

© 2014 Rui Guo

A STEPWISE TEST CHARACTERISTIC CURVE METHOD TO  
DETECT ITEM PARAMETER DRIFT

BY

RUI GUO

THESIS

Submitted in partial fulfillment of the requirements  
for the degree of Master of Arts in Psychology  
in the Graduate College of the  
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Master's Committee:

Professor Hua-Hua Chang, Chair  
Professor Jeff Douglas

# ABSTRACT

An important assumption of item response theory (IRT) based equating is that the item parameters should be invariant over different testing occasions. Sometimes, however, item parameters do not remain invariant due to factors other than sampling error, and this is termed item parameter drift (IPD). Several methods have been proposed to detect drifted items. However, most of the existing methods aim at detecting the drift in individual items, which may not be ideal when only the overall test characteristic curve (TCC) is of interest to the users. One such occasion in common practice is IRT-based true score equating, where the goal is to create a conversion table to make the two TCCs as close as possible. This paper introduces a stepwise test characteristic curve (Stepwise TCC) method to dynamically detect item parameter drift based on TCC without requirement to set any critical values. Comparisons were made between the new method and two commonly used existing methods under the three-parameter logistic model. Results show that the new method performed well in IPD detection.

*To Shuo*

# ACKNOWLEDGMENTS

I am heartily thankful to my adviser, Hua-Hua Chang, whose encouragement, guidance and passionate support from the initial to the final. I would also like to thank Yi Zheng, who have helped me and supported me in any respect during the completion of the project.

Lastly, I offer my regards and appreciation to my family, who has helped me through the process of earning my graduate program.

# TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION . . . . .	1
CHAPTER 2	BACKGROUND . . . . .	2
2.1	Unidimensional item response model . . . . .	2
2.2	Existing methods for IPD detection . . . . .	2
CHAPTER 3	METHODOLOGY . . . . .	4
3.1	Using TCC to Detect IPD . . . . .	4
3.2	The Stepwise Test Characteristic Curve Method . . . . .	6
CHAPTER 4	SIMULATION STUDY . . . . .	9
4.1	Study I . . . . .	9
4.2	Study II . . . . .	13
CHAPTER 5	RESULTS . . . . .	14
CHAPTER 6	CONCLUSIONS AND DISCUSSION . . . . .	18
REFERENCES	. . . . .	20

# CHAPTER 1

## INTRODUCTION

*Linking and equating* are the procedures that put the test scores across different testing occasions on the same scale. They play an important role in large-scale assessments because after these procedures, examinees' performances are comparable across different testing occasions. Because the equating coefficients are usually obtained from a set of common items used across different administrations, the stability of the common items over these administrations, as reflected by the stability of their item parameters under the *item response theory* (IRT) framework, is crucial to the quality of the linking process. Any factor that may cause *item parameter drift* (IPD) across different administrations poses a threat to the validity of the IRT linking.

# CHAPTER 2

## BACKGROUND

### 2.1 Unidimensional item response model

When the *three-parameter logistic* (3PL) model is used, the probability of examinee  $i$  answering item  $j$  correctly (i.e.,  $Y_{ij} = 1$ ) given his/her ability  $\theta_i$  takes the following form:

$$P_j(\theta_i) = P(Y_{ij} = 1|\theta_i) = c_j + (1 - c_j) \frac{1}{(1 + \exp[-a_j(\theta_i - b_j)])} \quad (2.1)$$

where

$a_j$  is the discrimination parameter of the  $j^{th}$  item,

$b_j$  is the difficulty parameter of the  $j^{th}$  item,

$c_j$  is the guessing parameter of the  $j^{th}$  item, and

$\theta_i$  is the latent trait level of the  $i^{th}$  examinee.

Item parameter drift is defined as the change of item parameters (i.e.,  $a$ -parameter,  $b$ -parameter, and/or  $c$ -parameter) over different testing occasions (Goldstein, 1983).

### 2.2 Existing methods for IPD detection

Item parameter drift can occur for various reasons, such as disclosure and sharing of items, changes in the answer sheet design, or social background change, etc. Several methods have been developed to detect drifted items, including non-IRT methods such as the *Mantel-Haenszel method* (Holland & Thayer, 1986), and IRT-based methods such as the *Lord's chi-square statistic* (Lord, 1980), the signed and unsigned areas between two item response functions (Raju, 1990), the signed and unsigned closed-interval mea-



asures (S.-H. Kim & Cohen, 1991), the *compensatory differential item functioning* (CDIF) method, and the *non-compensatory differential item functioning* (NCDIF) method (Roju, Van der Linden, & Fleer, 1995). Many of the above-mentioned methods are based on the *item characteristic curve* (ICC), which is produced by the item response functions, such as Equation 2.1 for the 3PL model. However, the Stepwise TCC method proposed in this paper is based on the test characteristic curve (TCC), which is the summation of the ICCs of all items in the test.

# CHAPTER 3

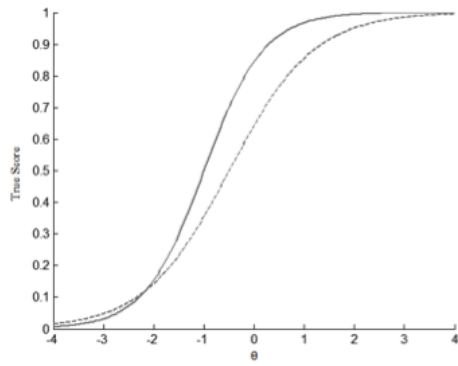
## METHODOLOGY

### 3.1 Using TCC to Detect IPD

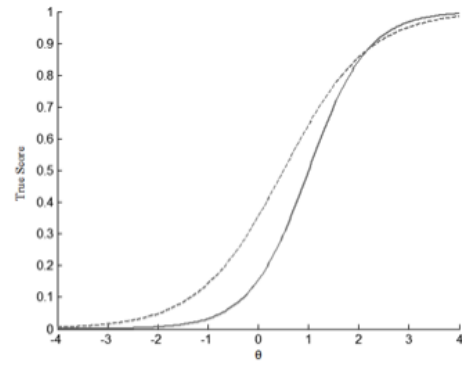
The idea of using TCC in IPD detection can be traced back to Roju et al. in 1995 (Roju et al., 1995), who proposed to use the differential test functioning (DTF) criterion based on TCCs in detecting drifted items. In their study, however, the DTF was applied only as a measure to assess the effectiveness of the IPD detection after the detection procedure has been finished by other ICC based methods. When drifted items have been flagged by methods such as the Mantel-Haenszel method, the signed and unsigned area measures, CDIF, and NCDIF, the DTF was then used as an index to compare the detection results. In other words, the use of TCC in the entire procedure of IPD detection is still underdeveloped.

Except Roju et al.'s use of DTF (Roju et al., 1995), most of the existing methods are solely based on examining individual items. This may not be necessary in the occasions when only the TCC of a test is of interest to the users. More specifically, a group of drifted items may not show scale drift collectively when they are inspected as a whole set. Therefore, if only the TCC is of interest, it may not be necessary to exclude those items as long as their resulting TCC is stable enough. One such occasion is *true score equating*, which is widely used in operational testing programs. In true score equating, the equating results will only be affected by the two TCCs in the two administrations, instead of individual ICCs.

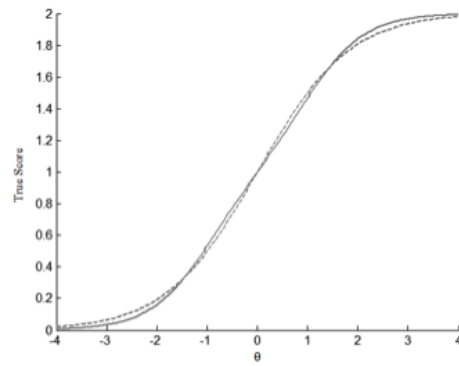
Figure 3.1 illustrates the case where the drift effects of two items are canceling out when the two items drift towards opposite directions. In Figure 3.1a, the original ICC of Item 1 is presented by the solid line, with its original  $a$ -,  $b$ -, and  $c$ -parameters being 1,  $-1$ , and 0, respectively. The dashed line represents the drifted ICC with a decrease of .3 in the  $a$ -parameter and a



(a) Comparisons of ICCs of Item 1



(b) Comparisons of ICCs of Item 2



(c) Comparisons of TCCs of Item 1 and 2 as a whole set

Figure 3.1: Comparisons of ICCs and TCCs

decrease of .5 in the  $b$ -parameter. Figure 3.1b shows the ICCs of Item 2, with its original parameters the same as in Item 1 and the drift being a decrease of .3 in the  $a$ -parameter and an increase of .5 in the  $b$ -parameter. The combined effect of Item 1 and 2, as illustrated by Figure 3.1c, creates an almost non-drifted item set when the two items are treated as a whole.

The goal of this paper is to present a dynamic method, that is the Stepwise TCC method, that detects drifted items 'stepwise' until an ultimate collection of items that exclusively causes the TCC drift is found. Inspired by the stepwise regression method (Efroymson, 1960) in Statistics, which selects the optimal combination of the predictive variables in a regression model, the proposed method iteratively removes some items that potentially cause TCC drift from the linking item set while bringing some excluded items back to the linking set. The process iterates until an optimal set of linking items is found. A simulation study was conducted and comparisons were made between the stepwise TCC method and two other existing methods: the  $d^2$  method and the Mantel-Haenszel method.

### 3.2 The Stepwise Test Characteristic Curve Method

Inspired by the stepwise regression, which aims to optimize the statistical significance of a regression model, the proposed stepwise TCC method attempts to minimize the area difference of the two TCCs between two test administrations. The Stepwise TCC method starts with an initial set of common items regarded as the initial linking set and updates it recursively. In each iteration, a new item is entered if the current linking set together with the new item generates a smaller area difference between the TCCs, and/or a current item is removed if the linking set excluded that item provides a smaller area difference. This iterative process is repeated until no more items are entered or removed. The final linking set is the optimal combination of common items that generates the smallest area difference of TCCs between test administrations.

Suppose there are  $n$  common items in both administrations, then examinee  $i$ 's expected proportions correct (referred to as the true score) of the common

items in the two administrations can be expressed as

$$T_1(\theta_i) = \sum_{j=1}^n P_j(\theta_i) \quad (3.1)$$

$$T_2(\theta_i) = \sum_{j=1}^n P_j(\theta_i) \quad (3.2)$$

where  $T_1(\theta_i)$  denotes the true score of the  $i$ 's examinee in the first administration and  $T_2(\theta_i)$  denotes his/her true score in the second administration. Then, the Stepwise TCC method includes the following steps:

1. The item parameters for each administration are calibrated separately.
2. An initial linking process is conducted using all of the common items.
3. The area differences of ICCs between two test administrations are calculated for all common items using Equation 3.3 and the values are arranged from low to high.

$$\text{ICC Difference} = \sum_{q=1}^Q |P_1(\theta_q) - P_2(\theta_q)| \quad (3.3)$$

where  $\theta_q$ 's are  $Q$  quadrature points from  $-4$  to  $4$  with equal interval.

4. The *initial linking set* is randomly sampled from the common items. The number of sampled items is also randomly chosen. The *remaining set* contains the remaining common items.
5. In each iteration, one item is entered into the *current linking set* and/or one item in the current linking set is removed. The linking procedure is implemented again using the updated *linking set*. Specifically,

*The enter step:*

- (a) A number of  $R$  *proposal linking sets* are formed by adding each of the  $R$  items in the *remaining set* into the *current linking set*. The TCC differences for the  $R$  *proposal linking sets* are calculated by Equation 3.4:

$$\text{TCC Difference} = \sum_{q=1}^Q |T_1(\theta_q) - T_2(\theta_q)| \quad (3.4)$$

- (b) Compare the TCC differences of the *R proposal linking sets*. The one with the smallest TCC difference is chosen as the *candidate set*. If the TCC difference of the *candidate set* is less than or equal to that of the *current linking set*, the *current linking set* is replaced with the *candidate set*; otherwise the *current linking set* remains unchanged.

*The remove step:*

- (c) A number of *L proposal linking sets* are formed by removing each of the *L* items in the *current linking set*. The TCC differences for all *proposal linking sets* are calculated.
- (d) The *proposal linking set* with the smallest TCC difference is chosen as the *candidate set*. If the TCC difference between the *candidate set* and the *current linking set* is less than an *error threshold* (explained below), the *current linking set* is replaced with the *candidate set*; otherwise the *current linking set* remains unchanged.

*Note:* The reason of using an *error threshold* is because some calibration error should be tolerated. The *error threshold* can be obtained using a simulation study under the non-drift condition. First, two response matrices using all common items are simulated for the two administrations, and the item parameters are calibrated separately based on the two response matrices; Then, the *error threshold* is computed by averaging the ICC differences, as given in Equation 3.3, of all common items. Multiple replications are implemented and the average of the *error threshold* values is obtained and used in the stepwise procedure.

- 6. Step 5 is repeated until no items are entered or removed according to the algorithms in both entering and removing steps.

# CHAPTER 4

## SIMULATION STUDY

Two simulation studies were conducted to compare the Stepwise TCC method with two existing methods: the  $d^2$  method and the Mantel-Haenszel method. One hundred replications were made in both studies.

### 4.1 Study I

Study I was simulated to mimic two administrations a year apart. The two administrations had 15 common items. Two sample sizes (i.e., number of simulees,  $N = 500, 1000$ ) were examined.

#### 4.1.1 Parameter Simulation

The response data was generated using the 3PL model. The logarithm of the  $a$ -parameters and the  $b$ -parameters were randomly sampled from a multivariate normal distribution, that is  $\begin{pmatrix} \log(a) \\ b \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}\right)$ . All items had the same  $c$ -parameter value,  $c \equiv 0.2$ , as described by Donoghue and Isham(1998).

For each administration, examinee trait levels ( $\theta$ 's) were distributed following  $N(\mu_k, 1)$ , where  $k$  indicates the  $k$ 's administration. The mean of the  $\theta$  distribution of the first administration,  $\mu_1$ , is 0, whereas the mean of the second administration,  $\mu_2$ , is .1. This represents a moderate increase in  $\theta$  from year 1 to year 2, and is similar in magnitude to the differences seen in large-scale assessments such as the National Assessment of Educational Progress (Campbell, Finn, Donahue, Educational Testing Service, & National Center for Education Statistics, 1998).

Drift Type	No Drift	Type 1	Type 2	Type 3	Type 4	Type 5
Proportion	$p_0 = .61$	$p_1 = .20$	$p_2 = .04$	$p_3 = .07$	$p_4 = .04$	$p_5 = .04$

Table 4.1: Proportion of Linking Items for Each Type of Drift

#### 4.1.2 IPD Simulation

According to a previous study of a large-scale state assessment (Chang et al., 2011), there are mainly five types of item parameter drift.

Type 1: Shifted  $c$ -parameter. This type of items have almost identical ICCs at the middle and higher end of the  $\theta$  span, but there is a large gap at the lower end. A possible reason of this type of drift is that the items may have been disclosed and shared by a considerable proportion of examinees, as mathematically derived by Veerkamp and Glas (Veerkamp & Glas, 2000).

Type 2: Shifted  $bc$ -parameter. This type of items have a shift in both  $b$ - and  $c$ -parameters. In the ICC plots, one curve is above the other over most part of the  $\theta$  span.

Type 3: Shifted  $b$ -parameter. This type of items have almost the same  $a$ - and  $c$ -parameters, with the  $b$ -parameter shifted only. This change in difficulty could be caused by changes in item appearance, position or social background.

Type 4: Shifted  $a$ -parameter. This type of items have almost the same  $b$ -parameter across years, but the  $a$ -parameter is shifted. This means the discriminating power of the item has changed.

Type 5: Extreme shift. This type of items shows an extreme difference in all of the  $a$ -,  $b$ -, and  $c$ -parameters. The reason could be their disclosure to a large proportion of students or error in other procedures (e.g., item matching, item parameter calibration, etc.). This pattern of the change in the ICCs is saliently different from the other types.

According to this large-scale test, the five types of drifted items took about 20%, 4%, 7%, 4%, and 4% of the common items, respectively. The proportions of the different types of drifted items in this paper's simulation design will be based on these findings, as shown in Table 4.1.



For the second year’s common items, the five types of IPD as mentioned above were artificially created. For items with shifted  $b$ -parameters, the  $b$ -parameters in year 2 differed by  $-0.4$  from those in year 1. Similarly, for items with shifted  $a$ -parameters, the  $a$ -parameters in year 2 differed by  $-0.5$  from those in year 1. For items with shifted  $c$ -parameters, the  $c$ -parameter in year 2 differed by  $+0.2$  from those in year 1. Similar magnitudes of drift have been observed on large-scale standardized tests in Stone and Lane(1991), Cohen(1992), and Wells, Subkoviak and Serlin(2002) (Stone & Lane, 1991; Cohen, 1992; Wells, Subkoviak, & Serlin, 2002).

### 4.1.3 Parameter Estimation

Item parameters were estimated using the *marginal maximum likelihood estimation* (MMLE) algorithm in Multilog (Thissen, Chen, & Bock, 2003). The item parameter estimates for the two administrations were then equated using the *Stocking-Lord test characteristic curve* method (Stocking & Lord, 1983) via the computer program STUIRT (S. Kim & Kolen, 2004). After the initial linking using all common items, different methods were implemented to detect IPD.

### 4.1.4 IPD Detection

The Stepwise TCC method is conducted and compared with two other existing methods: the  $d^2$  method and the Mantel-Haenszel method. Following are brief descriptions of the two methods. The  $d^2$  method is based on empirical critical values and the Mantel-Haenszel method is based on hypothesis testing.

*The  $d^2$  method.* The  $d^2$  method is an empirical application of the NCDIF method proposed by Roju (Roju et al., 1995) and is one of the most commonly used methods by researchers (Meade, Lautenschlager, & Hecht, 2005)) and practitioners such as Pearson Assessments & Information. For example, Murphy et al. used the  $d^2$  method to study the impact of scale drift on equating results and student performance (Murphy, Little, Fan, Lin, & Kirkpatrick, 2010) . The  $d^2$

index is defined as the weighted sum of the squared deviation between the ICCs across administrations, as shown in Equation 4.1:

$$d_j^2 = \sum_{i=1}^n [P_{1j}(\theta_i) - P_{2j}(\theta_i)]^2 * g(\theta_i) \quad (4.1)$$

where  $P_{1j}(\theta_i)$  and  $P_{2j}(\theta_i)$  denote the probability of examinee  $i$  answering item  $j$  correctly in the two administrations and  $g(\theta_i)$  denotes the weights for  $\theta_i$ . The  $d_j^2$  values of all common items are reviewed, and the 95<sup>th</sup> percentile of these values is chosen as the critical value for flagging drifted items.

*Mantel-Haenszel Method.* The Mantel-Haenszel procedure is based on the chi-square test of contingency table data. It is used under the *classical testing theory* framework. The details can be found in Holland's work (Holland & Thayer, 1986).

For the  $d^2$  method and the Mantel-Haenszel method, the same significance level  $\alpha=0.05$  was used. For the Stepwise TCC method, no  $\alpha$  value is needed.

#### 4.1.5 Evaluation Indices

The results of IPD detection can be studied by comparing the following indices.

*Overall False Rates.* This includes the *False Positive* (FP) rate and the *False Negative* (FN) rate. They measure the rates of false classifications of the common items, as defined in equations 4.2 and 4.3 below. In the equating setting, FN rate is more severe than FP rate because failing to detect drifted items may cause more problems during the linking procedure than over-detecting non-drifted items.

$$\text{FP Rate} = \frac{\# \text{ of nondrifted items flagged as drifted}}{\text{total } \# \text{ of nondrifted items}} \quad (4.2)$$

$$\text{FN Rate} = \frac{\# \text{ of drifted items classified as nondrifted}}{\text{total } \# \text{ of drifted items}} \quad (4.3)$$

*False Negative Rates for Each Type of Drift.* As claimed in the paper that the Stepwise TCC method can detect items with  $c$ -parameters drift effectively, it is of interest to look at the FN rates for each type of drift, especially for the first type. The reason only FN rate is of interest is because it is more important than PF rate.

$$\text{FN Rate for Type } k \text{ Drift} = \frac{\# \text{ of drifted items classified as nondrifted in type } k}{\text{total } \# \text{ of drifted items in type } k} \quad (4.4)$$

*The TCC Difference.* The TCC difference between the two administrations is another evaluation index. It is calculated by Equation 3.3.

## 4.2 Study II

Study II has the same setting as study I with one exception: a multi-directional item drift was simulated rather than a single-directional drift as in study I. Specifically, for items with shifted  $b$ -shift, the  $b$ -parameters in the second administration differed by  $\pm 0.4$  from the value in the first administration. Similarly, for items with  $a$ -shift, the  $a$ -parameters differed by  $\pm 0.5$ . For items with  $c$ -shift, the  $c$ -parameters differed by  $\pm 0.2$ . The direction of drift in each case was randomly chosen with equal probability, and the magnitudes and the proportions of each type of parameter drift were the same as in Study I.

# CHAPTER 5

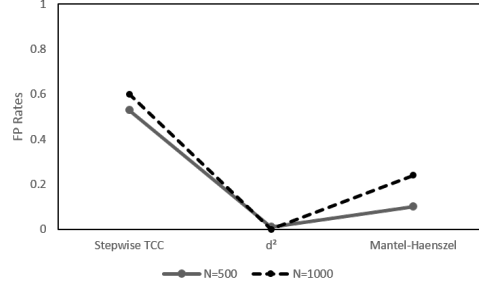
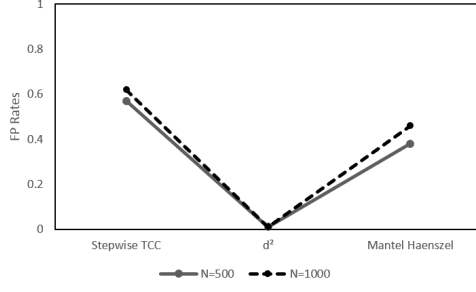
## RESULTS

Simulation study results show that on overall the Stepwise TCC method was more effective in detecting drifted items, especially those with shifted  $c$ -parameter, than the two compared existing methods. The detailed results are described below.

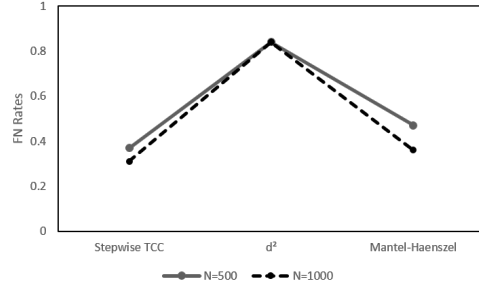
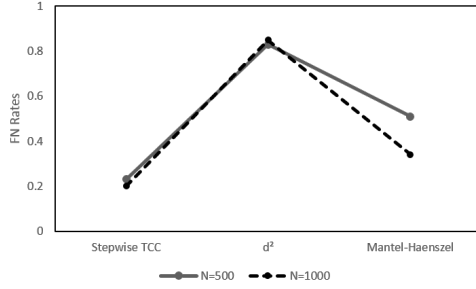
Figure 5.1 shows the False Rates of the linking items in the two studies. Generally, the Stepwise TCC method generated much lower FN rates though it has slightly higher FP rates. In terms of FN rates, the Stepwise TCC method performed best, followed by the Mantel-Haenszel method, and then the  $d^2$  method. In terms of FP rates, the Mantel-Haenszel outperformed the Stepwise TCC method by a small margin. The  $d^2$  method has the smallest FP rates, which almost reached 0%. Note that there is a well-known trade-off between FN and FP rates. In the IPD detection scenario, FN rate is usually more severe than FP rate. FP rate occurs when a non-drifted item is incorrectly flagged as drifted but these flagged items can still be discussed by content experts or be recalibrated to determine whether they are actually drifted. However, FN occurs when drifted items are not detected. As a result, the procedure of equating and linking will be affected seriously.

Furthermore, as the sample size (i.e., the number of simulees) increases, FN rates generally decreases for most of the methods. It confirms the intuitive expectation that a larger sample size can improve IPD detection accuracy. However, this pattern is not that obvious when the  $d^2$  method is used because an empirical 95<sup>th</sup> is used.

Figure 5.2 provides detailed comparisons of the FN rates for each type of parameter drift under the two studies. The Stepwise TCC method can effectively detect most types of parameter drift, which is the reason it leads to lower overall FN rates. Especially, the Stepwise TCC method is best at detecting items with shifted  $c$ -parameters, which takes a large proportion among all types of drift as shown in Table 4.1. The Mantel-Haenszel method

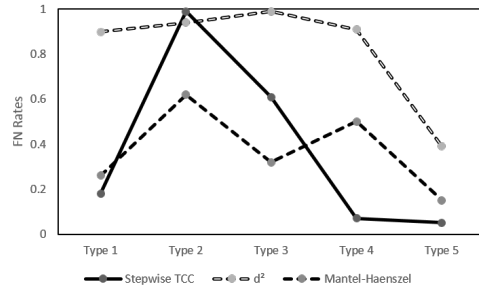
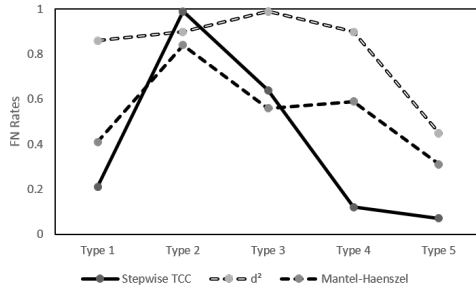


(a) Comparisons of FP rates in Study I (b) Comparisons of FP rates in Study II

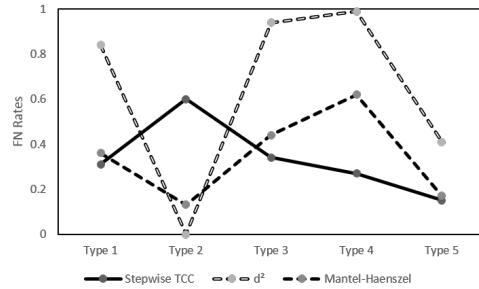
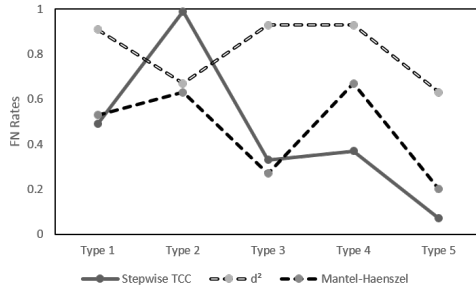


(c) Comparisons of FN rates in Study I (d) Comparisons of FN rates in Study II

Figure 5.1: Comparisons of FP and FN rates in Study I and Study II

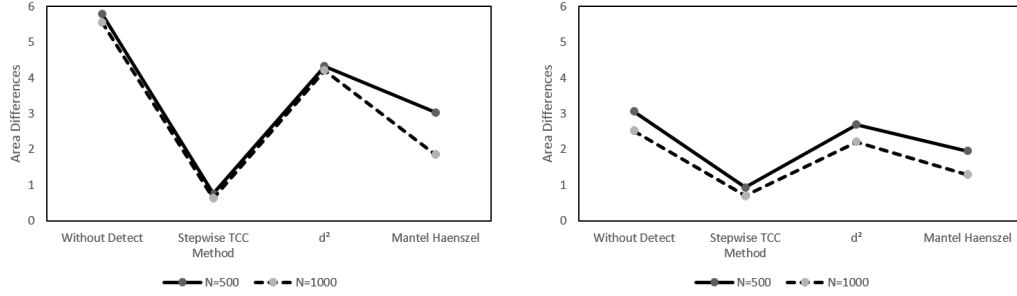


(a) FN rates for each drift type in Study I, N=500 (b) FN rates for each drift type in Study I, N=1000



(c) FN rates for each drift type in Study II, N=500 (d) FN rates for each drift type in Study II, N=1000

Figure 5.2: FN rates for each drift type in Study I and Study II



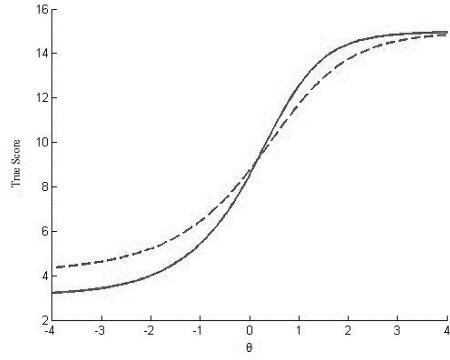
(a) Comparisons of TCC differences under study I (b) Comparisons of TCC differences under study II

Figure 5.3: Comparisons of TCC differences under both studies

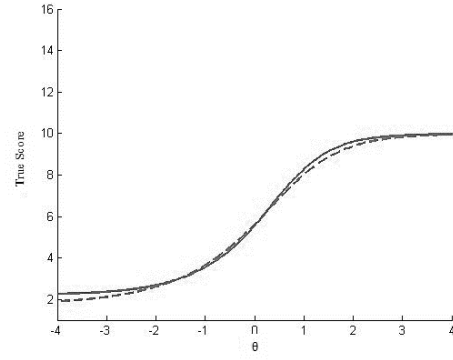
performed slightly better than the stepwise TCC method in Type 2 and Type 3 drift, but worse in the remaining types. The  $d^2$  method is the least effective one in all types except Type 2 drift.

Figure 5.3 compares the TCC differences in both studies. A smaller TCC difference between years indicates a more accurate conversion table when true score equating is conducted. According to the figures, the Stepwise TCC method generated the smallest TCC difference among all of the compared methods. In addition, no matter how many common items, the Stepwise TCC method always provides a small TCC difference and that difference becomes smaller when the sample size increases. In contrast, the  $d^2$  method always generates relatively large TCC difference and that difference does not decrease as the sample size increases. The effectiveness of the Mantel-Haenszel method is in between the Stepwise TCC and the  $d^2$  method.

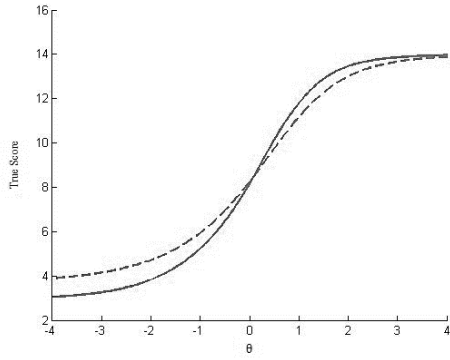
Figure 5.4 provides an example in study I that illustrates the TCC curves of the linking items from the first and second administrations. Figure 5.4a compares the TCCs of all 15 common items without IPD detection in study I and shows a relatively big TCC difference. Figure 5.4b shows the TCCs of the linking items excluding items flagged by the Stepwise TCC method, indicating a relatively smaller TCC difference between two administrations. Figure 5.4c shows the TCCs of the linking items excluding items flagged by the  $d^2$  method. The TCC difference in the lower end of the  $\theta$  span is still large because of the difficulty in detecting the shifted  $c$ -parameter items by this method. The effectiveness of the Mantel-Haenszel is shown in Figure 5.4d.



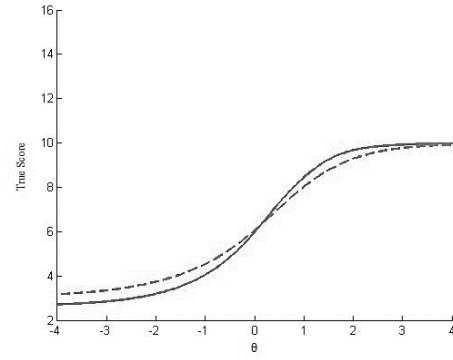
(a) TCC difference without IPD detection, N=500



(b) TCC difference by Stepwise TCC method, N=500



(c) TCC difference by  $d^2$  method, N=500



(d) TCC difference by Mantel-Haenszel method, N=500

Figure 5.4: TCC differences by different IPD detection methods

## CHAPTER 6

### CONCLUSIONS AND DISCUSSION

Scale stability is an important quality for large-scale testing programs and should be maintained across multiple administrations in a long term. An effective method to detect drifted items can improve the quality of tests in such aspect. The Stepwise TCC method is similar to the stepwise variable selection method in linear regression. The former compares the TCC difference in each iteration to determine whether to enter one item into the current linking set or remove one item from the current linking set; whereas the latter conducts decisions whether a variable should be included into or excluded from the regression model in each iteration. Our simulation study results showed that, the Stepwise TCC method generally provides lower FN rates as well as a smaller TCC difference than the two compared existing methods.

The Stepwise TCC method has several advantages. First, it provides a dynamic algorithm to detect IPD and the algorithm terminates automatically when a certain stopping criterion has been met. Second, the Stepwise TCC method has no arbitrary critical value. Therefore, no matter how few or how many drifted items actually exist, they will be detected more effectively compared with the methods with an arbitrary cut-off criterion. Third, the conversion table provided by the Stepwise TCC method is more precise than that provided by the existing methods because the Stepwise TCC method is designed to minimize the TCC difference by nature. Hence, true score equating will be much more accurate using the Stepwise TCC method.

Some issues of the Stepwise TCC method still need to be further studied. First, current studies of item IPD detection are all based on the uni-dimensional 3PL model. When the latent construct being measured is actually multi-dimensional, the uni-dimensionality assumption will be violated. Therefore, how to incorporate the Stepwise TCC method to detect item drift under multi-dimensional IRT framework can be considered in the fu-



ture studies. In addition, since the computer based testing (CBT) is gaining increasing attention, it is promising to generalize the Stepwise TCC method in item drift detection into the CBT framework.

# REFERENCES

- Campbell, J., Finn, K., Donahue, P., Educational Testing Service, & National Center for Education Statistics. (1998). *NAEP 1996 trends in academic progress*. U.S. Dept. of Education, Office of Educational Research and Improvement.
- Chang, H., Ryan, K. E., Zheng, Y., Ali, U. S., Wang, C., & Lin, H. (2011). *Scale stability: an empirical study of ISAT linking* (Research Report No. 10). Champaign, IL: University of Illinois at Urbana-Champaign. Illinois Assessment Accountability Project.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155.
- Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22(1), 33–51.
- Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, 1, 191–203.
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20(4), 369–377.
- Holland, P., & Thayer, D. (1986). Differential item performance and the mantel-haenszel procedure. In *the Annual Meeting of the American Educational Resesasrch Association*. San Francisco, CA.
- Kim, S., & Kolen, M. (2004). STUIRT [Computer software]. Iowa City, IA: The University of Iowa.
- Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied psychological measurement*, 15(3), 269–278.
- Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Lawrence Erlbaum.
- Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, 5(3), 279–300.
- Murphy, S., Little, I., Fan, M., Lin, C., & Kirkpatrick, R. (2010). The impact of different anchor stability methods on equating results and student performance. In *the Annual Meeting of the National Council of Measurement in Education*. Denver, CO.
- Raju, N. S. (1990). Determining the significance of estimated signed and

- unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2), 197–207.
- Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). Irt-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19(4), 353–368.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, 7(2), 201–210.
- Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education*, 4(2), 125–141.
- Thissen, D., Chen, W., & Bock, R. (2003). Multilog (version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25(4), 373–389.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26(1), 77–87.