Differential Item Functioning Analysis without A Priori Information on Anchor Items: Scree Plots and Graphical Test

Ke-Hai Yuan
University of Notre Dame
Hongyun Liu and Yuting Han
Beijing Normal University

CONTENTS

01 Introduction

Introduction

- A fair test needs all its items to have measurement invariance:

$$P(Y \mid W = w, V = v) = P(Y \mid W = w), \qquad (1)$$

where P(.) denotes probability, Y is the item score, w is the trait to be measured, and v is group membership.

- Differential item functioning (DIF) occurs if equation (1) does not hold, and it makes the test invalid.

Introduction

Common methods for detecting DIF

IRT methods:

Lord's chi-square test

The likelihood ratio test (LRT)

Item-specific Wald test

Non-IRT methods:

Mantel-Haenszel (MH)

Logistic regression (LR)

The simultaneous item bias test (SIBTEST)

Other methods:

Bayesian methods

Mixture modeling

Multiple indicators multiple causes (MIMIC)

IRT tree model

IRT models with covariates

Introduction

Dilemmas in obtaining DIF-free anchor items for Traditional Methods：

1. The definition of DIF involves matching ability, which needs DIF-free items when calibrating item parameters. But in practice, it is attempting to drag oneself up by one's own bootstraps (Doebler, 2018).

2. Most existing DIF detection methods are item-by-item approaches, which may lead to Type I error inflation due to multiple testing, especially in the test level.

Introduction

Differential item pair functioning (DIPF) method by Bechger and Maris (2015):

1. They clarified the concept regarding what can be tested for DIF analysis but also critically reviewed the problems with the existing methods, and a color map was used to inspect the clusters of DIF pairs visually.
2. The DIPF method cannot identify individual items that may favor a particular group; color map to classify items may encounter difficulties in operation; the results of the test for overall DIPF may not be consistent with the results of z-statistics with item pairs.

Introduction

Five contributions to the DIF literature of our Methods:

1. Developing two graphical tools to facilitate the identification of DIF-free items.
2. Using Monte Carlo test to conduct DIF analysis
3. Classifying items according to clusters of points displayed in the D-scree plot or RCD-scree plot.
4. Developing a confidence interval approach for visualizing DIF as well as a graphical test.
5. Synchronizing DIF analysis at item level with the overall test.

02

Rationale: DIF Detection Based on the True Null Hypothesis

- The Rash DIF model
- Relative Change of Difficulty Difference
- D-scree plot and the internal reference points
- RCD-scree plot
- Graphical test with RCD confidence interval (RCD-CI)

DIF Detection Based on the True Null Hypothesis

1.The Rasch DIF model:

$$P\left(Y_{ij}^{(g)}=1\middle|\theta_i^{(g)}\right)=\frac{\exp\left(\theta_i^{(g)}-b_j^{(g)}\right)}{1+\exp\left(\theta_i^{(g)}-b_j^{(g)}\right)},\qquad \theta^{(1)}\sim N(0,\sigma_1^2)\ and\ \theta^{(2)}\sim N(\mu,\sigma_2^2),$$

To identify the model and to compare the estimated item parameters, traditional DIF analysis determines the impact μ by assuming a certain number of items to be DIF-free (anchor items), which places the estimated item parameters onto a common scale. However, if the anchor contains DIF items, which is referred to as contamination (Finch, 2005; Wang, Shih, & Sun, 2012; Woods, 2009), the construction of a common scale for the item parameters is flawed.

DIF Detection Based on the True Null Hypothesis

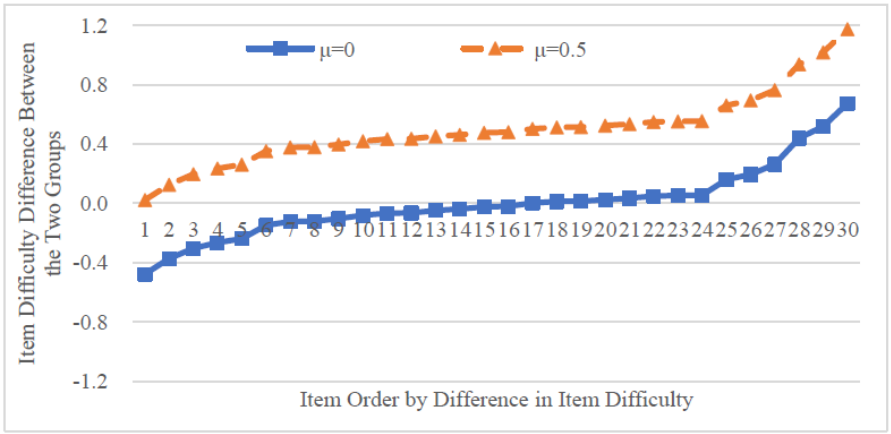
2. Relative Change of Difficulty Difference (RCD):

$$d_j = \Delta b_j = b_j^{(2)} - b_j^{(1)} \quad \Rightarrow \quad \delta_{jl} = d_j - d_l \ \ (j = 1, 2, \ldots, M)$$

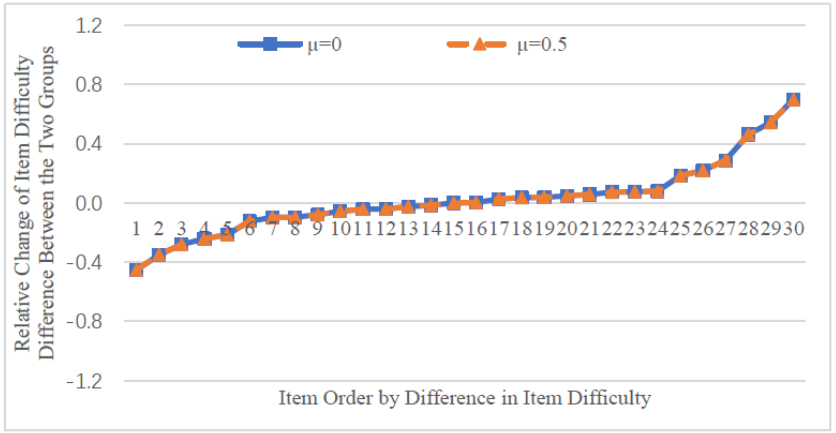
Depending on the impact μ $\Rightarrow$ Not depending on the impact μ

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(M)} \qquad \delta_{(j)} = d_{(j)} - \bar{d}_{(ref)}$$

The plots of $d_{(j)}$ and $\delta_{(j)}$ with different impact μ:



(a) Item difficulty difference $d_{(j)}$ between two groups



CD $\delta_{(j)}$ between two groups when $d_{(15)}$ is chosen as the internal reference point

Depending on the impact μ

Not depending on the impact μ

DIF Detection Based on the True Null Hypothesis

3. D-scree plot and the internal reference points：

Step 1: Analyze the empirical data. Estimate the parameters as $\widehat{b}_j^{(1)}$, $\widehat{b}_j^{(2)}$, $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. Compute the values of $\widehat{d}_j$ and $\widehat{d}_{(j)}$.

Step 2： Use Monte Carlo method to approximate the null distribution of $\widehat{d}_{(j)}$.

Step 3： Selecting internal reference points and forming clusters using D-scree Plot

Step 4: Calculate the RCD.

Violation of DIF-free by a single item will cause the scree plot to have a systematic departure from the **45-degree** line.

DIF Detection Based on the True Null Hypothesis

4. RCD-scree Plot

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(M)} \qquad \delta_{(j)} = d_{(j)} - \bar{d}_{(ref)}$$

5. Graphical test with RCD confidence interval (RCD-CI) :

a) if $\hat{\delta}_{(j)} < L_{\hat{\delta}_{(j)}}^{(H_0)}$, then all the items with RCD smaller than $\hat{\delta}_{(j)}$ (i.e., $\hat{\delta}_{(1)}, \hat{\delta}_{(2)}, \ldots, \hat{\delta}_{(j-1)}$) will be identified as possessing DIF that favors group 2;

b) if $\hat{\delta}_{(j)} > U_{\hat{\delta}_{(j)}}^{(H_0)}$, then all the items with RCD greater than $\hat{\delta}_{(j)}$ (i.e., $\hat{\delta}_{(j+1)}, \hat{\delta}_{(j+2)}, \ldots, \hat{\delta}_{(M)}$) will be identified as possessing DIF that favors group 1.

03

Illustrations the Application of Scree Plots and Graphical Test

- Illustrative Datasets
- The application of D-scree plot
- The application of graphical test

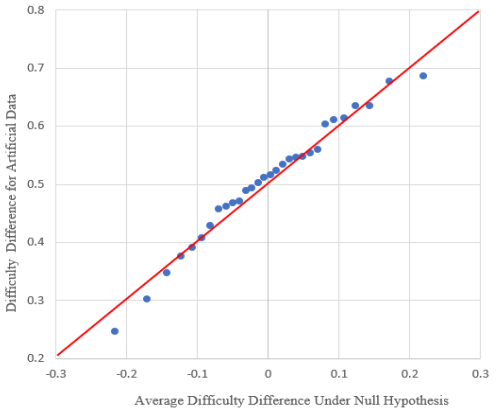
Illustrations the Application of D-Scree Plots and Graphical Test

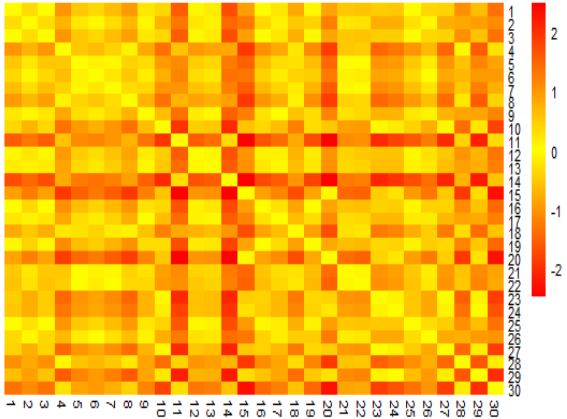
1. Illustrative Datasets (M=30, N1=N2=1000, C1-C8 Cases)

Item number	C1	C2	C3	C4	C5	C6	C7	C8
1	0	-1.0	-0.4	-2.0	-0.4	0	-1.5	-0.5
2	0	-0.9	0	-1.9	-0.4	0	-1.4	-0.5
3	0	-0.8	0	-1.8	-0.4	0	-1.3	-0.5
4	0	-0.7	0	-1.7	-0.4	0	-1.2	-0.5
5	0	-0.6	0	-1.6	-0.4	0	-1.1	-0.5
6	0	-0.5	0	-1.5	-0.4	0	-1.0	-0.5
7	0	-0.4	0	-1.4	-0.4	0	-0.9	-0.5
8	0	-0.3	0	-1.3	-0.4	0	-0.8	-0.5
9	0	-0.2	0	-1.2	-0.4	0	-0.7	-0.5
10	0	-0.1	0	-1.1	-0.4	0	-0.6	-0.5
11	0	0	0	-1.0	-0.4	0	-0.5	-0.5
12	0	0	0	-0.9	-0.4	0	-0.4	-0.5
13	0	0	0	-0.8	-0.4	0	-0.3	-0.5
14	0	0	0	-0.7	-0.4	0	-0.2	-0.5
15	0	0	0	-0.6	-0.4	0	-0.1	-0.5
16	0	0	0	-0.5	-0.4	0	0.1	-0.5
17	0	0	0	-0.4	-0.4	0	0.2	-0.5
18	0	0	0	-0.3	-0.4	0	0.3	-0.5
19	0	0	0	-0.2	-0.4	0	0.4	-0.5
20	0	0	0	-0.1	-0.4	0	0.5	-0.5
21	0	0.1	0	0	0	0.4	0.6	-0.5
22	0	0.2	0	0	0	0.4	0.7	-0.5
23	0	0.3	0	0	0	0.4	0.8	-0.5
24	0	0.4	0	0	0	0.4	0.9	-0.5
25	0	0.5	0	0	0	0.4	1.0	-0.5
26	0	0.6	0	0	0	0.4	1.1	-0.5
27	0	0.7	0	0	0	0.4	1.2	-0.5
28	0	0.8	0	0	0	0.4	1.3	-0.5
29	0	0.9	0	0	0	0.4	1.4	-0.5
30	0	1.0	0	0	0	0.4	1.5	-0.5

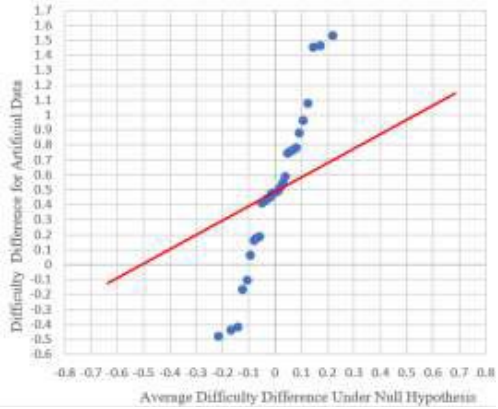
D-Scree Plots

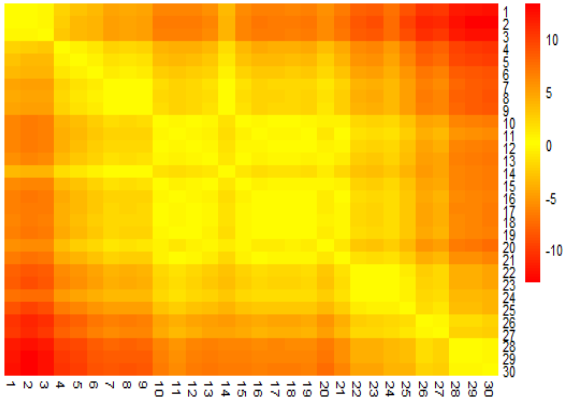
C1





C2

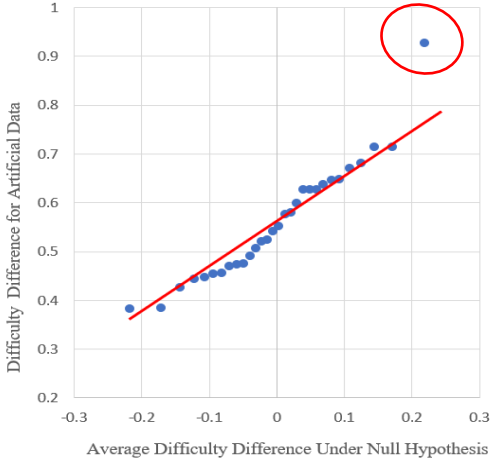


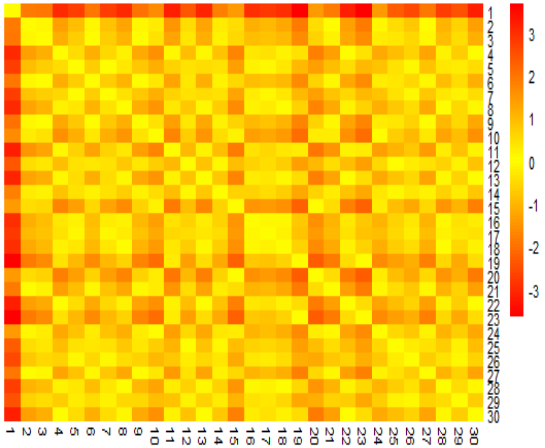


Item number	C1	C2
1	0	-1.0
2	0	-0.9
3	0	-0.8
4	0	-0.7
5	0	-0.6
6	0	-0.5
7	0	-0.4
8	0	-0.3
9	0	-0.2
10	0	-0.1
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
21	0	0.1
22	0	0.2
23	0	0.3
24	0	0.4
25	0	0.5
26	0	0.6
27	0	0.7
28	0	0.8
29	0	0.9
30	0	1.0

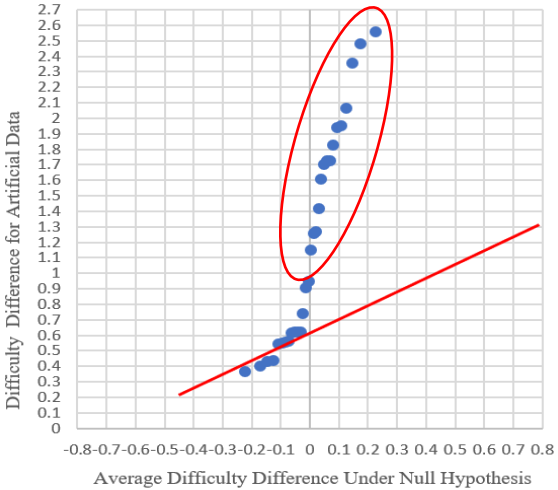
D-Scree Plots

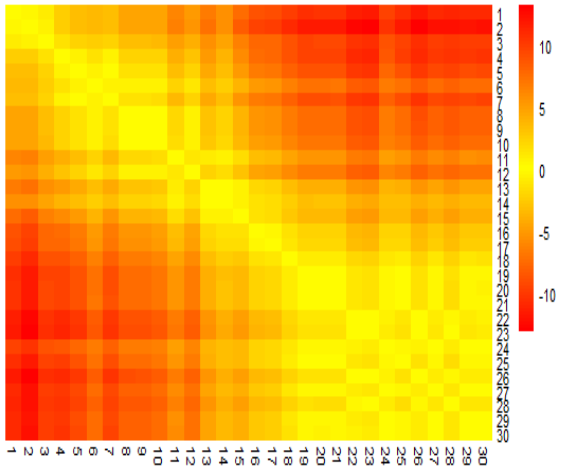
C3





C4

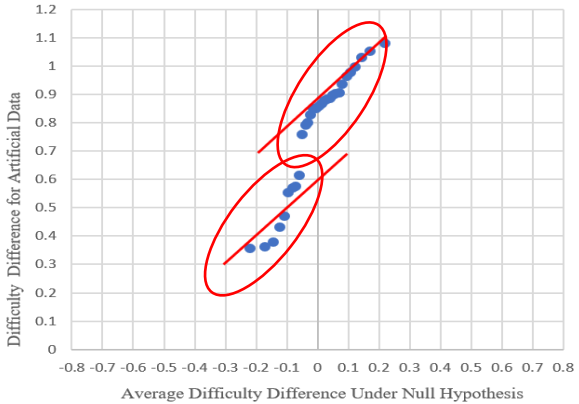


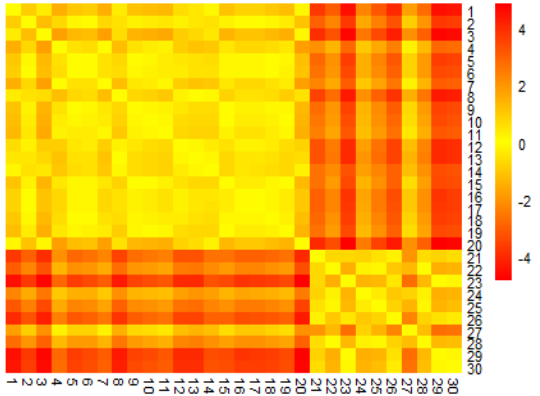


Item number	C3	C4
1	-0.4	-2.0
2	0	-1.9
3	0	-1.8
4	0	-1.7
5	0	-1.6
6	0	-1.5
7	0	-1.4
8	0	-1.3
9	0	-1.2
10	0	-1.1
11	0	-1.0
12	0	-0.9
13	0	-0.8
14	0	-0.7
15	0	-0.6
16	0	-0.5
17	0	-0.4
18	0	-0.3
19	0	-0.2
20	0	-0.1
21	0	0
22	0	0
23	0	0
24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	0
30	0	0

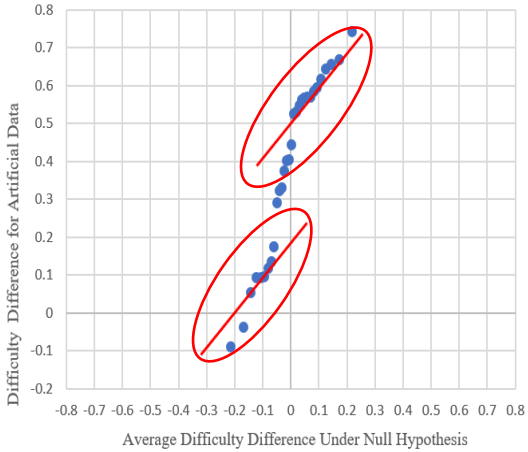
D-Scree Plots

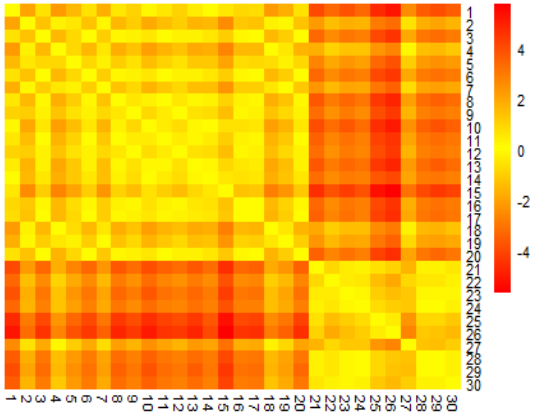
C5





C6

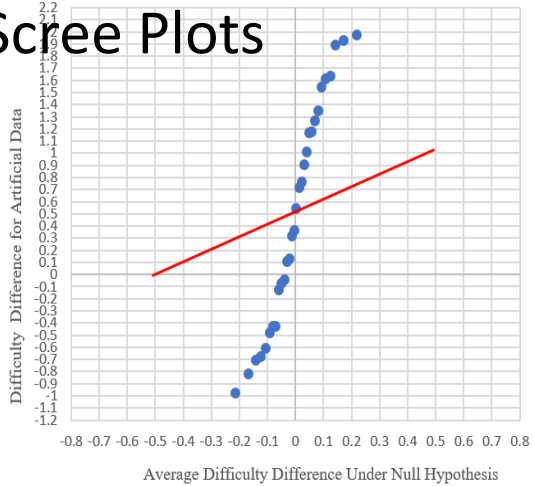


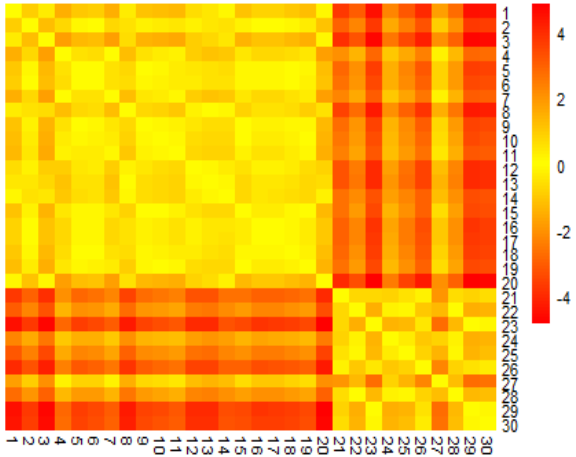


Item number	C5	C6
1	-0.4	0
2	-0.4	0
3	-0.4	0
4	-0.4	0
5	-0.4	0
6	-0.4	0
7	-0.4	0
8	-0.4	0
9	-0.4	0
10	-0.4	0
11	-0.4	0
12	-0.4	0
13	-0.4	0
14	-0.4	0
15	-0.4	0
16	-0.4	0
17	-0.4	0
18	-0.4	0
19	-0.4	0
20	-0.4	0
21	0	0.4
22	0	0.4
23	0	0.4
24	0	0.4
25	0	0.4
26	0	0.4
27	0	0.4
28	0	0.4
29	0	0.4
30	0	0.4

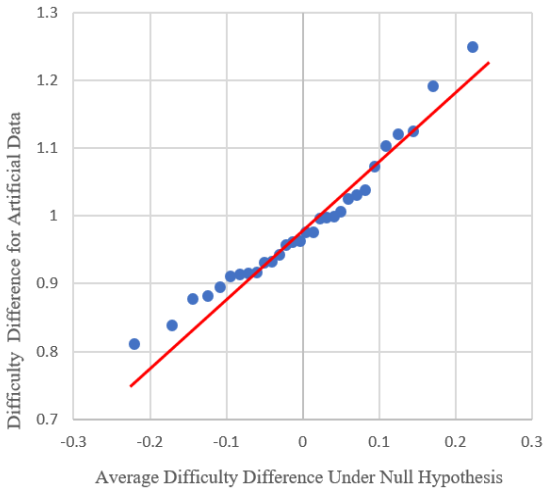
D-Scree Plots

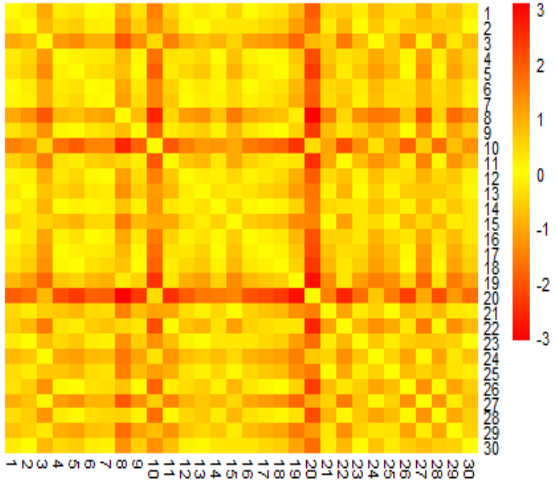
C7





C8





Item number	C7	C8
1	-1.5	-0.5
2	-1.4	-0.5
3	-1.3	-0.5
4	-1.2	-0.5
5	-1.1	-0.5
6	-1.0	-0.5
7	-0.9	-0.5
8	-0.8	-0.5
9	-0.7	-0.5
10	-0.6	-0.5
11	-0.5	-0.5
12	-0.4	-0.5
13	-0.3	-0.5
14	-0.2	-0.5
15	-0.1	-0.5
16	0.1	-0.5
17	0.2	-0.5
18	0.3	-0.5
19	0.4	-0.5
20	0.5	-0.5
21	0.6	-0.5
22	0.7	-0.5
23	0.8	-0.5
24	0.9	-0.5
25	1.0	-0.5
26	1.1	-0.5
27	1.2	-0.5
28	1.3	-0.5
29	1.4	-0.5
30	1.5	-0.5

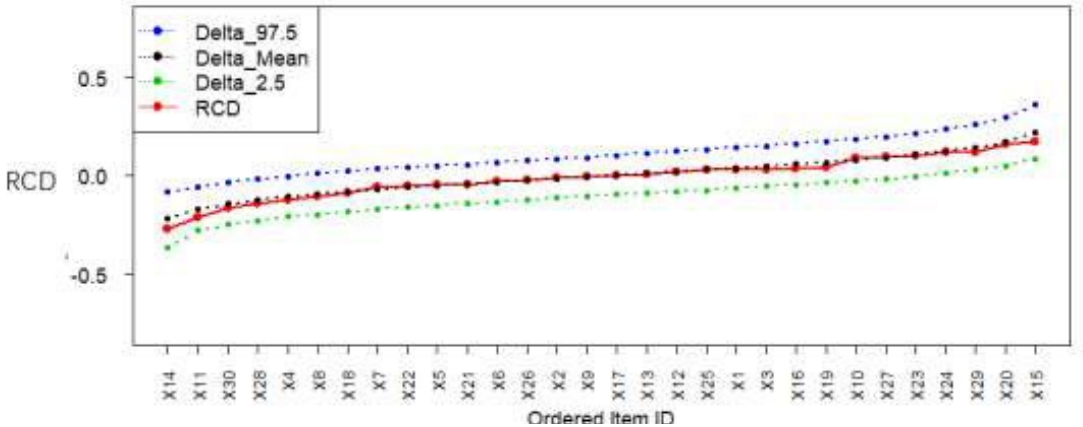
Illustrations the Application of Scree Plots and Graphical Test

The D-scree plot can be used to:

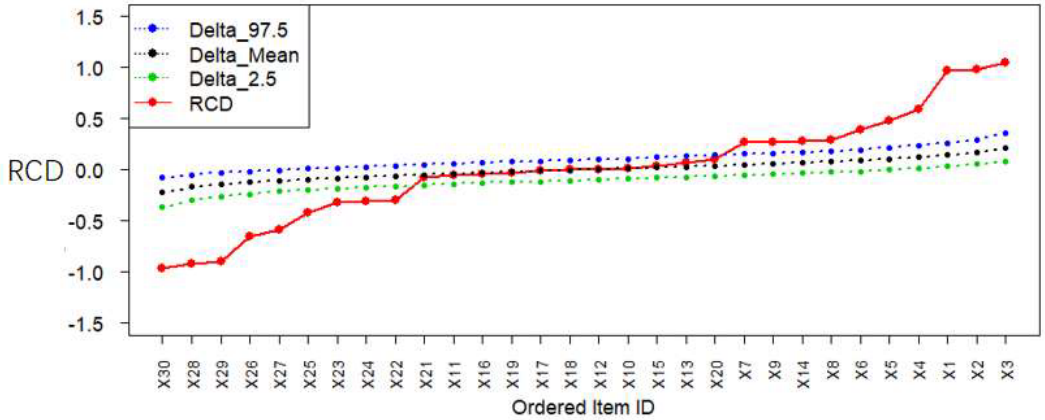
- It can help us to select the internal reference points (DIF-free items)
- The different clusters of items can be identified visually, and the overall DIF sizes of the test can also be judged roughly.
- It can help us to inspect the overall DIF size on the test level intuitively.

Graphical Test: RCD vs. MH

C1



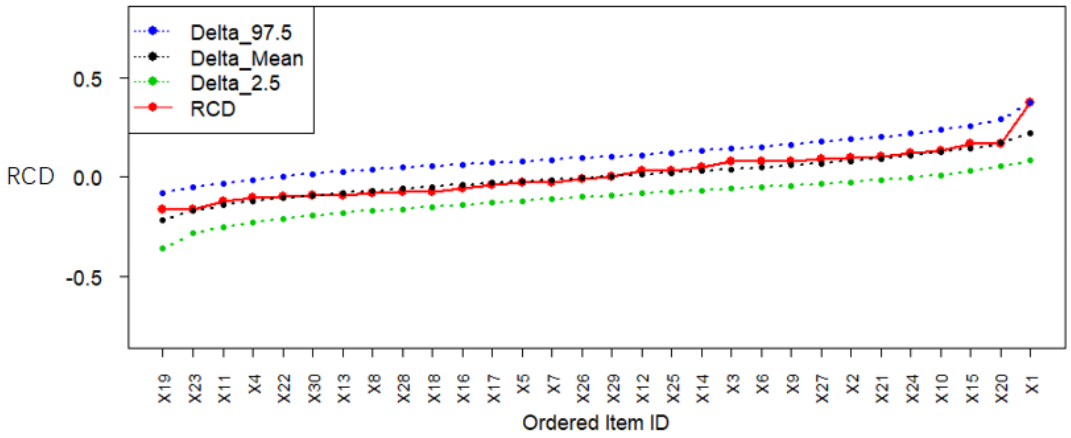
C2



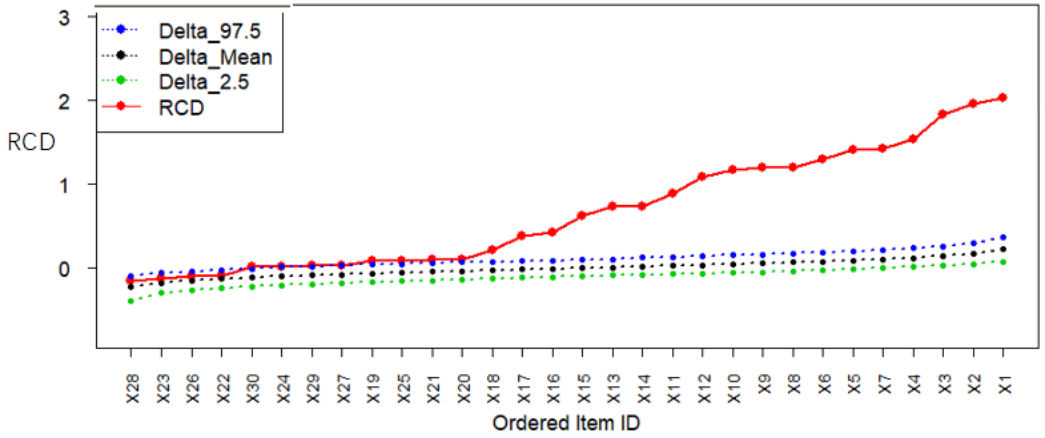
Item number	C1	C2	MH-C1	MH-C2
1	0	-1.0	N	Y^C
2	0	-0.9	N	Y^C
3	0	-0.8	N	Y^C
4	0	-0.7	N	Y^B
5	0	-0.6	N	Y^B
6	0	-0.5	N	Y^A
7	0	-0.4	N	N
8	0	-0.3	N	N
9	0	-0.2	N	N
10	0	-0.1	N	N
11	0	0	Y^A	N
12	0	0	N	N
13	0	0	N	N
14	0	0	Y^A	N
15	0	0	N	N
16	0	0	N	N
17	0	0	N	N
18	0	0	N	N
19	0	0	N	N
20	0	0	N	N
21	0	0.1	N	N
22	0	0.2	N	N
23	0	0.3	N	Y^A
24	0	0.4	N	Y^A
25	0	0.5	N	Y^A
26	0	0.6	N	Y^C
27	0	0.7	N	Y^B
28	0	0.8	N	Y^C
29	0	0.9	N	Y^C
30	0	1.0	N	Y^C

Graphical Test: RCD vs. MH

C3



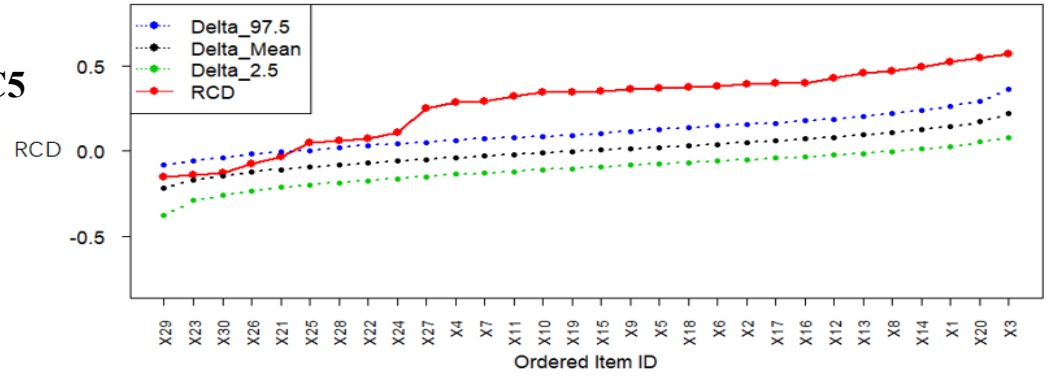
C4



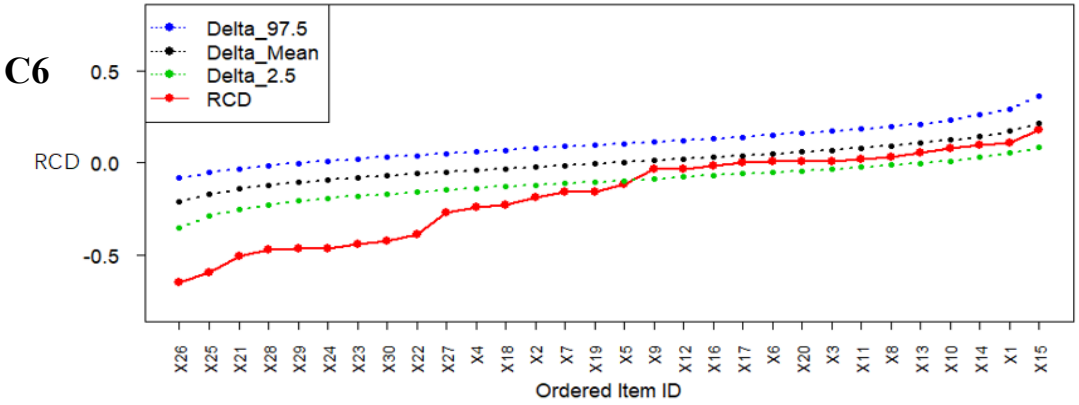
Item number	C3	C4	C3	C4
1	-0.4	-2.0	Y^A	Y^C
2	0	-1.9	N	Y^C
3	0	-1.8	N	Y^C
4	0	-1.7	N	Y^C
5	0	-1.6	N	Y^C
6	0	-1.5	N	Y^C
7	0	-1.4	N	Y^C
8	0	-1.3	N	Y^C
9	0	-1.2	N	Y^C
10	0	-1.1	N	Y^C
11	0	-1.0	N	Y^A
12	0	-0.9	N	Y^B
13	0	-0.8	N	A
14	0	-0.7	N	A
15	0	-0.6	N	A
16	0	-0.5	N	A
17	0	-0.4	N	A
18	0	-0.3	N	Y^A
19	0	-0.2	N	Y^B
20	0	-0.1	N	Y^B
21	0	0	N	Y^B
22	0	0	N	Y^C
23	0	0	N	Y^C
24	0	0	N	Y^B
25	0	0	N	Y^B
26	0	0	N	Y^B
27	0	0	N	Y^B
28	0	0	N	Y^C
29	0	0	N	Y^B
30	0	0	N	Y^B

Graphical Test: RCD vs. MH

C5



C6



Item number	C5	C6	C5	C6
1	-0.4	0	N	N
2	-0.4	0	N	N
3	-0.4	0	N	N
4	-0.4	0	N	N
5	-0.4	0	N	N
6	-0.4	0	N	N
7	-0.4	0	N	N
8	-0.4	0	N	N
9	-0.4	0	N	N
10	-0.4	0	N	N
11	-0.4	0	N	N
12	-0.4	0	N	N
13	-0.4	0	N	N
14	-0.4	0	N	N
15	-0.4	0	N	N
16	-0.4	0	N	N
17	-0.4	0	N	N
18	-0.4	0	N	N
19	-0.4	0	N	N
20	-0.4	0	N	N
21	0	0.4	Y^A	Y^B
22	0	0.4	N	Y^A
23	0	0.4	Y^B	Y^B
24	0	0.4	N	N
25	0	0.4	Y^A	Y^B
26	0	0.4	Y^B	Y^B
27	0	0.4	N	N
28	0	0.4	N	Y^A
29	0	0.4	Y^B	Y^B
30	0	0.4	Y^B	Y^A

04 Simulation Study

- Design
- DIF detection using the RCD-scree plot
- Results of RCD-CI: Type I error and Power

模拟研究

1.Design：

Study 1：

(1) DIF-free Test;

(2) Sample size: 500, 1000;

(3) μ = 0 and 0.5

Study 2：

(1) Sizes of DIF d=0.2, 0.4, 0.6, 和 0.8

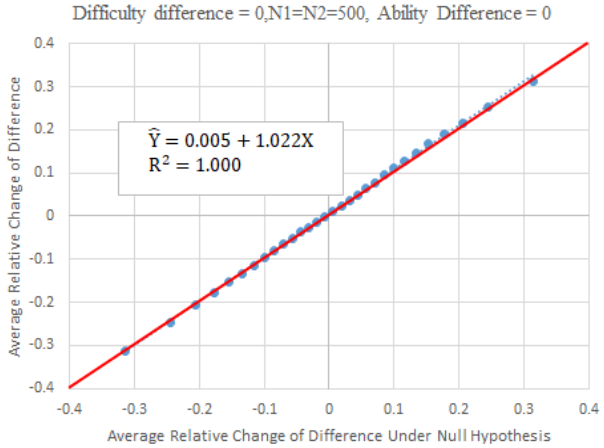
(2) DIF Pattern: Balanced vs. Unbalanced

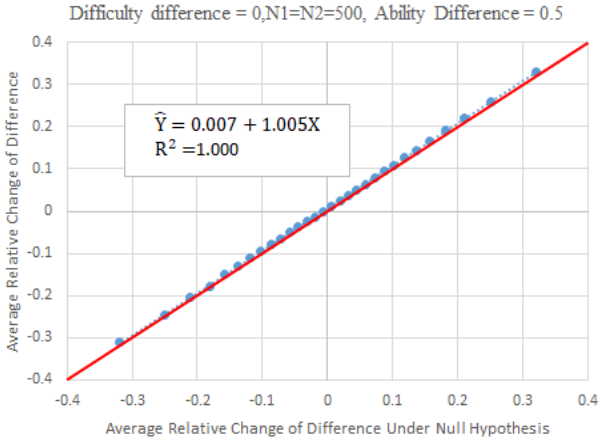
(3) Number if DIF items: 5

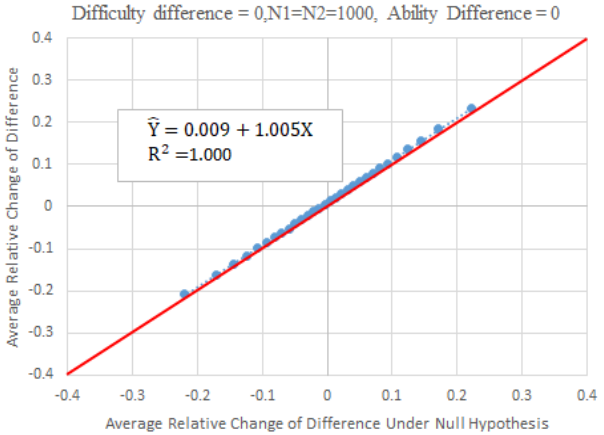
(4) Sample size: 500, 1000;

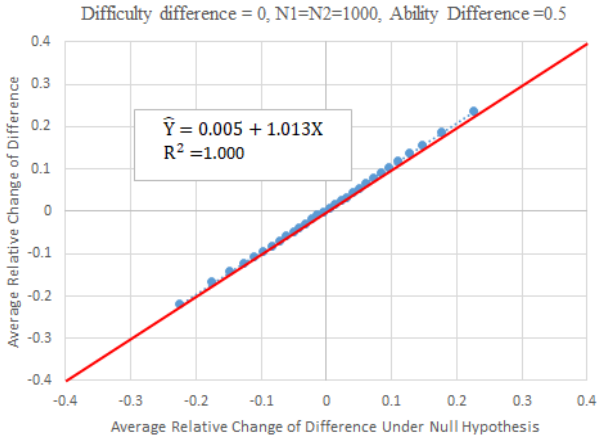
(5) μ = 0 and 0.5

模拟研究



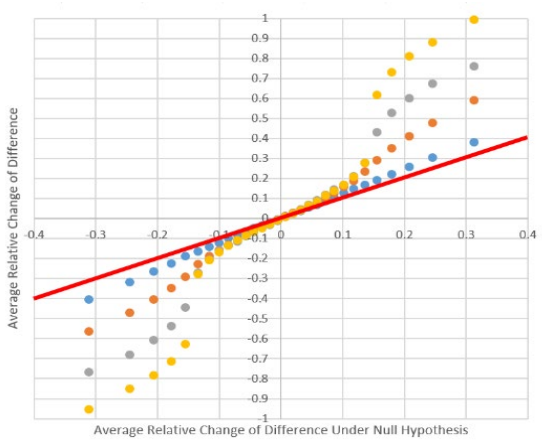




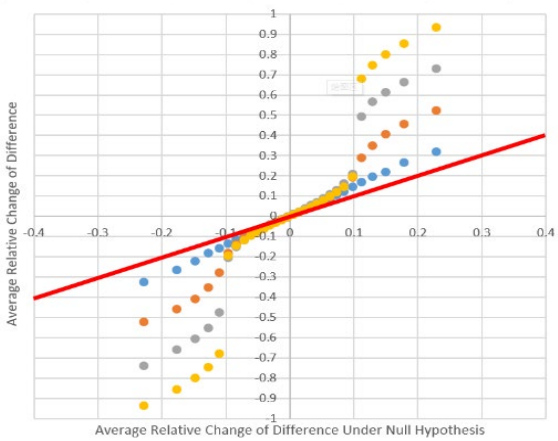


2. DIF detection using the RCD-scree plot:

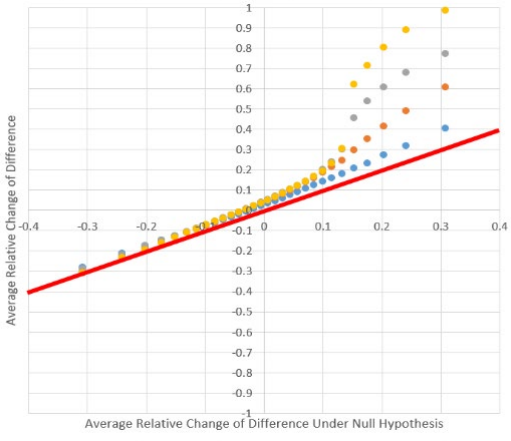
The figure contains the four scree plots of average RCD when all the items in the test are DIF-free, corresponding to the four conditions in study 1. The sample size or the value of the impact has essentially no effect on the plots.



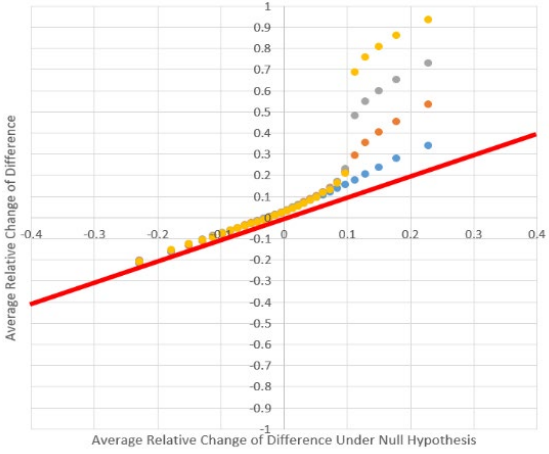
(a) Balanced $N_1=N_2=500$



(b) Balanced $N_1=N_2=1000$



(c) Unbalanced $N_1=N_2=500$



(d) Unbalanced $N_1=N_2=1000$

Figure. RCD-scree plots under different conditions when some items in the test are DIF.

Results: Study 1-Type I Error

Type I error (%) at the item and test levels when the null hypothesis is true

	Sample Size	Ability Difference	MH	MH-Holm	DIPF[a]	DIPF-Holm	**RCD-DIF**
Item Level	500	0	4.1	0.2	3.8	0.1	**0.5**
		0.5	4.2	0.2	4.1	0.1	**0.3**
	1000	0	4.4	0.2	3.0	0.1	**0.5**
		0.5	4.3	0.1	3.2	0.1	**0.4**
Test Level	500	0	72.2	4.6	1.4	15.2	**6.0**
		0.5	72.2	4.4	1.2	23.8	**4.0**
	1000	0	72.6	6.0	0.4	11.4	**4.0**
		0.5	73.4	3.8	0.6	12.2	**4.0**

Results: Study 2-Type I Error

Sample Size	Ability Difference	DIF SIZE	Balanced					Unbalanced				
			MH	MH-Holm	DIPF	DIPF-Holm	RCD-DIF	MH	MH-Holm	DIPF	DIPF-Holm	RCD-DIF
500	0	0.4	4.5	0.2	4.5	0.1	5.5	4.7	0.2	4.2	0.1	1.5
		0.6	4.5	0.2	4.2	0.2	7.0	4.5	0.2	4.2	0.1	2.3
		0.8	4.9	0.3	4.5	0.2	8.7	4.1	0.1	4.3	0.2	2.7
	0.5	0.4	4.8	0.3	4.4	0.2	7.9	4.1	0.1	4.0	0.1	2.2
		0.6	4.4	0.2	4.2	0.1	8.4	4.3	0.2	4.1	0.1	2.9
		0.8	4.1	0.2	4.7	0.2	7.5	4.6	0.2	4.2	0.1	2.8
1000	0	0.4	4.5	0.2	3.2	0.1	6.7	4.4	0.2	3.4	0.1	3.2
		0.6	4.6	0.3	3.1	0.1	7.9	4.3	0.2	3.0	0.1	3.4
		0.8	4.6	0.2	3.2	0.1	6.9	4.3	0.2	3.0	0.1	2.7
	0.5	0.4	4.3	0.2	3.1	0.1	6.7	4.4	0.2	3.1	0.1	2.9
		0.6	4.2	0.2	3.1	0.1	7.9	4.4	0.2	3.3	0.1	2.6
		0.8	4.6	0.3	3.3	0.1	7.6	4.7	0.2	3.2	0.1	2.3

Results： Study 2-Power

Sample Size	Ability Difference	DIF SIZE	Balanced					Unbalanced				
			MH	MH-Holm	DIPF	DIPF-Holm	RCD-DIF	MH	MH-Holm	DIPF	DIPF-Holm	RCD-DIF
500	0	0.4	68.9	29.1	44.5	13.0	72.5	66.3	25.0	33.6	4.8	43.8
		0.6	95.2	74.7	79.8	40.7	97.7	94.6	73.4	74.8	28.2	95.6
		0.8	99.1	93.9	96.7	78.2	100	99.5	95.4	96.0	70.4	99.6
	0.5	0.4	67.0	26.8	44.2	12.8	68.2	62.6	22.9	33.7	4.9	59.4
		0.6	93.8	71.0	78.9	40.2	95.6	92.2	70.5	74.2	28.4	94.2
		0.8	99.0	93.7	96.1	76.9	99.9	98.5	92.7	95.0	68.3	100
1000	0	0.4	93.5	70.1	79.2	41.1	96.9	93.7	68.1	75.7	32.2	91.8
		0.6	99.7	97.5	98.6	88.2	100	99.7	98.1	98.1	85.1	100.0
		0.8	100.0	99.7	100.0	99.5	100	100	100.0	100.0	99.1	100.0
	0.5	0.4	93.6	67.9	78.6	39.7	95.4	92.9	65.2	75.3	31.8	93.0
		0.6	99.7	97.4	98.3	87.9	100	99.5	96.6	97.4	82.6	99.8
		0.8	100.0	99.6	100.0	99.4	100	100	99.7	99.9	98.7	100.0

05 Application

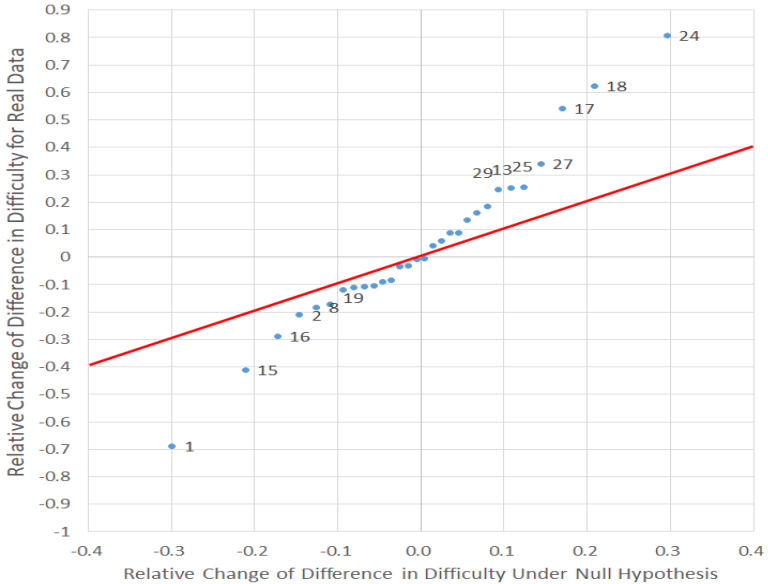
Application

Dataset

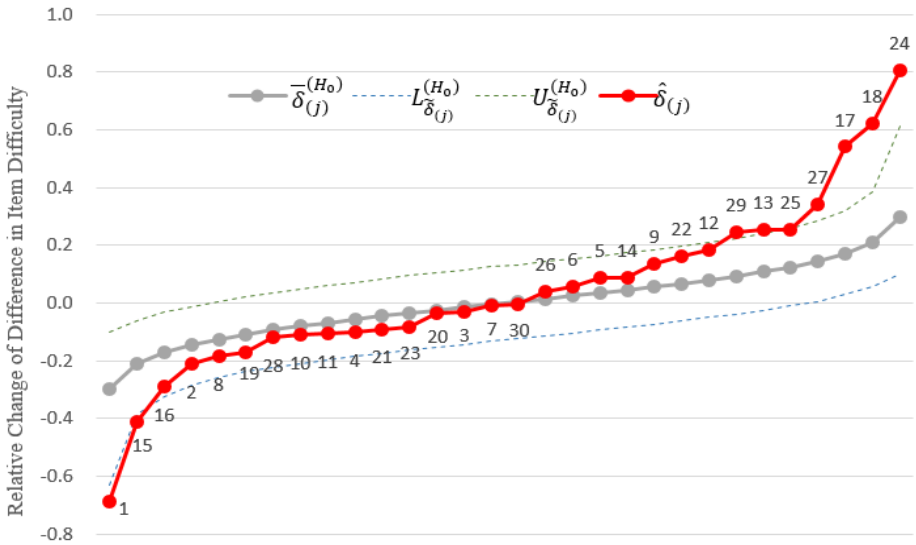
- PISA 2012 Math Test
- Test Length： 30
- 1081 students from Australia and 950 students from the United Kingdom.
- Methods: RCD, MH, DIPF

Item	DIPF	MH				RCD		
		Stat.	deltaMH	Class	$\bar{\delta}_{(j)}^{(H_0)}$	$\hat{\delta}_{(j)}$	Deviance	Class
t1	26	10.566*	1.711	Y^C	-0.329	-0.689	-0.360	Y^C
t15	20	10.000*	0.840	Y^A	-0.217	-0.411	-0.194	Y^A
t16	12	4.229	0.616	N	-0.174	-0.289	-0.115	N
t2	10	2.071	0.404	N	-0.146	-0.211	-0.065	N
….	….	….	….	….	….	….	….	
t12	9	2.432	-0.472	N	0.065	0.163	0.098	N
t22	8	2.895	-0.492	N	0.077	0.186	0.109	N
t29	14	4.788	-0.674	N	0.090	0.245	0.155	Y^A
t25	14	9.995*	-0.947	Y^A	0.104	0.253	0.149	Y^A
t13	14	11.253*	-0.991	Y^A	0.120	0.256	0.136	Y^A
t27	15	12.334*	-1.082	Y^B	0.141	0.341	0.200	Y^A
t18	19	26.800***	-1.238	Y^B	0.167	0.542	0.375	Y^B
t17	26	14.572**	-1.296	Y^B	0.210	0.624	0.414	Y^B
t24	27	65.179***	-2.232	Y^C	0.314	0.807	0.493	Y^C

Application



(a) RCD-scree Plot of PISA 2012 Math Test



(b) Graphical Test

Discussion

- RCD-DIF does not require pre-specification of anchor items. The DIF test at the item level is synchronized with that at the test level, and there is no issue of inconsistency.
- The RCD-DIF method provides a one-step classification of DIF and non-DIF items and does not need item purification.
- RCD-DIF can control Type I error both the item level and test level without adjustment procedures for multiple testing or iterative process for purification.
- Plots promote better understanding and conveniently communications in presenting and sharing the results.
- Our method can better deal with difficulty cases, especially when the majority of the items contain DIF.

Thank You!