Jurnal Penelitian dan Pembelajaran IPA JPPI
Vol. 5, No. 2, 2019, p. 94-111 p-ISSN 2477-1422 e-ISSN 2477-2038

Using Rasch Model to Detect Differential Person Functioning and Cheating Behavior in Natural Sciences Learning Achievement Test

(Received 24 July 2019; Revised 28 November 2019; Accepted 29 November 2019)

Purwo Susongko[1*], Mobinta Kusuma[2], Heru Widiatmo[3]

[1,2]Department of Science Education, Faculty of Teacher Training and Education,
Universitas Pancasakti Tegal, Tegal, Indonesia
Corresponding Author: *purwosusongko@upstegal.ac.id

[3]American College Testing, Iowa, United States

DOI: 10.30870/jppi.v5i2.5945

Abstract

The existence of abberant response showed the inaccuracy of measurement which, in turn, threatens the test validity. This study aims at: (1) Discovering the proportion of students who were having Differential Person Functioning (DPF) in final term assessment test of natural sciences course for 8th grade in odd semester of academic year 2016/2017 in Tegal Regency, Indonesia; (2) Identifying the students suspected of cheating during the final term assessment test of natural sciences course for 8th grade in odd semester of academic year 2016/2017 in Tegal Regency, Indonesia. This research involved 1011 student responses to final term assessment test of natural sciences course for 8th grade in odd semester of academic year 2016/2017 in Tegal Regency, Indonesia. The data were taken from four junior high schools; SMPN I Dukuhturi, SMPN I Suradadi, SMPN 2 Slawi and SMPN 2 Dukuhwaru. The scoring was done using Rasch model and the person fit index used Sijtsma's Ht person fit statistic (Ht). The result showed: (1) 14% of the students attending final term assessment of natural sciences course for 8th grade in the odd semester of academic year 2016/2017 in Tegal Regency, Indonesia were detected of having DPF. From the four junior high schools involved in the research, three junior high schools have a proportion of students having DPF at a range from 9.6% to 23%, and only 1.1% of the other one's students were having DPF; (2) Nine pairs of students were suspected of cheating during the final term assessment of natural sciences course for 8th grade in the odd semester of academic year 2016/2017 in Tegal Regency, Indonesia.

Keywords: Detection, Differential Person Functioning, Science Achievement Test

INTRODUCTION

One of the main functions of learning achievement test is to determine students' final competence achievement and the result could be used to consider to promote them to higher class, to declare their graduation of a school, to grant them a certain certificate. This fact causes the tests used are of high-stake nature such as final term assessment, final year assessment and national examination. Therefore, all parties related to the organization of these tests must ensure that the information obtained from the tests are acurate and fair. This is because the test score accuracy could affect the students and teachers' lives and brings significant consequences. The test developers are responsible for figuring out any possible threat to its validity while making and using the assessment.

Sireci (2007) concluded some fundamental aspects of validity: (1) Validity is not owned by a test. On the contrary, validity refers to the use of a test for a certain purpose, (2) To evaluate the use and feasibility of a test for a certain purpose, many sources of evidence are needed, (3) If the use of a test should be maintained for a certain purpose, an adequate evidence should be proposed, (4) Evaluating a test's validity is not a one-time, static event; It is a continuous process.

Messick(1996) argued that validity is a single concept which is expressed as a construct validity consisting of six elements each: (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external and (6) consequential. The substantive aspect is associated with the substance of the content aspect. This is achieved by finding it out empirically to ensure that the test takers actually really involve the ability in the field being measured in answering the test items. For example, in a multiple choice test, the test takes chosing the wrong answers (distractors) actually have a low ability. When it is found that some test participants with low ability can answer the high-level items then it can be said that the construct validity of substantive aspect is hampered.

The consistency of students in answering these items constitutes the measurement of construct validity of substantive aspect. In Rasch model, the student's consistency measurement in answering the test items is called Person fit statistic. The measurement in education assessment using Rasch model will have the same quality as the measurement made in the physical dimension in physics (Sumintono, & Widhiarso, 2014; Sumintono, 2018). In the modern test theory measurement, Rasch model is viewed as the most

objective measurement model. The concept of objective measurement in social sciences and education assessment according to Mok and Wright (2004) should have five criteria, they are: (1) Providing a linear measurement with the same interval, (2) Doing the right estimation process, (3) Finding the items which are not right (*misfits*) or unusual (*outliers*), (4) Dealing with the lost data, (5) Producing a replicablemeasurement (independent from the parameter being studied). Of these five requirements, so far only Rasch model is capable of fulfilling them. Wu and Adams (2007) showed that the use of Rasch model in educational measurement has its strengths in its high specific objectivity and item parameter estimation stability. Rasch model connects the chance to answer correctly each item (P(θ)) as a function of ability (θ) to the item difficulty level constant (b) through a relations as in equation 1.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (1)$$

This Rasch model has been developed further separately from IRT.It has even been developed far wider in polytomous scoring. By involving only one item parameter, the item parameter estimation or participant in Rasch model requires less data in its estimation than other models. The application of Rasch model in learning achievement since its introduction by its inventer Georg Rasch in 1960 has now been expanding to include not only the education world, but also the medical and public health fields (Luet al, 2013;Smithet al, 2010; Ayele et al, 2014). Validity Messick(1996) if applied with the use of Rasch model can be explained in Susongko (2016).

Based on Rasch model, students will only be able to answer the item with the same maximum difficulty level as the students' ability. The deviation from this is shown by their "abberant" or "inconsistent" responses. An abberant pattern of responses is the pattern of students' responses which does not match the expected model (Perkins, 2013). This occurs if the students successfully answer an item with a difficulty level above their ability and, on the contrary, fail to answer the item with a difficulty level lower than their ability. Abberant response can be caused by cheating, careless responding, creative responding, lucky guessing and random responding (Karabatsos, 2003, Meijer, 1996). The abberant response caused by cheating has been the main attention in test quality (Belov & Armstrong, 2010). The existence of abberant response shows the inaccuracy of measurement which, in turn, threatens

the test validity (Karabatsos, 2003; Scherbaum, 2003).

Abberant response causes the data to not match the model used. To what extent this data unmatching with the model could be seen from the existence of differential item functioning (DIF) and differential person functioning (DPF) (Johanson & Alsmadi, 2002; Engelhard Jr, 2009). DIF is a concept well-known in educational assessment studies (Gierl et al, 1999; Susongko,& Mardapi, 2000; Kalaycioğlu, & Berberoğlu, 2011;Feuerherd et al, 2014; Strobl et al, 2015; Dewi & Prasetyo, 2016; Luo et al, 2017; Hays, 2018). DIF is defined by Clauser and Mazor (1998) as the probability of different success in an item among groups. DIF is associated with the different item function in two groups of test takers. For example, an item is considered harder in a group than in another one, thus the chance of answering it correctly in both groups with similar ability becomes different.

DPF can be defined as the unexpected difference between the observed and expected performance from an individual in doing the test or a set of items (Engelhard Jr, 2009; Alsmadi& Alsmadi, 2009). DPF studies basically are the test to item invariance in measurement. DPF is said to be present if there is a difference in the pattern of responses from two groups of different items given to the same students (Scherbaum, 2003). Several methods are available to detect DPF. Karabatsos (2003) detected DPF by testing Person Fit index and Emons et al (2005) detected it by considering the person-response functions (PRF). Engelhard Jr (2009) used four approaches in determining DIF and DPF, namely: (1) Main-effects Model, (2) Condition- Group interaction, (3) item-group –condition interaction, (4) Person Fit.

Person fit with Rasch modelling is expressed with Outfit Mean Square (MNSQ) and Infit Mean Square (MNSQ) indices (Meijer & Sijtsma, 2001; Petridou & Williams, 2007). Outfit MNSQ and Infit MNSQ give a quantitative of to what extent an individual deviates from the expected model. However, Person Fit is different from DPF since DPF analysis does not merely detect Person Fit, rather it also gives a more comprehensive explanation regarding the causes of students' abberant response (Perkins,2013).

Final Term Assessment is an activity a junior high school administers to measure their students' competence achievement by the end of semester at Indonesian schools since the 2013 Curriculum was introduced. The scope of final term assessment includes all indicators which represent the whole

Basic Competence (KD) during the said period. Final Term Assessment serves important functions since in addition for a report of final term assessment, it is also considered in the criteria for passing to a higher grade. For this reason, Final Term Assessment is also called as high-stake test. Final Term Assessment in Tegal Regency is administered simultaneously using the same test instrument. This test instrument wasmade by Course Teacher Forum in natural sciencescourse test. During the administration of high-stake test, there is an extremely high possibility of cheating behavior, careless responding, creative responding, lucky guessing and random responding shown by abberant responses. To discover further the construct validity of substantive aspect of Final Term Assessment administered in Tegal Regency so far, detecting DPF in participants needs to be done to figure out the cause of abberant responses from the students.

It is also expected that the results of this research could detect the number of students who are cheating in doing their Final Term Assessment. Cheating is a despicable action in education process, thus during the teaching and learning process students must be prevented from cheating. This is in line with the policy of the Ministry of Education of the Republic of Indonesia in relation to Strengthening Character Education to students at all education levels. The teaching and learning in all of its aspects, including assessment, should be capable of teaching such values as being religious, honest, tolerant, disciplined, hard-working, creative, independent, democratic, curious, nationalistic, patriotic, appreciative to achievement, communicative, peace-loving, fond of reading, caring for environment, socially caring, and responsible (Ministry of Education and Culture, 2018). Through test at Final Term Assessment students can at least learn such values as being honest, hard-working, disciplined, independent, appreciative to achievement and responsible. Of course, these values would be implanted within students if during the administration the test, students work on their tests in a sportive manner and one of the psychometric indicators which could be detected is whether or not if the students have cheating behavior. The initial step in preventing cheating behavior is to make an examination system which would not allow students to cheat. Furthermore, there is a need to identify the cheaters both through direct observation and analysis of students' responses. The identification of abberant response combined with information on

ability through Rasch model application can be used as the basis in determining the students suspected of cheating during the test.

To answer these problems, this research aims at: (1) Discovering the proportion of students who are having Differential Person Functioning (DPF) in final term assessment test of natural sciences course for 8thin odd semester of academic year 2016/2017 in Tegal Regency, (2) Identifying the students suspected of cheating during the final term assessment test of natural sciences course for 8th grade in odd semester of academic year 2016/2017 in Tegal Regency, Indonesia.

METHOD

Emons et al (2005) detected DPF through three steps, those are: (1) Global analysis, (2) Graphical Analysis and (3) Local Analysis. Global analysis is used to identify Person Fit. Graphical analysis is used to determine pattern from Person Respons Function (PRF) and Local analysis is used to determine statically DPF appearing. Karabatsos (2003) compared 36 index to test Person Fit baik empirically or simulation. Research result showed that: (1) Ht index (Sijtsma'sHt person-fit statistic) is the best index among the other person fit, (2) Ht index is the best index in cheating detecting. Sijtsma& Meijer (1992) and Tendeiro& Meijer (2014) also proved that Ht index is the best measurement for Person Fit.

Rupp (2013) explained detection of abberant responds comprehensively must be able to answer five levels include: (1) How many persons respond aberrantly? (2) What kinds of persons respond aberrantly? (3) How do they respond aberrantly to selected items? (4) How many selected items do they respond aberrantly? (5) What kinds of items do they respond aberrantly? To reach them, researcher must use many techniques involved quantitative and qualitative aspect.

Detection of DPF in this study is confined in two questions as explained by Rupp (2013). To reach research objectives, it had been done some steps those: (1) detection of students proportional which have abberant respond experience in junior high school natural sciencesFinal Term Assessment 8th grade in odd semester of year academic 2016/2017 in Tegal regency, (2) detection of students which have DPF experience in in junior high school natural sciencesFinal Term Assessment 8th grade in odd semester of year academic 2016/2017 in Tegal regency, (3) Detection of cheating possibility in students who have the same scores in junior high school natural sciencesFinal Term Assessment 8th grade in odd semester of year academic 2016/2017 in

Tegal regency. All steps use measurement based on Rasch Model with used person index of Ht. To detect the possibility of cheating behavior on students used some criteria: (1) it is happened to students who sit close each other by looking at their test numbers and the same scores; (2) Copier defined by looking at students Ht index under cut off boundary from Person Fit Score/ PFS and considering appearing of PRF, (3) Source of cheating sheets defined by looking at students Ht index above cut off boundary of PFS and considering appearing.

Person response function (PRF) is the function which connect item with the probability someone answer correctly an item (Tendeiro et al, 2016). As consistent test participant, if they are given some dikotomis test items, they are expected able to answer correctly for easy item and they are failed to answer difficult item. More items is lower the probability to answer them correctly. This is the general principle of PRF. Deviation of general principle indicates there is significant abberant responds and finally it can show DPF existence. Figure 1 shows ideal PRF where test participants answer consistently while Figure 2 shows PRF with DPF.

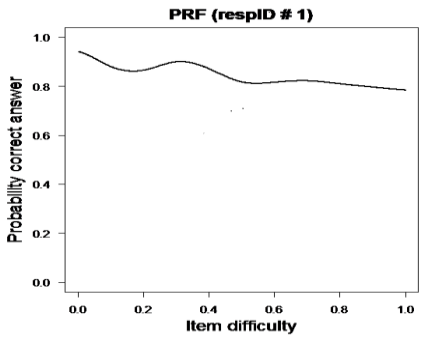

Figure 1. PRF for Consistent Response

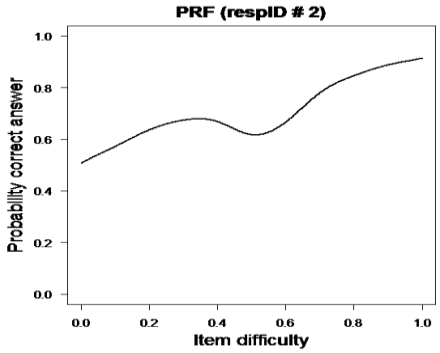

Figure 2: PRF for Abberant Response

1011 responses were given by eight graders from four schools involved in this research as shown in Table 1. The sampling technique used is convenience sample, in which respondents are selected based on convenience and availability (Creswell, 2010). All students are given sequence numbers from one school to another. The test instrument of final term assessment of natural sciences course for 8th grade in the odd semester of academic year 2016/2017 is made by the Natural Sciences Course Teacher Forum within the Education Department of Tegal Regency. The natural sciences test instrument consists of 35 multiple choice items and 5 essay items. This research is limited to only investigating

the the multiple choice test. The test item analysis with Rasch modelling uses eRm package version 0.15-6 (Mair et al, 2016) and the analysis of Ht statistics and PRF uses PerFit package version 1.4.1 (Tendeiro & Tendeiro, 2016). Both packages are run using open source R program version 3.4.3 .

Table 1. Distribution of Students Involved in the Research

School Name	Respon-dent Number	Number of Respondents
SMPN 1 Dukuhturi	1-293	293
SMPN 1 Suradadi	294-480	187
SMPN 2 Slawi	481-816	336
SMPN 2 Dukuhwaru	817-1011	195
Total		1011

RESULTS AND DISCUSSION

The difficulty level of multiple choice test item in natural sciencesFinal Term Assessment odd semester of academic year 2016/2017 8th grade in Tegal regency by using Rasch model is in the range of -1.693 (item of number 1) until 1.041 (item of number 15), it is shown at Table 2. It also shows the difficulty level of test items is in the exact range based on test participant ability generally from -2 to +2 (Hambleton et al, 1991), it is shown in Figure 3. Test item of number 1 measures student's competence in menstruation cycle in women while test item measures student's competence in the principle of food oxide reaction.

Table 2. Item Difficulty Level of natural sciences Final Term Assessment Test for 8th Grade in the Odd Semester of Academic Year 2016/2017 in Tegal Regency

Item No	Diffi-culty level	Diffi-culty level after being sequenced	Item No
1	-1.639	-1.693	1
2	0.909	-1.511	11
3	0.073	-1.155	33
4	-0.137	-0.919	35
5	- 0.353	-0.631	9
6	-0.525	-0.525	6
7	0.010	-0.458	14
8	-0.086	-0.413	17
9	-0.631	-0.403	12
10	-0.050	-0.353	5
11	- 1.511	-0.227	32
12	-0.403	-0.152	19
13	0.951	-0.137	4
14	- 0.458	-0.086	8
15	1.041	-0.054	29
16	-0.036	-0.050	10
17	- 0.413	-0.036	16
18	0.371	-0.018	27
19	-0.152	-0.008	26
20	0.951	0.010	7
21	0.700	0.073	3
22	0.407	0.136	31
23	0.345	0.176	28
24	0.407	0.314	34
25	0.975	0.345	23
26	-0.008	0.371	18
27	-0.018	0.407	22
28	0.176	0.407	24
29	- 0.054	0.700	21
30	1.008	0.909	2
31	0.136	0.951	13
32	-0.227	0.951	20
33	- 1.155	0.975	25
34	0.314	1.008	30
35	-0.919	1.041	15

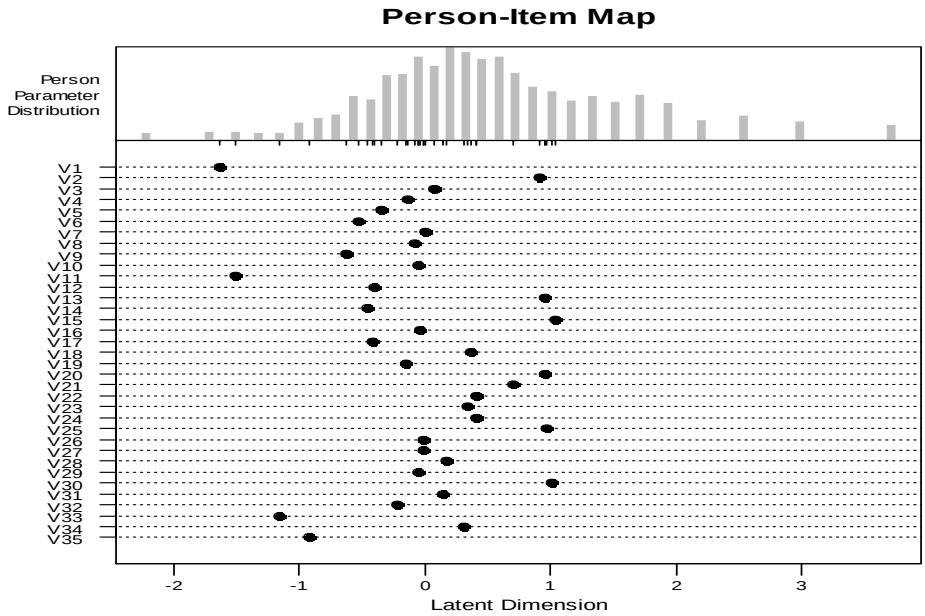

Figure 3. Person-Item Map of Natural Sciences Final Term Assessment Test Instrument for 8th Grade in the Odd Semester of 2016/2017 Academic Year in Tegal Regency

With person Fit analysis using Ht Statistics, 142 students or 14 % of all students are identified having DPF with a cut off score from PRF score of 0.0168. The student's identity numbers (student's sequence numbers) detected as having DPF are shown in Table 4. For example, the PRF for students detected as not having and having DPF could be seen in Figures 4 and 5.

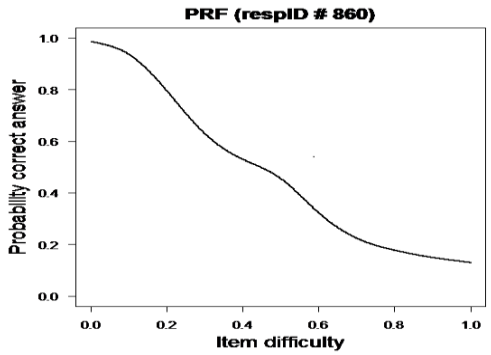

Figure 4.PRF of students detected as not having DPF

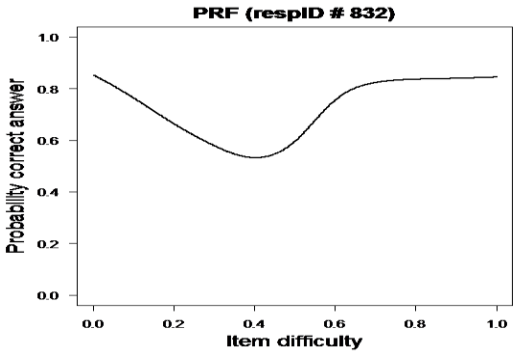

Figure 5. PRF of students detected as having DPF

It can be seen from Figure 3 that students detected as not having DPF are consistent, meaning that the higher the difficulty level of the items is, the less likely for them to answer them correctly. On the other hand, in those students detected as having DPF, the pattern of PRF is inconsistent and even shows a tendency that the higher the difficulty level of an item is, the more

likely for them to answer correctly. Likewise, 73 groups of students are detected as having equal ability and sitting next to each other as shown in Table 3. The list of student groups sitting next to each other with equal ability and detected as having DPF could be shown in Table 4. Based on the information in Tables 3 and 4, the great possibility for cheating behavior to occur between students can be determined. The basis is two students with equal ability and sitting next to each other and having different pattern of responses. The students detected as having DPF can be declared as the copier since they have a relatively inconsistent pattern of responses and the students detected as not having DPF serves as the source since they have a more consistent pattern of responses. If the student pairs share the same total score, yet they show no status difference in terms of DPF indication, then whether or not a cheating behavior occurs cannot be determined. This is as shown in Table 5 and the PRF of student pairs as the source and the copier could be shown in Figures 6, 7, 8 and 9.

Table 3. List of Students Detected as Having DPF

No	School Name	Student Identity	Number
1	SMPN 1 Dukuhturi	2 12 19 34 38 41 42 46 51 52 54 59 72 100 107 130 150 170 171 179 188 198 214 225 265 279 284 292	28 (9.6%)
2	SMPN 1 Suradadi	413 438	2 (1.1%)
3	SMPN 2 Slawi	513 519 521 531 544 545 546 548 576 631 646 647 648 651 653 654 656 658 661 662 663 665 668 670 671 672 673 674 675 676 677 678 679 683 684 687 688 689 690 692 693 697 699 700 701 702 704 706 708 709 710 712 713 716 724 725 728 730 731 733 735 738 739 740 742 745 779 796 798 799 800 801 803 808 809 811 812 815	78 (23.2 %)
4	SMPN 2 Dukuhwaru	825 827 829 831 832 837 854 855 856 857 859 862 863 864 866 867 896 906 907 908 911 934 935 940 941 981 982 994 999 1000 1002 1007 1009 1011	34 (17.4%)
	Total		142

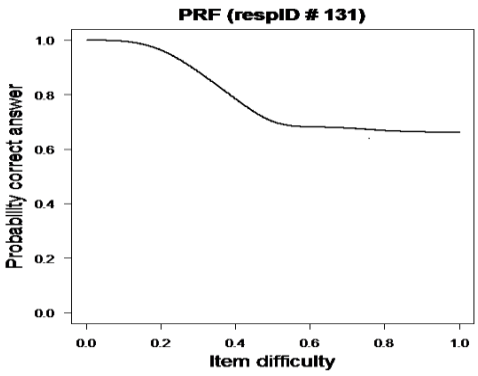

Figure 6. PRF for student no. 131 as the Source

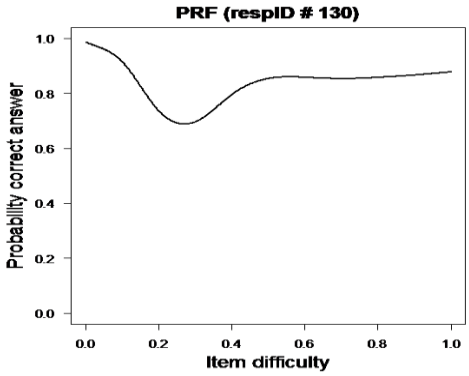

Figure 7. PRF for student no. 130 as the Copier

Table 4. List of Students with Equal Ability and Sitting Next to Each Other

No	Ability	Group of Students with Equal Ability and Sitting Next to Each Other	Number of Student Groups	Number of Students
1	-0.9988	300-301, 917-918	2	4
2	-0.8497	307-308, 365-366,383-384, 389-390	4	8
3	-0.5704	318-319,403-405, 615-616, 762-763	4	9
4	-0.4376	768-769, 904-905	2	4
5	-0.3079	968-969,988-989	2	4
6	-0.0540	323-324, 418-419,449-450, 708-709, 862-863, 928-930	6	13
7	0.0716	500-501, 550-552, 877-878	3	7
8	0.1976	327-328,341-342, 415-416, 584-585, 627-628, 644-645, 787-788, 822-823	8	16
9	0.3247	106-107, 385-386, 553-554, 703-704, 802-801,858-859, 1001-1002	7	14
10	0.4536	588-589, 602-603, 749-750, 791-792,794-795, 947-948	6	12
11	0.58522	4-5, 257-258, 568-571, 699-700, 798-799, 831-832. 854-855,906-907, 958-959	9	20
12	0.7207	334-335, 488-489, 532-533,686-687, 873-874, 912-913	6	12
13	0.8612	29-30, 510-511,991-992,994-995	4	8
14	1.0081	562-563	1	2
15	1.1636	78-79, 200-201, 224-225	3	6
16	1.3300	482-484,	1	2
17	1.5107	131-130	1	2
18	1.7108	60-61, 66-67,88-89	3	6
19	3.7157	198-199	1	2
	Total		73	151

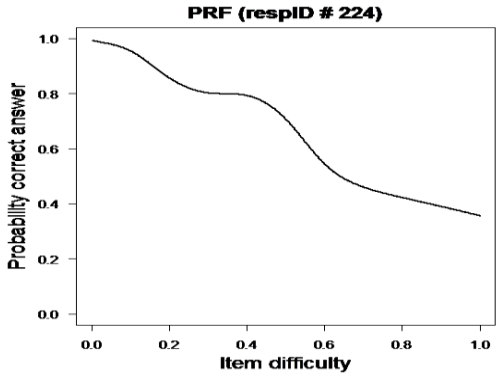

Figure 7. PRF for student no 224 as the Source

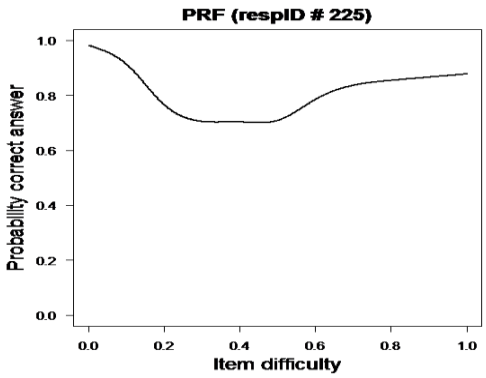

Figure 8. PRF for student no 225 as the Copier

Table 5: List of Students Allegedly Indicated of Cheating

No	Classical score	Ability	Students as the Source	pF score	Students as the Copier	PF Score
1	20	0.3247	106	0.0985	107	-0.0136
2	28	1.5107	131	0.1457	130	-0.0420
3	34	3.7157	199	0.1022	198	-0.0700
4	26	1.1636	224	0.1405	225	0.0058
5	23	0.72075	686	0.0446	687	0.0016
6	20	0.3247	802	0.0168	801	0.0099
7	20	0.3247	858	0.0168	859	0.0051
8	24	0.8612	995	0.1091	994	0.0078
9	20	0.3247	1001	0.0168	1002	0.0051

The number of students detected as having DPF between schools as shown in Table 4 turns out fairly varied. In SMPN 1 Suradadi, for example, it is found that only 1.1% of their students are having DPF, far lesser than what happens in SMPN 2 Dukuhwaru, SMPN 1 Dukuhturi and SMPN 2 Slawi. This rises many questions for further studies, particularly regarding the causes of DPF in students. DPF basically indicates to what extent an abberant response occurs and as explained by Karabatsos (2003) and Meijer (1996), abberant response can be caused by cheating, careless responding, creative responding, lucky guessing and responding. Heckler et al (2010) conducts an investigation which connect abberant response to the process of misconception in science learning. Likewise, Yih & Lin (2010) perform a study which associates abberant response to the formation process of concept structure owned by students in mathematics learning. Meanwhile, Tsai-Wei & Pei-Chen (2013) uses abberant response as the basis of misconception through a diagnostic test in mathematics learning.

These studies at least provide a tentative answer of the cause of abberant response which is quite massive in SMPN 2 Slawi, SMPN 2 Dukuhwaru and SMPN 1 Dukuhturi to over 10%. The abberant response occuring in the three junior high schools might be due to the misconception in students in addition to the classical reasons which have been studied so far (cheating, careless responding, creative responding, lucky guessing and random responding). Students who are not having misconceptions will surely fulfill the assumption of Rasch model where they can successfully answer the items with a difficulty level below their ability. However, in reality, some students fail to answer the items with a difficulty level below their ability. This could be interpreted that some students are having misconception and to prove it, a deeper investigation needs to be done using qualitative approach.

The abbberat response which quantitatively expressed as DPF can later be used as a detection of existence of misconception and can also be used in terms of its sensitivity to the misconception identification models which have been studied so far (Redhana et al, 2017; Rahmawati et al, 2017; Wijaya, & Muhardjito, 2016; Gurel et al, 2015; King, 2010). In relation to this research finding, the teaching and learning of natural sciences in SMPN 1 Suradadi at least needs to be studied more intensively, particularly from such aspects as teaching and learning model and teacher's ability and, eventually, for comparison with the three other junior high schools in this research.

The inconsistency of student's performance and the model can also be interpreted as alleged cheating behavior, particularly if the students sit next to those students with equal ability and detected as not having DPF. In both empirical and simulative studies, Karabatsos (2003) proves that Ht index is the best index in detecting cheating. By taking into account the students' status in relation to being detected as having or not having DPF, then an alleged cheating can be determined as in Table 5 where nine pairs of students are suspected of cheating. This suspicion is only made to those student pairs with equal ability and sitting next to each other and having different DPF status. The students detected as having DPF is declared as the copier and students detected as not having DPF is declared as the source. Of these nine student pairs, four of them are in SMPN I Dukuhturi, two in SMPN 2 Slawi and three in SMPN 2 Dukuhwaru. Suradadi

as with the number of students having DPF, in SMPN I Suradadi no students are detected of cheating.

Despite the foregoing, a weakness is still found in the method of determining cheating as used in this research. The weakness is that the researcher does not check the answer distribution between student pairs suspected of cheating. It is this matter that renders the researcher unable to justify the cheating, rather it is merely a suspicion based on the students' consistency in responding to the test. Some cheating detection method is implemented by considering the pattern of wrong answers in student pairs (Sotaridona & Meijer, 2002 ; Sotaridona, 2003; Sotaridona et al, 2006; Widiatmo, 2009). Salim (2016) conducts a studeny on cheating during the National Examination by considering the patter of wrong answers and to determine the role of each student pair, he uses Nominal Response Model (NRM). However, Salim (2016) research has a weakness in that it does not involve the student seats due to the limited access to the national examination data.

At least, this research pioneers a new method for detecting cheating behavior by applying Rasch modelling based on students' responses. Students' responses is extremely important and reliable for various studies on student score validity. In *Testing integrity symposium issues and recommendation for best practice 2013*[6], 3 methods which can be used to see alleged cheating in test administration are presented; they are (1) *ratio analysis / erasure analysis*, i.e. seeing the pattern of changes in answers from the wrong to the right answers; (2) *item-response patern analysis*, i.e. seeing the pattern of students' answers in a group with the same answers, and (3) *test–score analysis*, i.e. seeing the score achievement trend from one year to another, if the "*gains*" is great, the trend can be used as the basis for that cheating suspicion (US Department of Education, 2013).

CONCLUSION

14% of students participating in the final term assessment of natural sciences course for for junior high school 8th grade in the odd semester of academic year 2016/2017 in Tegal Regency, Indonesia were detected of having DPF. From the four junior high schools involved in this research, three have a range of proportion of students having DPF from 9.6% to 23%. Meanwhile, in the other school only 1.1% of their students were having DPF.

Nine student pairs during the administration of final term assessment of natural sciences course for junior high

School 8th grade in the odd semester of academic year 2016/2017 in Tegal Regency, Indonesia were suspected of cheating.

REFERENCES

Alsmadi, YM & Alsmadi, AA 2009,'Detecting differential person functioning in emotional intelligence',*Journal of Instructional Psychology*, vol. 36,no.4, pp. 284-88.

Ayele, DG, Zewotir, T, and Mwambi, H 2014, 'Using Rasch Modeling to Re-Evaluate Rapid Malaria Diagnosis Test Analyses', *International Journal of Environmental Research and Public Health ,* vol.11,no.7 ,pp. 6681-91.

Belov, DI & Armstrong, RD 2010, 'Automatic detection of answer copying via Kullback-Leibler divergence and K-index'*,Applied Psychological Measurement*, vol.*34*, no.6, pp. 379-92.

Clauser, BE & Mazor, KM 1998, 'Using statistical procedures to identify differentially functioning test items'*Educational Measurement: issues and practice*, vo. *17*, no.1, pp. 31-44.

Creswell, JW 2010, Research Design Pendekatan Kualitatif, Kuantitatif dan Mixed, A. Fawaid (Trans), Pustaka Pelajar, *Yogjakarta*

Dewi, NDL & Prasetyo, ZK 2016,'Pengembangan instrument penilaian IPA untuk memetakan critical thinking dan practical skill pesertadidik SMP',*Jurnal Inovasi Pendidikan IPA*, vol.*2* no.2, pp. 213-22.

Emons, WH, Sijtsma, K & Meijer, RR 2005, 'Global, local, and graphical person-fit analysis using person-response functions', *Psychological Methods*,vol.*10*, no.1, pp. 101.

Engelhard Jr, G 2009, 'Using item response theory and model—data fit to conceptualize differential item and person functioning for students with disabilities', *Educational and Psychological Measurement*, vo. 69, no.4,pp. 585-602.

Feuerherd, M, Knuth, D, Muehlan, H& Schmidt, S 2014, 'Differential item functioning (DIF) analyses of the Impact of Event Scale-Revised (IES-R): Results from a large European study on people with disaster experiences',*Traumatology*, vol. 20, no.4,pp. 313.

Gierl, M, Khaliq, SN &Boughton, K1999,'Gender differential item functioning in mathematics and science: Prevalence and policy implications', In*annual meeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec*.

Gurel, DK, Eryılmaz, A & McDermott, LC 2015,'A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science',*Eurasia Journal of Mathematics, Science & Technology Education*, vo. *11*, *no.*5, pp.989-1008.

Hambleton, RK, Swaminathan, H, & Rogers, HJ1991, *Fundamentals of item response theory*, Sage, California

Hays, RD, Calderón, JL, Spritzer, KL, Reise, SP, & Paz, SH2018, 'Differential item functioning by language on the PROMIS® physical functioning items for children and adolescents',*Quality of Life Research*, vol. 27, no.1, pp.235-47.

Heckler, AF, Scaife, TM, & Sayre, EC 2010,'Response times and misconception-like responses to science questions', In *Proceedings of the Annual Meeting of the Cognitive Science Society, vol.32,no.32*

Johanson, G & Alsmadi, A 2002,'Differential person functioning'*Educational and Psychological Measurement*, vol.*62*, no.3,pp. 435-43.

Kalaycioğlu, DB & Berberoğlu, G 2011,'Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey',*Journal of Psychoeducational Assessment*, vol.29, no.5, pp. 467-78.

Karabatsos, G 2003,'Comparing the aberrant response detection performance of thirty-six person-fit statistics', *Applied Measurement in Education*, vol.16,no.4, pp. 277-98.

Ministry of Education and Culture 2018, *Peraturan Menteri Pendidikan dan Kebudayaan No 20 tahun 2018 tentang Penguatan Pendidikan Karakter Pada Sekolah Formal*, Jakarta, Ministry of Education and Culture of Republic of Indonesia.

King, CJH2010,'An analysis of misconceptions in science textbooks: Earth science in England and Wales',*International Journal of Science Education*, vol.*32,no.*5, pp. 565-601.

Lu, YM, Wu, YY, Hsieh, CL, Lin, CL, Hwang, SL, Cheng, KI, & Lue, YJ 2013, 'Measurement precision of the disability for back pain scale-by applying Rasch analysis'**,** *Journal of Health and Quality of Life Outcome***s** , vol.11,no.1, pp. 119.

Luo, S, Liu, Y, Teresi, JA, Stebbins, GT, & Goetz, CG 2017,'Differential item functioning in the Unified Dyskinesia Rating Scale (udysrs)'. *Movement Disorders*, vol.32 no. 8, pp. 1244-49.

Mair, P, Hatzinger, R, Maier, MJ, Rusch, T, &Mair, MP2016, Package 'eRm', *R Foundation,Vienna, Austria*

Meijer, R R 1996, 'Person-fit research: An introduction',*Applied Measurement in Education*, vol.9, no.1, pp. 3-8.

Meijer, RR, & Sijtsma, K 2001, 'Methodology review: Evaluating person fit',*Applied Psychological Measurement*, vol.*25*, no.2, pp. 107-35.

Messick 1996, 'Validaty and washback in language testing', Languange Testing, vol.13 no.3, pp. 241-56.

Mok, M. and Wright, B 2004. 'Overview of Rasch*Model* Families', In Introduction to Rasch Measurement: Theory, Models and Applications, Jam Press, Minnesota

Perkins, A 2013, *Differential Person Functioning*, Ph.D thesis, Emory University, Atlanta

Petridou, A & Williams, J 2007, 'Accounting for aberrant test response patterns using multilevel models'*Journal of Educational Measurement*, vol.*44*, *no.*3, pp. 227-47.

Rahmawati, I, Sutopo, S, & Zulaikah, S 2017, 'Analysis of Students' Difficulties about Rotational Dynamic Topic Based on Resource Theory'*Jurnal Pendidikan IPA Indonesia*, vol.6,no.1, pp. 95-102.

Redhana, IW, Sudria, IBN, Hidayat, I & Merta, LM 2017, ' Identification of Chemistry Learning Problems Viewed from Conceptual Change Model',*Jurnal Pendidikan IPA Indonesia*,vol.6, no. 2, pp. 356-364.

Rupp, AA 2013, 'A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies'*Psychological Test and Assessment Modeling*, vol. *55*, no.1, pp. 3-38.

Salim, AN 2016, Perbandingan metode pendeteksian integritas hasil tes ujian nasional 2015. MPD thesis, Universitas Muhammadiyah Prof Dr. Hamka, Jakarta

Scherbaum, CA 2003, *Detecting intentional response distortion on measures of the five-factor model of personality: An application of differential person functioning*, Ph.D thesis, Ohio University, Ohio

Sijtsma, K, & Meijer, RR 1992, 'A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model',*Applied Psychological Measurement*, vol.*16*, no. 2, pp. 149-157.

Sireci, SG 2007, 'On validity theory and test validation',*Educational Researcher*, vol. *36*, no.8, pp. 477-481.

Smith, AB, Fallowfield, LJ, Stark, DP, Velikova, G, & Jenkins, V 2010, 'A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) – 12', Journal Health and Quality of Life Outcomes ,vol.8, no.45, pp. 45.

Sotaridona, LS & Meijer, RR 2002. 'Statistical properties of the K-index for detecting answer copying'. *Journal of Educational Measurement*, vol.*39*, no.2, pp. 115-32.

Sotaridona, LS, van der Linden, WJ, & Meijer, RR 2006,'Detecting answer copying using the kappa statistic',*Applied Psychological Measurement*, *vol.30*, no.5, pp. 412-31.

Sotaridona, LS 2003, 'Statistical Methods for the Detection of Answer Copying on Achievement Test', Ph.D thesis, Twente University Netherlands, Holland

Strobl, C, Kopf, J &Zeileis, A 2015, 'Rasch trees: A new method for detecting differential item functioning in the Rasch model'',*Psychometrika*, vo.*80*, *no.*2, pp. 289-316.

Sumintono, B 2018, 'Rasch Model Measurements as Tools in Assesment for Learning', In *1st International Conference on Education Innovation 2017*, Atlantis Press.

Sumintono, B & Widhiarso, W 2014, *Aplikasi model Rasch untuk penelitian ilmu-ilmusosial*. Trim Komunikata Publishing House, Bandung.

Susongko, P 2016, 'Validation of science achievement test with the Rasch model', *Jurnal Pendidikan IPA Indonesia*, vol.5, no.2, pp. 268-77.

Susongko, P & Mardapi, D. 2000, 'Keberfungsian Butir Diferensial Perangkat Tes Ebtanas Kimia Sekolah Menengah Umum di Jawa Tengah', *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 3, no.4, pp. 1-14.

Tendeiro, JN & Meijer, RR 2014, 'Detection of invalid test scores: The usefulness of simple nonparametric statistics', *Journal of Educational Measurement*, vol.*51*, no.3, pp. 239-59.

Tendeiro, JN, & Tendeiro, MJN 2016, Package 'PerFit', viewed 23 June 2019, http://www.est.colpos.mx/R-mirror/web/packages/PerFit/PerFit.pdf

Tendeiro, JN, Meijer, RR,&Niessen, A S M2016, 'PerFit: An R package for person-fit analysis in IRT', *Journal of Statistical Software*, vol.*74, no.*5, pp. 1-27.

US Department of Education 2013, Testing integrity symposium Issues and Recommendations for Best Practice, US Department of Education, *Institute of Education Sciences National Center for Education Statistics 2013*, Washington DC, United States.

Tsai-Wei, H, & Pei-Chen, W 2013, Classroom-based cognitive diagnostic model for a teacher-made fraction-decimal test, *Journal of Educational Technology & Society*, vol.*16*, no.3, pp. 347-61.

Widiatmo, H 2009,'Metode untuk mendeteksi penyontekan jawaban pada tes pilihan ganda: studi kasus SMP di Kabupaten Garut'. *Pusat Penelitian Pendidikan, Balitbang Diknas*, pp. 219-26.

Wijaya, CP, & Muhardjito, M 2016,'The diagnosis of senior high school class x mia b students misconceptions about hydrostatic pressure concept using three-tier'. *Jurnal Pendidikan IPA Indonesia*, vol.*5*, no.1, pp. 13-21.

Wu, M, & Adams, R 2007, *Applying the Rasch model to psycho-social measurement: A practical approach*, Educational Measurement Solutions, Melbourne

Yih, JM& Lin, YH 2010, 'Concept structure based on response pattern detection of SP chart with application in algebra learning', *Learning*, vol.100, no.8, pp.847-56.