International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Differential Item Functioning Detection Methods: An Overview

Solomon Chukwu Ohiri[1], *Momoh, Osi Christopher*[2], *Ikeanumba, Chukwuemeka Benedict*[3]

[1] *Directorate of Academic Planning, Alvan Ikoku University of Education, Owerri, Nigeria E-mail:* ohiri.chukwu2018@gmail.com
[2] *Teaching Practice Unit, Alvan Ikoku University of Education, Owerri, Nigeria*
[3] *Dept. of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Nigeria E-mail:* chukwuemekaikeanumba@gmail.com

DOI: https://doi.org/10.55248/gengpi.5.0224.0505

ABSTRACT

With the increasing concerns over the fairness of educational and psychological test items, differential item functioning (DIF) approach has been widely used to find out items that are biased. In the educational and psychological measurement settings, a test item is recognized as differentially functioning across groups when the probability of an examinee's response to it depends on group membership. This paper explored differential item functioning as the statistical characteristic of an item that shows the degree to which the item might be measuring different abilities for members of distinct subgroups. The framework of DIF has become an essential part of test validation methodology and the study of test fairness. Differential item functioning is typically identified using inferential DIF detection methods. Eight DIF detection methods were explained: Logistic Regression, Mantel-Haenszel, Item Response Theory, Lord's Chi-Square, Likelihood-ratio test method, b-parameter method, Simultaneous item bias tests method (SIBTEST) and the Rasch model. For each method, a number of the most widely applied software were mentioned.

Key words: Differential Item Functioning, Test fairness, Attributes, Performance, Subgroups.

Introduction

Differential item functioning denotes to variances in the functioning of items across groups, many a time demographic, which are matched on the ability or more generally the characteristic being measured by the items or test. This means that when personal attributes, such as gender or ethnicity systematically affect examinee performance on an item, the result can be differential item functioning. This can be described as the statistical characteristic of an item that shows the degree to which the item might be measuring different abilities for members of distinct subgroups. If the factor bringing about such a difference is not part of the construct of focus in the test, then the test would be biased (Karami & Nodoushan, 2011). If, on the other hand, the differential performance of two groups can be attributed to a true difference in their ability levels, it is called impact rather than bias (Kamata & Vaughn, 2004).

Differential item functioning

Normally, it is expected that two individuals at the same level of ability, regardless of what group they belong to, will have the same probability of correctly or affirmatively responding to an item. If this is not true for an item, the item is said to be functioning differentially.

Differential item functioning (DIF) is an analysis of performance across groups on specific test items. It is a statistical technique that is used to identify differential item response patterns between groups of examinees such as male and female which helps in verifying potentially biased test items. DIF is of great interest to researchers and educators given that DIF poses a potential threat to test fairness. As an item analysis methodology different from comparing mean scores at test level, DIF plays an important role in detecting the items that function differentially in a test.

Differential item functioning occurs when individuals of the same ability level from separate groups have a different probability of answering an item correctly (Annan-Brew, 2020). DIF is an indicator of bias observed when test takers from different groups have different probability or likelihood of responding correctly to an item, after controlling for ability. Subgroups typically studied in DIF analyses are examinees' characteristics such as gender, school type, religion and socioeconomic status. DIF is a threat to comparability and occurs if an item is easier for one group of test takers than for another after controlling for overall ability.

The framework of DIF has become an integral component of test validation methodology and the study of test fairness. All DIF methods work by comparing how two groups perform on each item in an exam, after matching examinees on a criterion (Davidson, Ko, Wortzman & Li, 2021). That criterion can be the total score on ability, or some other variable. In the two-group case, the group that is concerned about items being biased against is called the focal group, and the other group, is the reference group. The focal group is the focus of the analysis, and the reference group serves as a basis for comparison for the focal group. When items composing a test behave differently for the reference group compared to the focal group, even after controlling for student proficiency, DIF is said to have occurred.

In the case of ability testing, DIF is defined as an item level performance difference between groups of examinees matched on ability. The definition of DIF for surveys (an attitude measure) is slightly different because respondents on surveys are matched on overall agreement level instead of ability (Dodeen, 2014).

Differential item functioning is typically identified using inferential DIF detection methods. Inferential DIF detection methods use a significance test to determine if an item possesses DIF. The numerical value obtained from the inferential DIF method indicates that an item is more difficult for a particular subgroup than originally intended. In the case of attitude item, DIF indicates that a particular subgroup responded more positively to an item than another group. Dodeen and Johnson (2013) stated that the correct answer in the cognitive context is similar to the positive affective of attitude towards the item.

Items on ability or attitude assessments may exhibit DIF for several reasons. The occurrence of DIF could be due to Type 1 error or item multidimensionality. Type 1 error occurs if the null hypothesis is rejected when it could have been retained. If the sources of DIF cannot be explained after the item is evaluated, Camilli and Shepard in Oratokhai (2021) stated that the item was likely flagged for DIF due to a Type 1 error of the statistical method used to detect DIF. Although Type 1 error can account for unexplained DIF, it is likely that other reasons could be hypothesized as to the reason the DIF could not be explained. To give example, DIF may go unexplained due to lack of understanding of the examinee population taking the assessment. Differential item functioning that can be explained, however, can be interpreted to be caused by item multidimensionality. Item multidimensionality occurs when an item simultaneously measures two or more constructs.

Shealy and Stout in Oratokhai (2021) posed a theory stating that DIF results from an item measuring an additional construct or dimension. The theory stated that examinees have the same ability or agreement level on the matching dimension (the primary dimension of interest), but differ in their ability or agreement level on the second dimension. The presence of DIF, however, is not always detrimental to a group of examinees or respondent's trait inferences, but for DIF to affect examinee or respondent inferences adversely, the second dimension must be caused by a specific factor. Two factors that cause item multidimensionality are a construct relevant factor and a construct irrelevant factor. A construct relevant factor has no adverse effect on item measurement, whereas a construct irrelevant factor changes the measurement intent of an item.

Forms of Differential Item Functioning

Generally, Differential item functioning is of two types, uniform and non-uniform. Uniform or unidirectional DIF exists when the probability of endorsing an item (answering an item correctly) is greater for one group than for the other group over all the levels of proficiency. The existence of non-uniform or crossing DIF demonstrates that the difference in probabilities of a correct response is not the same at all levels of proficiency between the two comparison groups. That is, the probability of correctly answering an item is higher for one group at some points on the scale, and higher for the other group at other points (Ibrahim, 2017).

DIF can also be represented graphically as the difference between the item characteristic curves of the focal and reference groups. If there is no DIF present, the two curves would be superimposed on each other. Plots of items presenting uniform and non-uniform DIF are presented in Figures 1 and 2 respectively. In uniform DIF, one group is advantaged throughout the range of the ability, that is, for the advantaged group, the probability of correctly responding to an item is consistently greater than for the other group. Items that show uniform DIF have different difficulty parameters for the groups meaning that the item is more difficult for one of the groups throughout the ability continuum. In the case when an item shows non-uniform DIF, members of one group find the item more difficult in one part of the ability continuum and then this is reversed for another part of the ability continuum. Items that function differently in a non-uniform way, have different discrimination parameters and potentially different difficulty parameters.

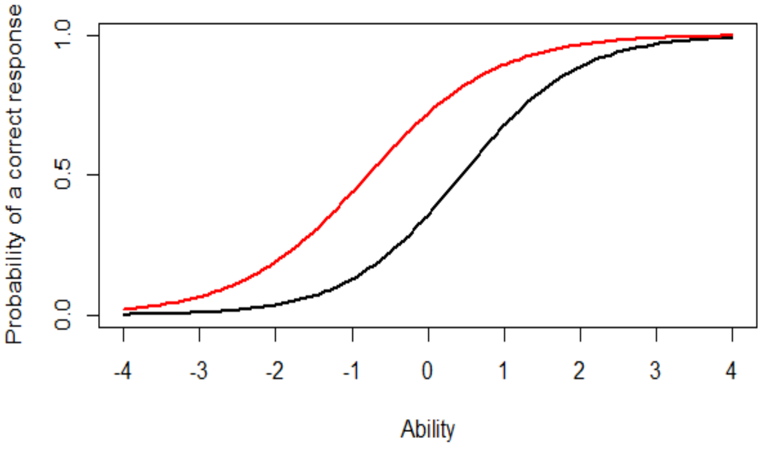

Figure 1. An item presenting uniform DIF

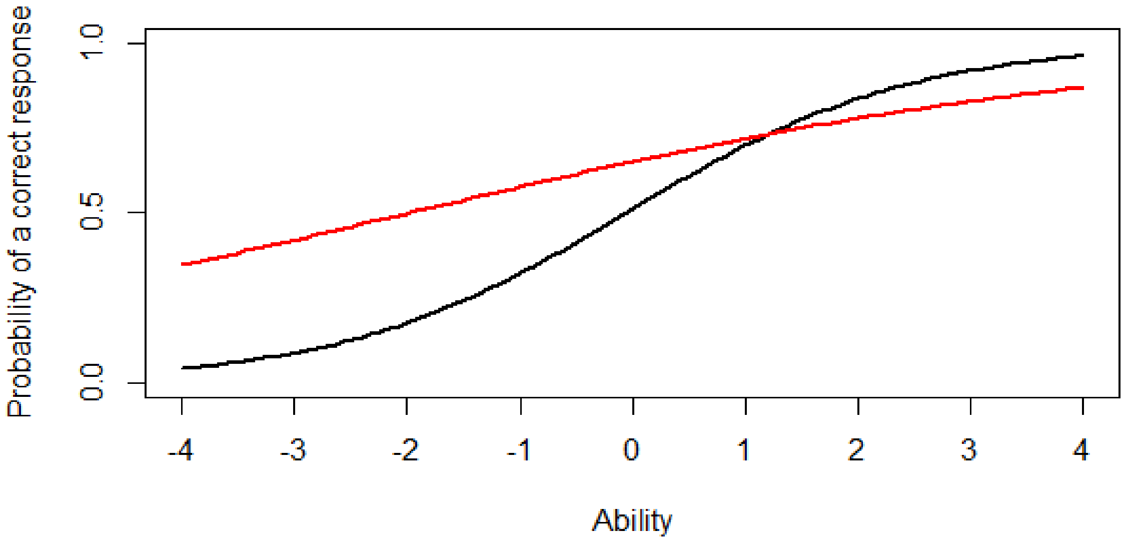

Figure 2. An item presenting non-uniform DIF. The ICCs of the focal and reference groups cross because the advantage is not uniform across the scale.

Methods of detecting differential item function

The presence of bias in test items is a cause for concern to test developers and education researchers. If some items in a test function differently for a specific subgroup than the majority of the population being tested, direct comparison of their performance on the items make no sense (Krishnan, 2013). This led to many item bias detection procedures or differential item functioning (DIF) to flag for possible item bias. A substantial number of DIF detection methods have been proposed based on different test statistics. Unfortunately, most of these approaches are based on the comparison of the item parameter estimates between two or more pre-specified groups of examinees, such as males and females, white or non-white, as focal and reference groups. Researchers tend to pre-specify the focal group and reference group in advance, (e.g., males versus females) in gender-related DIF studies.

There are two major psychometric theories in the study of measurement procedures, which are classical test theory and item response theory and their corresponding models have been currently used for addressing differential item functioning studies. Classical test theory (CTT) is a psychometric theory of assessment/measurement that purports that every individual has some innate or “true” ability for any given attribute, and that the attribute can be measured, and the process of measurement inherently has error (Wang, 2018). On the other hand, item response theory (IRT) is an area of test theory which provides probabilistic approach to overcome some of the limitations of classical methods (Ashraf & Jaseem, 2020). The item response theory (IRT), also known as the latent response theory refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations. It is a statistical technique involving models expressing the probability of a particular response to a scale item as a function of the ability of the subject.

Methods of detecting DIF which utilize classical test theory include logistic regression, correlation methods, log-linear analysis, analysis of variance, scheuneman’s chi-square, mantel-Haenszel method, transformed item difficulty or delta plot which is based on the difference between the difficulty parameter estimates obtained in each group. Other methods based on item response theory include b-parameter, Wald statistics, likelihood-ratio, Lord’s chi-square.

Logistic regression method

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. The logistic model computes the probability of the selected response as a function of the values of the predictor variables. If a predictor variable is categorical with two values, then one of the values is assigned the value 1 and the other is assigned the value 0. If a predictor variable is a categorical with more than two categories, then a separate dummy variable is generated to represent each of the categories except for one which is excluded. The value of the dummy variable is 1 if the variable has that category, and the value is 0 if the variable has any other category. If the variable has the value of excluded category, the entire dummy variable generated for the variables are zero.

The logistic regression procedure was proposed by Rogers and Swaminathan (1990) to detect differential item functioning between manifest groups (reference group and focal group). In this method, the outcome (dependent) variables can be identified as item responses (0 = incorrect, 1 = correct), whereas, the predictors (independent) can be defined as the total test score (matching criterion, K), manifest group membership (gender), and interaction between the total test and manifest group membership (kamala & vaughn in Sapmaz, 2019).

Logistic regression in general models a binary dependent variable and independent variables based on the logistic function. The independent variables can be binary or continuous. In the application to DIF detection, logistic regression models the binary response of an item as the dependent variable and test scores, group membership and interaction between test scores and group membership as the independent variables. The model for predicting the probability of giving the correct response according to Swaminathan and Rogers is given by

$$P\ (\mu=1/\theta) = \{e^{(\beta 0 + \beta 1\theta)}\} / \{1 + e^{(\beta 0 + \beta 1\theta)}\} \qquad (1)$$

where μ is the binary response to the item, θ is the observed ability and β_0, β_1 are the intercept and slope.

The logistic regression is performed first on an item using respondents in the reference group and then on the same item using respondents in the focal group, with the slopes and intercepts estimated. β_{0R}, β_{1R} for the first and β_{0F}, β_{1F} for the second. If there is no DIF, the probability of correctly answering the question should be the same for both groups and so the intercept and slope parameters should also be equal. Consequently, if any of the parameters are different then it indicates the presence of DIF. If the intercept parameters ($\beta_{0R}\ and\ \beta_{0F}$) are the same but the slope parameters differ, then the item exhibits non-uniform DIF because the logistic regression curves have different slopes and so could possibly cross. On the other hand, if the slope parameters ($\beta_{1R}\ and\ \beta_{1F}$) are equal and the intercept parameters differ then the item displays uniform DIF and the logistic regression curves are separate and parallel. In practice the two separate logistic regressions for the focal and reference groups are combined in a nested model.

In the course of this study, the logistic regression procedure will be used to detect DIF in terms of gender. The item response (0 or 1) of the participants will serve as the dependent variable, with grouping variable (dummy coded as 1 = reference, 2 = focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This method will provide a test of DIF conditionally on the relationship between the item response and the total score, testing the effects of group for uniform DIF, and the interaction of group and TOT to assess non-uniform DIF.

Logistic regression is a viable and flexible procedure for detecting DIF that does not require specific forms of item response function or large sample sizes (Narayanam & Swaminathan in Chen & Jin, 2018). They went further to say that it also demonstrates computational simplicity and is easily implementable using commercial software like SPSS, SAS or STATA, without additional effort or knowledge. Logistic regression provides a more precise description of differential item functioning compared to the MH procedure, and allows for distinguishing between two types of DIF: uniform and non-uniform DIF. However, based on differential item functioning analyses from classical test theory, the logistic regression technique is quite a powerful approach that has been used in identifying differential item functioning in dichotomous items.

Logistic regression approach to differential item functioning detection involves running a separate analysis for each item. The independent variables included in the analysis are group membership, an ability matching variable typically a total score, and an interaction term between the two. The dependent variable of interest is the probability or likelihood of getting a correct response or endorsing an item. Since the outcome of interest is expressed in terms of probabilities, maximum likelihood estimation is the appropriate procedure. The procedure can be formulaically presented as follows:

$$L_n \left[\frac{P_{mi}}{1 - P_{mi}} \right]$$

$$= b_0 + b_1 tot + b_2 group + b_3\ (tot \ x \ group) \qquad (2)$$

In the formula, b_0 is the intercept, $b_1$tot is the effect of conditioning variable which is usually the total score on the test, b_2 group is the grouping variable, and finally b_3 (tot x group) is the ability by grouping interaction effect. If the conditioning variable alone is enough to predict the item performance, with relatively little residuals, then no DIF is present. If group membership, $b_2$group, adds to the precision of the prediction, uniform DIF is detected. That is, one group performs better than another group and this is a case of uniform DIF. Finally, in addition to total scores and grouping, if an interaction effect, signified by b_3 (tot x group) in the formula, is also needed for a more precise prediction of the total scores, it is a case of non-uniform DIF (Zumbo, in Karami, 2012).

Also, the formula is based on logistic function denoted by

$$L_n \left\{ \frac{P_{mi}}{1 - \mathcal{P}_{mi}} \right\}$$

where P_{mi} is the probability of giving a correct answer to item i by person m and 1-P_{mi} is the probability of a wrong response. In simple words, it is the natural logarithm of the odds of success to the odds of failure (Karami, 2012).

Identifying DIF through logistic regression is similar to step-wise regression in that successive models are built up in each step entering a new variable to see whether the new model is an improvement over the previous one due to the presence of the new variable.

An advantage of logistic regression method is the ability to test for both uniform and non-uniform DIF. The presence of uniform DIF is evaluated by testing whether the regression coefficient of group membership (b_2) differs significantly from zero. A test of the interaction coefficient between group membership and ability (tot) (b_3) can be used to assess non-uniform DIF. Some researchers advocate first testing the presence of both uniform and non-uniform DIF simultaneously using a test of the null hypothesis that $b_2 = b_3 = 0$ (Roger & Swaminathan in Oratokhai, 2021).

Zumbo in Karami (2012) argued that logistic regression has three main advantages over other DIF detection techniques in that one:

- ✓ need not categorize a continuous criterion variable
- ✓ can model both uniform and non-uniform DIF
- ✓ can generalize the binary logistic regression model for use with ordinal item scores.

McNamara and Roever in Karami (2012) also state that logistic regression is useful because it allows modeling of uniform and non-uniform DIF, is non-parametric, can be applied to dichotomous and rated items, and requires less complicated computing than IRT-based analysis.

Mantel-Haenszel procedure

The Mantel-Haenzel (MH) procedure was first proposed for DIF analysis by Holland and Tayer (Kamata & Vaughn, 2004). The MH DIF procedure compares dichotomous item performance between two groups after matching respondents on overall scores. Respondents in the focal and reference groups are matched on total test scores by dividing respondents in both groups into defined strata on those scores (Liu & Bradley, 2021). The total scores are generated by summing item scores across all item. Estimates of the odds ratio for a given item can be calculated based on 2 x 2 x k contingency table with K representing the k-th group (K = 1,2,…,k). The item is not DIF if the odds of answering the item correctly are about the same across focal and reference groups, that is, if the odds ratio.

$$\alpha_k = \frac{A_k / B_k}{C_k / D_k} = \frac{A_k D_k}{B_k C_k} \quad \text{is close to 1} \qquad (3)$$

Table 1: A 2 x 2 x k Contingency

	Scores on the studied items		
	1(Right)	0(Wrong)	Total
Reference group	A_k	B_k	N_{rk}
Focal Group	C_k	D_k	N_{fk}
Total	M_{ik}	M_{ok}	N_k

The A_k, B_k, C_k and D_k denote the numbers of respondents in the cells. N_k represents the number of respondents in the K-th stratum. The cells A_k and C_k represent the total number of respondents who answered the item correctly in the reference and focal groups, respectively, within the matched subgroup K. B_k and D_k denote the total number of respondents who answered the item incorrectly in the reference and focal groups, respectively, within k-th group. M_{ik} and Mo_k denote the number of respondents who answered the item correct and incorrect, respectively, with K-th group (Liu & Bradley, 2021). To account for all total score levels simultaneously, the Mantel-Haenszel estimate of the odds ratio α_{MH}, can be calculated as the weighted average of the odds ratios across the score levels:

$$\alpha^2_{MH} = \frac{\sum_k \frac{A_k D_k}{N_k}}{\sum_k \frac{B_k C_k}{N_k}} \qquad (4)$$

If the item functions the same for different groups for all levels of total score, K, then the odds ratio α_{MH}, equals 1, indicating that there is no DIF, and for a DIF item that fovours the reference group, α_{MH}, will be greater than 1. When a DIF item favours the focal group, α_{MH}, will be less than 1.

To test for the presence of DIF, the Mantel-Haenszel χ^2_{MH} statistics can be calculated as

$$\chi_{MH} = \frac{\{|\sum_k [A_k - {}^{(A_k + B_k)(A_K + C_k)}]| - 0.5\}^2}{\dfrac{\sum_k (A_k + B_k)(A_K + C_k)(B_k + D_k)(C_K + D_k)}{N^2_k (N_k - 1)}} \qquad (5)$$

and compared with a critical value from the distribution with degree of freedom (df) = 1 to determine the p-value.

MH procedure is the most widely used procedure to detect DIF in practice since it is not only easy to understand and compute, it can provide both a significance test and estimate of the magnitude of DIF as well (Millsap in Lui & Bradley, 2021). Millsap went further to say that the major criticism of the MH procedure is the adequacy of using the total score as a substitute for the latent trait.

Item response theory

In item response theory, the goal is to examine how observable test performance relates to unobservable latent traits that are measured by the items in a test. IRT consists of methods and models that allow for analysis of particular items on a test and how they are responded to. The main difference between IRT DIF detection techniques and other methods including logistic regression and MH is the fact that in non-IRT approaches, "examinees are typically matched on an observed variable (such as total test score), and then counts of examinees in the focal and reference groups getting the studied item correct or incorrect are compared". That is, the conditioning or the matching criterion is the observed score. However, in IRT methods, matching is based on the examinees' estimated ability level or the latent trait, Θ (Karami, 2012).

Methods based on item response theory are conceptually elegant though mathematically very complicated. The building block of IRT is item characteristics curve (ICC). It is a smooth S-shaped curve which depicts the relationship between the ability level and the probability of correct response to the item. As it is evident from Fig. 3, the probability of correct response approaches 1 at the higher end of the ability scales, never actually reaching 1. Similarly, at the lower end of the ability scale, the probability approaches, but never reaches zero.

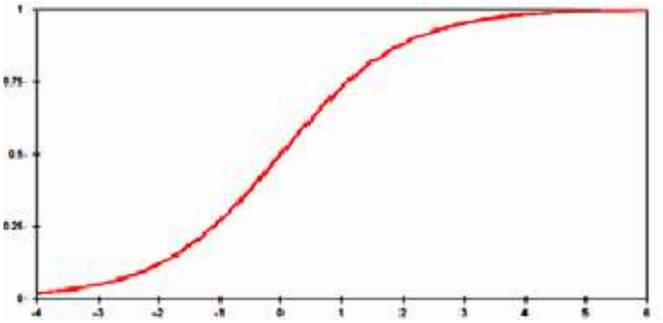

Fig. 3. A Typical ICC

IRT uses three features to describe the shape of the ICC: item difficulty, item discrimination, and guessing factor. Based on how many of these parameters are involved in the estimation of the relationship between the ability and item response patterns, there are three IRT models, namely one, two and three parameter logistic models. In the one parameter logistic model and the Rasch model, it is assumed that all items have the same discrimination level. The two parameter IRT model takes account of items difficulty and item discrimination. However, guessing is assumed to be uniform across ability levels. The three-parameter model includes a guessing parameter in addition to item difficulty and discrimination. The models provide a mathematical equation for the relation of the response to ability levels. The equation for the three-parameter model is

$$P_{(\Theta)} = C + (1 - C) \left(\frac{1}{1+e^{-\alpha}(\Theta\text{-}b)} \right) \tag{6}$$

Where b is the difficulty parameter, α is the discrimination parameter, C is the guessing or pseudo-chance parameter and Θ is the ability level.

The basic idea in detecting DIF through IRT models is that if DIF is present in an item, the ICCs of the item for the reference and the focal groups should be different. However, where there is no DIF, the item parameters and hence ICCs should be almost the same. It is evident that the ICCs would be different if the item parameters vary from a group to another. Thus, one possible way of detecting DIF through IRT is to compare item parameters in two groups. If the item parameters are significantly differently, the DIF is ensured.

IRT DIF can be computed using BILOG-MG for dichotomously scored items, PARSCALE and MULTILOG for the polytomously scored items. In addition, for small sample sizes, non-parametric IRT can be employed using the Test Graft software. Pea in Karami (2012) undertook a DIF study of examinees with different academic backgrounds sitting the English subtest of the Korean National Entrance Exam for Colleges and Universities. He applied the three parameter IRT using the MULTILOG.

According to Osterlind and Everson in Oratokhai (2021), three major advantages of using IRT in DIF detection are:

- ✓ Compared to classical test theory, IRT parameter estimates are not as confounded by sample characteristics.
- ✓ Statistical properties of items can be expressed with greater precision which increases the interpretation accuracy of DIF between two groups.

- ✓ These statistical properties of items can be expressed graphically, improving interpretability and understanding of how items function differently between groups.

Lord's Chi-Square (Wald Test)

Lord (1977-1980) proposed a χ^2 statistic to test for DIF detection under IRT, usually called the Wald test. To identify DIF items, the Wald test is used to compare vectors of IRT item parameters between groups (Gao, 2019). Initially, Lord proposed the evaluation of DIF for the location (difficulty) parameters only:

$$Z_j = \frac{b_{Fj} - b_{Rj}}{\sqrt{\{Var(\hat{b}_F) + Var(\hat{b}_R)\}}} \qquad (7)$$

Where $\hat{b}_{Fj}$ and $\hat{b}_{Rj}$ are the maximum likelihood estimates of the parameter b_j for each focal group and reference group, and Var (b_{Fj}) and Var (b_{Rj}) are the corresponding estimates of the sampling variance of $\hat{b}_{Fj}$ and $\hat{b}_{Rj}$, respectively (Lee, 2015). To test a single parameter, b, the difference between the estimated bs across groups is compared to its standard error;

$$SE(b_F - b_R) = \sqrt{Var(b_F) + Var\ (b_R)} \qquad (8)$$

Z^2 is a chi-square distribution with df = 1 for large samples. Lord extended this test to a generalized test of the joint differences between the vectors of discrimination and difficulty parameters across focal and reference groups. The test statistic, χ^2 is

$$\chi_j^2 = Vj\sum^{-1}V_j \qquad (9)$$

for 2PL model, where V_j is [$\hat{a}_{Fj} - \hat{a}_{Rj}$, b_{Fj} - b_{Rj}], and let $\sum^{-1}$ define the corresponding variance –covariance matrix. For large samples, the χ_j test statistic follows a chi-square distribution with df = 2 (Lee, 2015). In general, the df are the number of parameters per item j, which is being tested for DIF.

Likelihood-ratio test method

The likelihood-ratio test is another IRT based method for assessing DIF. This procedure involves comparing the ratio of two models. In general, the procedure begins with an omnibus test that examinees whether any item parameter for a given item (e.g. item difficulty, item discrimination, or both in a two-parameter logistic model) differs between the reference and focal groups (Bulut & Suh, 2017). The steps for conducting an omnibus test are as follows: first a compact (C) model, where all item parameters are constrained to be equal across the reference and focal groups, is estimated. Second an augmented (A) model, where all parameters of the item are allowed to vary across the two groups, is estimated. To test whether the item exhibits DIF, the likelihood test statistic, which is -2 times the difference in log likelihood from the compact and augmented models, is computed as follows.

$$LR = -2InLc - (-2InL_A) \qquad (10)$$

Where Lc is the log likelihood of the compact model and L_A is the log likelihood of the augmented model. The LR statistic is approximately distributed as chi-square (χ^2) distribution with degrees of freedom (df) equal to the difference in the number of parameters estimates between the two models (Bulut & Suh, 2017).

If the likelihood statistic from the omnibus test is significant, then follow-up tests should be performed to identify the type of DIF (Woods in Bulut & Suh, 2017). In the subsequent analyses, the discrimination parameter of the item is constrained to be equal but the difficulty parameter of the item is estimated for the two groups separately. This process results in a second compact model that can be compared against the same augmented model from the omnibus test. A significant likelihood statistic from this comparison indicates that the item should be flagged for non-uniform DIF. If this LR statistic is not significant, then the item should be tested for uniform DIF (Bulut & Suh, 2017).

Suh and Cho (2014) extended the IRT-LR test for multidimensional IRT (MIRT) models. Similar to the IRT-LR test for unidimensional IRT models, the IRT-LR test for MIRT models depend on the evaluation of model fit by comparing nested models. The IRT-LR test for MIRT models requires additional assumptions. To ensure metric indeterminacy across multiple latent traits in their application, the means and variances of two talent traits are fixed at o and 1s, respectively, in both groups. Also, the correlation between the two latent traits is fixed at 0 in the augmented model with freely estimated item parameters across the two groups. In addition, the discrimination parameter for the first item is fixed at 0 on the second latent trait for both groups to satisfy the model identification when the test has a non-simple structure.

b-parameter method

The b-parameter is an item response theory (IRT)-based index of item difficulty. As IRT models have become an increasingly common way of modeling item response data, the b-parameter has become popular way of characterizing the difficulty of an individual item, as well as comparing the relative difficulty levels of different items.

The b-parameter is used to detect differential item functioning when 1PL model is used. This means that the a-parameters are constrained to be equal for both groups leaving only the estimation of the b-parameters. After examine the ICCs, there is an apparent difference in b-parameters for both groups (Oratokhai, 2021). Using a similar method to a student's t-test, the next step is to determine if the difference in difficulty is statistically significant, under the null hypothesis.

Ho: $b_r = b_f$

Lord in Oratokhai (2021) provided an easily computed and normally distributed test statistic

$$d = (b_r – b_f) / SE(b_r – b_f) \quad (11)$$

The standard error of the difference between b-parameters is calculated by

$\sqrt{[SE(b_r)]^2 + \sqrt{[SE(b_f)]^2}}$

Where

d = Estimated difficulty parameter difference

br = Estimated difficulty parameter for males (reference group)

bf = Estimated difficulty parameter for females (focal group)

Simultaneous item bias tests method (SIBTEST)

Simultaneous item bias test (SIBTEST) employs the non-parametric multidimensional DIF model of Shealy and Stout in Alordiah (2015), which looks at the differences in probability of correct responses between focal and reference groups (beta index), after matching respondents on true ability score. Previous DIF detection procedures focus on each item separately but with SIBTEST method, multiple items can be tested to detect the amount of DIF in the entire subtest. To operate SIBTEST on standardized achievement test, the test items are divided into studied (suspect) subtest and the matching (valid) subtest. The studied subtest is comprised of the items believed to measure the primary and secondary dimensions based on the substantive analysis in the first stage (comprised of the items in the test that are suspected to exhibit DIF), whereas the matching subtest contains the items believed to measure only the primary dimension. The matching (valid) subtest is used as the internal matching criterion to control for the group differences in the "target ability" that is intended to be measured by the test in the detection of DIF. That is, it is used to place individual in the focal and reference groups at each score level. The estimate of unidirectional SIBTEST DIF index given by Atar in Alordiah (2015) is

$$B_u = \Sigma P_{fk} (Y_{rk} – Y_{fk}) \quad (12)$$

Where

K = number of score levels on the valid subtest

B_u = maximum score level on the valid subtest

P_{fk} = proportion of the focal group examinees that obtain a valid subtest score of K

Y_{rk} = mean suspect subtest score for reference group

Y_{fk} = mean suspect score for focal group at the Kth valid subtest score level.

The null hypothesis of no unidirectional DIF is

Ho: $\beta_u = 0$

The SIBTEST test statistic associated with the null hypothesis is

$$SIBTESTu = \beta_u / (\sigma\beta_u) \quad (13)$$

Where $(\sigma\beta_u)$ is the estimated error for unidirectional SIBTEST DIF index, β.

Roussos and Stout in Alordiah (2015) classified the strength of DIF as:

1) A- level DIF: the absolute value of beta index is less than 0.059

2) B- level DIF: the absolute value of beta index is between

3) C- Level DIF: the absolute value of beta index is equal or higher than 0.088. The A, B and C level of DIF is also categorized as negligible, moderate respectively. Sometimes group differences in the ability distribution might le estimate of the SIBTEST DIF index indicating the presence of DIF when there is no DIF. In this case, regression correction is used to compute an unbiased estimate of the SIBTEST DIF index.

Rasch model

The Rasch model focuses on the probability of endorsing item i by person m. In aiming to model this probability, it essentially takes into account person ability and item difficulty. Probability is a function of the difference between person ability and item difficulty. The following formula shows just this:

$$P\ (x = 1/\ \theta, \delta) = f\ (\theta n - \delta i) \qquad (14)$$

where θn is person ability and δi is item difficulty.

The formula simply states that the probability of endorsing the item is a function of the difference between person ability, θn, and item difficulty, δi. This is possible because item difficulty and person ability are on the same scale in the Rasch model. It is also intuitively appealing to conceive of probability in such terms. The Rasch model assumes that any person taking the test has an amount of the construct gauged by the test and that any item also shows an amount of the construct. These values work in the opposite direction. Thus, it is the difference between item difficulty and person ability that counts. The exact formula for the Rasch model is given as:

$$\text{Ln}\ \{P_{ni}\ /1 - P_{ni}\} = \theta_n - \delta_i \qquad (15)$$

The Rasch model provides us with sample independent item difficulty indices. Therefore, DIF occurs when invariance is not accrued in a particular application of the model (Karami, 2012). That is, the indices are dependent on the sample who takes the test. The amount of DIF is calculated by a separate calibration t-test approach first proposed by Wright and Stone (Smith in Karami, 2012). The formula is given below:

$$t = d_{i2} - d_{i1}/\sqrt{(S^2 - S^2})\qquad (16)$$

where d_{i1} is the difficulty of item i in calibration on group 1, d_{i2} is the difficulty of item i in calibration based on group 2, S^2 is the standard error of estimate for d_{i1}, and S^2 is the standard error of estimate for d_{i2}. Among the software for DIF using the Rasch model are ConQuest, Winsteps and Facets.

Sample Size and DIF Detection

Many DIF detection methods are based on inferential statistics, thus sample size can impact the statistical power of DIF detection. When small sample sizes are confronted in real testing situations, statistical power can be compromised in a way that the inferences made in DIF analyses could be unstable. Sufficient sample sizes are needed in both focal and reference groups when conducting DIF analysis to have enough power for DIF detection across groups. Sample sizes of 200 to 250 per group will likely have enough power to detect DIF using non-IRT methods (Liu, 2017). IRT-based methods for detecting DIF generally require larger sample sizes in order to estimate model parameters for both the reference and focal groups.

A substantial body of research has examined the impact of sample size on DIF detection using a variety of DIF detection approaches (Bernstein, Samuels, Woo & Hagge, 2013). There is wide agreement that DIF detection rates typically improve with larger sample sizes.

Wise cited in Liu (2017) conducted a simulation study using the M-H DIF detection model with two different sample sizes, 400 and 800. He found that the power of the M-H method increased substantially as sample size increased. It was also noted that detection rates were best with items with moderately high discrimination as opposed to items with lower or higher discrimination values.

Conclusion

Differential item functioning (DIF) ensues when items that are supposed to measure a latent trait are unfair, favouring one group of individuals over another. Hence, it is normal to assume that there are one or more variables, in addition to the target ability, that account for or explain a group difference in item performance. DIF analysis aims to detect items that differentially favour candidates of the same ability levels but from different groups. In educational measurement and quantitative psychology, the absence of DIF is regarded as an important aspect of test fairness by educational and psychological researchers. Detecting items having DIF in a test is important, as it helps in maintaining the tests' fairness and validity.

In terms of DIF detection methods, it is significant that all DIF detection methods available are designed to match the groups, either directly or indirectly, on the proficiency measured by the items (Angoff, 1993), and all DIF measurements investigate how different groups perform on individual test items to determine whether the test items are creating problems for a particular group (Zumbo, 1999). DIF detection methods are all based on the philosophy that if different groups of examinees (like, males versus females) have approximately the same level of ability, they should perform similarly on individual test items regardless of group membership.

References

Alordiah, R. P. (2015). *Comparison of index of differential item functioning under the methods of item response theory and classical test theory in mathematics*. Unpublished Ph.D. Dissertation, Department of Guidance and Counseling, Delta State University, Abraka.

Angoff, W. H. (1993) Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. New York: Routledge CRC Press.

Annan-Brew, R. (2020). *Differential item functioning of West African senior school certificate examination core subjects in Southern Ghana*. Unpublished Ph.D. Dissertation, Department of Education and Psychology, University of Cape Coast, Ghana.

Ashraf, Z. A. & Jaseem, K. (2020). Classical and modem methods in item analysis of test tools. *International Journal of Research and Review, 7*(5), 397-403.

Bernstein, I., Samuels, E., Woo, A. & Hagge, S. L. (2013). Assessing DIF among small samples with separate calibration t and Mantel-Haenszel χ^2 statistic in the Rasch model. *Journal of Applied Measurement, 14*(4), 389 – 399.

Bulut, O. & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators, multiple causes model, the item response theory likelihood ratio test and logistic regression. *Frontiers in Education, 2*(51), 43 – 54.

Chen, H. & Jin, K. (2018). Applying logistic regression to detect differential item functioning in multidimensional data. *Journal of Frontiers in Psychology, 9*(1), 48-57.

Davidson, M. J.; Ko, A. J.; Wortzman, B. & Li, M. (2021). Investigating item bias in a CS1 exam with differential item functioning. *SIGCSE 21*, Virtual Event, USA, 13th 20th March 2021.

Dodeen, H. (2014). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education*, *72*, 181-193.

Dodeen, H. & Johnson, R. R. (2013). On the consistency of individual classification using short scales. *Psychological Methods, 12*(1), 105-120.

Gao, X. (2019). *A comparison of six DIF detection methods.* Unpublished Master's Thesis, Department of Educational Psychology, University of Connecticut, USA.

Ibrahim, A. (2017). Relative effectiveness of generalized Mantel-Haenszel, simultaneous item bias test and logistic discriminate function analysis for detecting differential item functioning in ordinal test items. *Ife Psychologia (An International Journal), 25*(1), 104-132.

Karami, H. & Nodoushan, M. A. S. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies (IJLS), 5*(3), 133-142.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59-76.

Kamata, A. & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. Learning Disabilities*: A Contemporary Journal, 2*, 49 - 69.

Krishnan, V. (2013). The early child development instrument: An item analysis using classical test theory on Alberta's data. *Early Child Development Mapping Project (ECMap) Alberta,* Community- University Partnership, Faculty of Extension, University of Alberta, Canada.

Lee, S. Y. (2015). *Lord's Wald test for detecting DIF in multidimensional IRT models: A comparison of two estimation approaches.* Unpublished Dissertation, Department of Foundations of Education, the State University of New Jersey.

Liu, M. (2017). Differential item functioning in large scale mathematics assessment comparing the capability of the Rasch Tree Model to Traditional Approaches. Unpublished Ph.D Dissertation, Department of fundations of Education, University of Toledo.

Liu, R. & Bradley, K. D. (2021). Differential item functioning among English language learners on a large-scale mathematic assessment. *Front Psychol, 12*(1), 61-75.

Oratokhai, D. I. (2021). *Investigating differential item functioning in National Business and Technical Examination English language multiple choice test items.* Unpublished Ph.D. research seminar, Department of Educational Evaluation and Counseling Psychology, University of Benin, Benin City.

Sapmaz, Z. M. (2019). *Detection of gender-related differential item functioning (DIF) in the mathematics subjects in Turkey.* Unpublished Masters' Thesis, Department of Educational and Psychological Studies, University of South Florida.

Suh, Y. & Cho, S. J. (2014). Chi-square difference test for detecting differential functioning in a multidimensional IRT Model: A Monte Carlo study. *Applied Psychological Measurement, 33*, 184-199.

Wang, V. (2018). *Handbook of research on program development and assessment methodologies in K-20 education*. Chicago: IGI Global.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.