Research and Evaluation in Education Journal
Volume 1, Number 1, June 2015 (100-113)
Available online at: http://journal.uny.ac.id/index.php/reid

MODIFIED ROBUST Z METHOD FOR EQUATING AND DETECTING ITEM PARAMETER DRIFT

[1)]Rahmawati; [2)]Djemari Mardapi
[1)]Center of Educational Assessment, Indonesia; [2)]Yogyakarta State University, Indonesia
[1)]rahmapepuny2011@gmail.com; [2)]djemarimardapi@gmail.com

Abstract

This study is aimed at: (1) revising the criterion used in Robust Z Method for detecting Item Parameter Drift (IPD), (2) identifying the strengths and weaknesses of the modified Robust Z Method, and (3) investigating the effect of IPD on examinees' classification consistency using empirical data. This study used two types of data. The simulated data were in the form of responses of 20,000 students on 40 dichotomous items generated by simulating six variables including: (1) ability distribution, (2) differences of groups' ability between groups, (3) type of drifting, (4) magnitude of drifting, (5) anchor test length, and (6) number of drifting items. The empirical data was 4,187,444 students' response of UN SD/MI 2011 who administered 41 test forms of Indonesian language, mathematics, and science. Modified Robust Z method was used to detect IPD and the IRT true score equating method was used to analyze the classification consistency. The results of this study show that: (1) the criterion of 0.5 point raw score TCC difference leads to 100% consistency on passing classification, (2) the modified Robust Z is accurate to detect the b and ab- drifting when the minimal length of anchor test is 25%, (3) IPD occurring on empirical data affected the passing status of more than 2,000 students.

Keywords: *Robust Z Method, Item Parameter Drift, IRT True Score Equating*

Introduction

The use of multiple test forms which is considered as parallel is widely implemented recently. Multiple test forms are used due to the test security, and to prevent the examinees from cheating easily to others. The other reason of designing parallel test forms is minimalizing the chance of practicing the test. If the administration of the test can be taken twice or more by a particular examinee, then using similar test form would kame the item get exposed frequently, the examinee may recall and practice the items.

Although the test is designed to be parallel, it is so hard to have the multiple test forms are perfectly parallel. Different item will have different level of difficulty, regardless similar resources of item's specification. The difference level of items' difficulties can raise unfair issues. The less difficulty test form will advantage the examinee who took the form, while examinee who took the more difficult item will get less score not caused by less ability. Thus, comparing the score between groups who took different test forms will lead to a bias result.

Non Equivalent Anchor Test (NEAT) design is a way to design parallel test forms, so that the difference of difficulty levels also the difference of groups' ability can be adjusted. The adjustment of differences is determined bu ancor items. Example of national test that using NEAT design is National Exam (NE) for elementary schools (ES) and *Madrasah Ibtidaiyah*/MI (Islamic-based elementary school) which is familiarly named as NE ES/MI. UN SD/MI items are constructed by provincial item writing team. All province used the same test specification and items' indicators. Each province then has their own test which differ from one province to others. To maintain the function of the test as a national measurement tool, 25% of the items were removed and replaced by national anchor items. The national ancor items were place in the same order, and preserve exactly the same content, format, even layout. No changes on national anchor items were allowed. All provinces had to make sure similarity of the anchor items.

The anchor items have a very important role. The accuracy of test form's difficulty level and the accuracy of examinee's ability estimation depend on the quality of anchor items. The score on anchor items defines the difference of groups' ability. A group which gets higher score on anchor items is considered as having better ability. Based on the ancor items' property, the difference of test form' level of difficulty can be determined and used for scoring adjustment (Cook & Eignor, 1991). Regarding its importance, the anchor items' parameter should satisfy the measurement invariance assumption. The assumption is that the parameter's value may shift around the bound of sampling error. Instead of being stable, anchor items' parameter are not uncommon shifting accross subsample, test administration, or location. These shifting conditions are known as item parameter drift (IPD) and may cause bias on ability estimation.

Keller and Wells (2009, p. 6) investigated the impact of drifting anchor items(IPD) on the accuracy of examinees' ability estimation. The study found that the difference of groups' ability defines the magnitude of IPD's impact. Even only one moderate drifting anchor item could give a bias ability estimation.

Robust Z method (Hyunh & Meyer, 2010) is a method for detecting drifting items and for fitting linking constants A and B which will be used in scaling process. Robust Z method applies a simple algoritm, yet still presents linking constant that is close to linking constant of the Stocking Lord method. The weaknesses of Robust Z method are its over-sensitivity and the absence of clear cut off criteria (Arce & Lau, 2011). The Robust Z method often detects undrifting anchor items as drifting. The criteria which are used are based on the probability of occurance in a hypothetic distribution; flagging an item as statistically significant IPD does not always mean that the impact of drifting ancors is practically significant.

Regarding the criteria problem, thus, modification of Robust Z method is necessary. The modification which is made is aimed at practically detecting meaningful IPD. Only anchor items which caused significant practical impact will be excluded from scaling process. The modification can give consideration to make decision for either retaining or refining the anchor items. An example of practically meaningful impact is changes on examinees' classification decision; passing to failing or failing to passing.

This study is aimed at: (1) revising the criterion which is employed in Robust Z Method so that the detection of item parameter drift (IPD) can be related to a practically meaningful criterion, (2) identifying the strengths and weaknesses of the modified Robust Z Method in various conditions, and (3) investigating the effect of IPD on the examinees' classification consistency in real life situation by implementing the modified Robust Z method on empirical data.

Research Method

Type of Research

The research is categorized as a descriptive study. The study described the strengths of modified Ribust Z method, compared to the original version. The study also described the weaknesses of the modified Robust Z method and identified the test's characteristics which were potential for having 'practically meaningful' IPD. The descriptions of IPD's impact on examinees' classification in real life situation were also revealed. The real life situation was illustrated by analyzing empiric data using the modified Robust Z method.

Time and Location

The research took place at Yogyakarta State University, Indonesia, the center of educational assessment, and a province that held item writing workshop for constructing NE ES/MI in the academic year of 2013. The research was conducted in 11 months, starting from March 2013 until February 2014.

Population and Sample

The population of this study was all students who were enrolled as examinee of NE ES/MI in the academic year of 2011 who took the main tests among all provinces in Indonesia. The main tests are defined as the tests which are administered on the main schedule of NE ES/MI. Students who took repeated session or make up session were excluded from the population. According to the population definition, the total number of the students in the research is 4,187,444.

Sample selection in this study was based on the result of cheating validation process. A school is considered as a cheating school if at least one item were identified as being responded identically incorrect by at least 90% students in the school. Identification of cheating school resulted exclusion of all students' responses of the identified school from the database. This cheating validation process eliminated about 40% of responses and the number of responses remained in the database were: 2,509,646 for bahasa Indonesia test, 2,509,517 for mathematics test, and 2,509,751 for science test.

Technical Steps on Modifying Robust Z Method

In order to improve the criteria of Robust Z method, the principle of the Difference that Matter (DTM) which is proposed by Brennan (2008, p. 108) at a topic of 'population invariance' was used. A way of considering an item as a drifting item is not only a statistical significance but an impact which is caused by the drifting items. How significant the impact is is determined by the researchers. The researchers set the practical impact which was considered as meaningful. In this study, the practical impact which was used to determine wether a drifting item was meaningful or not was the changes on classification consistency. If the detected drifting items made the score test equating changes significantly so causes any examinee classify differently, then the items considered as a practically meaningful IPD. It is suggested to exclude the practically meaningful IPD from scaling process, otherwise the decision of examinee classification may disadvantage both the examinee and the user.

The Robust Z method consists of several algoritms which, in the end, give the linking constant of A and B. These constants were then used in the scaling process to transform the scale of anchor and non anchor items' parameter from a focal test form into the same scale as the reference test form. The transformation of the items' parameter were used to plot the Test Characteristic Curve (TCC). The linking of point to point between TCC focal test and TCC transformed focal test became the conversion table for equating test score. The equated test score was then used to decide wheter an examinee passes or fails in the test.

In order to evaluate the IPD impact on modified Robust Z method, Wyse and Reckase (2011) formula was adapted. The formula was used to see the significant difference between TCC total and TCC refinement. TCC total is TCC of transformed focal test that used all anchor items for scaling process. TCC refinement is TCC of transformed focal test that using only non drifting anchor items for scaling process. If the difference between the two TCCs is small, then the impact of IPD on classification consistency can be waived. On the other hand, when the difference is big, then the IPD is practically meaningful and suggested to be excluded from the scaling process. In tis study, the cut off value of 0.5 point 'raw score' was used as the maximum difference between the two TCCs. This cut off ensured a hundred percent of classification consistency.

Equation (1), (2), (3), and (4) are the formulas which were used in modifying Robust Z method's citeria.

$$\Delta = Max\left| \sum_{iCV=1}^{n} P_{iCV}(\theta) - \sum_{iTOTAL=1}^{n} P_{iTOTAL}(\theta) \right| \tag{1}$$

$$P_{iCV}(\theta) = \frac{1}{1+\exp\left[-1{,}7a_{Y*CV}(\theta - b_{y*CV})\right]} \tag{2}$$

$$P_{iTOTAL}(\theta) = \frac{1}{1+\exp\left[-1{,}7a_{Y*TOTAL}(\theta - b_{y*TOTAL})\right]} \tag{3}$$

$$a_{Y*CV}=a_{Y}/A_{CV} \text{ dan } b_{Y*CV}=A_{CV}*b_{Y}+B_{CV} \tag{4a}$$

$$a_{Y*TOT}=a_{Y}/A_{TOT} \text{ dan } b_{Y*TOT}=A_{TOT}*b_{Y}+B_{TOT} \tag{4b}$$

Equation 4a and 4b are formulas which were used to calculate the linking constant of A and B in two different conditions: without refining IPD items(A_{tot} and B_{tot}) and by refining IPD items(A_{cv} and B_{cv}). Both A and B linking constants were used to scale both anchor items and non anchor items' parameter. The two kinds of A and B linking constants also lead to two kinds of TCC plots: TCC without refinement (ΣPi_{total}) and TCC by refining IPD (ΣPi_{cv}). The maximum absolute value of the difference between two TCCs was then compared to the DTM cut off value to find out the summary of practically meaningful IPD.

Data, Instrument, and Data Collection

Empirical data which were used in this research were collected by documentation process. The NE ES/MI of the year of 2011 data were copied from Center for Educational Assessment database. This concludes that the type of the data which was used was secondary data. The collected data were raw responses on the 41 test forms of bahasa Indonesia test, 41 test forms of mathematics test, and 41 test forms of science test. The key of each test form was also collected to complement the raw responses data sets.

The instruments which were used in this research was analysis software. There were 5 softwares which were used in this study, namely: WinGen, Bilog-MG, Winstep, R program, and Robust Z Modif. The software functions are: generating response data, validating responses, estimating item parameter, detecting IPD, constructing conversion table, and equating test scores.

Data Analysis

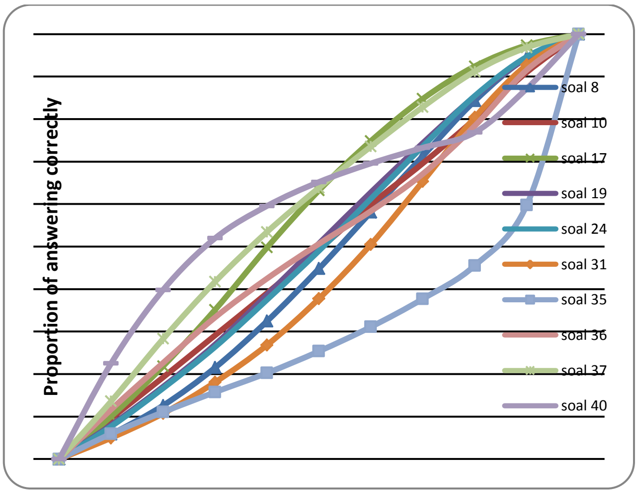

Figure 1. Curve of Proportion of Correct Responses on Mathematics Test Using 2.5 Million Responses of NE ES/MI 2011 Examinees

The analysis was started by determining the Item Response Theory (IRT) model that would be used. To find out the most suitable model, curves of raw score againts the proportion of students within each group that respond correctly on particular items were manually plotted. Figure 1 is an example of anchor items curve for mathematics test.

After deciding the IRT model which was used, simulation study data were generated using WinGen (Han, 2007) software. Each dataset generated was represented responses of 20,000 examinees on 40 dichotomus items. There are six manipulated variables: (1) The percentage of anchor items relative to total number of items (15%, 25%, and 40%); (2) the percentage of drifting items relative of total number of anchor items (15%, 30%, and 45%); (3) the magnitude of drifting. There are two kinds of drifting: the a-parameter drifting (no drifting, moderate drifting of 0.3, and large drifting of 0.7); the b-parameter drifting (no drifting, moderate drifting of 0.5, and large drifting of 0.8); (4) the direction of IPD (symmetrical two direction, one direction); (5) the ability distribution shape (normal and negatively skewed); and (6) comparison of the ability distribution between groups (similar ability distribution and different ability distribution).

In total, there are 188 conditions. Each manipulated condition was replicated 50 times for both the reference and the focal groups which resulted analysis of 18,800 datasets. The percentage occurance of manipulated drifting items detected as an IPD named as power rate, the percentage occurance of non manipulated drifting items detected as an IPD named as type I error rate, and the percentage occurance of TCCs differences larger than the cut off value named as DTM rate. The expected results from this study are combination of a high power rate, a low type I.

The analysis of empirical data was started with calibration of national anchor items using national responses. The parameter estimated from the national responses was then used as references for calibrating non anchor items in each province. The method which was used to calibrate provincial items is known as fixed item parameter calibration. The similarity of mean and standard deviation between non anchor test and anchor test was used to select the reference test form for equating process. After the reference test form was selected, equating score test of each provincial main test form can be conducted. For each provincial main test form, there are two equating processes: using all anchor items regardless the drifting and using only non drifting anchor items. Based on the two equating processes, each examinee will be classified two times. The classification consistency analysis categories examinees into four groups as follows: (1) passing and keep passing, (2) passing then failing, (3) failing then passing, and (4) failing and keep failing.

For each group, the proportion of examinees relative to total number of examinees was calculated. Classification consistency is the sum of proportion of examinees at groups of 'passing and keep passing' and 'failing and keep failing'. The analysis of empirical data also determined the frequency of each anchor item which was detected as an IPD accross 41 test forms. This frequency was named as IPD rate. The anchor item that has high IPD rate needs detail analysis on source of drifting. The expected results from the empirical data are a high percentage of classification consistency and a low IPD rate.

Findings and Discussion

Results of Simulation Study

The result of analysis power rate based on the type of ability distribution is presented in figure 2. The pattern of power rate of normal distribution is similar with the pattern of skewed distribution. Accross different level of drifting magnitude, the type of ability distribution does not present different results. It indicates that the performance of modified Robust Z method is similar with the two types of ability distribution.

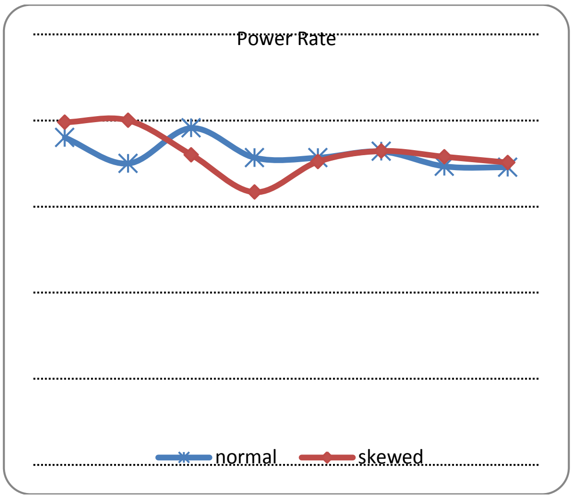

Figure 2. Power Rate Graph of Type of Ability Distribution accross Different Level of IPD's Magnitude

The modified Robust Z method is accurate when the ability of examinees in one group differs from the other group. Figure 3 and figure 4 are graphs of power rate and type 1 error rate IPD detection on interaction between number of anchor condition and difference of ability among group condition. Figure 3 shows that the modified Robust Z method is accurate when the number of anchor items is 40% and the groups are different in ability. A 100% of power rate means that the modified Robust Z method can detect manipulated drifting items accross all replications. A type 1 error rate close to 0% means that the occurance of detecting IPD incorrectly is almost close to zero.

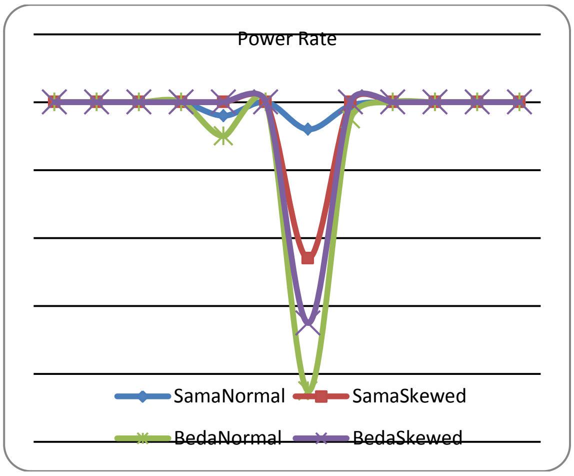

Figure 3. Power Rate Graph of Interaction between Type of Distribution and Ability Differences among Groups, Accross Number of Anchor Items and Type of IPD

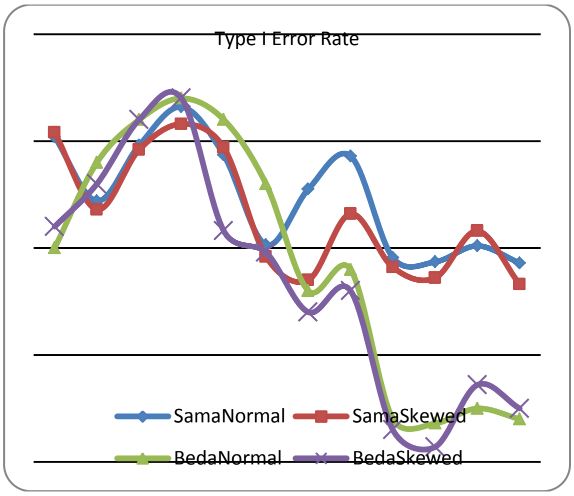

Figure 4. Type 1 Error Rate Graph of Interaction between Type of Distribution and Ability Differences Among Groups, Accross Number of Anchor Items and Type of IPD

The results presented in figure 5 shows that using 40% anchor items can mimimalize the impact of IPD on the classification consistency. The DTM rate for condition of 40% anchor items is close to 0%, not only for the type a-drift but also tyoe b-drift, for both moderate and large level of drifting magnitude. It concludes that designing multiple test forms using 40% of anchor items anticipates the impact of IPD that may arise. Although the anchor test may have an IPD, at least the impact of the IPD to classification consistency can be minimalized.

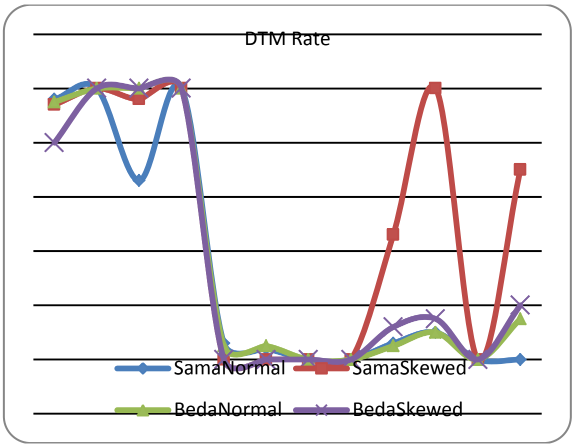

Figure 5. DTM Rate Graph of Interaction between Type of Distribution and Ability Differences Among Groups, Accross Number of Anchor Items and Type of IPD

Table 1. Power rate, Type I error rate, and DTM Rate Based on Anchor Test Length, Number of Drifting Items, and IPD Direction

Anchor Test Length	Number of Drifting	Power rate
15%	10% (one way)	100.0
	25% (symmetric)	100.0
	25% (one way)	98.3
	40% (one way)	17.1
25%	10% (one way)	91.4
	25% (one way)	99.4
	40% (symmetric)	95.5
	40% (one way)	38.6
40%	10% (symmetric)	100.0
	10% (one way)	100.0
	25% (symmetric)	90.0
	25% (one way)	97.1
	40% (symmetric)	100.0
	40% (one way)	9.5

The IPD detection rate accross different proportion of drifting items shows the weakness of modified Robust Z method as presented in Table 1. Table 1 shows that the power rate of modified Robust Z method is less than 20% in condition number of drifting items is 40% out of total number of anchor items. This finding summarizes that modified Robust Z method is not powerful to detect IPD when the proportion of drifting items in anchor test is big. Large proportion of drifting items makes the anchor items be distributed evenly around the fitting regression line, hiding the facts that many items were drifting. Overall, everything seemed normal and no outlier in the distribution. The modified Robust Z method failed to identify which anchors are drifting and which anchors are not.

Table 1 shows that the modified Robust Z method is still accurate in detecting many drifting items as long as the direction of drfiting is symmetric. A symmetric direction means that some items are drifting more difficult, while some others are drifting less difficult. It is shown that when the drifting items number is 40% of anchor test length, the power rate of one way direction is 9.5%, while the power rate of symmetric direction increases dramatically into 100.

Figure 6, figure 7, and figure 8 illustrate power rate, type 1 error rate, and DTM rate when direction of IPD distributions are one way and symmetrically two opposite direction. The results show that the modified Robust Z method perfoms better in looking the impact of IPD in test level not only in item level particularly. The practical impact of consistency classification is identified by modified Robust Z method as aggregate of items in test level. Even the number of drifting items were great, but when drifting in an opposite direction, the effect will cancel out and the practical impact can be waived.

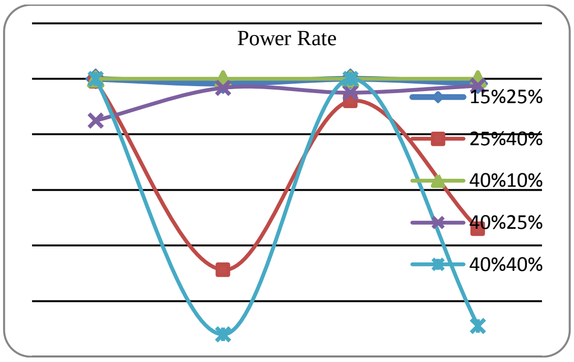

Figure 6. Power Rate Graph of Interaction Anchor Test length Condition, Number of IPD Condition, Ability Distribution, and IPD Direction.

The simulation study results show that the modified Robust Z method improves the performance of original Robust Z method specifically on test level. The original version cannot give conclusion on the impact caused by some detected drifting items as a part of a test. The original Robust Z method only justifies whether an item is drifting or not. The modified version adds information about the impact of all drifting items on the test score equating. This is similar to complement the analysis of differential item functioning (DIF) with differential test functioning (DTF) analysis. Many DIF items at the end can be waived if the DTF analysis performs no difference.

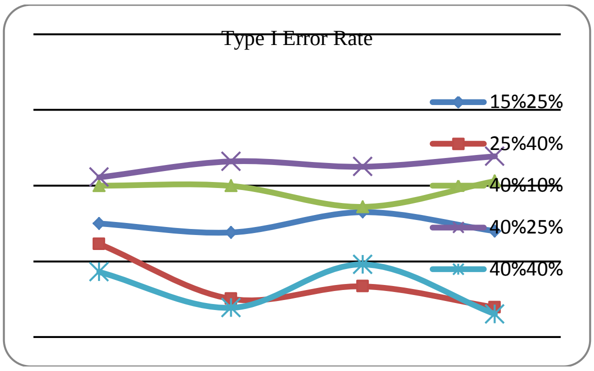

Figure 7. Type 1 Error Rate Graph of Interaction Anchor Test length Condition, Number of IPD Condition, Ability Distribution, and IPD Direction.

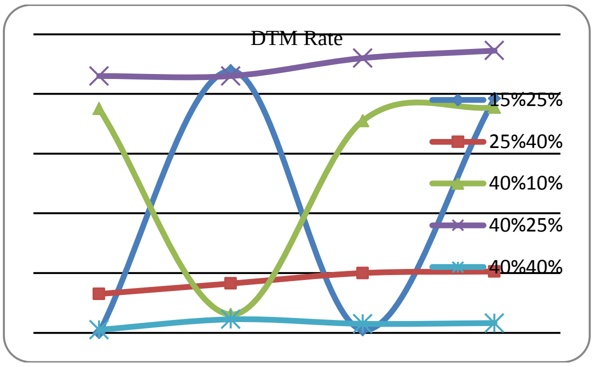

Figure 8. DTM Rate Graph of Interaction Anchor Test length Condition, Number of IPD Condition, Ability Distribution, and IPD Direction.

Results of Empirical Study

Table 2, table 3, and table 4 present the item parameter estimation for bahasa Indonesia test, mathematics test, and science test after they were calibrated using the national data. Each table represents the parameter of one test data which were conducted.

Table 2. Anchor Items Parameter for Bahasa Indonesia Test

Item Code	Location Parameter	Slope Parameter
BIN 18	-2.442	1.791
BIN 20	-2.035	0.829
BIN 21	-2.752	0.898
BIN 22	-2.173	1.147
BIN 23	-1.917	1.372
BIN 25	-2.640	1.307
BIN 27	-2.756	1.309
BIN 31	-1.796	0.705
BIN 32	-2.438	1.913
BIN 35	-1.999	1.304
BIN 36	-0.997	0.811
BIN 37	-2.186	0.942
BIN 40	-2.776	1.164

Table 3. Anchor Items Parameter for Matematics Test

Item Code	Location Parameter	Slope Parameter
MAT8	-0.722	1.261
MAT10	-1.622	0.635
MAT17	-1.328	1.350
MAT19	-0.939	1.145
MAT24	-0.888	1.227
MAT31	-0.632	1.283
MAT35	1.479	0.328
MAT36	-0.451	0.856
MAT37	-1.552	1.144
MAT40	-7.492	0.093

Table 4. Anchor Items Parameter for Science Test

Item Code	Location Parameter	Slope Parameter
IPA2	-2.869	0.974
IPA3	-2.611	1.102
IPA9	0.027	0.466
IPA10	-1.506	0.768
IPA18	-1.475	1.337
IPA23	-0.536	0.519
IPA27	1.204	0.257
IPA29	-2.225	0.843
IPA32	-0.889	0.951
IPA38	-2.348	1.132

The parameter of anchor test was used to calibrate non anchor items (fixed item parameter calibration) to select the best reference for the test form. IPD detection was implemented using modified Robust Z method in over 41 test forms for each subject. Table 5 presents the IPD rate for each anchor item.

Table 5. IPD Rate of Each Anchor Items over 41 Test Forms

ID	% IPD	ID	% IPD	ID	% IPD
Bin 18	**3**	Mat 8	**18**	Ipa 2	**45**
Bin 20	**8**	Mat 10	**90**	Ipa 3	**15**
Bin 21	**35**	Mat 17	**10**	Ipa 9	**93**
Bin 22	**13**	Mat 19	**13**	Ipa 10	**85**
Bin 23	**15**	Mat 24	**8**	Ipa 18	**0**
Bin 25	**48**	Mat 31	**28**	Ipa 23	**5**
Bin 27	**8**	Mat 35	**53**	Ipa 27	**46**
Bin 31	**58**	Mat 36	**20**	Ipa 29	**18**
Bin 32	**0**	Mat 37	**15**	Ipa 32	**3**
Bin 35	**68**	Mat 40	**100**	Ipa 38	**18**
Bin 36	**60**				
Bin 37	**20**				
Bin 40	**3**				
DTM	73	**DTM**	95	**DTM**	93

The results show that in bahasa Indonesia test, there is one anchor item which was detected as IPD, more than 60% anchor items which was detected in more than 85%, while science test has 2 anchor items detected as IPD in more than 85% provinces. The simulation study prooved that the modified Robust Z method has an accurate IPD detection. Then, the result of 85% IPD rate in empirical data means the item is truely drifting items.

The anchor items which were detected as drifting items were then taken into consideration while performing scaling process. The drifting items impact determined whether it is practically meaningful or not. Empirical data analysis considers the examinee as passing the test if the score of each subject is at least 4.00. Scoring process was conducted twice: in refinement condition and without refinement condition. For each subject, the examinee will have two passing statuses. Table 6, Table 7, and Table 8 present the examinee status proportion based on the scoring processes. Tabel 6 for bahasa Indonesia subject, Table 7 for mathematics subject, and Table 8 for science subject.

Table 6 summarizes analysis results of bahasa Indonesia test's passing status. Eleven out of 41 test forms used show that IPD does not make the difference of TCCs bigger than DTM criteria's cut off value. Careful examination on the eleven test forms proved that when the difference is less than DTM cut off value, the classification consistency is 100%. No examinee changes the passing status over two scaling conditions. It concludes that cut off criteria of 0.5 point raw score guarantee 100% classification consistency.

Table 7 shows that only one drifting item with large magnitude such as mat 40 has a large impact on classification consistency. The DTM rate for mathematics test is very close to 100%. The number of inconsistent classification at the national level is also very huge, about 25.58 %. This number is equal to 621,600 students regarding the numerous students for Indonesia population. This is a very huge number and significant result. These 621,600 students represent student population in East part of Indonesia.

The smallest percentage of inconsistent classification which is persented in Table 8 is 0.05%. This persentage seems small, but considering Indonesian huge population, this percentage is equal to 2050 students that enrolled in NE ES/MI 2011. If those 2050 students are assumed to continue their study in Junior High School/ *Madrasah Tsanawiyah* (Islamic-based JHS) which has capacity of 100 student, it means that 20 JHS/MTs will have under-quality students to be JHS/MTs students and passed the test just because of the measurement error.

Table 6 . Percentage of Classification Consistensy of Passing Status Based on Bahasa Indonesia Test

Test Form	Number of students	Pass/Pass	Pass/Fail	Fail/Pass	Fail/Fail	DTM
BIN_01_P01	61,195	99.74	0	0	0.26	No
BIN_01_P02	72,992	99.62	0.07	0	0.31	Yes
BIN_02	398,178	99.51	0.09	0	0.40	Yes
BIN_03_P01	116,156	99.84	0	0	0.16	No
BIN_03_P2/3	289,772	99.88	0.03	0	0.09	Yes
BIN_04_P01	43,573	99.95	0	0	0.05	No
BIN_05_P01	448,289	99.61	0.08	0	0.31	Yes
BIN_06_P01	45,432	97.54	0	0	2.46	No
BIN_07_P01	80,311	98.08	0	0.41	1.50	Yes
BIN_08_P01	69,929	99.79	0	0	0.21	Yes
BIN_09_P01	78,401	99.57	0.09	0	0.34	Yes
BIN_10_P01	16,133	98.76	0	0.23	1.01	Yes
BIN_11_P01	71,833	99.03	0	0	0.97	Yes
BIN_12_P01	98,296	99.44	0.14	0	0.42	Yes
BIN_13_P01	66,601	98.41	0.36	0	1.23	Yes
BIN_14_P01	25,581	99.19	0	0	0.81	No
BIN_15_P01	55,073	99.33	0.16	0	0.51	Yes
BIN_16_P01	55,640	99.47	0.11	0	0.42	Yes
BIN_17_P01	5,058	45.06	50.04	0	4.90	Yes
BIN_18_P01	11,002	98.85	0	0	1.15	Yes
BIN_19_P01	19,011	98.81	0	0	1.19	No
BIN_19_P02	12,944	99.79	0	0.06	0.15	Yes
BIN_20_P01	6,461	98.64	0	0	1.36	Yes
BIN_21_P01	5,692	97.86	1.35	0	0.79	Yes
BIN_22_P01	25,854	99.97	0	0	0.03	Yes
BIN_23_P01	45,069	98.35	0	0.32	1.33	Yes
BIN_24_P01	12,966	93.95	0	0	6.05	No
BIN_25_P01	16,592	92.19	0	0	7.81	Yes
BIN_26_P01	15,232	99.65	0	0	0.35	Yes
BIN_28_P01	17,629	99.93	0.02	0	0.05	Yes
BIN_29_P01	10,941	99.12	0	0	0.88	No
BIN_30_P01	127,834	99.21	0	0	0.79	No
BIN_31_P01	21,547	99.94	0.01	0	0.05	Yes
BIN_32_P01	10,401	97.66	0	0	2.34	No
BIN_33_P01	8,488	94.62	0	0	5.38	No
National	**2,500,100**	**97.42**	**1.32**	**0.03**	**1.23**	

Table 7 . Percentage of Classification Consistensy of Passing Status Based on Mathematics Test

Test Form	Number of students	Pass/Pass	Pass/Fail	Fail/Pass	Fail/Fail	DTM
MAT_01_P01	61,194	21.78	68.96	0.00	9.26	YES
MAT_01_P02	72,990	31.30	61.24	0.00	7.47	YES
MAT_02	398,119	85.47	9.41	0.00	5.12	YES
MAT_03_P01	116,119	92.60	0.00	1.62	5.78	YES
MAT_03_P02	137,498	94.63	0.00	0.00	5.37	NO
MAT_03_P03	152,318	89.60	2.44	0.00	7.96	YES
MAT_04_P01	43,569	97.13	0.00	1.41	1.46	YES
MAT_05_P01	448,303	92.48	0.00	1.40	6.12	YES
MAT_06_P01	45,403	62.37	29.90	0.00	7.73	YES
MAT_07_P01	80,314	66.51	23.06	0.00	10.43	YES
MAT_09_P01	78,393	0.79	93.84	0.00	5.37	YES
MAT_10_P01	16,133	68.62	21.38	0.00	10.00	YES
MAT_11_P01	71,822	66.89	28.01	0.00	5.10	YES
MAT_12_P01	98,266	60.83	27.88	0.00	11.29	YES
MAT_13_P01	66,600	32.40	28.92	0.00	38.68	YES
MAT_14_P01	25,581	62.54	31.11	0.00	6.36	YES
MAT_15_P01	55,078	72.25	7.49	0.00	20.26	YES
MAT_16_P01	55,636	69.45	14.52	0.00	16.02	YES
MAT_17_P01	5,058	22.18	57.85	0.00	19.97	YES
MAT_18_P01	11,004	69.43	21.21	0.00	9.36	YES
MAT_19_P01	19,015	74.55	14.18	0.00	11.26	YES
MAT_19_P02	12,949	88.96	9.17	0.00	1.87	YES
MAT_19_P03	25,767	78.74	12.93	0.00	8.32	YES
MAT_20_P01	6,456	69.05	20.12	0.00	10.83	YES
MAT_21_P01	5,692	80.32	16.30	0.00	3.37	YES
MAT_22_P01	25,854	97.51	1.54	0.00	0.96	YES
MAT_23_P01	45,073	66.86	22.40	0.00	10.73	YES
MAT_24_P01	12,966	4.90	26.02	0.00	69.08	YES
MAT_25_P01	16,551	45.77	32.56	0.00	21.67	YES
MAT_26_P01	15,231	73.29	13.35	0.00	13.35	YES
MAT_27_P01	8,252	37.71	36.40	0.00	25.88	YES
MAT_28_P01	17,629	14.95	78.37	0.00	6.68	YES
MAT_29_P01	10,942	77.19	13.53	0.00	9.29	YES
MAT_30_P01	127,831	82.37	14.18	0.00	3.46	YES
MAT_31_P01	21,548	25.37	60.24	0.00	14.40	YES
MAT_32_P01	10,407	62.27	22.30	0.00	15.43	YES
MAT_33_P01	8,488	6.95	81.73	0.00	11.32	YES
NATIONAL	**2,430,049**	**62.20**	**25.58**	**0.12**	**12.10**	

Table 8 . Percentage of Classification Consistensy of Passing Status Based on Science Test

Test Form	Number of students	Pass/Pass	Pass/Fail	Fail/Pass	Fail/Fail	DTM
IPA_01_P01	61,195	95.82	2.83	0.00	1.35	Yes
IPA_01_P02	72,988	96.07	2.63	0.00	1.30	Yes
IPA_02	398,196	92.70	5.98	0.23	1.09	Yes
IPA_03_P01	116,123	99.30	0.00	0.00	0.70	No
IPA_03_P02	137,504	99.65	0.12	0.00	0.22	Yes
IPA_03_P03	152,321	99.46	0.00	0.00	0.54	Yes
IPA_04_P01	43,570	99.86	0.05	0.00	0.10	Yes
IPA_05_P01	448,309	98.70	0.00	0.32	0.97	Yes
IPA_06_P01	45,444	97.33	0.00	0.71	1.96	Yes
IPA_07_P01	80,309	96.16	0.00	1.13	2.71	Yes
IPA_08_P01	69,932	99.11	0.30	0.00	0.59	Yes
IPA_09_P01	78,421	98.48	0.00	0.77	0.75	Yes
IPA_10_P01	16,133	98.76	0.00	0.00	1.24	Yes
IPA_11_P01	71,848	99.13	0.00	0.00	0.87	Yes
IPA_12_P01	98,279	98.00	0.00	0.00	2.00	Yes
IPA_13_P01	66,598	93.94	0.00	0.00	6.06	No
IPA_14_P01	25,580	98.32	0.00	0.00	1.68	Yes
IPA_15_P01	55,080	94.96	0.00	1.49	3.55	Yes
IPA_16_P01	55,639	96.74	0.00	0.00	3.26	Yes
IPA_17_P01	5,058	35.11	59.15	0.00	5.73	Yes
IPA_18_P01	11,003	96.95	0.79	0.00	2.26	Yes
IPA_19_P01	19,015	98.85	0.00	0.00	1.15	Yes
IPA_19_P02	12,948	99.86	0.05	0.00	0.08	Yes
IPA_19_P03	25,739	98.41	0.50	0.00	1.09	Yes
IPA_20_P01	6,457	98.44	0.53	0.00	1.04	Yes
IPA_21_P01	5,692	98.84	0.67	0.00	0.49	Yes
IPA_22_P01	25,855	99.82	0.00	0.00	0.18	Yes
IPA_23_P01	45,073	97.60	0.00	0.71	1.70	Yes
IPA_24_P01	12,966	93.38	1.61	0.00	5.01	Yes
IPA_25_P01	16,595	90.32	0.00	0.00	9.68	Yes
IPA_26_P01	15,232	99.38	0.00	0.00	0.62	Yes
IPA_27_P01	8,253	92.88	3.55	0.00	3.57	Yes
IPA_28_P01	17,628	98.79	0.96	0.00	0.25	Yes
IPA_30_P01	127,839	99.65	0.00	0.00	0.35	Yes
IPA_31_P01	21,547	94.10	5.43	0.00	0.48	Yes
IPA_32_P01	10,408	97.53	1.40	0.00	1.07	Yes
IPA_33_P01	8,494	94.28	0.00	0.00	5.72	No
NATIONAL	**2,489,271**	**95.59**	**2.34**	**0.14**	**1.93**	

A deep attention must be put to answer the results of classification consistency. The table shows that inconsistent classification is mostly in categories of passing, while in fact, the status is failing. This means that thousand even hundred thousands students are decided as passing the test, while in fact, their competencies are still below the standard. This inconsistency has a big influence because the NE score is then used as a selection tool for ebtering secondary schools. The starting point of learning process cannot be in the right starting point. The students need to repeat or remedy what their lack of for their primary school's competencies before continuing to a higher level of competency.

Summary and Suggestion

Summary

The analysis proved that external criteria of 0.5 point raw-score TCC difference for modifying Robust Z method can make the modified Robust Z method able to give information about the consequencies of IPD to classification consistency. If the difference of TCC is less than 0.5 point raw-score, then the classification of consistency will be 100%.

The modified Robust Z method perfoms better on looking the impact of IPD in test level not only particularly in item level. The practical impact of consistency classification is identified by modified Robust Z method as aggregate of items in test level. Even the number of drifting items were great, but when drifting in an opposite direction, the effect will cancel out and the practical impact can be waived.

The implementation of modified Robust Z method in empirical data shows that the impact of IPD was very significant for NE ES/MI 2011 examinees. At least 2000 students were classified as passing, while in fact, their competencies were not sufficient to pass the exam and continue to secondary education.

Suggestion

The use of multiple test forms is more frequent. Score test equating process has to be performed. The heterogenity of ability accross provinces in Indonesia is also potential for the occurrence of drifting items, which in the study has a big impact on the classification consistency. Regarding to those facts, then the modified Robust Z method is suggested to be used for both detecting drifting items and equating test score, especially when the design of the multiple test form employs set of ancor items, passing classification.

The analysis shows that in order to minimalize the effect of IPD on classification consistency, it is suggested to have 40% of anchor test length. This proportion has quite big risk both from the security of anchor items from being too exposed and also less variance items accross provinces. The rule of thumb of anchor test length is 20% (Hambleton, Swaminathan, & Rogers, 1991). To have better prevention of drifting items yet still maintain the item exposure and variability accross provinces, the 40% anchor test length can be constructed in matrix sampling design. Split the anchor tests into several clusters. One cluster to others shares overlapped items.

This study also has limitations. The condition simulated in this study is too few to represent all variance of conditions in a real life situation. Then it is suggested to extend this study using broader condition so that the strengths and weaknesses of modified Robust Z method can be comprehensively analyzed.

This study also only estimates the impact of drifting items on classification consistenty, and there is no analysis performed to see the performance of modified Robust Z method on ability estimation accuracy or scaling equation accuracy. Thus, a study which employs similar method but focuses on the consequences of ability estimation accuracy is very suggested. The results of ability estimation accuracy or scaling constant accuacy will complement the rsults of this study.

References

Arce, A. J. & Lau, A. C. (2011). *Statistical properties of 3PL Robust Z: An investigation with real and simulated data sets.* Paper presented in the Annual Meeting of the National Council on Measurement in Education, in New Orleans, Lousiana.

Brennan. (2008). A discussion of population invariance. *Applied Psychological Measurement.* Volume 32 (1), pp. 102-114.

Cook, L. L. & Eignor, D. R. (1991). IRT equating methods. *Educational*

Measurement: Issues and Practice, 10, pp. 37-45.

Hambleton, R. K., Swaminathan. H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Han, K. (2007). WINGEN: Windows software that generates IRT parameter and item responses. *Applied Psychological Measurement,* 31, pp. 457–459.

Huynh & Meyer. (2010). Use of Robust Z in detecting unstable items in item response theory models: Practical assessment. *Research and Evaluation Electronic Journal,* 15 (2).

Keller & Wells. (2009). *The effect of removing anchor items that exhibit differential item functioning on the scaling and classification of examinees*. Paper presented in the annual meeting of NCME, in Denver.

Wyse & Reckase. (2011). A graphical approach to evaluating equating using test characteristic curve. *Applied Psychological Measurement*, 35 (3), pp. 217-231.