The Relationship between Differential Distractor Functioning (DDF) and Differential Item Functioning (DIF): If DDF Occurs, Must DIF Occur?

Jiayi Deng

Educational Psychology and Research

University of Kansas

Submitted to the graduate degree program in Educational Psychology and Research and to the Faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the master's degree

Thesis Committee:

Chair: Bruce Frey, Ph. D

Neal Kingston, Ph. D

Vicki Peyton, Ph. D

Date of Defense Meeting, April 19th 2020

The thesis committee for Scholar Jiayi Deng

certifies that this is the approved version of the following these:

The Relationship between Differential Distractor Functioning (DDF) and Differential Item Functioning (DIF): If DDF Occurs, Must DIF Occur?

Chair: Bruce Frey, Ph. D

Date Approved: 19 April 2020

Abstract

Differential Distractor Functioning (DDF) and Differential Item Functioning (DIF) are two critical ways for detecting potential test fairness issues. The current study aims to illustrate the relationship between the existence of DDF and that of DIF by using multiple-choice items in PIRL2016 achievement test. Multinomial logistic regression and binary logistic regression were used for DDF and DIF detection, respectively. Correlation test and binomial test were used to explore the relationship between DDF and DIF. The results show no relationship between DDF and DIF was detected. In addition, there was no evidence for the association between DIF and DDF detection and the test content.

Table of Contents

List of Tables

Introduction

Test fairness is a popular topic nowadays and attracts a lot of attention. In *2014 Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), which is the latest version, test fairness has been firstly added as one of the critical test foundations in addition to reliability and validity. One way to identify test fairness issues was the absence of measurement bias, which occurs when some unrelated factors (e.g. gender, cultural background and examinees cognitive health), other than what test intends to measure, have influence on the test results. In this case, the items with potential measurement bias favors some subgroups of test takers with particular identifications, causing the test scores mean differently to examinees from different subgroups.

Differential item functioning (DIF) and differential distractor functioning (DDF) are two widely used approach to flag items with potential psychometric bias. Though the relationship between DIF and DDF has been studied before, they all automatically considered the association as cause-and-effect and conducted the analysis of DDF based on the results of DIF analysis. Very scant literatures truly focused on their relationship, or the possibility of independent relationship, between DIF and DDF (e.g. whether DDF could existed without the occurrence of DIF). This study explored the relationship between DDF and DIF, which may affect the item analysis and distractor analysis model design in future investigations.

Literature Review

Differential Item Functioning

Differential Item Functioning (DIF) occurs when test takers with equivalent ability show unexpected different performance in answering the test items (Dorans & Schimitt, 1991; Dorans & Holland, 1992; Zumbo, 1999; Mapuranga et al., 2008). In other words, after controlling for examinees proficiency in test domain, if the test takers with different characteristics answer the item correctly in unexpected different rates, the item shows DIF, representing that there are irrelevant factors impacting test scores. DIF analysis has been regarded as an important approach to investigate test fairness (Martinková et al., 2017). In previous DIF studies, examinees' ability was controlled as matching variables, which was utilized to build the comparison of examinees' performance between different subgroups. It was mainly represented by two types of scores: observed test scores or unobserved latent score. There were usually two or more subgroups (e.g. male and female, students with and without disability, and the Black and the White) in DIF analysis, which compose the grouping variable (e.g. gender, cognitive health status and culture), in which focal group and reference group were usually used. The focal group was legally protected group and was selected to investigate if it was disadvantaged by the item, while the reference group was corresponding comparison group. Based on the relationship between the grouping variable and the matching variable in DIF analysis, two different types of DIF were identified: uniform DIF and non-uniform DIF. When the response pattern of an item varies across subgroups with equivalent ability, if the different pattern is constant in each level of matching variable, the item shows uniform DIF; if the different pattern also varies across matching variable groups, it demonstrates non-uniform DIF (Mellenberg, 1982). In terms of statistical relationship, non-uniform DIF occurs when the interaction between the grouping variable and the matching group exists (Swaminathan & Rogers, 1990).

Methods for DIF Detection

According to the classification of DIF analysis approaches conducted by Mapuranga, Dorans and Middleton in 2008, there are four main methods implementing to detect DIF: (1) expected item score methods, (2) nonparametric odds ratio methods, (3) generalized linear model methods, and (4) IRT-based methods. Each of them makes different assumptions of null DIF and builds various models for DIF detection.

Expected Item Score Methods. In expected item score methods, the comparison of proportions of correct responses over total number of responses was conducted at each level of matching variable for each subgroup of grouping variable. The assumption of null DIF is that the correct proportions are equal at significant level across each level of matching variable. Thus, DIF would be detected if the difference of proportions between subgroups existed, representing the expected performance in this item was different for subgroups after accounting for their ability.

Standardization method is an example located in this category. It was developed by Dorans and Kulick (1983) in DIF investigation for SAT items operated in 1977 based on gender. Conditional correct response probability was utilized for DIF analysis, in which the condition meant for controlling the score level (test takers ability). Root Mean Weighted Squared Difference (RMWED) was calculated for each item to serve as an index for unexpected performance difference. The rationale behind is that taking the conditional probability of responding correctly to the item of base (i.e. reference) group as the expected conditional probability of correct response to the item of study (i.e. focal) group. The difference (D_f) between the observed conditional probability of successful performance in focal group and the expected conditional probability of correct response for focal group (i.e. observed probability of correct response in reference group) was calculated and weighted for a final standardized $RMWSD_f$ (the subscript f represents focal group). Null DIF was assumed to exist when the D_f equaled to 0, representing no difference between observed and expected probability of correct response for the subpopulation existed given the examinees' ability. In this study, unexpected performance difference was identified for several items among which one item was out of acceptable range and contributed to the test bias. Standardization was also used to examine DIF based on ethnicity for SAT-verbal tests (Schmitt & Bleistein, 1987), in which the Black students were found to have worse performance than the White students at the same ability level.

Nonparametric odds ratio methods. Mantel-Haenszel (MH) procedure is a representative approach of nonparametric odds ratio methods and has been extended and developed for various analysis (Clauser & Mazor, 1998). It was developed by Mantel and Haenszel in 1959 and then used as a DIF detection procedure by Holland and Thayer in 1988. The odds ratio is employed as the index of measurement bias of this item. It is calculated through using a 2x2 contingency table, in which two rows represent grouping variable (e.g. male and female) while two columns are used for outcome variable (i.e. correct response and incorrect response). In using MH method, the matching condition (e.g. ability) is required to be coded into a categorical variable. For instance, if the total score is used as the proxy of matching variable (i.e.examinees' ability), it should be coded into several levels, such as high, medium and low ability level. In order to control the ability being measured by the test (i.e. matching variable), a 2x2 contingency table would be made for each matched level of it. An odds ratio is calculated across the contingency table and is used for DIF detection.

The null DIF assumption is that the odds ratio equal to one, representing the item has no preference to any group given examinees' ability. DIF was detected based on the comparison of odds ratio of preference frequencies between reference group and focal group, after controlling for the matching variable. When the odds ratio is smaller than one, it represents that the item prefers the reference group, and when the odds ratio is larger than one, it indicated that the item favors the focal group. Due to the asymmetry property (Clauser & Mazor, 1998) and the indirect meaning (Stoneberg, 2004) of the results, ETS developed it through the transformation into a delta metric (Zieky, 1993), which has used as an index to identify the magnitude of DIF (Monahan et al., 2007; Zwick, 2012). MH has been widely implemented for its inexpensive and simple operations, and application of chi-square test as significance test (Holland & Thayer, 1988). Besides, another advantage of MH method over the Standardization method is the utilization of optimal weight with stronger statistical power (Dorans, 1989). However, the limitation of MH method is that it is unable to identify uniform and non-uniform DIF (Swaminathan & Rogers, 1990) since the statistic interaction relationship is not allowed.

Generalized Linear Model Methods. In utilizing generalized linear model methods, unlike two method categories above, the assumption of null DIF is met when the grouping variable does not serve as an efficient predictor to the outcome variable. DIF is detected if the grouping variable can be used to predict the probability of selecting the right answer, after accounting for the matching variable.

Logistic regression model, created by Swaminathan and Rogers (1990), is one of the most widely used methods which located in this category. In DIF analysis built upon the logistic regression model, for any dichotomous item, the probability of test takers' selection of correct answer is the dependent variable while the matching standards (e.g. observed total scores), groups of interests (e.g. male and female), and statistical interaction relationship between them are entered as independent variables. In other words, logistic regression model is used to predict probability of responding correctly to a dichotomous item by the matching variable, the grouping variable and their interaction. The detection of DIF was determined upon the significance of contribution made by grouping variable. Meanwhile, the significance of contribution made by interaction between matching variable and grouping variable can be used in addition to the significance of contribution made by grouping variable to distinguish uniform or non-uniform DIF (Clauser & Mazor, 1998). The magnitude of DIF can also be described by logistic regression model through the size of variance in probability responding correctly that can be explained by the grouping variable. Logistic regression model has been recommended as having best effect (Zumbo, 1999) for investigating DIF for direct utilization of continuous matching condition (e.g. observed total scores) and the ability to identify uniform and non-uniform DIF.

IRT-based methods. IRT-based methods utilize estimation of latent ability as matching variable and implement item characteristic curves (ICCs) to represent the item functioning, expressing the probability of responding correctly by examinees estimated latent ability. There are three critical parameters in IRT-based methods that can be identified in ICCs and were widely used for DIF detection. The parameter *a*, expressed by the slop of the ICC, represents the discrimination ability of the item. The parameter *b*, the interception of the ICC, indicates the item difficulty. The parameter *c* is the lower asymptote of ICC, representing the probability the test taker with extremely low ability is able to respond correctly. The model of IRT-based methods can be conducted by the comparison of the areas between ICCs for reference group and focal group (Wang, & Su, 2004) or the investigation of one or more parameter differences between two groups (Gómez-Benito & Navas-Ara, 2000). The assumption for null DIF is that the ICCs are exactly the same for reference group and focal group. If the ICCs of one item are not same for different groups (Clauser & Mazor, 1998), in other words, if there is at least one parameter different for focal group and reference group, this item shows DIF. There are evidences that IRT-based methods are effective and accurate in various conditions (Startk, Chernyshenko & Drasgow, 2006) though they are more complicated for implementation than other methods in detecting DIF.

Differential Distractor Functioning

Extended from DIF, Differential Distractor Functioning (DDF) investigate the item invariance of incorrect options (i.e. distractors) (Green, Crone & Folk, 1989). Unlike the DIF, whose attention was put on the probability of successful performance, DDF only investigate incorrect responses. It occurs when the test participants with equivalent ability respond to distractors differently. Distractors are of great importance for multiple-choice items thought it cannot provide rewards for examinees. They will distract test takers unequally from the correct answer if they function differently for the subgroups. Therefore, they are also critical for ensuring the fairness of a test item. Though DDF analysis shifts focus from the successful response versus unsuccessful responses to distractors versus each other, a great deal of components in DDF analysis is kept the same with that in DIF analysis. In DDF analysis, the influence of examinees' ability on their performance is controlled by the matching variable, while the groups of interests (e.g. gender, ethnicity, etc.) are regarded as the grouping variable. Determined by the relationship between the matching variable and the grouping variable, there are also uniform DDF, occurring when the performance patterns of each distractor varies for different subpopulation but is constant in every levels in matching variable (i.e. the matching variable and grouping variable are independent), and non-uniform DDF, referring to the different patterns of distractor selection in the subgroups also vary across each level in matching variable (i.e. there are interaction existed between the matching variable and group variable).

Methods for DDF Detection

Most methods used in DIF detection can be developed or adjusted for the investigation of distractors. Koon (2010) extended the classification of DIF detection approach (Mapuranga et al., 2008) for DDF study. Expected item score methods examine the proportion of each distractor response and assumes null DDF as all distractor selection proportions are equal given the examinees' ability. Nonparametric odds ratio methods utilize the odds of a distractor to the reference distractor as the DDF index given ability level and provide the null hypothesis that each incorrect option attracts examinees equally (i.e. the odds ratio is 1). Generalized linear methods make null DDF assumption that the performance regarding distractors cannot be predicted by examinees' subgroup (grouping variable) when the ability (matching variable) is controlled.

Green et al. (1989) conducted a distractor analysis for SAT Verbal test to study whether the distractors in multiple-choice items held different attractiveness to Black, Hispanic and White students. They utilized log-linear method for investigating the impact of ethnicity on respondents' performance on distractors, in which investigators made use of three-way contingency tables with distractor selection, the ethnic subgroups and ability levels. The matching variable, which was represented by observed total score, instead of continuous proxy, was grouped into categorical levels. The frequency related to the combination of the distractor selection, ethnicity subgroups and examinees ability levels was produced for each cell and was entered into the model as outcome variable. Log-linear method was favored for its simple operation and its allowance for both main effect and interaction effect. However, the usage of log-linear method for investigation of DDF cannot be conducted at the option level so that the particular distractor that functions differently across subgroups cannot be identified.

Abedi et al. (2008) examined the distractors fairness of Stanford Achievement Test (Stanford 9) for students with and without disabilities through analysis of DDF, in which logistic regression method was utilized. As mentioned before, the outcome variable in logistic regression should be binary. Thus, the author grouped three distractors in each multiple-choice item into two categories: most commonly selected distractor and the other two less popular distractors. Students cognitive status (with or without disabilities) was the main interests of this study and was treated as the grouping variable. Students' ability was represented by the standardized total score in the test. In order to take control of the ability before investigating the existence of distractor preference, the ability proxy was entered into the multi-step logistic regression before the grouping variable and the interaction effect. Three models, with ability proxy only, with ability proxy and grouping variable, and with the interaction effect as an additional variable, were conducted and compared by goodness-of-fit of the model. The distractors were finally found to be more attractive to students with disabilities in Grade 9. In particular, the students with and without disabilities were attracted by different distractors. The logistic regression utilized in this study held the advantage over the log-linear model. It is able to provide information in option level. However, due to the limitation of outcome variable in logistic regression that it must be binary, the distractor functioning analysis failed to be specified to each single distractor.

One way for solving this limitation is multinomial logistic regression because it allows the outcome variable to contain more than two categories. Kato, Moen and Thurlow (2009) applied multi-step multinomial logistic regression in the study of item bias for two statewide reading instruments and realized the analysis for each single option. In their analysis, the correct and incorrect options were entered into the model as the outcome variable for DIF and DDF detection simultaneously. Standardized test total scores, students' cognitive status (with and without disabilities) and their interaction were entered into the model step by step so three models were created for each multiple-choice item. The significant increase in model fit was used to flag items with potential DIF or DDF. The flagged items were further analyzed by mean absolute difference of response characteristic curves to identify DDF or DIF. Multinomial logistic regression has been preferred in distractor analysis for its permission of analysis at option level, which is able to provide specific information regarding measurement bias of each option.

Relationship between DDF and DIF in Previous Studies

DDF has been considered as a cause of DIF and thus it is used to explore the exists of DIF. There do have studies hold the notion that DDF analysis supplements and exemplifies the DIF findings (Middleton & Laitusis, 2007; Park, 2007; Mapuranga et al., 2008; Penfield, 2008; Suh & Bolt, 2011) so that they only conduct analyses of DDF for items that showed DIF (Middleton & Laitusis, 2007; Tsaousis, Sideridis & Al-Saawi, 2008). However, it is possible for items to demonstrate DDF without displaying DIF. In this case, a distractor or distractors of an item may show preference to a particular subgroup of population, but it does not affect the subgroup's ability to response correctly overall. As a result, items which function differently across different subgroups may be ignored if DDF analysis is only built for items demonstrating DIF. Moreover, Kato et al. (2009) studied DDF and DIF at the same time and utilized largest Mean Absolute Difference (MAD) of response characteristic curves as the evidence for detection of DIF and DDF. To some degree, they ignored the situation in which an item had DDF and DIF at the same time. In other words, they did not discuss the situation in which the difference of the MAD of a distractor and the MAD or the correct answer was very small. There was no criterion discussed in the study for the difference of MAD of RCC curve to detect only DIF or DDF, or both.

Present Study and Significance

In present study, generalized linear model was implemented for DIF and DDF analysis. To be specific, binary logistic regression and multinomial logistic regression were implemented to detect DIF and DDF, respectively, for PIRLS achievement assessment. In terms of statistical perspectives, first of all, logistic regression function was not developed from any test theory, but it could be connected to item response test theory and classical test theory equally (Mapuranga et al., 2008), which made the selection of matching variable—test takers' ability proxy—less limited. Either observed test scores or latent (unobservable) ability scores can be used as the index of examinees' ability in logistic regression model. Besides, statistical significance of variable membership in logistic regression model can be provided by Chi-square test of significance in likelihood ratio test. Moreover, the results of logistic regression supply the statistic relationship between matching variable and grouping variable, making it powerful to distinguish between uniform and non-uniform DIF and DDF (Swaminathan & Rogers, 1990). With regarding to practical perspectives, binary logistic regression and multinomial logistic regression are both relatively simple because no complicated assumption will be required. Meanwhile, the manipulation procedure is also be easily completed by SPSS software.

The results of this study could provide clear instruction for item development and revision. If an item showed DIF, but there are no distractors functions differently among the groups with equivalent ability, there is sufficient evidence to indicate that only the correct answer should be revised because it is the only option of this item functioning differently among these groups. If an item showed DDF without DIF, that means there exists at least one incorrect option that functions differently among groups with equivalent ability. If this occurs, there is sufficient evidence for suggesting further studies to identify the specific distractor that has different function among the groups and make modification to it.

Statement of Problem

The primary purpose of this study was to investigate the relationship between differential item analysis (DIF) and differential distractor analysis (DDF). To be more specific, whether the DDF must indicate the existence of DIF of that item or not. In the other words, whether the incorrect option(s) with differential functioning across groups with various characteristics must indicate the existence of the correct option's preference to specific group.

Therefore, the following research question guided the conduction and report of this study: if distractors (i.e. incorrect options) are chosen differently, will the right answer be chosen differently?

Methodology

Instruments

Data from achievement instrument in Progress in International Reading Literacy Study (PIRLS) in 2016 was used in this study. There were 16 booklets, containing multiple-choice items and constructed response items, in PIRLS 2016. As for statistical properties, PIRLS 2016 demonstrated satisfactory reliability. The median of the reliability coefficient for all countries was 0.83 (Foy, Martin, Mullis & Lin, 2017). For the two countries, United States and Macau SAR (Special Administration Region), whose data were used in this study, the reliability coefficients were above the median, with Cronbach alpha 0.90 and 0.87, respectively (Foy, Martin, Mullis & Lin, 2017). All data were open and available for public use in IEA (International Association for the Evaluation of Educational Achievement) PIRLS International Study Center Website. Only the 86 multiple-choice items were used in DDF and DIF analysis in present study.

Participants

In order to ensure there would be enough sample size for each item to conduct both DDF and DIF study, two ethnicity groups were selected from the middle ability group (i.e. achievement test scores locate between intermediate benchmark and high benchmark). In this study, two cultural groups, the U.S. and Macau SAR, were selected. They located at the same ability level (medium level) in PRILS 2016 and can represent two different culture, American and Chinese culture, respectively. There were 4425 U.S. participants, including 2217 girls (50.1%) and 2208 boys (49.9%), and they took English version achievement instrument, while there were 4059 Macau SAR students, containing 1990 girls (49%) and 2069 (51%), and they took traditional Chinese, English or Portuguese version. The total number of participants for present study was 8484, among which 4270 were girls（49.6%) and 4277 were boys (50.4%). All these students were at fourth schooling year, which was fourth grade in both U.S. and Macau SAR.

Data Analysis Procedure

In this study, there were two steps. In the first step, DDF and DIF analysis was conducted to count the DIF and DDF magnitude and to flag DIF and DDF items. In the second step, two different analysis were conducted based on the results in step one to reach the answer for proposed research question.

Step 1: DDF and DIF Analysis

Generalized linear models—multinomial logistic regression and logistic regression—were implemented in DDF and DIF detection respectively, which were described in detail below. There are two critical assumptions for logistic regression and multinomial logistic regression: non-multicollinearity and independence among the dependent variable options (Starkweather &

Moske, 2011). The assumption of non-multicollinearity violation occurs when the independent variables are highly correlated with each other, while the assumption of independence among the outcome variable options gets violated when the categories in outcome variable highly correlate with each other, representing that the occurrence of one impacts the existence of others. The former assumption was tested by variance inflation factor (VIF) of two independent variables when putting them into regression model but excluding any interactions, and there was sufficient evidence to conclude that assumption of multicollinearity was not violated since VIF value for both two independent variables were smaller than 2 and very close to 1. The second assumption was not violated either since each test takers can only select one options from all given choice, which meant the selection of one option is independent of the selection of any others. Multinomial logistic regression and logistic regression were implemented, in DDF and DIF analysis respectively, for all multiple-choice questions (N = 86) in achievement instrument of PIRLS 2016 data.

Multinomial logistic regression was used for DDF analysis because the dependent variable included three categories. After the DDF analysis (i.e. the first step), all multiple-choice items were divided into two categories based on the results of DDF analysis: items showing DDF and the items not displaying DDF. In DIF analysis, since the independent variable was dichotomous, binary logistic regression was employed for all 86 multiple-choice items. After that, all items were divided into two groups: items with DIF and items without DIF. All analysis mentioned above was conducted by SPSS software version 26.

DDF Analysis

In DDF analysis, the purpose was the detection of DDF with the magnitude for all items. As a result, (1) items showing uniform DDF or non-uniform DDF were all categorized into the group of items that demonstrate DDF; (2) all correct answers for each item were coded as 0 and then be coded as missing data because only distractors (i.e. incorrect answers) were included in this step. DDF was investigated through multinomial logistic regression, in which the incorrect options were the dependent variable while ability proxy and culture were independent variables in the model. Taken the correct answer out, since all multiple-choice question in PIRLS 2016 achievement test had four options, there were three distractors for each item, thus, the dependent variable (i.e. distractors) was a categorical variable and there were three different categories of it. For example, if the key of one item is B, the distractor should be A, C, and D. Ability proxy was represented by the total score of all 86 multiple-choice items. The studied item was included in the matching criterion for the analysis of this item to reduce the bias in the estimation of DIF (Holland & Thayer, 1988; Tan, Xiang, Dorans & Qu, 2010). The other independent variable in this study was the culture where students were educated and took the PIRLS test in 2016. The U.S. and Macau SAR were selected to represent two cultural groups. Therefore, independent variable culture was a categorical variable with two different categories. The U.S. was coded as 0 and Macau SAR was coded as 1.

The distractor of each multiple-choice items was analyzed with the following multinomial logistic regression model:

For the kth item:

Model 1: $Y_i = \log\frac{P(Y=i \mid X_1)}{P(Y=j \mid X_1)} = a_i + b_{i1}X_1$ ……………………………………….(1)

Model 2: $Y_i = \log\frac{P(Y=i \mid X_1, X_2)}{P(Y=j \mid X_1, X_2)} = a_i + b_{i1}G_1 + b_{i2}X_2$ …………………………...(2)

Model 3: $Y_i = \log\frac{P(Y=i \mid X_1, X_2)}{P(Y=j \mid X_1, X_2)} = a_i + b_{i1}G_1 + b_{i2}X_2 + b_{i3}(X_1 * X_2)$ …………(3)

Where:

k = 1…k denotes the specific item analyzed;

i = 1…j denotes the distractor categories;

j denotes the reference (base) level (category);

X_1 represents the ability proxy;

X_2 represents the culture;

$X_1 * X_2$ represent the interaction between ability proxy and culture;

Y is the logarithm of the odds ratio (i.e. the probability of selecting ith distractor over the probability of selecting jth distractor).

For the kth item, there were three models in DDF analysis. Only ability proxy was entered as independent variable in model 1 based on the equation (1). This model demonstrated whether ability proxy was an effective predictor to the students' choice of distractors. In model 2, which based on equation (2), an additional variable, culture, was entered. This model was used to identify the existence of uniform DDF. The interaction between culture and ability proxy was entered into the last model, which based on equation (3), to investigate the existence of non-uniform DDF.

Based on the research question of this proposed study, DDF analysis was conducted on item level. In the other words, DDF was identified for each item but not for each single option. As a result, instead of testing which distractor showed DDF, the focus of this study was whether DDF occurred in each item. Items demonstrating DDF, no matter it was uniform or non-uniform DDF, were considered as DDF items.

DDF analysis was conducted on the two comparisons of the goodness-of-fit (i.e. model fit) between models. It was examined by likelihood ratio tests in which -2 log likelihood was compared between models by chi-squared test. The existence of non-uniform DDF was determined by the comparison of model fit statistics between model 2 and model 3. If the mode fit improved significantly from model 2 to model 3 (i.e. after adding the variable of the interaction between ability proxy and culture) at 0.01 level of significance and the Nagelkerke R^2 change was larger than 0.003 (Abedi et al., 2008), there would be sufficient evidence to conclude the interaction (i.e. non-uniform DDF variable) made significant contribution to model 3 and the item would be considered as demonstrating non-uniform DDF. The occurrence of uniform DDF was detected by comparing the model fit statistics of model 1 and model 2. If there was significant improvement of model fit from model 1 to model 2 at 0.01 level of significance and the Nagelkerke R^2 change was larger than 0.003 (Abedi et al., 2008), the item would be considered as exhibiting uniform-DDF because entering the variable culture significantly changed the mode fit (i.e. variable culture makes significant contribution to this model). The item was considered as not showing DDF if either none of the model fit comparison above showed significant improvement or Nagelkerke R^2 change was not reach 0.003. All items flagged as DDF item, no matter uniform or non-uniform DDF, were considered as items demonstrating DDF in this study. For flagged DDF items, the model fit change, which will be represented by Nagelkerke R^2 difference between model 3 and model 1, was considered as the DDF magnitude. Even though the DDF magnitude for uniform DDF items should be Nagelkerke R^2 difference between model 2 and model 1, there were two considerations: (1) in order to ensure the consistency of calculating DDF count for all items; and (2) for uniform DDF items, the Nagelkerke R^2 change from model 2 to model 3 was very tiny, which made little impact on total DDF count.

DIF Analysis

In DIF analysis, the aim of this step was the detection of DIF for all multiple-choice items. Thus, (1) items showing uniform DIF or non-uniform DIF will all be flagged as items that exhibit DIF; (2) all options, including correct answers and incorrect responses, were included in this step. Each item was coded as dichotomous, whose response was either correct or incorrect. All correct answers for each item were coded as 1 while all incorrect answers (distractors) were coded as 0. DIF was studied through binary logistic regression model, in which the result of a student's response to the item (i.e. right or wrong) was dependent variable while ability proxy and culture were independent variables in the model. Thus, the dependent variable was a dichotomous categorical variable. Two independent variable, ability proxy and culture, kept the same with the first step to ensure the results of DIF analysis were comparable with that of DDF analysis.

The DIF analysis was conducted and whether students' responses correct or not was analyzed with the following binary logistic regression models:

For the kth item:

Model 1: $Y = \log\frac{P(Y=1 \mid X_1\ )}{P(Y=0 \mid X_1)} = a_i + \ b_{i1}X_1$……………………………………..(4)

Model 2: $Y = \log\frac{P(Y=1 \mid X_1,X_2)}{P(Y=0 \mid X_1,X_2)} = a_i + \ b_{i1}X_1 + b_{i2}X_2$ …………………………...(5)

Model 3: $Y = \log\frac{P(Y=1 \mid X_1,X_2)}{P(Y=0 \mid X_1,X_2)} = a_i + \ b_{i1}X_1 + b_{i2}X_2 + b_{i3}(X_1 * \ X_2)$ …………(6)

Where:

k = 1…k denotes the specific item analyzed

X_1 represents the ability proxy;

X_2 represents the culture;

$X_1 * X_2$ represent the interaction between ability proxy and culture;

Y is the logarithm of the odds ratio (i.e. the probability of selecting correct answer over the probability of selecting distractors).

Ability proxy was the only independent variable to be entered into the first model in DIF analysis, which was based on the equation (4). This model exhibited if ability was a significant predictor for that whether students gave correct or incorrect response to the item. In the second model, which is conducted on equation (5), cultural groups were entered in addition to ability proxy to investigate whether it made significant contribution to the model. It was used to test the existence of uniform DIF. In the last model, interaction effect was entered over ability proxy and culture to identify the occurrence of non-uniform DIF.

Chi-square test with degree of freedom of two was utilized for likelihood ratio test, in which the statistical significance of the difference of the -2 times of log likelihood between model 3 and model 1 was tested. This was the simultaneous test for uniform and non-uniform DDF (Zumbo, 1999). R square change also demonstrated the existence of DIF. Nagelkerke R square, which covered full range from 0 to 1, was selected as pseudo R square in this proposed study. An item was flagged as demonstrating DIF if two followings occurred simultaneously: Nagelkerke R square change was greater than 0.003 and the likelihood ratio test was significant at 0.01 level of significance (Abedi et al., 2008). The item, in which Nagelkerke R square change was observed as significant between model 2 and model 3, was flagged as non-uniform DIF. When Nagelkerke R square change was significant between model 1 and model 2, these items were flagged as uniform DIF. Both items flagged as uniform and non-uniform DIF were considered as showing DIF. Thus, all item from DDF analysis in step one were categorized into two groups: DIF items and non-DIF items. For flagged DIF items, the model fit change

(Nagelkerke R square) between mode 3 and model 1, was considered as the DIF magnitude. The consideration was the same as mentioned in DDF analysis.

Step 2: Investigation of The Relationship between DIF and DDF

In this step, two analysis were developed: correlation test and the test of equal proportion. Firstly, the relationship between the strength of DIF and that of DDF was explored by correlating DIF magnitude and DDF magnitude of all 86 items according to the Pearson correlation coefficient. In second analysis, test of equal proportion was conducted between two proportions by binomial test to figure out if there is any difference between the number of items which also showing DIF and the number of items which does not show DIF. The first proportion was obtained by the number of items showing DIF over the total number of DDF items. This proportion indicated how many items' correct answer was chosen differently if the distractors were chosen differently. The other proportion was calculated by dividing the number of its item not showing DIF by the number of flagged DDF items, which represented among all item with distractors selected differently, how many of them the correct option selected equally. The result of test of binomial test was compared with the result of correlation test to answer the proposed research questions. The cognitive process associated with all items in these two types will be discussed for better understanding the results.

Limitation

Since only 86 multiple-choice items were studied in this study, it was possible that only very few items (less than five) were detected with DDF or DIF in step one or step two. In other words, it was possible that either or both one of the two item groups whose proportion would be compared in the third step had less than five items. This led to small sample size for test of equal proportion in step two, which may decrease the power of the statistical test and decline the accuracy of the test result.

Results

The purpose of this study was to investigate when distractors were selected differently, will the correct answer be chosen differently. Using the PIRLS 2016 achievement test data, the differential distractor functioning (DDF) analysis were conducted by multinomial logistic regression to flag DDF items while differential item functioning (DIF) analysis were developed by binary logistic regression to flag DIF items. DDF and DIF magnitude for each item were also obtain. Pearson correlation coefficient and test of equal proportion were implemented to investigate their relationship from practical perspective and research perspective, respectively.

Step 1: DDF and DIF Analysis

DDF Analysis

In DDF analysis, multinomial logistic regression was utilized, in which only the distractors (i.e. incorrect answers) were taken for analysis. There were two items, however, in which only two distractors were selected. In other words, there was one distractor in each of these two items was not selected by any participants (i.e. fourth-grade students from the U.S. and from Macao). Given that there should be more than two categories in dependent variable for conducting multinomial logistic regression, these two items were dropped from DDF analysis, but their item cognitive process will be provided in discussion chapter. Thus, excluding these two items, DDF analysis were conducted for 84 MC items. According to the power analysis conducted by Hsieh, Bloch, and Larsen (1998), for multinomial logistic regression, given the significance level of .01, minimum requirement of statistic power of 0.8 was 47. The smallest and the next smallest sample size of these 84 MC items was 29 and 42, which were the only two items that did not met the requirement. Sample size of other items were all satisfactory with the 302 as the medium, promising 0.8 statistic power.

The result showed among these 84 MC items, 38 items were flagged as DDF items in total. Out of these 38 DDF items, 1 item displayed non-uniform DDF item because the result of likelihood ratio test between model 3 and model 2 was statistically significant and model fit change was at least 0.003. The rest 37 items demonstrated uniform DDF items because the likelihood ratio test between model 3 and model 2 was non-significant but that between model 2 and model 1 was statistically significant, and Nagelkerke R^2 change was greater than 0.003. The results of multinomial logistic regression, including results of likelihood ratio test, DDF magnitude and DDF conclusions were shown in Table 1.

Table 1

Multinomial Logistic Regression Results for DDF Analysis

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DDF Conclusion
R11F01M	0.675	**9.233****	0.003	**0.049**	Uniform DDF
R11F02M	1.00	8.329*	0.005	0.045	Non-DDF
R11F03M	4.189	**9.536****	0.022	**0.051**	Uniform DDF
R11F04M	2.133	5.401	0.012	0.031	Non-DDF
R11F05M	1.734	**46.377****	0.003	**0.095**	Uniform DDF
R11F11M	1.890	**82.482***	0.004	**0.196**	Uniform DDF
R11F13M	3.436	6.913*	0.019	0.041	Non-DDF

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DDF Conclusion
R41O01M	1.648	3.530	0.011	0.024	Non-DDF
R41O06M	1.086	**23.824****	0.005	**0.073**	Uniform DDF
R41O11M	0.459	4.582	0.001	0.011	Non-DDF
R41O12M	4.681	**15.067****	0.014	**0.050**	Uniform DDF
R21Y01M	4.656	7.665*	0.015	0.026	Non-DDF
R21Y02M	**9.229****	3.240	**0.038**	0.013	Non-uniform DDF
R21Y04M	0.459	2.077	0.001	0.006	Non-DDF
R21Y05M	0.887	**38.293****	0.003	**0.129**	Uniform DDF
R21Y06M	1.624	**14.705****	0.004	**0.044**	Uniform DDF
R21Y07M	2.184	1.411	0.008	0.006	Non-DDF
R21Y08M	2.486	0.687	0.014	0.005	Non-DDF
R21Y11M	4.232	**43.885****	0.012	**0.134**	Uniform DDF
R31M01M	0.380	0.445	0.002	0.003	Non-DDF
R31M03M	3.309	**14.394****	0.008	**0.039**	Uniform DDF
R31M05M	0.452	**80.085****	0.001	**0.193**	Uniform DDF
R31M06M	1.065	**13.623****	0.003	**0.041**	Uniform DDF
R31M07M	0.408	2.070	0.003	0.010	Non-DDF

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2Δ$	$R^2Δ$	DDF Conclusion
R31M08M	2.425	7.613*	0.011	0.033	Non-DDF
R31M11M	2.950	**13.834****	0.015	**0.070**	Uniform DDF
R31M12M	2.272	**23.021****	0.006	**0.062**	Uniform DDF
R31M13M	1.443	3.569	0.016	0.039	Non-DDF
R31M14M	0.491	5.386	0.002	0.026	Non-DDF
R31M15M	0.054	4.990	0.000	0.019	Non-DDF
R41H01M	3.685	1.848	0.036	0.019	Non-DDF
R41H02M	2.183	5.251	0.013	0.032	Non-DDF
R41H05M	4.528	1.209	0.019	0.006	Non-DDF
R41H07M	5.239	**35.633****	0.009	**0.061**	Uniform DDF
R41H09M	0.077	**79.929****	0.000	**0.150**	Uniform DDF
R41H10M	0.663	**12.248****	0.002	**0.047**	Uniform DDF
R41H11M	0.806	2.957	0.006	0.019	Non-DDF
R41H12M	0.739	2.230	0.004	0.012	Non-DDF
L21B02M	1.512	0.388	0.038	0.010	Non-DDF
L21B03M	0.025	6.193*	0.001	0.094	Non-DDF
L21B05M	2.320	1.506	0.023	0.015	Non-DDF

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DDF Conclusion
L21B06M	1.898	5.000	0.011	0.028	Non-DDF
L21B10M	3.531	**11.902****	0.012	**0.040**	Uniform DDF
L21B11M	0.081	0.530	0.001	0.007	Non-DDF
L21B12M	5.072	11.324	0.123	0.368	Non-DDF
L21B14M	6.467*	2.700	0.027	0.011	Non-DDF
R21K03M	0.552	1.387	0.001	0.005	Non-DDF
R21K04M	4.195	3.037	0.009	0.007	Non-DDF
R21K06M	4.755	1.827	0.018	0.008	Non-DDF
R21K08M	0.447	8.172*	0.001	0.021	Non-DDF
R21K09M	0.776	**11.132****	0.003	**0.051**	Uniform DDF
R21K11M	2.681	6.370*	0.008	0.018	Non-DDF
R41I02M	8.195*	**71.098****	0.012	**0.119**	Uniform DDF
R41I05M	1.639	**61.268****	0.003	**0.146**	Uniform DDF
R41I06M	**6.844****	2.764	**0.014**	0.008	Non DDF
R41I08M	2.057	4.498	0.012	0.026	Non-DDF
R41I10M	1.079	**16.459****	0.002	**0.031**	Uniform DDF
R41I12M	1.729	2.34	0.004	0.005	Non-DDF
R11L02M	0.611	**57.315****	0.001	**0.100**	Uniform-DDF

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DDF Conclusion
R11L05M	0.416	**28.651****	0.001	**0.079**	Uniform DDF
R11L07M	3.028	0.928	0.005	0.002	Non DDF
R11L09M	0.69	**17.455****	0.005	**0.114**	Uniform DDF
R11L11M	3.552	**29.717****	0.01	0.079	Uniform DDF
R31W03M	0.836	4.789	0.003	0.019	Non DDF
R31W05M	2.466	**9.574****	0.006	**0.022**	Uniform DDF
R31W06M	4.954	2.22	0.034	0.016	Non DDF
R31W08M	6.261*	5.817	0.026	0.024	Non DDF
R31W09M	0.389	**13.016****	0.001	**0.023**	Uniform DDF
R31W10M	1.138	**9.413****	0.002	0.02	Uniform DDF
R31W12M	0.49	**22.691****	0.001	**0.046**	Uniform DDF
R41T01M	5.781	1.377	0.039	0.009	Non DDF
R41T05M	1.052	6.484*	0.002	0.016	Uniform DDF
R41T09M	2.7	**14.937****	0.004	**0.028**	Uniform DDF
R41T12M	0.821	7.002*	0.004	0.036	Uniform DDF
R41T13M	0.992	4.198	0.001	0.008	Non DDF
R41T15M	1.563	**15.514****	0.005	**0.050**	Uniform DDF
R41T16M	8.16*	3.435	0.034	0.014	Non DDF

	Likelihood Ratio Test (dfΔ=2)		DDF Magnitude		
	Non-Uniform (M3-M2)	Uniform (M2-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DDF Conclusion
L21E03M	0.094	$\mathbf{12.292^{**}}$	0.000	0.057	Uniform DDF
L21E04M	0.225	$\mathbf{20.167^{**}}$	0.001	0.088	Uniform DDF
L21E05M	1.798	2.148	0.044	0.055	Non DDF
L21E08M	4.52	$\mathbf{11.506^{**}}$	0.020	0.054	Uniform DDF
L21E09M	0.663	$\mathbf{21.856^{**}}$	0.001	0.045	Uniform DDF
L21E11M	0.496	0.448	0.005	0.005	Non DDF
L21E17M	0.926	$\mathbf{17.750^{**}}$	0.009	0.058	Uniform DDF

Note. M1: Model 1; M2: Model 2; M3: Model 3; -2LLΔ: -2 times of the log likelihood difference; dfΔ: degrees of freedom difference; $R^2\Delta$: Nagelkerke R^2 changes; statistically significant likelihood ratio test and significant Nagelkerke R^2 change was highlighted in bold.

$^{*}p < .05$; $^{**}p < .01$.

DIF Analysis

In DIF analysis, binary logistic regression was used. Test examinees’ performance to each item were all included in data analysis and were classified into two groups: correct or incorrect. In order to keep consistent dataset with the DDF analysis, two items dropped in DDF analysis were also excluded in DIF analysis. According to the power analysis conducted by Hsieh, Bloch, and Larsen (1998), given the significance level of .01, in order to reach statistic power of 0.8, the minimum requirement of the binary logistic regression was 47. So, the samples sizes of each item, with 1081 as the smallest, was large enough and substantially met the requirement.

According to the results of DIF analysis, there were 45 items flagged with DIF in total as the likelihood ratio test between model 3 and mode 1, which was the simultaneous uniform and non-uniform DIF test, showed significant test results. Among these 45 items, 4 items showed significant Nagelkerke R^2 change between model 3 and model 2, and they were flagged as non-uniform DIF, while 41 items displayed tiny Nagelkerke R^2 change from model 2 to model 3, but they showed substantial Nagelkerke R^2 change from model 1 to model 2, thus they were detected as uniform DIF. Nagelkerke R^2 change for all these 45 DIF items were greater than 0.003. The results of binary logistic regression, including results of likelihood ratio test, DIF magnitude and DIF conclusions, were shown in Table 2.

Table 2

Logistic Regression Results for DIF Analysis

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DIF Conclusion
R11F01M	**88.089****	0.002	**0.076**	Uniform DIF
R11F02M	0.312	0.000	0.000	Non-DIF
R11F03M	**15.793****	**0.013**	0.003	Non-uniform DIF
R11F04M	**13.538****	**0.007**	0.006	Non-uniform DIF
R11F05M	4.045	0.001	0.002	Non-DIF
R11F11M	**66.954****	0.004	**0.055**	Uniform DIF
R11F13M	**27.746****	0.010	**0.021**	Uniform DIF
R41O01M	0.667	0.000	0.001	Non-DIF
R41O06M	6.036*	0.003	0.002	Non-DIF
R41O11M	0.674	0.001	0.000	Non-DIF
R41O12M	2.523	0.001	0.002	Non-DIF
R21Y01M	**23.440****	0.008	**0.011**	Uniform DIF
R21Y02M	2.313	0.000	0.002	Non-DIF
R21Y04M	**39.701****	0.008	**0.022**	Uniform DIF
R21Y05M	1.319	0.001	0.000	Non-DIF

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DIF Conclusion
R21Y06M	**14.032****	0.0000	**0.012**	Uniform DIF
R21Y07M	5.598	0.0030	0.003	Non-DIF
R21Y08M	2.071	0.0010	0.001	Non-DIF
R21Y11M	4.944	0.0000	0.004	Non-DIF
R31M01M	6.188*	0.0000	0.006	Non-DIF
R31M03M	**11.460****	0.0000	**0.009**	Uniform DIF
R31M05M	**91.580****	0.0070	**0.071**	Uniform DIF
R31M06M	**34.662****	0.0010	**0.027**	Uniform DIF
R31M07M	**16.04****	0.0060	**0.009**	Uniform DIF
R31M08M	3.188	0.0020	0.001	Non-DIF
R31M11M	8.400*	0.0000	0.008	Non-DIF
R31M12M	6.265*	0.0020	0.003	Non-DIF
R31M13M	1.810	0.0030	0.000	Non-DIF
R31M14M	**28.876****	0.0090	**0.015**	Uniform DIF
R31M15M	**102.111****	0.0130	**0.070**	Uniform DIF
R41H01M	**13.761****	0.0010	**0.020**	Uniform DIF
R41H02M	1.178	0.0000	0.001	Non-DIF

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DIF Conclusion
R41H05M	**31.254****	0.0000	**0.029**	Uniform DIF
R41H07M	**25.582****	0.0010	**0.019**	Uniform DIF
R41H09M	**34.394****	0.0000	**0.029**	Uniform DIF
R41H10M	**22.951****	0.0020	**0.017**	Uniform DIF
R41H11M	**25.461****	0.0000	**0.028**	Uniform DIF
R41H12M	8.981*	0.0040	0.005	Non-DIF
L21B02M	**16.816****	0.0060	**0.039**	Uniform DIF
L21B03M	2.456	0.0040	0.001	Non-DIF
L21B05M	5.764	0.0010	0.007	Non-DIF
L21B06M	**73.573****	0.0010	**0.082**	Uniform DIF
L21B10M	**139.335****	0.0010	**0.113**	Uniform DIF
L21B11M	1.070	0.0020	0.000	Non-DIF
L21B12M	3.304	0.0270	0.003	Non-DIF
L21B14M	**29.835****	0.0010	**0.027**	Uniform DIF
R21K03M	**14.624****	0.0030	**0.010**	Uniform DIF
R21K04M	2.429	0.0010	0.001	Non-DIF
R21K06M	**40.442****	0.0030	**0.035**	Uniform DIF

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	R^2Δ	R^2Δ	DIF Conclusion
R21K08M	**33.075****	**0.0140**	0.014	Non-uniform DIF
R21K09M	5.847	0.0060	0.000	Non-DIF
R21K11M	4.472	0.0020	0.002	Non-DIF
R41I02M	1.910	0.0010	0.000	Non-DIF
R41I05M	2.543	0.0000	0.002	Non-DIF
R41I06M	2.557	0.0010	0.001	Non-DIF
R41I08M	**39.873****	0.0000	**0.040**	Uniform DIF
R41I10M	1.675	0.0010	0.001	Non-DIF
R41I12M	0.213	0.0010	0.000	Non-DIF
R11L02M	**37.761****	0.0000	**0.032**	Uniform DIF
R11L05M	**205.314****	0.0010	**0.168**	Uniform DIF
R11L07M	2.005	0.0010	0.000	Non-DIF
R11L09M	**63.675****	0.0020	**0.076**	Uniform DIF
R11L11M	**174.480****	0.0020	**0.141**	Uniform DIF
R31W03M	**42.199****	0.0010	**0.035**	Uniform DIF
R31W05M	**17.441****	0.0020	**0.012**	Uniform DIF
R31W06M	**12.165****	0.0060	**0.009**	Uniform DIF

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	R^{2}Δ	R^{2}Δ	DIF Conclusion
R31W08M	7.277*	0.0000	0.007	Non-DIF
R31W09M	**14.583****	0.0030	**0.008**	Uniform DIF
R31W10M	2.671	0.0000	0.002	Non-DIF
R31W12M	1.342	0.0010	0.000	Non-DIF
R41T01M	4.435	0.0000	0.005	Non-DIF
R41T05M	**28.087****	0.0020	**0.022**	Uniform DIF
R41T09M	**26.073****	0.0000	**0.019**	Uniform DIF
R41T12M	5.777	0.0000	0.007	Non-DIF
R41T13M	**30.208****	0.0020	**0.022**	Uniform DIF
R41T15M	**38.846****	0.0020	**0.038**	Uniform DIF
R41T16M	5.062	0.0050	0.001	Non-DIF
L21E03M	**9.476****	0.0050	**0.005**	Uniform DIF
L21E04M	**50.922****	0.0080	**0.044**	Uniform DIF
L21E05M	**11.110****	**0.0260**	0.004	Non-uniform DIF
L21E08M	**68.762****	0.0000	**0.069**	Uniform DIF
L21E09M	**26.748****	0.0010	**0.023**	Uniform DIF
L21E11M	1.801	0.0010	0.001	Non-DIF

	Likelihood Ratio Test (dfΔ=2)	DIF Magnitude		
	Simultaneous Test of Uniform and Non-Uniform DIF (M3-M1)	Non-Uniform (M3-M2)	Uniform (M2-M1)	
Item	-2LLΔ	$R^2\Delta$	$R^2\Delta$	DIF Conclusion
L21E17M	1.046	0.000	0.001	Non-DIF

Note. M1: Model 1; M2: Model 2; M3: Model 3; -2LLΔ: -2 times of the log likelihood; dfΔ: degrees of freedom difference; $R^2\Delta$: Nagelkerke R^2 change; statistically significant likelihood ratio test and significant Nagelkerke R^2 change was highlighted in bold.

$^{*}p < .05$; $^{**}p < .01$.

Step 2: Investigation of The Relationship between DIF and DDF

Correlation Test

Two correlation tests were conducted to explore the relationship between DIF magnitude and DDF magnitude. The first correlation coefficient was calculated for all 84 items, while the other one was calculated for only 23 items flagged DIF and DDF simultaneously. The results of correlation tests were displayed in Table 3 and Table 4, respectively. Neither Person correlation coefficient was significant. However, the *p* value of the correlation for 23 items (r = .30, *p*=.086) was very close to .05. Its non-significance may be caused by the small sample size.

Table 3

Result of Correlation Test for 84 Items

	DIF Magnitude	DDF Magnitude
DIF Magnitude	1	.176
DDF Magnitude	.176	1

Table 4

Result of Correlation Test for 23 Items with Both DIF and DDF

	DIF Magnitude	DDF Magnitude
DIF Magnitude	1	.300
DDF Magnitude	.300	1

Test of Equal Proportion

In order to answer the research question of present study, test of equal proportion was implemented by binomial test to exam, among all detected DDF items, to exam if the proportion of flagged DIF items and the proportion of flagged non-DIF items equals or not. The frequency and the proportion of DIF and non DIF items among DDF items was shown in Table 5. Out of all 38 detected DDF items, 23 were also identified with DIF, so the observed proportion of DIF item among DDF item was .61. This proportion was tested with expected proportion, which is .5. The binomial test indicated that the proportion of DIF items among DDF items of .61 was equal to the expected .5, p = .256 (two-sided).

Table 5

Frequency and Proportion of DIF and Non-DIF Items among DDF Items

	Frequency	Observed Proportion
Item with DIF	23	.61
Item without DIF	15	.39
Total	38	1

Cognitive Processes

In PIRLS 2016 achievement test, there were two independent cognitive processes considered in creating the MC items: the purposes of reading and the processes of comprehension. The purposes of reading included (1) Acquire and Use Information and (2) Literary Experience, while the processes of comprehension contained (1) Evaluate and Critique Content and Textual Elements, (2) Focus on and Retrieve Explicitly Stated Information, (3) Interpret and Integrate Ideas and Information, and (4) Make Straightforward Inferences. The number of DIF and DDF items of each reading purpose category and each comprehension process level were summarized in Table 6 and Table 7, respectively. The two items, which were excluded in data analysis process because only two distractors were chosen by participants, were both designed for the same cognitive process—Acquire and Use Information in terms of purpose of reading and Focus on and Retrieve Explicitly Stated Information in terms of processes of comprehension. As two levels of item content were independent with each other, their relationship with the item flag (i.e. DIF, DDF, both or neither) was explored separately. Chi-square test of independent was performed to investigate the relationship between reading purpose and item flag. The Fisher's Exact test was utilized to examining the association between comprehension process and the item flag, as there were 9 cells (56.3%) had expected count less than 5 with minimum expected count is 1.96. The Chi-square results demonstrated the relationship between reading purpose and item flag, $X^2(3, N = 84) = 7.49, p = .862$, was non-significant. The results of Fisher's Exact test showed the non-significance relationship between comprehension process and item flag ($p = .166$). Thus, there is no evidence that item content impacts DDF or DIF detection.

Table 6

Proportion of Different Items in Each Reading Purpose

	R1	R2	Total
Item with DIF only	.237	.283	22
Item with DDF only	.184	.174	15
Item with DIF and DDF	.316	.239	23
Neither DIF nor DDF	.263	.304	24
Total	38	46	84

Note. R1: Acquire and Use Information; R2: Literary Experience.

Table 7

Proportion of Different Items in Each Comprehension Process

	C1	C2	C3	C4	Total
Item with DIF only	.133	.348	.182	.286	22
Item with DDF only	.200	.087	.455	.143	15
Item with DIF and DDF	.467	.174	.182	.286	23
Neither DIF nor DDF	.200	.391	.182	.286	24
Total	15	23	11	35	84

Note. C1: Evaluate and Critique Content and Textual Elements; C2: Focus on and Retrieve Explicitly Stated Information; C3: Interpret and Integrate Ideas and Information; C4: Make Straightforward Inferences.

Discussion

The research question that drove this study was whether there is a relationship between differential distractor functioning (DDF) and differential item functioning (DIF). Specifically, if DDF occurs, must DIF occur? In answering this research question, multiple-choice items from the PIRLS 2016 achievement test were used. Those items have one correct answer and three distractors. Two samples were used sample from the U.S. and Macao made up of more than 8000 4th grade students.

The analysis found 60 items (71.4% of total multiple-choice) with DDF, DIF or both. In other words, more than 70% MC items demonstrated potential measurement bias for test takers with American or Chinese backgrounds. To be more specific, out of these identified items, 22 (26.2%) demonstrated DDF only, indicating that though their distractors were selected differently, their correct answers were selected equally by American students and Macao students. 15 (17.9%) items displayed DIF only, indicating that difficulty varied based on nationality, but distractors were chosen equally. 24 (28.6%) items showed both DDF and DIF, which means their distractors and correct answers were both selected unequally by American and Macao examinees. For more than 70% MC items in PIRLS 2016 achievement test, examinees' performance was impacted by factors other than ability.

A series of analyses based on correlations and comparisons of proportions found no relationship between DDF and DIF. In previous study, DDF analysis were only conducted for the items which demonstrated DIF, which presumed that DIF was the precondition of DDF. In other words, it assumed that DDF only occurred when DIF appeared. This, however, might ignore the items demonstrating DDF without DIF. Tables 3 and 4 displayed the correlation test for all 84 items and for only 23 items with both DIF and DDF, respectively. Neither correlation coefficient was significant, suggesting that counts of DIF and DDF were independent of each other. The magnitude of one did not impact that of the other.

The binomial test, which results were shown in Table 5, further investigated the relationship between DIF and DDF. Two proportion were calculated for all identified DDF items, representing the precondition that DDF had already occurred. The results indicated that, out of all DDF items, the proportion of DIF item equaled to that of non DIF items, thus, when incorrect answers were chosen differently, the proportion of items whose correct answer was chosen differently was equal to that of items whose correct answer was selected equally. It provided a valuable suggestion that, in order to supply sufficient information for improving measurement functioning of MC items, the investigation of DIF and DDF should be independent because their presence was found to be unassociated.

To further explore these results, item content information, provided by the IEA with the PIRLS data, was compared with DDF and DIF presence. Before the analyses, there were two items deleted because each of them contained a distractor that was not selected by U.S. and Macau participants at all. It was really interesting that these two items were created for the same level regarding both reading purpose and comprehension process. These two items were designed for measuring students’ ability of information acquire and use and focus on and retrieve explicitly stated information. The reading purpose of acquiring and using information was associated with reading to learn in PIRLS 2016. For more details, items whose purpose was information acquire and use were designed for assessing the information contained in the passages. Therefore, it was not hard to explain that one incorrect option in these items was not selected at all. In doing these items, the only task for participants was to compare the option with the passage content. The option not selected might be obvious unrelated to the passage. In terms of comprehension process, the level—focus on and retrieve explicitly stated information—required little direct inference or interpretation but only asked examinees to locate the related information and understand the words or sentences in the passage. This might explain why there is one distractor left unselected.

Caveats

There were several issues that limited the interpretation of the results. First of all, both USA and Macao were placed at medium proficiency level. There were relatively less students located on the higher and lower end of the ability continuum and most of students were with medium reading ability, making it harder to detect the different performance patterns across different proficiency groups. This may explain why there was only small number of non-uniform DIF and non-uniform DDF detected. Next, there are three different language versions used in assessment for Macao students. This multiple language usage made the interpretation of results more difficult. In addition, there were two items whose sample sizes were not met the minimum requirement to provide .8 statistical power, so the results of these two items should be interpret carefully. Besides, there was a limited sample size of multiple-choice items and a limited number of cultural subgroups. There were only 84 items that could be used in these DDF and DIF analysis, which is not particularly powerful for finding a relationship. Only two cultural groups—American and Chinese culture—were considered in this study. As PIRLS was an international large-scale assessment, the DDF and DIF analysis could be extended to multiple cultures to provide more specific advice for item improvement.

Conclusion

The present study fills a gap, as few previous studies examine the relationship between existence of differential distractor functioning (DDF) and differential item functioning (DIF), and those that do not use the same approach as used here. This study provides useful suggestions for improving test fairness and reducing potential test bias for multiple-choice items which is to identify the potential bias associated with correct answer and incorrect options independently of each other.

Reference

Abedi, J., Leon, S., & Kao, J. C. (2008). Examining Differential Distractor Functioning in Reading Assessments for Students with Disabilities. CRESST Report 743. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice*, *17*(1), 31-44.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel--Haenszel method. *Applied Measurement in Education*, *2*(3), 217-233.

Dorans, N. J., & Holland, P. W. (1992). DIF DETECTION AND DESCRIPTION: MANTEL-HAENSZEL AND STANDARDIZATION 1, 2. *ETS Research Report Series*, *1992*(1), i-40.

Dorans, N. J., & Kulick, E. (1983). ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE OF FEMALE CANDIDATES ON SAT AND TSWE FORMS ADMINISTERED IN DECEMBER 1977: AN APPLICATION OF THE STANDARDIZATION APPROACH 1. *ETS Research Report Series*, *1983*(1), i-14.

Dorans, N. J., & Schmitt, A. P. (1991). CONSTRUCTED RESPONSE AND DIFFERENTIAL ITEM FUNCTIONING: A PRAGMATIC APPROACH 1. *ETS Research Report Series*, *1991*(2), i-49.

Foy, P., Martin, M. O., Mullis. I. V., & Yin. L. (2017). Reviewing the PIRLS item statistics. In M. O. Martin, I. V. Mullis, & M. Hooper (Ed.), *Methods and Procedures in PIRLS 2016* (pp. 10.1-10.26). Boston, MA: International Association for the Evaluation of Educational Achievement.

Gómez-Benito, J., & Navas-Ara, M. J. (2000). A Comparison of $\chi 2$, RFA and IRT Based Procedures in the Detection of DIF. *Quality and Quantity*, *34*(1), 17-31.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*(2), 147-160.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine*, *17*(14), 1623-1634.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Kato, K., Moen, R., & Thurlow, M. (2009). Examining DIF, DDF, and omit rate by discrete disability categories. *Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.*

Koon, S. (2010). Comparison of Methods for Detecting Differential Distractor Functioning.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning. *ETS Research Report Series*, *2008*(2), i-32.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distracter choices. *Journal for Research in Mathematics Education*, 325-336.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of educational statistics*, *7*(2), 105-118.

Middleton, K., & Laitusis, C. C. (2007). Examining test items for differential distractor functioning among students with learning disabilities. *ETS Research Report Series*, *2007*(2), i-34.

Monahan, P. O., McHorney, C. A., Stump T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*, 92-109.

Park, M. (2017). *Investigating differential options functioning based on multinomial logistic regression with widely used statistical software* (Doctoral dissertation, University of British Columbia).

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247-269.

Schmitt, A. P., & Bleistein, C. A. (1987). FACTORS AFFECTING DIFFERENTIAL ITEM FUNCTIONING FOR BLACK EXAMINEES ON SCHOLASTIC APTITUDE TEST ANALOGY ITEMS 1. *ETS Research Report Series*, *1987*(1), i-46.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292.

Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. *Consulted page at September 10th: http://www. unt. edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011. pdf, 29,* 2825-2830.

Stoneberg Jr, B. D. (2004). A Study of Gender-Based and Ethnic-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi Square Test. *Online Submission.*

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*(2), 188-205.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Tan, X., Xiang, B., Dorans, N. J., & Qu, Y. (2010). The value of the studied item in the matching criterion in differential item functioning (DIF) analysis. *ETS Research Report Series*, *2010*(1), i-27.

Tsaousis, I., Sideridis, G., & Al-Saawi, F. (2018). Differential Distractor Functioning as a Method for Explaining DIF: The Case of a National Admissions Test in Saudi Arabia. *International Journal of Testing*, *18*(1), 1-26.

Wang, W. C., & Su, Y. H. (2004). 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*(2), 113-144.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i-30.