The Effectiveness of an Iterative Randomized Anchor Selection Strategy in DIF Detection

YIP WANG

EdD

THE HONG KONG INSTITUTE OF EDUCATION

2014

The Effectiveness of an Iterative Randomized Anchor Selection Strategy in DIF Detection

By

YIP WANG

A Thesis Submitted to

The Hong Kong Institute of Education

in Partial Fulfilment of the Requirement for

the Degree of Doctor of Education

November 2014

STATEMENT OF ORIGINALITY

I, YIP Wang, hereby declare that I am the sole author of the thesis and the material presented in this thesis is my original work except those indicated in the acknowledgement. I further declare that I have followed the Institute's policies and regulations on Academic Honesty, Copy Right and Plagiarism in writing the Thesis and no material in this thesis has been published or submitted for a degree in this or other universities.

_____

YIP Wang

November 2014

Thesis Examination Panel Approval

Members of the Thesis Examination Panel approve the thesis of YIP Wang defended on 28 October 2014.

Supervisors

WANG Wen-Chung
Chair Professor
Department of Psychological Studies,
The Hong Kong Institute of Education

MOK Mo-Ching, Magdalena
Chair Professor
Department of Psychological Studies,
The Hong Kong Institute of Education

Examiner

HUANG Xiao-ting
Associate Professor
Graduate School of Education,
Peking University

YAN Zi
Assistant Professor
Department of Curriculum and Instruction,
The Hong Kong Institute of Education

Approved on behalf of the Thesis Examination Panel:

Chair, Thesis Examination Panel
LAM Tak Ming, Lawrence
Professor
Programme Director of EdD
The Hong Kong Institute of Education

ABSTRACT

The Effectiveness of an Iterative Randomized Anchor Selection Strategy in DIF Detection

by YIP Wang

The Hong Kong Institute of Education

Abstract

Differential Item Functioning (DIF) is an important topic in educational testing and psychometrics, which refers to the phenomenon in which test-takers having identical abilities that a particular test item is designed to measure, have different probabilities of correctly answering the item. To achieve the goal of setting an unbiased test, DIF assessment is therefore mandatory. A common metric is often established and serves as a matching variable for assessing whether an item in a target test exhibits DIF. A usual approach is to derive the common metric from an anchor set comprising carefully identified items from the test. It is highly desirable to have an accurate anchor set which is DIF-free, as the purity of an anchor set will significantly affect the accuracy of the item and person calibrations, which in turn affect the success of DIF assessment or detection. Normally, the more accurate the anchor set is, the higher the power and the better controlled the Type I error rate in the DIF detection

process will be.

This thesis proposes a new anchor selection method for improving the accuracy of DIF detection. The new method, abbreviated as IRCI, is based on an Iterative Randomized Constant Item selection process coupled with scale purification that repeatedly identifies DIF items using randomized short anchor sets and filters the DIF items from the candidate anchor items. A computer simulation program is implemented to serve as a platform for comparing the performance of the new IRCI method against other existing anchor selection methods including the AOI (All-Other-Item), AOI-SP (All-Other-Item with Scale-Purification) and CI (Constant-Item). The methods are evaluated upon different test data with different parameter settings (e.g., number of items, DIF contamination rate, etc.). Our simulation results show that the new IRCI method improves the anchor selection accuracy in a number of parameter settings. In general, the power of the IRCI method in DIF detection is higher than that of the AOI, the AOI-SP and the CI methods, and the Type I error rate is better controlled. The new method is particularly effective under high DIF contamination scenarios. When the number of DIF items is 30% or above, the new method performs better than the AOI, the AOI-SP and the CI methods with a bigger margin. Furthermore, when the sample sizes of the reference and focal groups are smaller, the new method also demonstrates improvement over the AOI and AOI-SP methods.

ACKNOWLEDGEMENTS

I would like to thank Prof. WANG Wen-chung and Prof. MOK Mo-ching, Magdalena for their guidance and support throughout my study at The Hong Kong Institute of Education. Apart from my personal curiosity in the subject, the expertise of Prof. WANG and Prof. MOK in the realm of educational assessment is the main reason why I chose the HKIEd for pursuing my doctoral degree study.

Pursuing a degree as a part-time student with a demanding full-time job needlessly to say is a difficult task, I therefore very much indebted to the unfailing support from my family members. I have been spending much time on classes and also on the thesis work instead of accompanying my family on Saturdays and Sundays. I am so grateful and thankful to my wife, Loretta, and also my lovely kids, Tiffany and Cyrus, who tolerate, accommodate and support my aspiration of doing what I myself am interested in. I therefore wish to dedicate this piece of work to them.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AOI	All other item
AOI-SP	All other item with scale purification
CI	Constant item
DIF	Differential item functioning
IRCI	Iterative randomized constant item
IRT	Item response theory

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Numerous educational or intelligence tests with different purposes have been established, reflecting that human societies require a tool, such as academic tests, to obtain a more accurate estimation on the intellectual ability of fellow human beings, so as to fulfil certain social needs such as selecting the suitable candidates to fill up vacancies in various social institutions. In the past century, big advancement has been made in the field of psychological, educational testing as well as psychometrics. Simply put, one can view psychological and educational testing as a means to achieve the ends of measuring the inner or latent psychological trait of a person. Unlike physical properties such as mass, distance or temperature which has already been measured objectively by different precise measurement tools, a scientifically well-defined unit of measurement for the degree of intellectual ability—at least from our latest knowledge on neuropsychology and brain functionality—has yet been established.

Psychometrics is an academic discipline which aims at developing good tools and scales to measure psychological traits such as intelligence and personality and also developing relevant theories to account for the results of the measurement. The issues related to the design of assessment tools such as tests and questionnaires, quantification of the results of the tests and

questionnaires, interpretation of the results, and building of different models in fitting the results or vice versa, are the major research areas of psychometrics.

A critical prerequisite in any post-test analysis in psychometrics is that the items in the test must be valid in measuring the right underlying trait of the target test takers. For instance, if a test is targeting at assessing the candidates' understanding on geometry, then all the items should be related to concepts of geometry but not other unrelated attributes. However, even when a test is designed to measure a specific ability, there could be other factors, or traits, or actually the background knowledge related to some attributes of the test takers such as the gender, race and socio-economical status, which could have significant influences on the test performance. In such cases, the test is regarded as a biased test. This kind of problem was first discovered a few decades ago (Cleary & Hilton, 1968; Angoff & Ford, 1971).

Item bias or test bias, as the terms coined, refers to situations where an item or a test apparently favours a particular group of test takers. According to some early pioneer studies on investigating the possible cause of the performance difference between races (Cleary & Hilton, 1968; Angoff & Ford, 1971), researchers first found that race had a role to play in the test performance for those high-stake tests such as the Scholastic Aptitude Test (SAT) conducted in the United States. Most early studies focused on finding the difference in the test performance among different ethnic groups such as White and Black/Hispanic. There was curiosity among researchers on whether the

relatively worse performance of Black/Hispanic test takers in cognitive tests was due to genuine difference in the ability being assessed or actually related to other factors such as difference in background knowledge among different races or cultural backgrounds. The later possibility was a very sensitive issue in the United States in the 1960s as civic rights and equal opportunities were being highly sought. Therefore, it would be normal and reasonable to request the removal of biased items which might contain some latent knowledge that was not within the sphere of minority's culture—especially for those high-stake tests used for selection purposes.

As the issue of item bias attracted more interests in academic discussions and more results were gathered, researchers reckoned that the term 'item bias' itself was also biased in certain sense. Researchers generally agreed that statistical finding of the aberrant behaviour of responses to an item could be explained by various reasons and some of such reasons might not necessarily related to unfair discrimination against group memberships. As a result, a new term—Differential Item Functioning (DIF)—was coined to directly report the fact that statistical evidence of group membership having an effect on the probability of correctly answering a given item under the prerequisite of comparing the comparable could be found. Judgment on the cause of the aberrant behaviour of the item was to be made at a later stage with more deliberations on the actual content of the item.

Detecting DIF items in a test is gaining increasing focuses due to its profound

impact on psychometric analysis. As more statistical procedures are developed in the area, more tools and results are available for researchers to compare and contrast different models and methods of DIF detection. On the one hand, psychometricians know more about the statistical behaviour of responses to items with DIF and what each statistic can tell. On the other hand, research studies on how to apply different DIF detection methods and investigations on the effectiveness of these methods have been started. As almost all theories eventually need to serve practical applications, DIF detection is now maturing in the sense that more emphasis is directed to the application side. On the application side, one of the major concerns is which DIF method is to be applied and how well the chosen method will work in practical situations. There have been substantial research studies on this aspect of DIF application and many useful results have been obtained (Hooland & Thayer, 1988; Shih and Wang, 2009; Wang, 2004, Wong, 2008, Wang & Yeh, 2003; Wang, Shih and Yang, 2009; Woods, 2009). The results of these studies generally indicated that there are many feasible methods in detecting DIF items in a test and each of these methods has its own merits. No single method is proven to be universally outperforming other methods and there are some trade-off factors involved in DIF detection method design. In the realm of Item Response Theory (IRT), DIF detection method typically involves an item calibration process to determine the ability of the test takers and the difficulties of the test items. One of the critical factors in determining the effectiveness of DIF detection method is how anchor items are selected for the item calibration process (Wang, 2004; Wang, 2008; Woods, 2009). In practice, target anchor

items are DIF-free items taken from the same test under investigation. With reference to the results gathered in earlier DIF studies, there are still rooms to further refine some established DIF detection methods—in particular in the aspect of anchor purification.

In this thesis, we propose a novel method for DIF detection for psychometric analysis. Specifically, we propose a new algorithm in purifying an anchor set for DIF detection. Our method adopts the Rasch model which is a simplified special case of the Item Response Theory (IRT) in which an item response function takes into account only 2 parameters, namely, person ability and item difficulty (Rasch, 1960; Bond & Fox, 2001). The Rasch model has been considered a simple but robust IRT model in such a way that most DIF methods are first designed to base upon the Rasch model and are further extended, if necessary, to other more complicated IRT models with straightforward adaptations in general. In this sense, the Rasch model provides a simple platform which enables focused analysis and understanding of DIF detection methods.

We also adopt a simulation based approach, in which raw data with varying person abilities and item difficulties, as well as item responses for each test taker, are generated with well-controlled parameters based on the Rasch model. DIF items are introduced to the test set, and DIF detection methods are executed on the simulation data and being assessed based on how well they can recover the DIF items. Two quantitative measures, Type-I error rate and

power, are calculated for each method for comparisons. It is desirable to have a DIF detection method with well-controlled Type-I error rate and a high power.

The commonly adopted All-Other-Item with scale purification (AOI-SP) method uses all the items other than the item under question (i.e., for DIF assessment) as anchors (Holland & Thayer, 1988). The scale purification process of iteratively removing probable DIF items from an anchor set is the essence of the method. On the other hand, the Constant Item (CI) method (Wang, 2008) which uses a constant number of items in the anchor set suggests that the use of a short anchor set (i.e., one with fewer anchor items) can boost the performance of DIF detection. Inspired by the gist of these two methods, our main idea is to combine the scale purification approach with short anchor selection. We start with modifying the AOI-SP method by selecting only a subset of its resulting anchor items to serve as a final short anchor set. We compare the use of different lengths of the subset and evaluate their performances. It is found that there is slight performance gain both in terms of the Type-I error rate and the power. There are a few useful observations obtained from this preliminary trial. Based on these insights, we further devise our new DIF detection method, which we call the Iterative Randomized Constant Item (IRCI) method, that involves a randomized selection of short anchor sets coupled with a scale purification scheme. Our experiment results show that the IRCI method has obtained encouraging performance gain over other methods, acquiring well-controlled Type-I error rate and high power, even when the reference and focal groups are of smaller sample sizes. The new

method therefore provides a more accurate tool for practitioners to detect DIF items in daily applications.

The subsequent chapters of this thesis are organized as follows. In CHAPTER 2, we first present some background and related work in the areas of psychometrics, with focuses in particular on DIF, DIF detection and anchor selection. We then detail in CHAPTER 3 our first attempt in adapting the AOI-SP method, the evaluation of the modified method, as well as the insights gained from this adaptation. In CHAPTER 4, we present the new IRCI method proposed by this thesis, and its performance against other existing methods in various settings. Finally, a summary and a conclusion of this research study are given in CHAPTER 5.

CHAPTER 2

BACKGROUND

In this chapter, we introduce some basics in psychometrics. We in particular put our emphasis on the topics of the basis of DIF, DIF detection as well as anchor selection, which are the major focuses of this work.

2.1 Item Response Theory

In modern psychometrics, the Classical Test Theory and the Item Response Theory are the two most prominent theoretical models regarding psychometric analysis. These two models were developed in different times and have different assumptions on test scores. It is arguably true that almost all psychometrics researches on test score manipulation or interpretation would either touch upon these two theories or adopt one of these two as the basis of their studies. The choice of a particular model has subtle influences on the quantitative analysis of the test results and subsequently the qualitative accounts. Here, we will give a very brief introduction to the Classical Test Theory while focusing more on the Item Response Theory which is the underlying theoretical model that our method adopts.

The Classical Test Theory (CTT), as indicated by its name, has a relatively longer history than the Item Response Theory. Also known as the True Score Theory, it assumes that each observed score comprises two parts: a true score part and a random error part. The relationship between the observed score obtained by a test taker and the theoretical true score is expressed by the simple equation $E = X - T$, where X is the observed score, T is the true score, and E is the random error of measurement which can be either positive, negative, or zero. There is an underlying assumption in CTT that the error of observed scores shall reside around the true score and the sum of all the error terms shall be zero, and that there is no correlation between E and T as E is a random error term. The model represented by CTT is simple and easy to comprehend. However, there are a number of shortcomings in the model, in which arguably the most fatal one, is the lack of mechanism for conducting sophisticated statistical analysis on items and test takers individually.

Item Response Theory (IRT), on the contrary, is based on a more sophisticated mathematical model which computes statistically what coined as item and person characteristics. IRT enables the estimation of a test taker's ability from any set of items given to him and also the assessment of how effective each item is at measuring each ability level (Lord, 1980, p.11-12)—neither of these two can be accomplished by the CTT model. IRT adopts a probabilistic approach in modelling the response to an item for a test taker with a given ability level, which is given by the Item Response Function as:

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}} \quad (1)$$

In Equation (1), $P(\theta)$ denotes the probability of correctly answering an item with person ability θ, and a, b and c, respectively, denote the item discriminating power, item difficulty and guessing factor. Discrimination power reflects the extent the item response function varies as the person ability varies, and is proportional to the slope of the item response function at the inflection point as shown in a typical item characteristic curve in Figure 1 below. Item difficulty is a measure of how hard an item can be correctly answered. The inflection point of the curve is where the item difficulty equals person ability, i.e., $b=\theta$. Guessing factor refers to the probability that a test taker can correctly answer an item when his ability approaches negative infinity. Equation (1) is a 3-parameter logistic function which monotonically increases as the person ability θ increases with a given set of a, b, and c (see Figure 1).

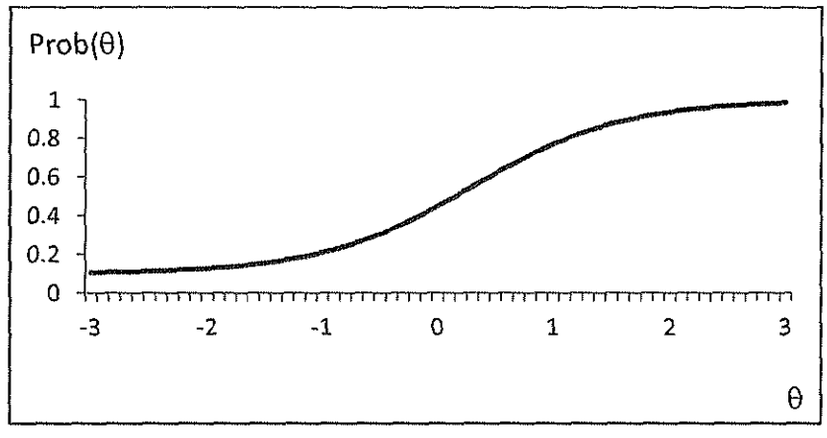


Figure 1. A typical item response function under the Item Response Theory.

It can be seen from the item response function in Figure 1 that the probability of correctly answering an item increases non-linearly as the ability of the test taker increases. The rate of growth or decline approaches zero when approaching the two extreme ends. The probability of correctly answering an item approaches 0 and 1, respectively, when the ability approaching negative infinity and positive infinity. This particular feature of the item response function concurs with the intuition that the chance of answering an item correctly should not be increasing linearly as the ability increases because ability premium is unlikely to bring about any significant gain if it is in much excess. Moreover, the probability being rarely zero or one is also a reasonable modelling of the reality since no matter how competent or incompetent a test taker is, he always has a slim chance of answering an item incorrectly or correctly.

Figure 2 below shows two item response functions with different difficulties while having all other parameters being identical. The solid curve is the same function as shown in Figure 1 and the dashed curve corresponds to another item with a larger difficulty value. The probability of correctly answering the more difficult item is lower for test takers with the same ability value of θ. The difference in the probability is large for mid-range abilities and is small at the two extremes. In general, a difficult item poses challenge to test takers in all range of abilities and it particularly helps a test to differentiate persons in the high ability spectrum.

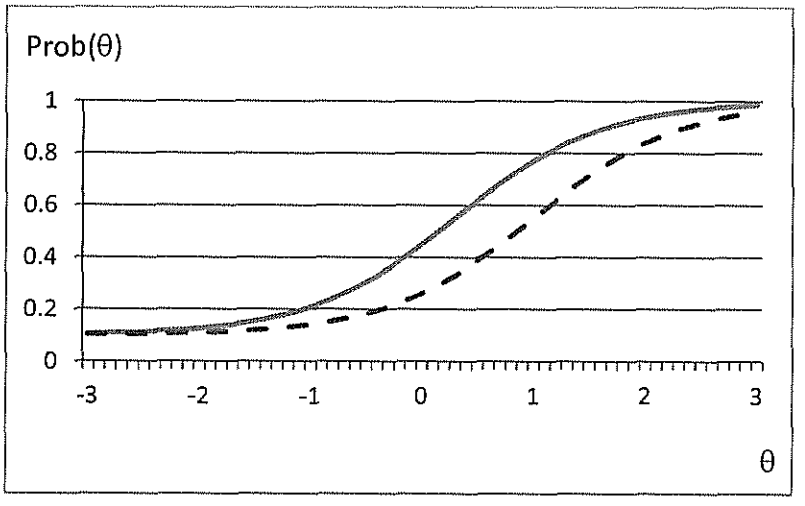


Figure 2. Two item response functions corresponding to items with different difficulties.

Apart from the 3-parameter IRT model given by Equation (1), there are also 1-parameter and 2-parameter IRT models, which can be viewed as special cases of the 3-parameter model. The 2-parameter IRT model contains all the parameters in the 3-parameter IRT model except the guessing factor. This model assumes that only person ability, item difficulty and discrimination factor have roles to play in interacting with each other.

The 1-parameter IRT model, also known as the Rasch model, is the simplest one in the family of IRT models and places emphasis only on item difficulty and person ability. Georg Rasch developed this model in early 1960s and it was designed to analyse response to tests or questionnaires that capture categorical data. Without modelling the extra parameters of item discriminating power and guessing factor, the Rasch model provides a simpler and more straightforward way for psychometricians to study the responses to a

test. As such, we also adopts the Rasch model in the simulation and analysis for our methods.

Both CTT and IRT have their own merits and effective areas in application. However, it is apparently the trend that IRT is attracting more research focuses and uses in recent decades. IRT has been applied in large scale tests both locally in Hong Kong ("*Grading Procedures and Standards-referenced Reporting in the HKDSE Examination*", 2011) and in worldwide assessments such as TOEFL (Tang, 1996). In the Hong Kong Diploma of Secondary Education examination, IRT has been used as a tool to maintain standards of the grading across successive years of examinations ("*Grading Procedures and Standards-referenced Reporting in the HKDSE Examination*", 2011). The application of theory into practice provides researchers with more information and insight about how a model fits the data and vice versa. Subsequently, it has aroused numerous research interests in both the theoretical and practical aspects to further refine the methodologies associated with the models.

2.2 Differential Item Functioning

Differential item functioning (DIF) is by nature a statistical finding that refers to when an analysis is conducted individually for each group of test takers on an item, the statistics would show whether test takers of the same ability but from different groups have different probabilities of endorsing (i.e., correctly

answering) the item. From this definition, in the analysis of DIF, our goal essentially is to estimate the difficulty of a test item, given the responses and group membership of the test takers as inputs. However, caution must be taken in interpreting this simple functional relationship—it is in many cases that the overall ability of the groups to be compared is actually different. This bring in an idea about comparing the comparable, that is, it is only fair to compare groups of test takers with comparable ability or at least to compare a subset of test takers from different groups with roughly the same ability. Simpson's paradox (Simpson, 1951) is a very classic example of ignoring the importance of comparing the comparable. Therefore when conducting DIF analysis, one must take into account the group difference, a.k.a. the impact factor, which primarily indicates whether one group of test taker is with higher ability than the other or not. For instance, suppose we know that there are 6 takers with pre-known ability in the mastery of geometry and their ability values and test scores are shown in Table 1 below, and the average scores of the gender groups of male and female are 60 and 70, respectively. This scenario may be mistakenly labelled as a biased test, since the test is considered favouring the female group. However, it is obvious from the mapping of ability and test scores that there is no evidence of mismatching of scores from ability and no sign of gender bias at all. Therefore, the early method of the direct use of average raw scores is not an appropriate one in conducting analysis for DIF and appropriate measures have to be taken to make sure the ability difference, if any, is rightly compensated before checking for DIF.

Table 1. Ability and test scores of 6 test takers

Person	Gender	Ability in geometry	Geometry test score
1	Female	High	82
2	Female	High	88
3	Female	Mid	70
4	Male	Mid	74
5	Male	Mid	68
6	Male	Low	38

In practice, there can be more than two groups in a classification scheme. For example, the social economic background can be divided into three or more discrete groups. Ethnic group can be divided into Black, White, Asians and Hispanic in some ethnically diverse societies like the United States. However, in the investigation of DIF, a simple grouping system dividing test takers into only two groups is sufficient to serve the purpose of assessing the effectiveness of different DIF detection method. Studies involving 3 or more groups can be generally built on methods designed for 2-group cases with some adaptations (Wang, 2008). In this thesis, DIF analysis is performed based on a 2-group classification.

In a simple 2-group classification, the group under investigation is usually referred to as the focal group and the other group is referred to as the reference group. Although it makes no difference which group is labelled as the focal group or reference group, it is the convention that the minority group is usually classified as the focal group as its performance is often the target of the interest in many education settings.

2.2.1 DIF Detection

Various DIF detection methods have been proposed by researchers with different underlying assumptions (Lord, 1980; Hooland & Thayer, 1988; Thissen, Steinberg & Wainer, 1993). Some of these methods adopt the non-IRT based approaches in which the raw scores of the test are used to classify test takers into different ability groups, while some others adopt the IRT-based approaches in which the estimation of person ability is used instead. Both approaches are adopted in practice. As this research only touch upon the IRT model, discussions hereafter are therefore mainly focusing on IRT-based methods.

Under the IRT models, person ability estimation would, by the model nature itself, takes into account the performance difference in easy and difficult questions. The person ability values—if it is correctly estimated—have already provided a solid ground for comparing persons of similar abilities in the focal and reference groups. With the provision of the estimates of item difficulty, it is easy to determine whether the presence of DIF has any relationship with the hardness of questions. As such, the models under IRT (e.g., the Rasch Model) suits almost perfectly for the studies in relation to DIF analysis.

As previously mentioned, the Rasch Model (Rasch, 1960; Bond & Fox, 2001) is a one-parameter (1-PL) model in the IRT family. Item difficulty and person

ability are estimated in the Rasch Model based on a probabilistic model given by:

$$P(X_{ni}=1)=\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}, \tag{2}$$

where $P(X_{ni})$ is the probability of correctly answering a dichotomous item i with item difficulty β_n for a person with ability δ_i. Both the item difficulty and the person ability can be represented using the same scale known as log odds, or simply logit, which is the natural logarithm of the ratio between the probabilities of correctly answering a question over that of incorrectly answering a question. By Equation (2), the probability of incorrectly answering the item is:

$$P(X_{ni}=0)=1-\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}=\frac{1}{1+e^{(\beta_n-\delta_i)}} \tag{3}$$

The logit is therefore given by:

$$\ln(\frac{P(X_{ni}=1)}{P(X_{ni}=0)})=\beta_n-\delta_i\,. \tag{4}$$

In a simple 2-group scenario consisting of one reference group (Group 1) and one focal group (Group 2), an item will have an individual difficulty value for each group. Algebraically, the logits of an item of the two groups are:

$$\ln(\frac{P(X_{n1i}=1)}{P(X_{n1i}=0)})=\beta_{n1}-\delta_i\text{, and} \tag{5}$$

$$\ln(\frac{P(X_{n2i}=1)}{P(X_{n2i}=0)})=\beta_{n2}-\delta_i\,. \tag{6}$$

The difficulties of item i for person n in Group 1 and Group 2 are β_{n1} and β_{n2} respectively. If the item is not a DIF item, then $\beta_{n1} = \beta_{n2}$. On the other hand, if the item is indeed a DIF item, then $\beta_{n1} \neq \beta_{n2}$. Since there are errors in the estimation of parameters, statistical significance test is required to confirm the DIF analysis when the two beta values differ.

Evaluation criteria. To compare different DIF detection methods fairly, one must establish criteria for assessing their performances. The task of setting up a fair assessment criteria for DIF detection is a straightforward one since there are only four possible outcomes when the analysis is conducted on item level: (i) DIF item is correctly identified as DIF; (ii) DIF item is incorrectly identified as non-DIF; (iii) non-DIF item is correctly identified as non-DIF; and (iv) non-DIF item is incorrectly identified as DIF. Since both the information for outcomes (ii) and (iii) can be deduced from the outcomes (i) and (iv), therefore the assessment criteria covering outcomes (i) and (iv) shall be sufficient. The effectiveness of each approach is usually measured in the terms of *Type-I error rate* and *power* obtained in the eventual DIF detection process.

Type-I error rate is defined as the percentage of non-DIF items that has been incorrectly identified as DIF items. Following the common practice, the percentage of replications that DIF-free items are wrongly classified as DIF is recorded as the Type-I error rate. This measure shows how much error a DIF detection method would make in the detection process. This assessment

criterion is another important measure to understand how possible it is for an 'innocent' item to be wrongly classified as a 'suspect' item. Therefore for this assessment criterion, the lower is generally the better with reference to the statistical significance pre-defined.

Power is defined as the percentage of DIF items out of all the DIF items that have been correctly identified in the DIF detection process. Similar to that of the Type-I error rate, the average percentage of hit over replications is recorded as the power of the overall DIF detection. This measure corresponds to outcome (i) mentioned above. Power is an important assessment criterion to evaluate the performance difference among different DIF detection methods. It shows how effective the method is in successfully identifying those DIF items existed in a test. Since it is a percentage measurement, therefore the result must ranges from 0 to 1, in which 1 means all the DIF items have been successfully found and 0 means none of the DIF items has been found. Therefore for this assessment criterion, a higher value is the better. In case there are no DIF items in a test, this assessment measurement could not be applied.

Simulation Testing. Regarding the test cases to be used for DIF detection method comparisons, the straightforward approach is to make use of real test data (Dodeen & Johanson, 2003; Cauffman & MacIntosh, 2006). This approach is applicable when the study is aimed at finding DIF item in real tests, but is not practical for comparing different methods since the ground truth for

DIF item identification is unknown. On the other hand, computer simulation is an effective approach that can be used to assess the performances of different DIF detection method or anchor selection strategies, with the aid of artificially generated test data with controlled parameters. A number of computer simulation research studies have been conducted on DIF detection and item anchoring (Finch, 2005; Wang, 2008; Shih & Wang, 2009; Wang, Shih & Yang, 2009; Woods, 2009). In our work, computer simulation is also adopted since we focus on assessing the effectiveness of DIF detection methods instead of finding DIF items in real tests.

Table 2. Typical parameter settings for a DIF simulation study

Parameter	Remarks
Reference group ability	● Ability distribution ● Ability mean and spread
Focal group ability	● Ability distribution ● Ability mean and spread
DIF contamination rate	Percentage of items that exhibits DIF
DIF size	The logit difference of DIF items
DIF pattern	Uniform or non-uniform DIF
DIF item distribution	DIF is one-sided or both-sided
Person	● Total number of test takers ● Distribution of test takers between the reference and focal groups

It has been shown that parameter setting of the simulations is an influencing factor to the performance of DIF detection methods. It is therefore extremely important to understand the assumptions or parameter setting behind each simulation studies. In a common simulation study, Table 2 above lists the parameters that can be manipulated which in turn will significantly influence the overall performance of DIF detection methods.

Some methods are effective in identifying DIF items under certain DIF simulation settings. Among the various parameter settings, it is worth noting that DIF contamination rate greatly influences the performance of DIF detection methods. When DIF contamination rate is high, most existing DIF detection methods have a much lower power and a much higher Type-I error rate (Finch, 2005; Wang, 2008; Shih & Wang, 2009; Wang, Shih & Yang, 2009; Woods, 2009). This result is undesirable both theoretically and in actual practice. Effort should be made to make improvements such that the DIF detection would be more accurate regardless of the DIF parameter settings.

2.2.2 Anchor Selection

Despite the fact the aforementioned mathematical ground of DIF detection under the Rasch Model is straightforward and readily implementable, there is one very critical issue that needs attention—the person ability and item difficulty must be correctly estimated. In other words, we must find a set of items, also known as the anchor items, which can serve the purposes of estimating person abilities and item difficulties. Without the anchor items, the mathematical processes described in the previous sub-section could not proceed for DIF detection. In this section, we give a detailed introduction of other research studies in the area of anchor selection, and discuss whether it is possible to streamline some existing anchor selection methods for a better

performance.

There are generally two strategies in finding items to form an anchor set: one is using the test result itself and the other is by referring to external sources of items. When the test result itself is used as the means to calibrate person ability, it is generally known as calibration by the use of internal matching variables. On the other hand, when the other sources of information are used to calibrate person ability, it is known as calibration by the use of external matching variables. It must be noted that the external sources of information or the tests used must be assessing the same dimension of knowledge or skill as the test under DIF study.

When internal matching variables are used, the best scenario will be that all DIF-free items are included as the anchor items and none of the DIF items are included in the anchor set. In such best scenarios, person abilities would be most accurately estimated than in those cases when some of the anchor items are actually DIF. This is because DIF items possess different difficulty values for test takers in the reference and focal groups. If DIF items are included in the anchor set, the calibration process for estimating person abilities and item difficulties would be based on an invalid assumption that items are non-DIF, and subsequently both person abilities and item difficulties would be incorrectly estimated. When external matching variables are available, then the person calibration of this test can be used as the fixed parameter. Unfortunately this kind of clean person ability information is rarely available in reality.

Generally, using internal matching variable is more practical and less restrictive than using external matching variables. As a result, many practical DIF detection methods employ calibration using internal matching variables. Commonly used internal matching variables include the total score of the target test (Osterlind & Everson, 2009), or the scores of a subset of test items. This research work focuses only on the use of internal matching variables for anchor formation.

There are substantial amount of publications regarding DIF detection methods and anchor selection strategies. Wang (2008) and Kopf, Zeileis and Strobl (2013) provided a good starting point for the discussion on DIF detection methods, particularly in the area of anchor selection that this thesis is focused on. Researchers have recommended different approaches (Wang, 2004; Wang, 2008; Wan & Yeh, 2003; Woods, 2009) in locating DIF-free items from a test filled with both DIF and DIF-free items. Wang (2008) summarized that there are generally three major methods in DIF detection and finding internal matching variables (anchors): (1) The Equal-Mean-Difficulty (EMD) Method; (2) The All-Other-Items (AOI) Method; and (3) The Constant Item (CI) Method. The EMD method assumes equal mean item difficulties across groups. While each item under study may have different difficulty values for different groups, the average item difficulty values of all the items are identical to all the groups under study. This method performs perfectly when either the test is DIF-free, or some items favour the reference group and some favour the focal group and

the difference exactly cancels out each other. On the other hand, the AOI method and its variant AOI-SP method, as well as the CI method are two frequently used families of anchor selection strategy, which attracted more researches on them. Our method is inspired by these two strategies, which will be discussed in more details below.

AOI and AOI-SP methods. The AOI method is probably one of the earliest anchor selection strategies for DIF detection purposes. When determining whether an item exhibits DIF or not, it treats all the other items in the test as the anchor. The obvious problem of this method is that whenever there is a DIF item in the test, the calibration of item difficulty and person ability is destined to be erroneous. In other words, it works perfectly for all items only when there are no DIF items in the test. To ease this obvious hazard, the AOI with scale purification (AOI-SP) method was proposed. Using the AOI method as the starting point, there is an iterative process built in the AOI-SP method for purifying the anchor set. At the beginning of the purifying process, all items are regarded as DIF-free and a DIF analysis is conducted to identify the first batch of DIF items. This first batch of DIF items is then excluded from the anchor set for the calibration of the next iteration of purification. The second batch of DIF items are then identified with the new set of anchor items. The process repeats until no more DIF items can be identified. The advantages of this purification procedure include the simplicity of its design and implementation, and also its effective computations in locating DIF items. However, one of its major shortcomings is that the criteria and the procedure to

label an item as DIF are not sufficiently sophisticated and therefore some DIF-free items may wrongly be classified as DIF items in the purification procedure—particularly under highly DIF contaminated scenarios. The selection procedure is not sufficiently sophisticated in the sense that the whole procedure is a single-test based process in which the result of one test is used to determine the likelihood of whether an item is DIF or not. In other words, an DIF-free item could be labelled as DIF or an DIF item could be labelled as DIF-free with the result of a single DIF analysis only. Therefore it is particularly error prone when the target test contains a high percentage of DIF items.

CI method. The initial idea of the CI method (Wang, 2008) is to use not all other items but only one single item as the anchor for DIF detection. The method treats each item in turn as the anchor and then finds out the DIF statistic of all the other items. When this process is over, the sum of DIF statistic values over iterations for each item is computed. The heuristic used in the CI method is to select those items with the lowest DIF sum values to form an anchor set with a pre-determined length. The underlying assumption of this heuristic is that when an DIF-free item is tested against each of the other items in the test, its DIF statistic value sum is likely to be smaller than those of the DIF items in the same test. An intuitive advantage of the CI method is its conservativeness in including items in an anchor set, which makes it less likely to bring DIF items into the final anchor set. The disadvantages of the CI method are its relatively more intensive computation needed in determining the

likelihood, and also the potentially lower precision in calibration as it takes fewer items for estimating item difficulties and person abilities.

As a related remark, the aforementioned anchor selection strategies are available from some well-developed tools. EMD method is implemented in the ConQuest software and the AOI method is adopted in the WINSTEPS software. The CI-method is similar to the AOI-method as both of them use a subset of items from the target test to serve as the anchor. As a result, the CI method can also be implemented using WINSTEPS.

There are previous simulation studies on assessing the effectiveness of anchor selection strategies (Finch, 2005; Wang, 2008; Shih & Wang, 2009; Wang, Shih & Yang, 2009; Woods, 2009). These results show that generally the AOI method is with a high Type-I error rate and moderate power, the AOI-SP method with a better Type-I error rate and high power, and the CI method with an excellent Type-I error rate and low power. Hence, there is an obvious area for improvement: a new anchor selection strategy with a better-controlled Type-I error rate as well as a high power when comparing with the existing ones.

In general, the selection of anchor is a matter of choosing items which are most likely to be DIF-free, or in other words, of finding a most appropriate method to rank the items according to the perceived probability of being DIF-free. The AOI and the CI methods actually represent the two ends on a continuum of the

number of items to be included in the final anchor set. For instance, in a test with 40 items, it is possible to use 1 up to 39 items to form the anchor set in the DIF analysis for any single item. In such an example, the AOI method uses 39 items and the CI method may use 1 item only depending on the anchor length specified. In fact, almost all methods related to anchor selection focus on finding a right balancing point between including all the other items and using just one item.

One recent study by Woods (2009) has brought anchor selection into further consideration and some positive results has been obtained. In her study, Woods used IRT-LRT test to detect DIF items and the strategy was to compute and to rank the LR to f (where f is the number of free parameters) ratio of the items. Those items with smallest LR/f ratio, as claimed by Woods, would likely to be DIF-free and be served as the anchor. Woods did not conduct any significance testing to classify DIF or non-DIF item while only LR/f ratio was used as the metric for selecting the anchor. The Type-I error rate obtained in her study was extremely low, which was probably due to the fact that short anchors were employed (at most 20% of the total number of items). Furthermore, the sample size was also relatively high at R500/F1500, which also probably led to high power in her result. This research study hints that there is a room for further streamlining different variations of the AOI-SP method. One thing that Woods did not investigate in her study is the case when the sample size is smaller, which is the practical scenario in most class-based and level-based test environments in a school.

From the above discussions, it is evident that more research work is warranted in the area of anchor selection strategy so as to improve the overall DIF detection effectiveness and preciseness. The ultimate goal is to keep Type-I error rate within control and boost the power at the same time. Previous research studies in this field bring an important clue for the direction of further improvement, which is to combine the robustness of keeping a well-controlled Type-I error of the CI method with the high power of the AOI-SP method.

CHAPTER 3

ALL-OTHER-ITEM SCALE PURIFICATION WITH CONSTANT ITEM ANCHOR SELECTION

As discussed in the preceding chapter, previous research studies on DIF detection have laid a solid ground for test developers to have better confidence in using different DIF detection methods, and they also create a space for researchers to further investigate how different DIF detection methods can be refined or in what the directions may new methods be pursued or developed. In this research study, before reaching a final proposal of a new anchor selection strategy, some preliminary ground work has been done in making adaptation to some existing methods. We first set up a platform for making comparison possible by implementing the AOI, AOI-SP and CI methods with a computer simulation program. We then assess whether it is rewarding in terms of DIF detection performance by directly making some changes to the AOI-SP and CI methods.

3.1 Adaptation of Existing Methods

For the AOI-SP method, since the power of AOI-SP is generally high whereas

the Type-I error rate is not well controlled in some simulation settings, our adaptation to the AOI-SP method is to further shorten the anchor length on the basis that doing so can lower the Type-I error rate in many cases. The selection criterion for the anchor set is based on the absolute DIF value contrast of an item between two groups after the scale purification process has been done—items with smaller DIF contrasts are of lower priority to be excluded from the anchor set. The assumption behind this selection criterion is that larger DIF difference could be a manifestation of a higher chance of being DIF. We name this strategy as AOI-SP with Anchor Selection, abbreviated as AOI-SP-AS or AS thereafter. Different proportions of anchors are selected from the purified set of anchor items. We consider five levels of anchor length, ranging from 20% to 80% of the length of the original purified anchor set obtained by the original AOI-SP method. We name the respective methods as AS-*z*, where *z* is the anchor length in percentage. For instance, AS-60 refers to an anchor set containing 60% of the items which are of the lowest DIF contrast in the original purified anchor set.

On the other hand, the CI method, which uses a short anchor, produces a well-controlled Type-I error rate though its power is not high. Hence, our adaptation to the CI method is to lengthen its anchor set instead. According to Wang (2008), the CI method selects the item(s) with the lowest DIF amount estimates after running the iterative procedure in computing the DIF amount. In this experiment with a simulated test of 20 items, four different anchor set lengths (2, 4, 8, 16 items) are considered, and we name the methods as CI-*x*

where x is the number of items in the anchor set.

3.2 Simulation Results

Our simulation program directly uses the implementation of the AOI-SP method provided by WINSTEPS, which follows the algorithm proposed by Lord (1980), to obtain the DIF detection result. For the CI method, the algorithm proposed by Wang (Wang, 2008) is implemented. All the DIF detection statistics are generated using WINSTEPS which supports item calibration with the Rasch Model.

We compare 9 DIF detection methods, AOI-SP, AS-80, AS-60, AS-40, AS-20, CI-2, CI-4, CI-8 and CI-16, under different settings with varying group sizes and DIF contamination rates. Three different group sizes are considered: R500/F500, F500/F100 and R1000/F500, where R denotes the reference group and F denotes the focal group. For each group size, we consider six levels of DIF contamination rate: 0%, 10%, 20%, 30%, 40%, and 50% of the total number of items, which determines the number of items to be designated as DIF. For those DIF items, different difficulties are assigned to the reference group and the focal group. Furthermore, we assume that there are 20 items in total for DIF detection for each setting and one hundred replications are executed for each individual simulation setting. Over that one hundred replications, the type-I error rate is computed by taking the average of false positive cases

counted for each DIF-free items, and, the power is computed by taking the average of the number of DIF-hit cases for each DIF item.

Table 3. Type-I error rates of DIF detection methods

Ref/ Foc	DIF item %	Type-I Error Rate								
		AOI-SP	AS-80	AS-60	AS-40	AS-20	CI-2	CI-4	CI-8	CI-16
R500/	0	0.06	0.07	0.06	0.07	0.06	0.04	0.06	0.07	0.07
F500	10	0.06	0.06	0.06	0.07	0.06	0.04	0.05	0.05	0.05
	20	0.06	0.07	0.06	0.05	0.06	0.03	0.06	0.05	0.06
	30	0.07	0.06	0.06	0.07	0.03	0.05	0.07	0.09	0.11
	40	0.10	0.09	0.08	0.06	0.03	0.05	0.11	0.18	0.33
	50	0.38	0.36	0.34	0.26	0.07	0.12	0.37	0.54	0.63
R500/	0	0.05	0.06	0.06	0.06	0.05	0.04	0.06	0.07	0.07
F100	10	0.06	0.07	0.06	0.06	0.05	0.03	0.04	0.05	0.07
	20	0.06	0.06	0.07	0.07	0.05	0.03	0.05	0.06	0.07
	30	0.08	0.08	0.08	0.07	0.04	0.03	0.08	0.10	0.09
	40	0.13	0.13	0.12	0.11	0.06	0.03	0.08	0.12	0.16
	50	0.28	0.25	0.24	0.21	0.09	0.05	0.17	0.21	0.24
R1000/	0	0.04	0.06	0.06	0.06	0.05	0.04	0.05	0.06	0.06
F500	10	0.06	0.06	0.06	0.06	0.06	0.03	0.05	0.05	0.05
	20	0.06	0.07	0.06	0.06	0.05	0.03	0.05	0.06	0.06
	30	0.07	0.07	0.08	0.08	0.04	0.05	0.09	0.10	0.10
	40	0.09	0.08	0.08	0.06	0.03	0.05	0.08	0.18	0.38
	50	0.28	0.29	0.24	0.18	0.04	0.15	0.43	0.60	0.72

Table 3 above lists the Type-I error rates of the DIF detection methods obtained in the simulation run. It shows that in the baseline cases, i.e., those with no DIF item, the Type-I error rate for these DIF-free datasets is around 0.04 to 0.07, and mostly in between 0.05 and 0.06. Under all the simulation settings, the Type-I error rate increases as the percentage of DIF items increases regardless of the ratios of the reference and the focal group sizes, and both methods lose

control on it when the DIF contamination rate is 30% and beyond. The CI method with short anchor generally has a better-controlled Type-I error rate. For the AOI-SP-AS method, the Type-I error rate also increases as the anchor size increases, which is similar to how the CI method behaves.

Table 4. Power of DIF detection methods

Ref/ Foc	DIF item %	Power								
		AOI-SP	AS-80	AS-60	AS-40	AS-20	CI-2	CI-4	CI-8	CI-16
R500/	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F500	10	0.98	0.98	0.98	0.96	0.90	0.60	0.89	0.97	0.98
	20	0.99	0.98	0.98	0.96	0.85	0.58	0.85	0.96	0.97
	30	0.98	0.99	0.98	0.94	0.70	0.59	0.88	0.96	0.97
	40	0.96	0.97	0.95	0.90	0.66	0.50	0.82	0.88	0.86
	50	0.67	0.69	0.68	0.64	0.46	0.35	0.54	0.59	0.61
R500/	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F100	10	0.70	0.73	0.73	0.68	0.52	0.23	0.49	0.65	0.70
	20	0.67	0.69	0.66	0.64	0.48	0.22	0.52	0.62	0.65
	30	0.59	0.64	0.62	0.58	0.37	0.21	0.45	0.55	0.57
	40	0.45	0.52	0.50	0.48	0.27	0.19	0.38	0.44	0.41
	50	0.23	0.30	0.31	0.31	0.22	0.14	0.23	0.28	0.26
R1000/	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F500	10	1.00	1.00	1.00	1.00	0.93	0.68	0.92	1.00	1.00
	20	0.97	0.97	0.97	0.96	0.87	0.65	0.91	0.96	0.96
	30	0.99	1.00	0.99	0.98	0.79	0.72	0.94	0.99	1.00
	40	0.99	0.99	0.99	0.96	0.78	0.66	0.91	0.97	0.93
	50	0.82	0.81	0.79	0.78	0.74	0.51	0.67	0.74	0.74

The power of each DIF detection method is shown in Table 4 above. It can be observed that AOI-SP performs well in locating the DIF items, particularly when the DIF contamination rate is low. When a subset of items is chosen from the AOI-SP anchor, the power generally drops as the anchor set becomes shorter. It is worth to note that in some scenarios, a shorter anchor set results

in a better power, though the magnitude is small. For instance in the R500/F100 case, the power of AS-80 and AS-60 is slightly higher than that of the AOI-SP method. It can be observed that AS-80 and AS-60 in many cases perform better than the original AOI-SP method. This provides important evidence that improvement is possible to the AOI-SP method when a suitable set of short anchor can be formed. In the case of the CI method, it can be observed that a longer anchor usually yields a better power. While comparing the power of AOI-SP-AS method settings with that of the CI method settings, the figures in Table 4 show that the former one has a better potential in achieving a higher power. However, it must be noted that the anchor length of AOI-SP-AS method is rarely identical to that of the CI method, therefore a direct comparison is not completely appropriate or fair in this sense.

3.5 Observations and Insights

From the simulation results obtained, it can be observed that there exists an association between the power and the Type-I error rate—when the power is high so is the Type-I error rate and vice versa. It seems that the space for further improving both the power and Type-I error rate is limited. However, the performance of AOI-SP-AS method over the AOI-SP method shed some light on maintaining high power but controlling the Type-I error rate at the same time. The key issue is about the accuracy of the anchor set, that is, the proportion of DIF-free items that are being included in the anchor set. Table 5

below shows the anchor selection accuracy of the DIF detection methods under different simulation settings.

Table 5. Anchor selection accuracy of DIF detection methods

Ref/ Foc	DIF item %	AOI-SP Avg Anchor Items	Anchor Selection Accuracy								
			AOI-SP	AS-80	AS-60	AS-40	AS-20	CI-2	CI-4	CI-8	CI-16
R500/	0	18.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F500	10	17.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	20	15.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.94
	30	13.09	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.97	0.82
	40	11.15	0.97	0.99	1.00	1.00	1.00	0.93	0.91	0.81	0.67
	50	9.52	0.65	0.66	0.65	0.66	0.66	0.46	0.47	0.48	0.50
R500/	0	18.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F100	10	17.56	0.97	0.98	0.99	1.00	0.99	0.98	0.99	0.98	0.96
	20	16.32	0.92	0.96	0.97	0.99	0.99	0.97	0.95	0.94	0.88
	30	15.34	0.84	0.90	0.94	0.95	0.94	0.88	0.90	0.86	0.75
	40	14.83	0.70	0.77	0.80	0.84	0.86	0.73	0.74	0.71	0.63
	50	14.91	0.52	0.52	0.53	0.55	0.58	0.48	0.48	0.49	0.49
R1000/	0	19.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F500	10	16.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	20	15.13	1.00	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.98
	30	13.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.84
	40	11.05	0.99	1.00	1.00	1.00	1.00	0.99	0.95	0.86	0.68
	50	8.97	0.80	0.77	0.77	0.77	0.90	0.53	0.53	0.52	0.50

It can be observed from Table 5 that when using the AOI-SP-AS method, the anchor selection accuracy maintains at a high level when compared with that of the AOI-SP method. In the scenarios when the DIF contamination rate is high, the anchor selection accuracy of the AOI-SP-AS outperforms the AOI-SP counterpart. However, when the DIF contamination rate reaches 50%, the anchor selection accuracy drops significantly, particularly if the sample size is

not large. The advantage of a high accuracy in selecting DIF-free anchors does not guarantee a rise in the power, but a drop in the Type-I error rate is envisaged. Anyhow, the power in most experimental settings is maintained to a comparable level with that of the AOI-SP method.

The results of the simulation studies conducted have provided promising empirical evidence that there are rooms for further improving the performance of certain well-established DIF detection methods. The possibilities lay in the area where we can make use of the advantage of high power of the AOI-SP method and the advantage of better controlled Type-I error of the CI method at the same time.

CHAPTER 4

ITERATIVE RANDOMIZED CONSTANT ITEM ANCHOR SELECTION

As pointed out by previous research studies mentioned in preceding chapters, and also indicated by the background investigation we conducted in CHAPTER 3, the most crucial process in DIF detection is the selection of the right items to form an anchor, which in turn is used for item and person calibrations under the Rasch Model. It has been discussed in details in CHAPTER 2 the prerequisites for different DIF detection methods to work perfectly well. However, the prerequisites, such as 'equal-mean-difficulty between groups' and 'all-other-items are all DIF-free', are unlikely to be the common case in reality. Therefore, efforts must be focused on how to select the right items to form an anchor even when the prerequisites cannot be satisfied.

The aims of this study are to (1) devise a new anchor selection strategy for DIF detection; and (2) assess the performance of the newly devised strategy by comparing it with some existing common anchor selection strategies for DIF detection. Devising a new anchor selection strategy is largely about finding a way to rank the items in a test according to their perceived probability of being DIF-free. This involves trying different approaches in filtering out potential DIF items and identifying candidate DIF-free items at the same time. In so doing, we seek a new anchor selection algorithm that is able to generate a

reasonably-long and accurate DIF-free set such that it can help to elevate the power and control the Type-I error rate in the final DIF detection process.

Our fundamental idea for devising a new anchor selection strategy for DIF detection is to use the AOI-SP method as the basis while incorporating also the concept of having a short anchor as featured in the CI method. A key issue is how to combine the strengths of these two methods effectively. It seems, nevertheless, unavoidable to make some wrong choices or judgements in the purification procedures. Even when it is affordable to exhaust all the possible combinations of partition of items, there will be no guarantee that DIF-free items will not be classified as DIF items in the purification process. Therefore, we aim at looking for a better heuristic which can lower the chance of picking the wrong anchor rather than finding out a solution that can work perfectly in all cases. In other words, this work investigates whether there exists an optimum figure in between 'one-item' and 'all-the-other-item' such that a better purified anchor can be formed.

4.1 The Algorithm

In view of the general phenomenon observed in the scale purification procedure of the AOI-SP method, it is desirable to further strengthen the scale purification procedure such that the chance of incorrectly labelling DIF-free items as DIF items can be lowered. It has been found in the simulation studies of the CI

method (Wang, 2008) that a series of examinations on the item set with different anchors could, to certain extent, increase the chance of locating the correct DIF items. It means that such a process may help in lowering the chance of excluding DIF-free items from the final anchor set. In the DIF item detection process of the CI method, it takes n-1 tests, where n is the number of items, to determine whether an item is likely to be a DIF item. Following that gist of the CI method, the purification procedure of the AOI-SP method can be modified in such a way that more than one single test is involved to determine which items are to be excluded from the anchor set.

We propose a novel anchor selection scheme, called the Iterative Randomized Constant Item (IRCI) method, for DIF detection. We adopt a multi-round purification process similar to the AOI-SP method such that a longer purified anchor set can be obtained. At the start of the algorithm, the candidate anchor set contains all the items in a test. In each iteration of the purification, an item with the highest DIF likelihood is identified and removed from the candidate anchor set. However, in deriving the DIF likelihood of an item, instead of using all other items as the anchors, we borrow the idea of the CI method and conduct multiple DIF test on an item against a selection of 2-item anchors. Specifically, suppose a candidate anchor set has n items, then we randomly obtain a collection of $\lfloor n/2 \rfloor$ groups of 2-item anchors from the candidate anchor set. Each item in the candidate anchor set is then subject to $\lfloor n/2 \rfloor$ DIF tests, each against one of the 2-item anchors. Hence, at the end of each iteration, each item has a total DIF-positive count ranging from 0 to $\lfloor n/2 \rfloor$

representing its likelihood of exhibiting DIF. To be conservative, the iterative anchor selection strategy excludes only the item with the highest DIF likelihood, i.e., the item having the largest DIF-positive count, from the candidate anchor set. The iterations continue until no more DIF item is identified and the resulting anchor set is thus obtained. The whole procedure is a computational demanding process. Its computational complexity is of $O(n^3)$ when using the big-O notation, where n is the number of items. The entire purification procedure of the IRCI method is stipulated in the flowcharts on pages that follow.

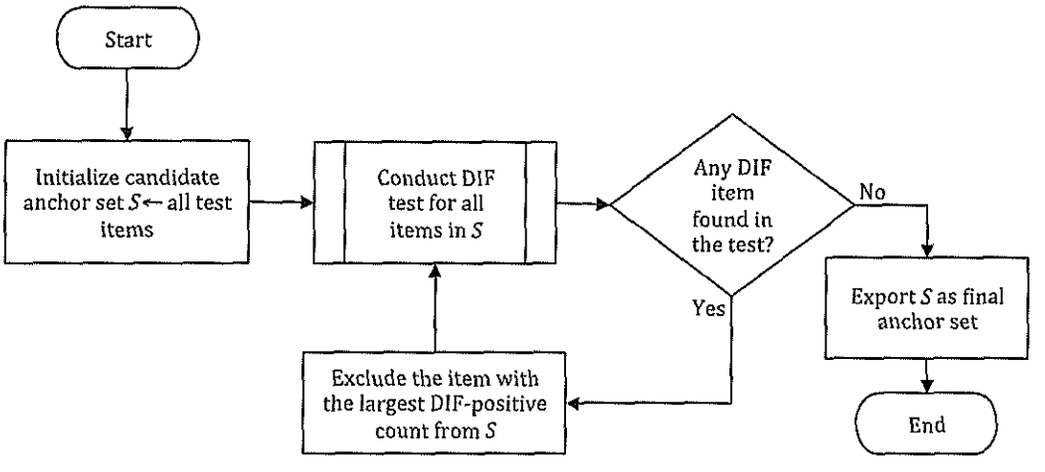


Figure 3(a). The main flowchart for the iterative randomized constant item method

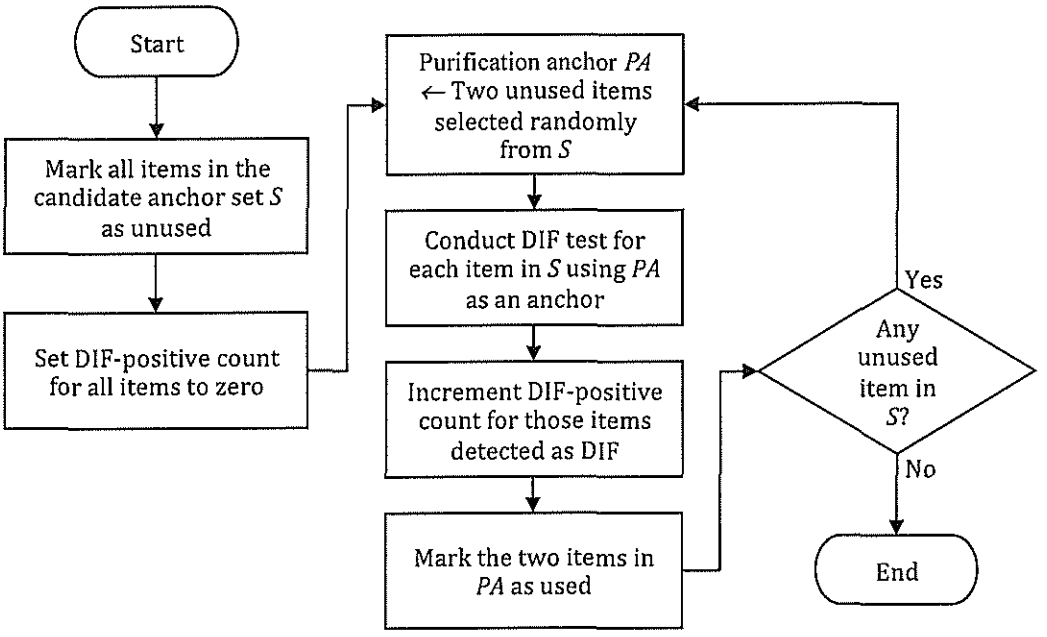


Figure 3(b). The sub-flowchart for "*Conduct DIF test for all items in S*" process in the main flowchart

As depicted in the flowchart in Figure 3(b) above, there is a level of iteration for multiple DIF testing of an item using randomly selected two-item anchor for the purification process. This is also the reason why the method is named as iterative randomized constant item (IRCI) anchor selection. In the end of this iterative process, each item in the current candidate anchor set is associated with a total DIF-positive count representing its likelihood of being DIF.

It can be seen that the IRCI method is a combination of the approaches taken by the AOI and the CI methods. We adopt the scale purification technique of the AOI method, while introducing multiple DIF testing with short anchors similar to that of the CI method for more reliable filtering of possible DIF items from the anchor set. Using too short an anchor increases the computational

complexity for the purification process (e.g. n-1 DIF tests are needed for each item when using a single-item anchor), and hence we opt to use a 2-item anchor for the purification process, as a balance between the computational complexity and the gain obtained in better controlling the Type-I error rate.

4.2 Simulation Environment

A computer program is written for generating test data and assessing the performance of the new anchor selection strategy for DIF detection. As discussed in the earlier chapters, by the merit of the scalability in generating different amount of test cases with different parameters, computer simulation is a very effective way in assessing the performance of different DIF detection methods. What is more crucial is that computer simulation is also fairer in comparing different DIF detection methods as identical test cases can be applied to each of these methods. In this computer program for the simulation studies, the AOI, the AOI-SP, the CI methods and the newly designed IRCI anchor selection strategy for DIF detection have been implemented. We use the same implementation for the AOI-SP and CI methods as in Chapter 3. Specifically, we use the AOI and AOI-SP methods provided by WINSTEPS and we implemented the algorithm proposed by Wang (Wang, 2008) for the CI method. The number of constant items is set to be 4 as suggested in Wang's paper (2008). The programming environment used for the development is Microsoft Visual Studio 2010 with Qt interface tools. WINSTEPS 3.70.0.2 is

used to assess the item DIF in all the simulations.

4.3 Simulation Test Generation

In our simulation, difficulty values of test items and the person ability values of test takers are randomly generated variables. Both of these two sets of values are in the logit scale defined under the Rasch Model. The artificial responses to the items are generated using also the Rasch model according to Equation (2) with the aid of pseudo-random number generating function provided in the standard library of the computer programming language C.

The test length is initially set at twenty items and these items are administered to both the reference group and focal group in the simulation runs. Two independent variables are manipulated in this research study: (1) the sample size of the reference group (R) and the focal group (F)—R500/F100, R500/F500, and R1000/F500; and (2) the percentage of DIF items in the test—0%, 10%, 20%, 30%, 40%, and 50%. We compare the newly designed anchor selection strategy, namely the Iterative Randomized Constant Item (IRCI) in anchor selection, with the AOI, AOI-SP and CI-4 (i.e., the CI method using 4 items as the anchor) DIF detection methods. The CI-4 method is chosen because it was recommended by Wang (Wang, 2008) that an anchor size of 4 items should be sufficient. To have a more reliable result for comparison, 100 replications of the simulation are conducted to each individual setting of

independent variables. The average result of the 100 simulations is computed to evaluate the performance of different anchor selection strategies for DIF detection. Item difficulty is ranged from -3 logit to 3 logit. Random normal sampling technique is applied in the item difficulty generation process. For the reference group, the sum of difficulties of all the items is constrained to be 0 in order to meet the system requirement of WINSTEPS. In this work, one-sided DIF pattern is studied primarily. DIF items are uniformly set to have a 0.8 logit advantage favouring the reference group with all the other items being DIF-free. The logit difference of 0.8 follows the simulation setting in a similar study (Wang, 2008) such that further comparison will be possible. Nevertheless the effect of different logit differences, 0.6 logit and 1.0 logit, on the strategy is also being studied in this research. We conduct extended studies with varying amount of DIF difference as well. On the other hand, person ability is also sampled from -3 logit to 3 logit according to the normal distribution. The probability of correctly endorsing an item follows Equation (2) mentioned in CHAPTER 2. To assess the significance of DIF, a *t*-test on the estimated item measures between the focal group and the reference group is conducted on each item. This statistics test is equivalent to the IRT-D2 test purposed by Lord (1980). The alpha level is set at 0.05 for the statistical significance level of classifying an item as with DIF.

4.4 Performance

Tables 6 and 7 below show the performance of the Iterative Randomized Constant Item (IRCI) method, in comparisons with the other scale purification methods.

Table 6. Type-I error rates of DIF detection methods

Ref/ Foc	DIF item %	Type-I Error rate				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
R500/ F500	0	0.053	0.053	0.053	0.057	0.066
	10	0.054	0.074	0.058	0.067	0.061
	20	0.052	0.178	0.064	0.062	0.060
	30	0.058	0.354	0.056	0.069	0.064
	40	0.049	0.542	0.063	0.077	0.052
	50	0.038	0.718	0.176	0.358	0.051
R500/ F100	0	0.047	0.047	0.049	0.052	0.061
	10	0.047	0.049	0.053	0.059	0.055
	20	0.047	0.092	0.061	0.064	0.064
	30	0.052	0.138	0.079	0.066	0.068
	40	0.064	0.233	0.134	0.090	0.074
	50	0.034	0.311	0.201	0.156	0.064
R1000/ F500	0	0.043	0.045	0.048	0.062	0.049
	10	0.038	0.070	0.052	0.054	0.055
	20	0.045	0.206	0.055	0.067	0.064
	30	0.051	0.436	0.058	0.054	0.066
	40	0.045	0.659	0.064	0.076	0.066
	50	0.048	0.852	0.304	0.529	0.096

Table 7. Power of DIF detection methods

Ref/ Foc	DIF item %	Power				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
R500/ F500	0	N/A	N/A	N/A	N/A	N/A
	10	0.985	0.980	0.990	0.930	0.990
	20	0.990	0.970	0.990	0.950	0.990
	30	0.990	0.945	0.988	0.965	0.990
	40	0.970	0.851	0.965	0.927	0.969
	50	0.974	0.750	0.912	0.793	0.973
R500/ F100	0	N/A	N/A	N/A	N/A	N/A
	10	0.775	0.715	0.740	0.590	0.750
	20	0.812	0.675	0.743	0.632	0.775
	30	0.815	0.540	0.672	0.573	0.777
	40	0.850	0.468	0.637	0.596	0.812
	50	0.805	0.309	0.352	0.434	0.736
R1000/ F500	0	N/A	N/A	N/A	N/A	N/A
	10	0.985	0.980	0.980	0.950	0.980
	20	1.000	0.980	0.998	0.970	1.000
	30	0.988	0.972	0.988	0.970	0.987
	40	0.976	0.917	0.976	0.963	0.975
	50	0.995	0.862	0.958	0.842	0.993

We have also included a baseline case which serves as a "ground truth" for comparisons. The baseline case tells what the performance would be if we deliberately include only all the DIF-free items in the anchor set for the simulation runs. This setting is actually identical to an ideal situation in which the scale purification process has successfully eliminated all the DIF items while keeping all the DIF-free items as the anchor items. It is therefore reasonable to assume that the baseline case in general has the best case performance and the other methods normally cannot perform as well as the baseline case.

Table 6 and Table 7 show the simulation results with three different parameter settings for the test taker counts of the focal group and reference group, namely,

R500/F500, R500/F100 and R1000/F500, respectively. Different settings are adopted in order to assess how each method performs when the data size varies. The data size is determined by the total number of test takers as well as how they are distributed between the focal group and the reference group. All three settings emulate a mid-size test in which there are around 600 to 1500 test takers in total. The distribution between groups varies from 50%-to-50% which is similar to gender division, to 16%-to-84% which is more similar to ethnic group division.

The results shown in Table 6 indicate that the AOI, AOI-SP and CI-4 methods perform worse than the baseline case and the IRCI method, in terms of the Type-I error rate. For the baseline case, the Type-I error rates are well controlled within 0.052 under different DIF contamination settings. This again confirms that well controlled Type-I error rate is achievable if the right set of DIF-free anchor are selected. For the IRCI method, the Type-I error rates range from 0.055 to 0.074. Although these figures are not as good as those of the baseline case, it is still a decent performance with false-positive cases being kept at a very well controlled level. On the other hand, the false-positive cases for the other three methods increase as the DIF contamination rate increases. The Type-I error rates rise to 0.311, 0.201 and 0.156, respectively, for the AOI, AOI-SP and CI-4 methods. Figure 4 below shows the performance of the DIF detection methods in terms of Type-I error rate for the R500/F100 case. It is evident that the AOI, AOI-SP and CI-4 methods are sensitive to DIF contamination rate while the baseline case and the IRCI method can maintain a

more or less constant performance regardless of the contamination factor. Overall speaking, when compared with the other three DIF detection methods, the IRCI method is more reliable in identifying the DIF items in a test as is reflected by its better controlled Type-I error rate.

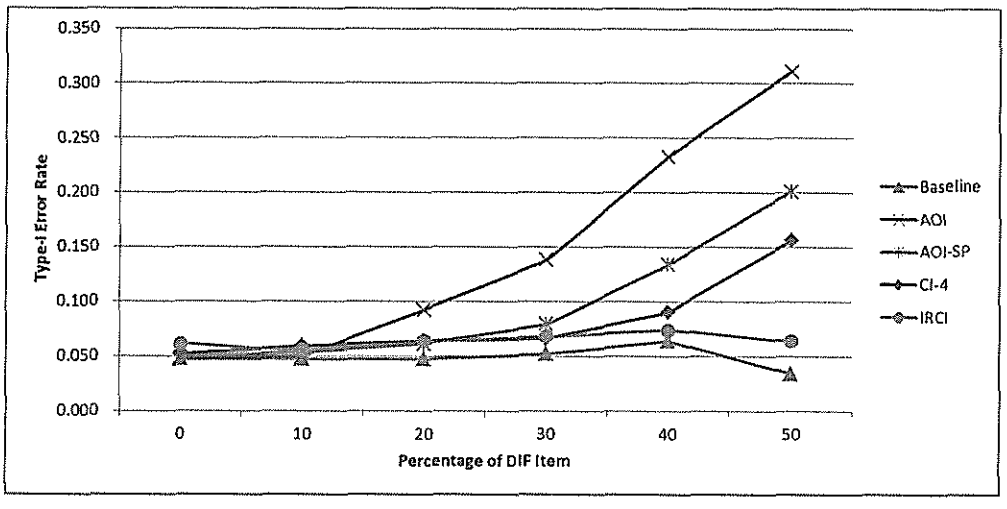


Figure 4: Type-I error rate of DIF detection of various methods (R500/F100)

It is shown in Table 7 that the power of DIF detection methods is higher when the data size is larger. This happens probably because with more data, on the one hand, data response can match more closely to the pre-defined item difficulties and person abilities; and on the other hand, it can also increases the accuracy and reliability of the statistical computations. The performance for the R500/F500 case is close to that of the R1000/F500 case. This may due to the fact that the data size is sufficiently large enough when there are 1000 test takers altogether or the distribution of half-to-half helps in providing better statistical results. Nevertheless, we mainly focus on the smaller data size setting since it is, firstly, more closely to school settings in which normally

there are only hundreds of students at a level; and secondly, there is a larger room for improvement on the performance for smaller data size scenarios.

Simulation results shown in Table 7 also reveal that as the DIF contamination rate increases, the powers of all the DIF detection methods generally decreases. For the baseline setting, the power maintains at a very consistent figure at around 0.98 for the two larger data sets of R1000/F500 and R500/F100. While for the target dataset of R500/F100, the power is much lower at around 0.80. This result shows that when the data size is smaller, the accuracy of detecting DIF items in a test will be lower even it is known in advance which items actually are DIF-free. Contrary to larger datasets, smaller datasets provide less data to confirm statistically whether an item exhibits DIF or not.

Among the performance of all four DIF detection methods shown in Table 7, the AOI method is the worst one while the other three methods with different approaches of scale purification could indeed boost the overall DIF detection power. As the DIF contamination rate increases from 10% to 50% of the total items in the R500/F100 setting, the power of the AOI method drops significantly from 0.715 to 0.309. In other words, in the dataset consisting of 20 items, there are approximately 2 to 7 DIF items which cannot be identified by the AOI method successfully under different contamination settings. The unsuccessful rate of the AOI method maintains high throughout low to high DIF contamination scenarios. This result reflects that if the anchor set contains too many DIF items, regardless of the relative proportion to the DIF-free items, the

ability of the AOI method to correctly identify DIF items in the test is severely hindered. The performance of the AOI-SP method is better than the AOI method by a 3% to 36% increase in the power of DIF detection. This result concurs with the theory that purified anchor sets with fewer DIF items can improve DIF detection performance. The performance of the CI-4 method, however, fluctuates under different contamination settings. The power of DIF detection ranges from 0.434 to 0.632, though the figures are slightly better the AOI method while generally not as good as the AOI-SP method. This results shows that there are drawbacks in terms of power when a fixed-length anchor is employed to DIF detection regardless of test length as well as the degree of DIF contamination. For the IRCI method, its performance is significantly better than the other three methods and is actually the closest to that of the baseline case. For the R500/F100 setting, the power of the IRCI method ranges from 0.731 to 0.812. Its performance is satisfactory under different DIF contamination settings and does not degrade significantly in highly-contaminated cases as observed in the other three methods.

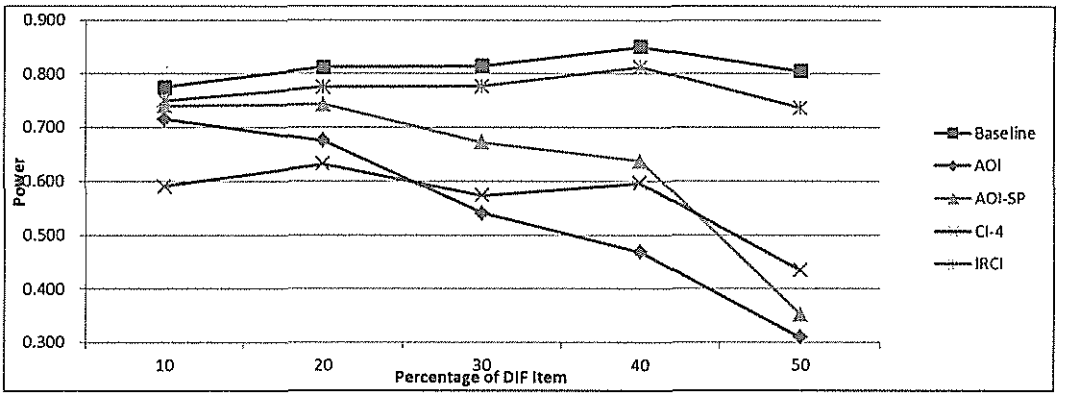


Figure 5: Power of DIF detection of various methods (R500/F100)

Figure 5 above shows graphically the power of different methods under various contamination settings in a particular data size setting, namely, the R500/F100 case. It can be seen easily that as more DIF items are present in the test data, the performance of the AOI, AOI-SP and CI-4 methods degrade while the baseline case and the IRCI maintain a comparably higher level of power regardless of the degree of contamination. Moreover, as the DIF contamination rate increases, the rates of degradation in performance for the methods differ. The AOI and AOI-SP methods degrade much faster than the other two methods as well as the baseline case. Note that as the DIF contamination rate reaches 50%, the power of the AOI, AOI-SP and CI-4 methods drastically drop to 0.45 or below. All in all, the simulation result shows that the baseline case and the IRCI method are less sensitive to the relative proportion of DIF items existed in a test and also can maintain a high level of power for DIF detection. This feature is a desirable one in practice since the DIF contamination rate is an unknown in authentic tests. The possible reason for the good performance of the IRCI method is that it is more effective in identifying the DIF-free items in the process forming the anchor set. This helps in lowering the Type-I error rate which could be largely ascribed to the use of a contaminated anchor set. Furthermore, the IRCI method is also more likely to form a relatively longer DIF-free anchor set, which in turn helps it to beat the CI-4 method with a higher power in DIF detection.

4.4.1 Variations in the Number of Test Items

One of the variables being manipulated in this study is the number of items in the simulation runs. This variable is worth investigating because test length varies greatly in practice among examinations as well as among school settings. The reasons for the existence of such a variation include the nature of the subject knowledge involved, the test duration limitation and the type of items being used, etc. Therefore, it is desirable to have an assessment on various DIF detection methods to verify how their performance may be affected as the test length varies. Tables 8 and 9 below show the performance of the DIF detection methods with simulated tests of different lengths. The DIF value and the group size remain constant at 0.8 logit and R500/F100, respectively, throughout the simulations concerned.

Table 8. Type-I error rates of DIF detection methods (R500/F100)

No. of items	DIF item %	Type-I Error rate				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
10	0	0.044	0.044	0.039	0.050	0.045
	10	0.046	0.057	0.056	0.049	0.058
	20	0.031	0.071	0.051	0.043	0.045
	30	0.044	0.123	0.069	0.064	0.073
	40	0.057	0.235	0.127	0.110	0.090
	50	0.058	0.322	0.226	0.174	0.116
20	0	0.047	0.047	0.049	0.052	0.061
	10	0.047	0.049	0.053	0.059	0.055
	20	0.047	0.092	0.061	0.064	0.064
	30	0.052	0.138	0.079	0.066	0.068
	40	0.064	0.233	0.134	0.090	0.074
	50	0.034	0.311	0.201	0.156	0.064
30	0	0.050	0.050	0.051	0.054	0.062
	10	0.048	0.061	0.054	0.057	0.061
	20	0.045	0.084	0.058	0.059	0.059
	30	0.045	0.150	0.066	0.062	0.055
	40	0.042	0.231	0.133	0.100	0.059
	50	0.050	0.100	0.063	0.065	0.063

It can be seen from Table 8 that as the number of test items increases, the Type-I error rate is largely within control for all methods when the DIF contamination rate is 30% or below, regardless of the number of items in a test. The only exception is for the AOI method, whose Type-I error rate raises significantly to over 10% when 30% of items in a test are with DIF. When the DIF contamination rate is 40% or above, the Type-I error rate of all the four methods rises to a different extent. For the AOI, AOI-SP and CI-4 methods, the raise is pretty significant which ranges from 0.09 to 0.322. This means that there could be up to 30% of items being wrongly classified as with DIF which are actually DIF-free. When comparing with the baseline case whose Type-I error rate is consistently maintained at around 0.05, the IRCI method is slightly inferior when the DIF contamination rate is over 30%. However, when

comparing with the AOI, AOI-SP and CI methods, IRCI is able to maintain a better controlled Type-I error rate—particularly in the highly DIF contamination settings. Figures 6 and 7 below show the performance comparison among DIF detection methods when the DIF contamination rates are at 10% and 40%, respectively, for varying number of test items. In a very mild contamination setting with 10% DIF items, there is no significant performance variation among the methods. In a severe contamination setting with 40% DIF items, the performance variation among different methods becomes conspicuous. The AOI method is of worst performance with the Type-I error rate skyrocketing. For the AOI-SP and CI-4 methods, their Type-I error rates maintain at a level in between 0.1 to 0.15. For the IRCI method, the corresponding level is in between 0.05 to 0.1. This shows that the IRCI method is effective in maintaining a good Type-I error rate in both mildly and severely DIF contaminated environments.

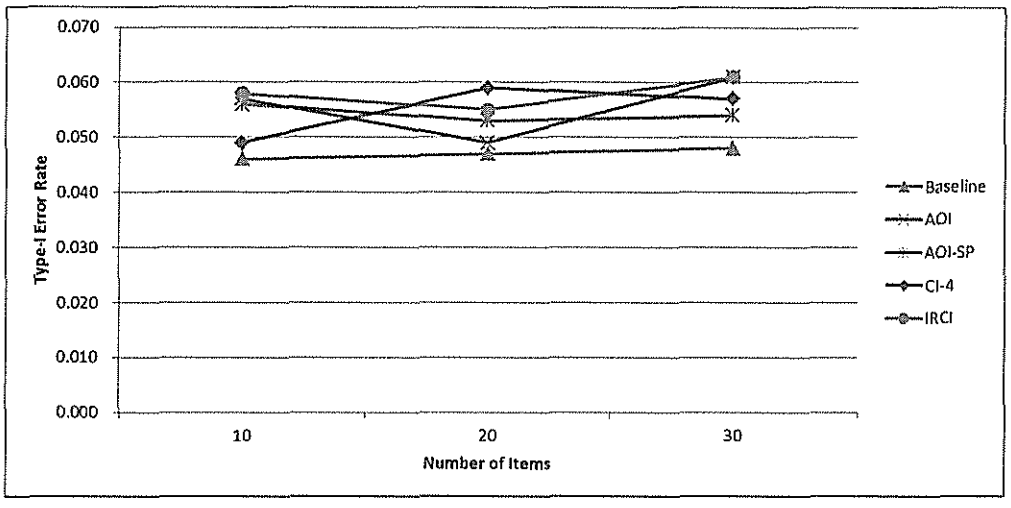


Figure 6: Type-I error rate of DIF detection methods (10% Contamination Rate)

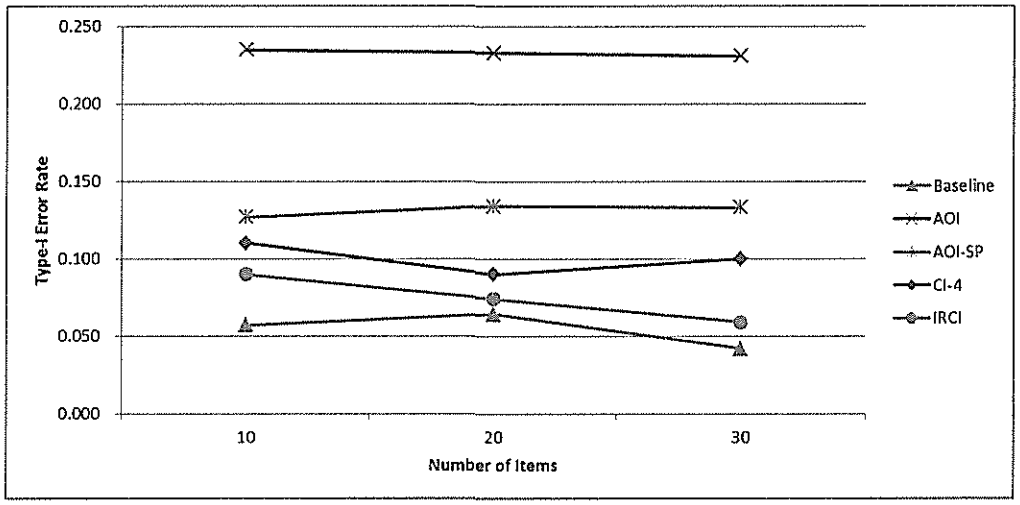


Figure 7: Type-I error rate of DIF detection methods (40% Contamination Rate)

Table 9. Power of DIF detection methods (R500/F100)

No. of items	DIF item %	Power				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
10	0	N/A	N/A	N/A	N/A	N/A
	10	0.770	0.720	0.720	0.670	0.710
	20	0.795	0.660	0.695	0.655	0.755
	30	0.783	0.523	0.597	0.567	0.687
	40	0.787	0.495	0.568	0.562	0.657
	50	0.748	0.368	0.422	0.382	0.640
20	0	N/A	N/A	N/A	N/A	N/A
	10	0.775	0.715	0.740	0.590	0.750
	20	0.812	0.675	0.743	0.632	0.775
	30	0.815	0.540	0.672	0.573	0.777
	40	0.850	0.468	0.637	0.596	0.812
	50	0.805	0.309	0.352	0.434	0.736
30	0	N/A	N/A	N/A	N/A	N/A
	10	0.790	0.707	0.757	0.597	0.783
	20	0.820	0.663	0.755	0.602	0.815
	30	0.841	0.593	0.750	0.655	0.816
	40	0.857	0.462	0.627	0.566	0.832
	50	0.832	0.668	0.737	0.687	0.820

From the simulation results shown in Table 9, with the exception of the CI-4

method, it is generally the case that the longer the test the higher the power of DIF detection. For the CI-4 method, its consistently low performance for all scenarios is probably because of its anchor set size being fixed with 4 items and therefore the power figures are high when comparing with other methods with a relatively longer anchor. Moreover, for the scenario with 10 items only, the CI-4 method takes too many DIF items in the anchor set as there are only 1 to 5 DIF items in total under various contamination settings. Therefore, its performance in terms of power is much hindered by the fixed anchor length requirement.

There are improvements in the power of DIF detection when more items are present in a test, but, the raise in the power is proportionally small. For instance, in the scenario where the DIF contamination rate is 10%, Figure 8 below shows that the increase in power is not significant as the number of item increases. For the scenario of 40% DIF contamination rate, Figure 9 below shows that the DIF detection power also does not increase proportionally to the increase in the number of items.

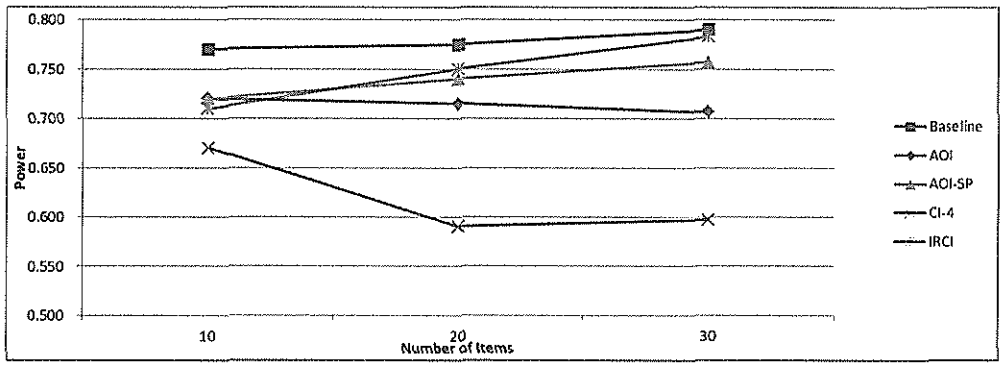


Figure 8: Power of DIF detection methods (10% DIF Contamination Rate)

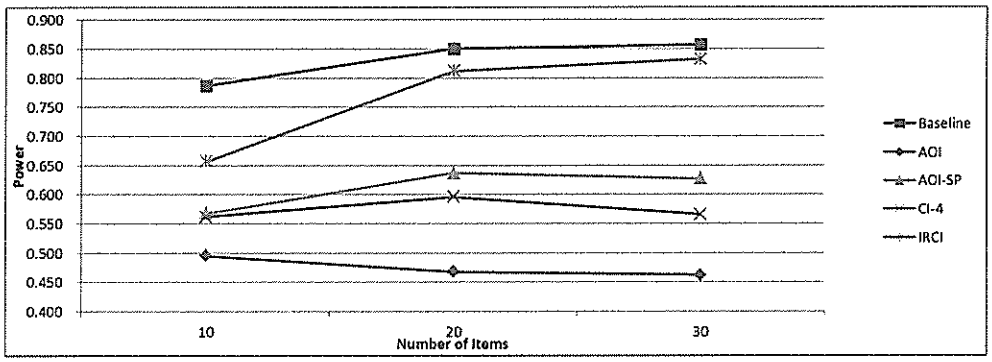


Figure 9: Power of DIF detection methods (40% DIF Contamination Rate)

For the two methods with scale purification—the AOI-SP method and the IRCI method, the power of DIF detection approaches the baseline case as more items are present in the test. One of the possible reasons for this phenomenon is that when the number of items is as low as 10, the effect on the power of DIF detection will be significant if one or two DIF items are included into the anchor set. On the other hand, when the number of items is as many as 30 or beyond, the effect on the power of DIF detection will be less significant even if there is an error in including a few DIF items in the anchor set. This finding suggests that scale purification is generally an effective strategy in handling tests with different amount of items as the strategy is more adaptive to length than does the CI-4 method.

As shown in Table 9 and the corresponding graphical aids provided in Figures 8 and 9, the IRCI method outperforms the other three DIF detection methods by a significant margin. Its DIF detection power is higher than the second best performing method by around 30% in some extreme cases. This result is

promising since the IRCI method is, based on the simulation study results, able to identify the majority of DIF items from a test with a better controlled Type-I error rate, without any adaption made to suit the variance in test length.

Generally speaking, when comparing with the other three DIF detection methods, the IRCI method is superior in terms of both the power and the Type-I error rate when the simulations are conducted with different number of items. When there are more items in the test, the performance of the IRCI methods approaches the corresponding figures obtained from the baseline cases.

4.4.2 Variations in the DIF Size

Another factor that may influence DIF detection performance is the size of DIF of those items exhibiting DIF. Since the DIF size is actually the difference in the odds ratio between the probabilities of correctly and incorrectly answering an item, therefore it is reasonable to assert that a smaller DIF size would probably make DIF detection more difficult and less accurate. In manipulating the DIF size, the group size and the number of items remain constant at R500/F100 and 20 items, respectively. The default DIF size in the previous experiment settings is 0.8 and in this simulation setting we consider also DIF sizes of 0.2 logit lower and 0.2 logit higher. Tables 10 and 11 below show the Type-I error rate and the DIF detection power with the different DIF sizes.

Regarding the Type-I error rate with different DIF sizes, the simulation results in Table 11 show that different methods except the AOI method generally maintain a flat false-positive rates in most DIF contamination scenarios. The AOI method is of the worst performance as its Type-I error rates increase to a very high level when the DIF contamination rate is over 20%. For the AOI-SP and CI methods, their Type-I error rates only increases significantly as the DIF size increases when the DIF contamination rate is at 50%. Whereas for the baseline case and the IRCI method, the Type-I error rates are generally maintain flat as the DIF size increases. The findings in simulations results on the variation of DIF size are aberrant. It seems that variations on DIF size should have no influence on those non-DIF items, which is the case for the baseline case and the IRCI method. However, the data obtained from the AOI, AOI-SP and CI methods does not concur with this intuition.

Table 10. Type-I error rates of DIF detection methods (R500/F100)

DIF size	DIF item %	Type-I Error rate				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
0.6	0	0.051	0.051	0.052	0.062	0.052
	10	0.053	0.054	0.057	0.051	0.054
	20	0.049	0.064	0.060	0.064	0.057
	30	0.038	0.092	0.068	0.070	0.063
	40	0.056	0.141	0.119	0.082	0.084
	50	0.046	0.211	0.212	0.122	0.106
0.8	0	0.047	0.047	0.049	0.052	0.061
	10	0.047	0.049	0.053	0.059	0.055
	20	0.047	0.092	0.061	0.064	0.064
	30	0.052	0.138	0.079	0.066	0.068
	40	0.064	0.233	0.134	0.090	0.074
	50	0.034	0.311	0.201	0.156	0.064
1.0	0	0.056	0.056	0.057	0.058	0.065
	10	0.041	0.053	0.048	0.050	0.049
	20	0.054	0.109	0.071	0.066	0.062
	30	0.048	0.202	0.076	0.073	0.063
	40	0.048	0.317	0.115	0.097	0.062
	50	0.056	0.474	0.407	0.279	0.079

Table 11. Power of DIF detection methods (R500/F100)

DIF size	DIF item %	Power				
		Baseline	AOI	AOI-SP	CI-4	IR-CI
0.6	0	N/A	N/A	N/A	N/A	N/A
	10	0.520	0.465	0.485	0.360	0.490
	20	0.603	0.422	0.502	0.440	0.525
	30	0.590	0.383	0.457	0.418	0.497
	40	0.590	0.273	0.318	0.345	0.459
	50	0.633	0.222	0.225	0.266	0.411
0.8	0	N/A	N/A	N/A	N/A	N/A
	10	0.775	0.715	0.740	0.590	0.750
	20	0.812	0.675	0.743	0.632	0.775
	30	0.815	0.540	0.672	0.573	0.777
	40	0.850	0.468	0.637	0.596	0.812
	50	0.805	0.309	0.352	0.434	0.736
1.0	0	N/A	N/A	N/A	N/A	N/A
	10	0.890	0.855	0.865	0.765	0.870
	20	0.915	0.835	0.887	0.815	0.895
	30	0.920	0.743	0.868	0.805	0.888
	40	0.921	0.624	0.816	0.749	0.890
	50	0.935	0.491	0.505	0.553	0.890

The results shown in Table 11 indicate that when the DIF size is larger, all the anchor selection strategies under study have a high power in DIF detection. Conversely when the DIF size is smaller, the power of DIF detections all drops. This result matches with the intuition that it is easier to detect DIF items whose aberrant behaviour is more obvious. Based on the simulation results obtained, the IRCI method is effective in both the low DIF size and high DIF size scenarios. All in all, the IRCI is a better choice for DIF detection regardless of the DIF size.

4.5 Usage Guidelines

From the results gathered in this simulation studies, the newly devised Iterative Randomized Constant Item method is effective in terms of both high power and better controlled Type-I error rate in DIF detection when comparing with the common anchor selection methods. The results indicate that the new method is worth considering at least from the performance point of view. Nevertheless, there are also practical considerations when one is going to apply this method in practice.

Since the Iterative Randomized Constant Item method is a computational intensive process, when the number of items is large, it may take hours to complete the whole DIF detection analysis. In the simulation conducted in this research study, for some larger data sets with 60 more items, it took around

6 to 8 hours to complete a simulation run (a notebook computer running Intel i7 1.8GHz processor with 4G RAM and 256GB SSD storage). However, if the analysis is a batch process without much press on the time for completing it, then the new approach is a better choice than the other ones tested in this study. Furthermore, the ever increasing in the processing power of modern computers can shorten the running time of this complex algorithm naturally.

On the other hand, it is also noticed that the performance of the new method only be superior to other existing methods when the DIF contamination rate is 30% or higher. In theory, DIF contamination rate is unknown in advance in real life. From this perspective, the new method is a good choice if the time needed for computation is affordable. While if the DIF contamination rates in real life tests are only low, the performance of the new method is similar to the existing methods when the sample size is large. But when the sample size is small, the new method can still outperform other methods. Furthermore, the new method seems be particularly effective for smaller sample size scenarios. This is one of the assets of the new method as practitioner can apply the new method in small classroom or school settings with a higher confidence on the accuracy of the DIF detection results. Its performance in larger sample size is also slightly better than the other methods. All in all, the new method is effective for DIF detection as its performance will be at least as good as other existing methods tested in this research.

CHAPTER 5

CONCLUSION

Fairness should be one of the key concerns in test development. Differential item functioning items, i.e., items exhibiting different difficulties to different groups of test takers with comparable abilities, is therefore an undesirable feature which should be minimized as far as possible if a total avoidance is not totally feasible.

Item bias has been attracting researchers' focus for decades. It refers to a statistical finding that certain items happen to be easier for a certain group of test takers. The grouping criteria usually relates to gender, ethnic group, social economic status, and etc. Item bias exists with two possibilities: (i) a certain group of test takers is actually more capable than the other groups, and (ii) the intrinsic properties of the item favour one particular group of test takers regardless of their abilities. Differential item functioning refers to the later possibility—test takers with similar abilities but in different groups perform significantly differently in an item exhibiting DIF. The main motive for conducting DIF analysis is to ensure test fairness by means of identifying and eliminating items favouring a particular group of test takers, given that the ability of the test takers from different groups are comparable. Spotting DIF items is an important process in item analysis, especially in high-stakes tests which usually possess both the selection and certification social functions.

Nevertheless, items exhibiting DIF are not necessary problematic or inappropriate. There are various interpretations on the causes of DIF and statistical results generally cannot tell the actual reasons behind the cause (Linacre, 2010, p.434).

There have been a lot of research studies in the areas of psychometrics trying to establish a good procedure to uncover those DIF items that exist in a test (Hooland & Thayer, 1988; Kop, Zeileis & Strobl, 2013; Shih and Wang, 2009; Wang, 2004, Wong, 2008, Wang & Yeh, 2003; Wang, Shih and Yang, 2009; Woods, 2009). Some positive results have been obtained in many of those studies. However, there are still plenty of rooms for further improving DIF detection techniques—particularly in the realms of anchor selection strategy and in those scenarios with high DIF contamination rate. All those studies mentioned above have provided a very useful foundation for the investigations conducted in this research.

In this thesis, a new iterative randomized constant item (IRCI) anchor selection strategy is proposed and assessed. The new strategy is largely based on the principles that have been adopted in the AOI-SP and CI methods. These two DIF detection methods are being widely used in practice nowadays. The advantages of the AOI-SP method including its high power in DIF detection, together with the advantages of the CI method including highly purified anchor and better controlled Type-I error rate, have been taken into considerations in designing the new anchor selection strategy. Besides that, another important

factor that has been considered is the sophistication of classifying an item as a potential DIF candidate. It is believed that instead of using one single test result for determining whether an item is with DIF, more independent tests should be conducted to make the final judgement more reliable. Also, the chance of wrongly judging a DIF-free item as a DIF item will likely to be smaller. The proposed IRCI strategy, to certain extent, combines the advantages of the AOI-SP and CI methods so that a reasonably long and better purified anchor item anchor can be generated.

Computer simulation runs have been conducted to assess the effectiveness of the new IRCI strategy. To have a fair comparison, the AOI, AOI-SP and CI methods have been implemented as well. A baseline case which shows the theoretical best performance that a DIF detection method may achieve is also included in the simulation. Based on the simulation results gathered, the new anchor selection strategy is effective in excluding DIF items from the anchor set, elevating the power, and controlling the Type-I error rate of the eventual DIF detection.

To have a more complete study on the effectiveness of the IRCI method, several variables for setting the simulation have been manipulated in order to assess how well the new method adapts to different settings. These settings include the number of test takers altogether and its distribution between the reference and focal groups, the percentage of DIF items in the test, the number of items in the test, and also the DIF size. In all the settings concerned, the new strategy

maintains a higher power and a better controlled Type-I error rate than those of the other three DIF detection methods implemented. In many cases, the performance of the new IRCI method is pretty close to the theoretical ceiling shown in the result obtained in the baseline case. The overall performance of the IRCI method is impressive as it works well in both mildly and severely DIF contamination scenarios.

Although the new method is effective in locating DIF items without too much sacrifice in making Type-I errors, its computation is more complex than the other three methods under study. In the simulations conducted, it may take hours to complete the algorithm of the anchor selection for a single experimental setting. In this perspective, one possible future research direction will be to further streamline the algorithm for simpler computation, with the baseline of not losing its effectiveness in DIF detection. Another possible future development on this research is to find an alternative way to group anchor items in a bid to further improve the performance of the method.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, Calif: Brooks/Cole Pub. Co.

Angoff, W.H. & Ford. S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10(2)*, 95-105.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, N.J: L. Erlbaum.

Cauffman, E. & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument. *Educational and Psychological Measurement*, *66(3)*, 502-521.

Cleary, T. A & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28(1)*, 61-75.

Diamond, J. M. (1998). Guns, germs, and steel: A short history of everybody for the last 13,000 years. London: Vintage.

Dodeen, H. & Johanson, G. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment and Evaluation in Higher Education*, *28*, 129-134.

Finch. H. (2005). The MIMIC model as a method for detecting DIF: comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, *29*, 278-295.

Grading Procedures and Standards-referenced Reporting in the HKDSE Examination (2011). Hong Kong, Hong Kong Examinations and Assessment Authority.

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

Kopf. J., Zeileis. A & Strobl, C. (2013). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Technical Report Number 150, 2013*, Department of Statistics, University of Munich.

Linacre, J. M. (2010). A user's guide to WINSTEPS: Rasch-model computer program [Computer program and manual]. Chicago: WINSTEPS.com.

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

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning (2nd ed.)*. Thousand Oaks, Calif.: SAGE.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogische Institut.

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

Shih. C.-L. & Wang. W.-C. (2009). Differential item functioning: detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199.

Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B), 13*: 238-241.

Tang, K. L. (1996). Polytomous Item Response Theory Models and Their Applications in Large-Scale Testing Programs: Review of Literature (ETS Research Memorandum: RM-96-08, TOEFL-MS-02). Princeton: Educational Testing Service.

Thissen, D., Steinberg, G. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

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

Wang. W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9(4)*, 387-408.

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

Wang. W.-C., Shih, C.-L. & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

Woods. C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42-57

APPENDIX A

PROGRAM CODE SEGMENT

The main program code of the IR-CI method (`SimRun::start_run_IRCI`)

```
void SimRun::start_run_IRCI()
{
    QString path,qs1,qs2,qs3,qs4,iw;

    // window initialization
    MainWindow *w = qobject_cast<MainWindow*>(parent());
    ui.progressBar->setValue(0);
    path.clear();
    QTextStream(&path) << w->sys->data_path << "/" <<
    ui.CB_ds->currentText();
    QDir dir=path;
    QStringList filters;
    filters << "*.dat";
    dir.setNameFilters(filters);
    QStringList datfiles = dir.entryList(QDir::Files);
    QString title;
    QTextStream(&title) << "IR-CI Execution Result (" <<
    ui.CB_ds->currentText() << "):";
    ui.groupBox_3->setTitle(title);
    reset_result_table();

    // load the simulation setting
    QFile qf1;
    qs1.clear();
    QTextStream(&qs1) << w->sys->data_path << "/" <<
    ui.CB_ds->currentText() << "/setting.txt";
    if(read_setting(qs1)){
        t31.total=ds.items;

        QTime time = QTime::currentTime();
        qsrand((uint)time.msec()); // randomize timer

        ui.TE_status2->clear();

        for(int h=0;h<datfiles.size();h++){

            ui.TE_status2->append(datfiles.at(h));
```

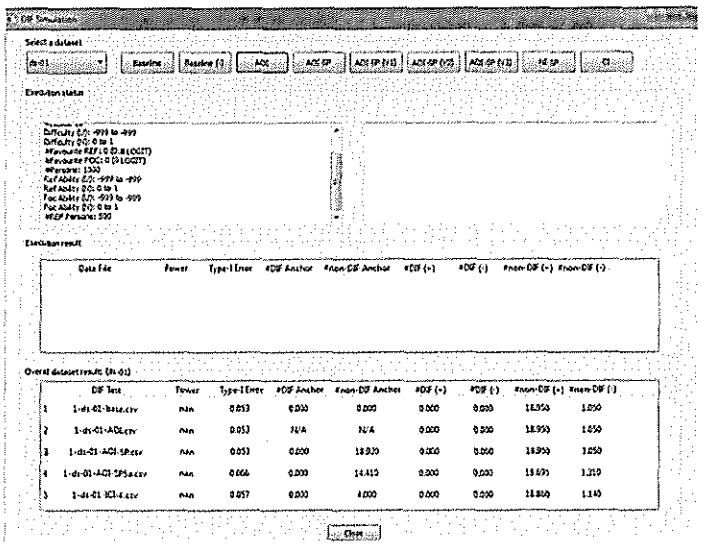
```
qs2=datfiles.at(h);

int ran[40];
for(int k=0;k<ds.items;k++)
    ran[k]=k;

// randomize the array for grouping anchors
for(int k=0;k<ds.items*2;k++){
    int t1=(int)(1.0*qrand()/RAND_MAX*ds.items);
    int t2=(int)(1.0*qrand()/RAND_MAX*ds.items);
    int tmp=ran[t1];
    ran[t1]=ran[t2];
    ran[t2]=tmp;
}

int size=2; // grouping size (2-item)
int matrix[40][40];
int removed=0;
bool re[40];
int sum[40];

// initialize result matrix
for(int k=0;k<ds.items;k++){
    for(int m=0;m<ds.items;m++)
        matrix[k][m]=-1;
    re[k]=false;
}

int iteration=0;
bool found=true;
int j=0;
do{
    // initialization
    iteration++;
    int b=-1;
    for(int k=0;k<ds.items;k++) sum[k]=0;

    // parameter setting for running WINSTEPS
    for(int k=0;k<(ds.items-removed)/size;k++){
        iw="IWEIGHT=* ";

        int m=0;
        int a=0;
        do{
            if(!re[ran[m]]){ // check removed?
                if(a<size && m>b){
                    iw.append(QString("%1,1
                    ").arg(ran[m]+1));
                    a++;
                    b=m;
                }
                else
                    iw.append(QString("%1,0
                    ").arg(ran[m]+1));
            }
            else{
```

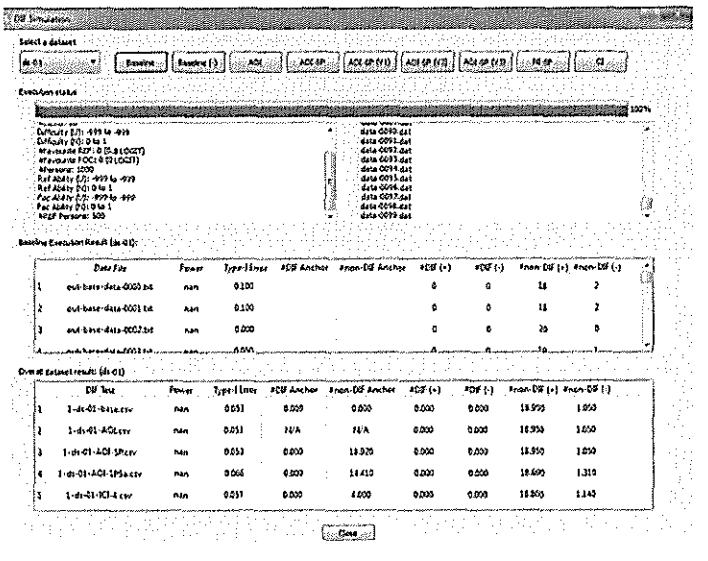
```
                iw.append(QString("%1,0
                ").arg(ran[m]+1));
            }
            m++;
        }while(m<ds.items);

        iw.append(" *");

        // run WINSTEPS with the current 2-item anchor
        run_sim(path,qs2,QString("out-5a-%2-%1.txt").arg
        (j+1).arg(qs2.left(qs2.length()-4)),iw);
        // show the result on screen
        check_result(path,QString("out-5a-%2-%1.txt").ar
        g(j+1).arg(qs2.left(qs2.length()-4)),false,false
        );

        j++;

        // update the result matrix after the run
        for(int m=0;m<ds.items;m++){
            if(t31.prob[m] < w->sys->p_value)
                matrix[k][m]=0;
            else
                matrix[k][m]=1;
            sum[m]+=matrix[k][m];
        }
    }

    // find the item with the highest DIF count
    bool done=false;
    int p,min;
    for(int k=0;k<ds.items && !done;k++){
        if(!re[k]){
            p=k;
            min=sum[p];
            done=true;
        }
    }
    for(int k=p+1;k<ds.items;k++){
        if(!re[k] && sum[k]<min){
            p=k;
            min=sum[k];
        }
    }

    // tie-breaker (randomly pick one)
    int min_count=0;
    for(int k=0;k<ds.items;k++){
        if(!re[k] && sum[k]==min){
            min_count++;
        }
    }
    int t3=(int)(1.0*qrand()/RAND_MAX*min_count);
    int t4=-1;
    bool min_found=false;
    for(int k=0;k<ds.items && !min_found;k++){
```

```
                    if(!re[k] && sum[k]==min){
                        t4++;
                        if(t4==t3){
                            p=k;
                            min_found=true;
                        }
                    }
                }

                if(min!=(ds.items-removed)/size){
                    // remove the most-likely DIF item
                    re[p]=true;
                    removed++;
                }
                else{
                    found=false;
                }


            }while(found);
            // running WINSTEPS with the final anchor set
            iw="IWEIGHT=* ";
            for(int k=0;k<ds.items;k++)
                if(re[k])
                    iw.append(QString("%1,0 ").arg(k+1));
                else
                    iw.append(QString("%1,1 ").arg(k+1));
            iw.append(" *");

            run_sim(path,qs2,QString("out-aoi-sp5a-%1.txt").arg(qs2.l
            eft(qs2.length()-4)),iw);
            check_result(path,QString("out-aoi-sp5a-%1.txt").arg(qs2.
            left(qs2.length()-4)),true,true);

            ui.progressBar->setValue((float)(h+1)/datfiles.size()*100
            );
            qApp->processEvents(QEventLoop::ExcludeUserInputEvents);
        }

        // save the result and show it on the screen
        result_summary(true);
        export_result(QString("%1/1-%2-IR-CI.csv").arg(path).arg(ds.c
        ode));
        show_ds_table();
    }
}
```

APPENDIX B

PROGRAM SCREEN CAPTURES





APPENDIX C

DETAILED SIMULATION RESULTS

The following tables show the detailed results obtained in the simulation runs in this research. Most, but not all, of the results related to some key parameter setting scenarios are listed below. The meaning of the heading symbols are as follows:

a: The number of DIF items in the final anchor set

b: The number of DIF-free items in the final anchor set

c: The number of DIF items successfully identified

d: The number of DIF items cannot be successfully identified

e: The number of DIF-free items successfully identified

f: The number of DIF-free items cannot be successfully identified

g: Type-I error rate

h: Power

1. Simulation Results (R500/F100, 20 items, 0% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	19.070	0.930	0.047	n/a
AOI	n/a	n/a	n/a	n/a	19.070	0.930	0.047	n/a
AOI-SP	n/a	19.010	n/a	n/a	19.020	0.980	0.061	n/a
CI-4	n/a	4.000	n/a	n/a	18.960	1.040	0.052	n/a
IR-CI	n/a	14.730	n/a	n/a	18.780	1.220	0.061	n/a

2. Simulation Results (R500/F100, 20 items, 10% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.550	0.450	17.160	0.840	0.047	0.775
AOI	n/a	n/a	1.430	0.570	17.120	0.880	0.049	0.715
AOI-SP	0.530	17.060	1.480	0.520	17.050	0.950	0.053	0.740
CI-4	0.020	3.980	1.180	0.820	16.950	1.050	0.059	0.590
IR-CI	0.340	13.790	1.500	0.500	17.010	0.990	0.055	0.750

3. Simulation Results (R500/F100, 20 items, 20% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.250	0.750	15.240	0.760	0.047	0.812
AOI	n/a	n/a	2.700	1.300	14.520	1.480	0.092	0.675
AOI-SP	1.010	15.000	2.970	1.030	15.020	0.980	0.061	0.743
CI-4	0.010	3.990	2.530	1.470	14.980	1.020	0.064	0.632
IR-CI	0.540	12.430	3.220	0.780	15.030	0.970	0.060	0.805

4. Simulation Results (R500/F100, 20 items, 30% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	4.890	1.110	13.270	0.730	0.052	0.815
AOI	n/a	n/a	3.240	2.760	12.060	1.940	0.138	0.540
AOI-SP	2.010	12.890	4.030	1.970	12.890	1.110	0.079	0.672
CI-4	0.310	3.690	3.440	2.560	13.080	0.920	0.066	0.573
IR-CI	0.910	10.870	4.660	1.340	12.040	0.960	0.068	0.777

5. Simulation Results (R500/F100, 20 items, 40% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	H
Baseline	n/a	n/a	6.800	1.200	11.230	0.770	0.064	0.850
AOI	n/a	n/a	3.740	4.260	9.200	2.800	0.233	0.468
AOI-SP	2.870	10.390	5.100	2.900	10.390	1.610	0.1134	0.637
CI-4	0.610	3.390	4.770	3.230	10.870	1.130	0.094	0.596
IR-CI	0.900	9.450	6.500	1.500	11.110	0.890	0.074	0.812

6. Simulation Results (R500/F100, 20 items, 50% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	8.050	1.950	9.663	0.337	0.034	0.805
AOI	n/a	n/a	3.089	6.911	6.891	3.109	0.311	0.309
AOI-SP	6.475	6.970	3.525	6.475	6.990	3.010	0.301	0.352
CI-4	1.871	2.219	4.337	5.663	8.436	1.564	0.156	0.434
IR-CI	1.832	8.228	7.356	2.644	9.356	0.644	0.064	0.736

7. Simulation Results (R500/F500, 20 items, 0% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	18.950	1.050	0.053	n/a
AOI	n/a	n/a	n/a	n/a	18.950	1.050	0.053	n/a
AOI-SP	n/a	18.920	n/a	n/a	18.950	1.050	0.053	n/a
CI-4	n/a	4.000	n/a	n/a	18.860	1.140	0.057	n/a
IR-CI	n/a	14.410	n/a	n/a	18.690	1.310	0.066	n/a

8. Simulation Results (R500/F500, 20 items, 10% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.970	0.030	17.030	0.970	0.054	0.985
AOI	n/a	n/a	1.960	0.040	16.680	1.320	0.074	0.980
AOI-SP	0.010	16.970	1.980	0.020	16.960	1.040	0.058	0.990
CI-4	0.000	4.000	1.860	0.140	16.800	1.200	0.067	0.930
IR-CI	0.000	13.000	1.980	0.020	16.830	1.170	0.065	0.990

9. Simulation Results (R500/F500, 20 items, 20% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.960	0.040	15.170	0.830	0.052	0.990
AOI	n/a	n/a	3.880	0.120	13.160	2.840	0.178	0.970
AOI-SP	0.050	14.940	3.960	0.040	14.970	1.030	0.064	0.990
CI-4	0.000	4.000	3.800	0.200	15.000	1.000	0.062	0.950
IR-CI	0.050	12.250	3.960	0.040	15.030	0.970	0.060	0.990

10. Simulation Results (R500/F500, 20 items, 30% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	5.940	0.060	13.190	0.810	0.058	0.990
AOI	n/a	n/a	5.670	0.330	9.050	4.950	0.354	0.945
AOI-SP	0.070	13.230	5.930	0.070	13.220	0.780	0.056	0.988
CI-4	0.000	4.000	5.790	0.210	13.030	0.970	0.069	0.965
IR-CI	0.050	10.770	5.940	0.060	13.100	0.900	0.064	0.990

11. Simulation Results (R500/F500, 20 items, 40% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	7.760	0.240	11.410	0.590	0.049	0.970
AOI	n/a	n/a	6.810	1.190	5.500	6.500	0.542	0.851
AOI-SP	0.280	11.270	7.720	0.280	11.240	0.760	0.063	0.965
CI-4	0.150	3.850	7.420	0.580	11.080	0.920	0.077	0.927
IR-CI	0.260	9.980	7.750	0.350	11.370	0.630	0.052	0.969

12. Simulation Results (R500/F500, 20 items, 50% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	H
Baseline	n/a	n/a	9.740	0.260	9.620	0.380	0.038	0.974
AOI	n/a	n/a	7.500	2.500	2.820	7.180	0.718	0.750
AOI-SP	1.090	8.440	9.120	0.880	8.240	1.760	0.176	0.912
CI-4	1.790	2.210	7.930	2.070	6.240	3.580	0.358	0.793
IR-CI	0.260	8.380	9.730	0.270	9.490	0.510	0.051	0.973

13. Simulation Results (R1000/F500, 20 items, 0% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	19.110	0.890	0.045	n/a
AOI	n/a	n/a	n/a	n/a	19.110	0.890	0.045	n/a
AOI-SP	n/a	19.040	n/a	n/a	19.050	0.950	0.048	n/a
CI-4	n/a	4.000	n/a	n/a	18.760	1.240	0.062	n/a
IR-CI	n/a	16.900	n/a	n/a	19.020	0.980	0.049	n/a

14. Simulation Results (R1000/F500, 20 items, 10% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.970	0.030	17.320	0.680	0.038	0.985
AOI	n/a	n/a	1.960	0.040	16.740	1.260	0.070	0.980
AOI-SP	0.040	17.060	1.960	0.040	17.060	0.940	0.052	0.980
CI-4	0.000	4.000	1.900	0.100	17.030	0.970	0.054	0.950
IR-CI	0.620	15.210	1.960	0.040	17.020	0.980	0.055	0.980

15. Simulation Results (R1000/F500, 20 items, 20% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	4.000	0.000	15.270	0.730	0.045	1.000
AOI	n/a	n/a	3.920	0.080	12.700	3.300	0.206	0.980
AOI-SP	0.010	15.130	3.990	0.010	15.110	0.890	0.055	0.998
CI-4	0.000	4.000	3.880	0.120	14.930	1.070	0.067	0.970
IR-CI	0.570	13.340	4.000	0.000	14.970	1.030	0.064	1.000

16. Simulation Results (R1000/F500, 20 items, 30% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	5.930	0.070	13.290	0.710	0.051	0.988
AOI	n/a	n/a	5.830	0.170	7.900	6.100	0.436	0.972
AOI-SP	0.070	13.180	5.930	0.070	13.190	0.810	0.058	0.988
CI-4	0.040	3.960	5.820	0.180	13.240	0.760	0.054	0.970
IR-CI	0.600	11.890	5.920	0.080	13.070	0.930	0.066	0.987

17. Simulation Results (R1000/F500, 20 items, 40% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	7.810	0.190	11.460	0.540	0.045	0.976
AOI	n/a	n/a	7.340	0.660	4.090	7.910	0.659	0.917
AOI-SP	0.270	11.340	7.810	0.190	11.230	0.770	0.064	0.976
CI-4	0.100	3.900	7.700	0.300	11.090	0.910	0.076	0.963
IR-CI	0.710	10.290	7.800	0.200	11.210	0.790	0.066	0.975

18. Simulation Results (R1000/F500, 20 items, 50% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	9.950	0.050	9.520	0.480	0.048	0.995
AOI	n/a	n/a	8.620	1.380	1.480	8.520	0.852	0.862
AOI-SP	2.390	8.880	9.580	0.420	6.960	3.040	0.304	0.958
CI-4	1.880	2.120	8.420	1.580	4.710	5.290	0.529	0.842
IR-CI	0.540	8.750	9.930	0.070	9.040	0.960	0.096	0.993

19. Simulation Results (R500/F100, 10 items, 0% DIF items, 0.8 logit DIF)

Method	A	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	9.560	0.440	0.044	n/a
AOI	n/a	n/a	n/a	n/a	9.560	0.440	0.044	n/a
AOI-SP	n/a	9.550	n/a	n/a	9.610	0.390	0.039	n/a
CI-4	n/a	4.000	n/a	n/a	9.500	0.450	0.045	n/a
IR-CI	n/a	8.600	n/a	n/a	9.550	0.450	0.045	n/a

20. Simulation Results (R500/F100, 10 items, 10% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	0.770	0.230	8.590	0.410	0.046	0.770
AOI	n/a	n/a	0.720	0.280	8.490	0.510	0.057	0.720
AOI-SP	0.280	8.500	0.720	0.280	8.500	0.050	0.056	0.720
CI-4	0.030	3.970	0.670	0.330	8.560	0.440	0.049	0.670
IR-CI	0.530	7.740	0.710	0.290	8.480	0.520	0.058	0.710

21. Simulation Results (R500/F100, 10 items, 20% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.590	0.410	7.750	0.250	0.031	0.795
AOI	n/a	n/a	1.320	0.680	7.430	0.570	0.071	0.660
AOI-SP	0.620	7.580	1.390	0.610	7.590	0.410	0.051	0.695
CI-4	0.090	3.910	1.310	0.690	7.660	0.340	0.043	0.655
IR-CI	0.700	7.200	1.510	0.490	7.640	0.360	0.045	0.755

22. Simulation Results (R500/F100, 10 items, 30% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	2.350	0.650	6.690	0.310	0.044	0.783
AOI	n/a	n/a	1.570	1.430	6.140	0.860	0.123	0.523
AOI-SP	1.240	6.460	1.799	1.210	6.520	0.480	0.069	0.597
CI-4	0.390	3.610	1.700	1.300	6.550	0.450	0.064	0.567
IR-CI	0.930	6.510	2.060	0.940	6.490	0.510	0.073	0.687

23. Simulation Results (R500/F100, 10 items, 40% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.150	0.850	5.660	0.340	0.057	0.787
AOI	n/a	n/a	1.980	2.020	4.590	1.410	0.235	0.495
AOI-SP	1.690	5.200	2.270	1.730	5.240	0.760	0.127	0.568
CI-4	0.840	3.160	2.250	1.750	5.340	0.660	0.110	0.562
IR-CI	1.320	5.380	2.630	1.370	5.460	0.540	0.090	0.657

24. Simulation Results (R500/F100, 10 items, 50% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.740	1.260	4.710	0.290	0.058	0.748
AOI	n/a	n/a	1.840	3.160	3.390	1.610	0.322	0.368
AOI-SP	2.850	3.860	2.110	2.890	3.870	1.130	0.226	0.422
CI-4	1.890	2.110	1.910	3.090	4.130	0.870	0.174	0.382
IR-CI	1.520	4.740	3.200	1.800	4.420	0.580	0.116	0.640

25. Simulation Results (R500/F100, 30 items, 0% DIF items, 0.8 logit DIF)

Method	A	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	28.500	1.500	0.050	n/a
AOI	n/a	n/a	n/a	n/a	28.500	1.500	0.050	n/a
AOI-SP	n/a	28.470	n/a	n/a	28.480	1.520	0.051	n/a
CI-4	n/a	4.000	n/a	n/a	28.380	1.620	0.054	n/a
IR-CI	n/a	19.610	n/a	n/a	28.150	1.850	0.062	n/a

26. Simulation Results (R500/F100, 30 items, 10% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	2.370	0.630	25.680	1.320	0.049	0.790
AOI	n/a	n/a	2.120	0.880	25.340	1.660	0.061	0.707
AOI-SP	0.730	25.540	2.270	0.730	25.540	1.460	0.054	0.757
CI-4	0.030	3.970	1.790	1.210	25.450	1.550	0.057	0.597
IR-CI	0.320	18.430	2.350	0.650	25.350	1.650	0.061	0.783

27. Simulation Results (R500/F100, 30 items, 20% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	H
Baseline	n/a	n/a	4.920	1.080	22.920	1.080	0.045	0.820
AOI	n/a	n/a	3.980	2.020	21.990	2.010	0.084	0.663
AOI-SP	1.470	22.620	4.530	1.470	22.610	1.390	0.058	0.755
CI-4	0.080	3.920	3.610	2.390	22.580	1.420	0.059	0.602
IR-CI	0.470	16.820	4.890	1.110	22.590	1.410	0.059	0.815

28. Simulation Results (R500/F100, 30 items, 30% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	7.570	1.430	20.050	0.950	0.045	0.841
AOI	n/a	n/a	5.340	3.660	17.840	3.160	0.150	0.593
AOI-SP	2.250	19.620	6.750	2.250	19.620	1.380	0.066	0.750
CI-4	0.190	3.810	5.890	3.110	19.710	1.290	0.062	0.655
IR-CI	0.870	14.930	7.340	1.660	19.850	1.150	0.055	0.816

29. Simulation Results (R500/F100, 30 items, 40% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	10.280	1.720	17.240	0.760	0.042	0.857
AOI	n/a	n/a	5.540	6.460	13.850	4.150	0.231	0.462
AOI-SP	4.470	15.590	7.520	4.480	15.600	2.400	0.133	0.627
CI-4	0.510	3.490	6.790	5.210	16.200	1.800	0.100	0.566
IR-CI	0.930	13.350	9.980	2.020	16.940	1.060	0.059	0.832

30. Simulation Results (R500/F100, 30 items, 50% DIF items, 0.8 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	12.670	2.330	14.080	0.920	0.061	0.845
AOI	n/a	n/a	5.380	9.620	9.860	5.140	0.343	0.359
AOI-SP	9.200	10.190	5.800	9.200	10.180	4.820	0.321	0.387
CI-4	1.900	2.100	6.540	8.460	12.170	2.830	0.189	0.436
IR-CI	1.624	11.218	11.802	3.198	13.941	1.059	0.071	0.787

31. Simulation Results (R500/F100, 20 items, 0% DIF items, 0.6 logit DIF)

Method	A	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	18.980	1.020	0.051	n/a
AOI	n/a	n/a	n/a	n/a	18.980	1.020	0.051	n/a
AOI-SP	n/a	18.960	n/a	n/a	18.970	1.030	0.052	n/a
CI-4	n/a	4.000	n/a	n/a	18.760	1.240	0.062	n/a
IR-CI	n/a	17.140	n/a	n/a	18.960	1.040	0.052	n/a

32. Simulation Results (R500/F100, 20 items, 10% DIF items, 0.6 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.040	0.960	17.040	0.960	0.053	0.520
AOI	n/a	n/a	0.930	1.070	17.030	0.970	0.054	0.465
AOI-SP	1.030	16.950	0.960	1.040	16.970	1.030	0.057	0.480
CI-4	0.010	3.990	0.720	1.280	17.080	0.920	0.051	0.360
IR-CI	0.870	15.030	0.980	1.020	17.030	0.970	0.054	0.490

33. Simulation Results (R500/F100, 20 items, 20% DIF items, 0.6 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	2.410	1.590	15.220	0.780	0.049	0.603
AOI	n/a	n/a	1.690	2.310	14.980	1.020	0.064	0.422
AOI-SP	2.000	15.050	2.010	1.990	15.040	0.960	0.060	0.502
CI-4	0.190	3.810	1.760	2.240	14.970	1.030	0.064	0.440
IR-CI	1.500	13.910	2.100	1.900	15.080	0.920	0.057	0.525

34. Simulation Results (R500/F100, 20 items, 30% DIF items, 0.6 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.540	2.460	13.460	0.540	0.038	0.590
AOI	n/a	n/a	2.300	3.700	12.710	1.290	0.092	0.383
AOI-SP	3.260	13.040	2.740	3.260	13.050	0.950	0.068	0.457
CI-4	0.420	3.580	2.510	3.490	13.020	0.980	0.070	0.418
IR-CI	2.250	12.480	2.980	3.020	13.120	0.880	0.063	0.497

35. Simulation Results (R500/F100, 20 items, 40% DIF items, 0.6 logit DIF)

Method	a	b	c	d	e	f	g	H
Baseline	n/a	n/a	4.20	3.280	11.330	0.670	0.056	0.590
AOI	n/a	n/a	2.180	5.820	10.310	1.690	0141	0.273
AOI-SP	5.450	10.560	2.540	5.460	10.570	1.430	0.119	0.318
CI-4	0.980	3.020	2.760	5.240	11.010	0.990	0.082	0.345
IR-CI	3.050	10.770	3.670	4.330	10.990	1.010	0.084	0.459

36. Simulation Results (R500/F100, 20 items, 50% DIF items, 0.6 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	6.330	3.670	9.540	0.460	0.046	0.633
AOI	n/a	n/a	2.220	7.780	7.890	2.110	0.211	0.222
AOI-SP	7.750	7.870	2.250	7.750	7.880	2.210	0.212	0.225
CI-4	1.860	2.140	2.660	7.340	8.780	1.220	0.122	0.266
IR-CI	3.830	9.040	4.110	5,890	8.940	1.060	0.106	0.411

37. Simulation Results (R500/F100, 20 items, 0% DIF items, 1.0 logit DIF)

Method	A	b	c	d	e	f	g	h
Baseline	n/a	n/a	n/a	n/a	18.880	1.120	0.056	n/a
AOI	n/a	n/a	n/a	n/a	18.880	1.120	0.056	n/a
AOI-SP	n/a	18.850	n/a	n/a	18870	1.130	0.057	n/a
CI-4	n/a	4.000	n/a	n/a	18.840	1.160	0.058	n/a
IR-CI	n/a	16.810	n/a	n/a	18.700	1.300	0.065	n/a

38. Simulation Results (R500/F100, 20 items, 10% DIF items, 1.0 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	1.780	0.220	17.260	0.740	0.041	0.890
AOI	n/a	n/a	1.710	0.290	17.050	0.950	0,053	0.855
AOI-SP	0.270	17.110	1.730	0.270	17.140	0.860	0.048	0.865
CI-4	0.020	3.980	1.530	0.470	17.100	0.900	0.050	0.765
IR-CI	0.750	15.350	1.740	0.360	17.120	0.880	0.049	0.870

39. Simulation Results (R500/F100, 20 items, 20% DIF items, 1.0 logit DIF)

Method	a	B	c	d	e	f	g	h
Baseline	n/a	n/a	3.660	0.340	15.130	0.870	0.054	0.915
AOI	n/a	n/a	3.340	0.660	14.260	1.740	0.109	0.835
AOI-SP	0.460	14.870	3.550	0.450	14.870	1.130	0.071	0.887
CI-4	0.080	3.920	3.260	0.740	14.940	1.060	0.066	0.815
IR-CI	0.810	13.760	3.580	0.420	15.000	1.000	0,062	0.895

40. Simulation Results (R500/F100, 20 items, 30% DIF items, 1.0 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	3.540	2.460	13.460	0.540	0.038	0.590
AOI	n/a	n/a	2.300	3.700	12.710	1.290	0.092	0.383
AOI-SP	3.260	13.040	2.740	3.260	13.050	0.950	0.068	0.457
CI-4	0.420	3.580	2.510	3.490	13.020	0.980	0.070	0.418
IR-CI	2.250	12.480	2.980	3.020	13.120	0.880	0.063	0.497

41. Simulation Results (R500/F100, 20 items, 40% DIF items, 1.0 logit DIF)

Method	a	b	c	d	e	f	g	h
Baseline	n/a	n/a	4.720	3.280	11.330	0.670	0.056	0.590
AOI	n/a	n/a	2.180	5.820	10.310	1.690	0.141	0.273
AOI-SP	5.450	10.560	2.540	5.460	10.570	1.430	0.119	0.318
CI-4	0.980	3.020	2.760	5.240	11.010	0.990	0.082	0.345
IR-CI	3.050	10.770	3.670	4.330	10.990	1.010	0.084	0.459

42. Simulation Results (R500/F100, 20 items, 50% DIF items, 1.0 logit DIF)

Method	a	b	c	d	e	f	g	H
Baseline	n/a	n/a	6.330	3.670	9.540	0.460	0.046	0.633
AOI	n/a	n/a	2.220	7.780	7.890	2.110	0.211	0.222
AOI-SP	7.750	7.870	2.250	7.750	7.880	2.120	0.212	0.225
CI-4	1.860	2.140	2.660	7.340	8.780	1.220	0.122	0.266
IR-CI	3.830	9.040	4.110	5.890	8.940	1.060	0.106	0.411