a generalizable significance test for NCDIF was a major impediment to its use.

Recently, a new significance test for NCDIF was proposed for the dichotomous case. The item parameter replication (IPR) method (Oshima, Raju, & Nanda, 2006) provides a means of deriving cutoff values that are tailored to a particular data set. The IPR method begins with estimates of item parameters for the focal group and the sampling variances and covariances of these estimates. Based on these initial estimates, a large number of replications of item parameters are then generated with the restriction that the expectation of the newly generated item parameters equals the initial estimates with the same sampling variance/covariance structure.

Because they are generated from the same distribution, any differences in the sets of estimates must be due to sampling error. Pairs of samples can then be used to compute DIF statistics. This produces an empirical sampling distribution of NCDIF under the null hypothesis that focal and reference groups have identical parameters. The resulting NCDIF values are then ranked and the cutoff is set at the percentile corresponding to the desired alpha level (e.g., the 99th percentile for $\alpha = .01$).

The IPR method has been implemented in the latest version of the DFIT software (Raju, Oshima, & Wolach, 2005).

Raju was working on extending the IPR method to the polytomous case when he passed away. Additional information on this work can be found in Raju et al. (2006).

Conclusions

In this instructional module, DFIT was introduced in the didactic tone as a psychometric tool for assessing measurement equivalence. It was our intention that DFIT is explained in the simplest language possible so that a wide audience of students and professionals in the educational measurement field would find it easy to understand and consider making use of DIF techniques in practice. This module is like a series of snapshots of DFIT and it is not meant to be comprehensive. Therefore, we strongly recommend that the interested reader obtain the original articles.

DFIT was one of the many psychometric contributions of Nambury Raju during his distinguished career. Dr. Raju was an accomplished mathematician, and he had a knack of simplifying a lengthy and complex problem to an amazingly simple equation. The DFIT framework is one of those operations.

Self-Test

True or False

1. The Mantal–Hanszel technique and DFIT are similar in the sense that they both utilize item parameter estimates from a parametric IRT calibration.
2. DFIT can be used both for dichotomous items and polytomous items.
3. DFIT can be used both for unidimensional IRT models and multidimensional IRT models.
4. DFIT can handle any sample sizes.
5. DFIT requires separate calibrations of IRT parameters for the focal group and the reference group.
6. If there is no DIF, the estimated item parameters from the focal group and the reference group would be identical.
7. If DTF is not significant, one can safely assume there is no DIF.
8. DFIT indices would be affected by the distribution shape of the focal group.
9. DFIT cannot identify nonuniform DIF.
10. The sum of NCDIF equals DTF.
11. If you remove significant CDIF items, DTF would not be significant.
12. NCDIF assumes that all the items but the studied item are DIF free.
13. NCDIF and CDIF can be calculated at the level of item bundles.
14. The recommended test for significance in DFIT is the χ^2 test.
15. Dr. Raju had a prolific career in psychometrics.

Key to Self-Test

1. F. The Mantal–Hanszel technique, as well as other popular techniques (logistic regression, SIBTEST, etc.), do not require item parameter calibration. DFIT requires IRT item calibrations using software, such as BILOG-MG3, Parscale, or NOHARM (Frasier, 1988).
2. T. This is one of the advantages for DFIT. In the dichotomous case, DFIT is calculated based on the difference between the ICFs of the focal group and the reference group. In the polytomous case, it is calculated based on the expected score (ES) difference. Therefore, any polytomous model which produces an ES can be used in the DFIT framework. Raju's original work in the polytomous case uses Samejima's GRM.
3. T. Except for the need to integrate over multiple ability dimensions, the DFIT framework stays the same regardless of the number of dimensions the test measures. DFIT has been developed for unidimensional dichotomous models, unidimensional polytomous models, and multidimensional dichotomous models so far. Work is still in progress for multidimensional polytomous models.
4. F. DFIT is not recommended for small sample sizes. DFIT is based on the item parameter estimates from an IRT calibration program. If items are not estimated well, then, the following DFIT analysis would be erroneous. (This is similar to the problem of measurement and statistics in research. That is, if variables are not measured well, then the subsequent statistical analyses would be erroneous.) The recommended sample sizes for DFIT (which are equal to the recommended sample sizes for calibrating various IRT models) can be found in the text above. Please note that the sample sizes listed there are required for EACH group (focal group and reference group).
5. T. For the current version of DFIT, one needs to calibrate items separately for the focal group and the reference group. Then, the items need to be put on a common scale through linking. The DFIT research has been conducted primarily using the linking procedure called the test characteristic curve method (Stocking & Lord, 1983). However, there are other linking methods that may be used. In the DFIT research, we have found that linking is a crucial part in the performance of DFIT. We also found that the two-stage linking is an essential part of DFIT, especially when the number of DIF items is large.
6. F. If there is no DIF, the "true" (not "estimated") item parameters from the focal group and the reference group would be identical. In practice, we do not know the true item parameters. Therefore, we will be dealing with two sets of item parameters (from the focal group and from the reference group) that are different even after linking. Then, we need to determine if the difference was within the sampling/estimation error or not. The new significance test (Oshima, Raju, & Nanda, 2006) addresses this issue.
7. F. Even if DTF is not significant, it is possible to get significant NCDIF items. For example, suppose one is conducting a DIF study based on gender. Say, three items favor males and three items favor females at about the same degree. Those six items would show significant NCDIF, but DTF would not be significant assuming all other items are DIF-free. On the other hand, when DTF is not significant, one can say there is no CDIF. Therefore, it is possible that an item shows significant NCDIF but not significant CDIF.
8. T. DTF is the average squared distance for two item response functions over the "focal" group. Therefore, theoretically, the ability distribution of the focal group would affect the DFIT indices. For example, one would get different values of DFIT indices if it is skewed as opposed to normally distributed. One might wonder why the focal group, and not the reference group. This question was asked of Dr. Raju many times. His response was that the focus of a DIF study should be on the focal group.
9. F. DFIT can identify both uniform and nonuniform DIF. The detection of DIF depends on the magnitude of DIF whether or not it is uniform or nonuniform.

10. F. The sum of CDIF equals DTF. The sum of NCDIF does not equal DTF.
11. T. In the DFIT framework, large CDIF items are removed one by one until DTF reaches nonsignificance. Those removed CDIF items are then considered significant. Therefore, after removing all those significant CDIF items, DTF should be nonsignificant.
12. T. CDIF, on the other hand, is affected by the presence of DIF on the other items on the test.
13. T. The bundle CDIF is simply the sum of item CDIF, just like the sum of CDIF was DTF. The bundle NCDIF is not the sum of item NCDIF. Instead, the bundle NCDIF is the DTF index computed on the items in the bundle.
14. F. Although Raju developed a significance test based on the χ^2 test in the early 1990s, it is not the recommended significance test today. The new significance test developed recently offers the cutoff score for each item based on the estimated sampling distribution of item parameters under the null (no-DIF) condition. Although the test is calculation intensive, the process is built in inside the new DFIT computer program. Therefore, from the user's end, the significance test is automatically conducted.
15. T. Dr. Raju "was a prolific writer and highly involved in the profession: the author of over 150 publications and presentations, member of more than 8 professional organizations, and editor or reviewer for more than 24 professional journals. . . . Dr. Raju supervised over 32 doctoral dissertations and 20 master theses and was held in the highest regard by all for his warm heart, strong intellect, and unflagging integrity." (PsychLink, 2006). As his colleagues and friends, we cannot agree more about his integrity and kindness to others. As many have said, he was a *great* man.

References

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *2*, 113–141.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential item functioning based on Mantel–Haenszel statistics. *Journal of Educational Measurement*, *34*, 123–139.

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

Chamblee, M. C. (1998). *A Monte Carlo investigation of conditions that impact type 1 error rates of differential functioning of items and tests*. Unpublished doctoral dissertation, Georgia State University, Atlanta.

Cohen, A. S., Kim, S., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*, 335–350.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, *23*, 309–326.

Frasier, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer program]. New South Wales, Australia: Center for Behavioral Studies, University of New England.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, *20*(1), 26–36.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*, 269–278.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lee, K., & Oshima, T. C. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement*, *20*, 230.

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

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*, 217–232.

Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago, IL: Scientific Software.

Oshima, T. C., Raju, N. S., & Domaleski, C. S. (2006, April). *Conditional DIF and DTF*. Paper presented at the Annual Meeting of American Educational Research Association, San Francsisco.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, *34*, 253–272.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. (1998). Differential bundle functioning (DBF) using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education*, *11*, 353–369.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*, 1–17.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*, 295–312.

Psychlink: The Newsletter of the Institute of Psychology. (2006). *Psychology mourns Dr. Nambury Raju*, *8*, 1–3. Chicago: Illinois Institute of Technology.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207.

Raju, N. S., Oshima, T. C., Fortmann, K., Nering, M., & Kim, W. (2006, February). *The new significance test for Raju's polytomous DFIT*. Paper presented at the New Directions in Psychological Measurement with Model-Based Approaches at Georgia Institute of Technology in Atlanta, GA.

Raju, N. S., Oshima, T. C., & Wolach, A. (2005). *Differential functioning of items and tests (DFIT): Dichotomous and polytomous* [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., van Der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement*, *19*, 353–368.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*, 133–144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *17* (monograph supplement).

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

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

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

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, *30*, 22–42.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG3* [Computer software]. St. Paul, MN: Assessment Systems Corporation.

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