Student Questionnaire Of Pisa 2006 And Rasch Analysis

Fitriati

ABSTRACT

PISA is one of the biggest international student assessments. The results of this study are frequently used to inform policy directions and to provide feedback to learning and teaching. However, due to country socioeconomic, cultural and language differences among the countries, the test instruments may not be funtioning in the same way in all culture. Those differences may cause non-equivalence of tests or or tests may not be fair among different cultures, which then challenges assumptions made about measurement equivalence. This study aims to examine the equivalence of motivation and self concept items on student questionnaires of PISA 2006 study whether fit Indonesia context or not. The study employed two analisis of Rasch model to seek the equivalence of the tests. The results revealed that there is no significant DIF on motivation scale. One important finding of DIF analysis using Indonesia data was that two items of self-concept scale have been shown favoring males. The study also found that all items on motivation and self concept scale fit the data set and are not dependent on upon the sample of Indonesian students. The test is fair among different groups and contexts. This study suggests that more scales analysis is required as it will provide more comprehensive findings about the equivalence of this survey test.

Key words:Test equivalence, Rasch model, DIF

Fitriati, Dose Prodi Pendidikan Matematika – STKIP Bina Bangsa Meulaboh, Aceh Barat, Email: firti_kindy@yahoo.com

Introduction

PISA 2006 is one of the biggest international student assessments. Achievement test was administered to over half million students over the world along with questionnaire for those students. Within this survey, students' affective value on science learning has been measured. This includes science enjoyment, science interest, science motivation, science value, science self-efficacy, science self-concept. The validity and reliability of the constructs are determined by international experts. In PISA 2006, categorical items from the context questionnaires were scaled using IRT modelling. Weighted likelihood estimates (logits) for the latent dimensions were transformed to scales with an OECD average of 0 and a standard deviation of 1 (with equally weighted samples). In order to develop valid and comparable latent measures Exploratory and Confirmatory Factor Analysis (EFA, CFA) were employed by PISA data analysis teams to validate indices of Likert type items. Validation of latent constructs was checked based on different fit indices such as root-mean square error of approximation (RMSEA), and comparative fit index (CFI). In short, all variables in this study meet construct validity requirements (OECD, 2009).Although theserequirements have been met, it is possible that the construct validity of the questionnaire items still be questioned as the items responses might vary among participants from different countries.

It is argued that country socioeconomic and cultural differences may affect different response on attitudinal survey, as both factors have proven to have influence on the students' attitude, in particular in learning science (Schibeci& Riley, 1986). These cultural differences may cause non-equivalence of the tests (Byrne, 2003). Additionally, it is widely recognized that language differences may have a powerful effect on equivalence (or non-equivalence) of test and questionnaire items (Schulz &Fraillon, 2009).Although PISA 2006 implemented rigorous translation verifications to achieve a maximum of "linguisticequivalence" and a set of test items is simple or context-free (OECD, 2005), these translated survey tests may not be functioning in the same way in all cultures. This is also argued that tests may not be equivalent or tests may not be fair among different cultures (Yildirim, 2006; Schulz &Fraillon, 2009).This consequently challenges assumptions made about measurement invariance. Therefore, as the persons and items are multifaceted in any measurement situation,the purpose of this paper is to examine the equivalence of motivation and self concept items on student questionnaires whether fit Indonesia context or not. As these items are polytomous, Rasch model (Rasch, 1960)will be employed to seek the equivalence of the items. The justification of why using this model is because analysis of affective scales with Rasch model allows for calibration of items and scales independently of the student sample and the

sample of items employed (Wright & Stone, 1999) and provide the response patterns of the individuals completing the survey and the amount of the attitude in the individual based on empirical evidence (Andrich, 1988; Krueger & Finger, 2001; Santor& Ramsay, 1998; Yildirim, 2006).Within Rasch model, two model analyseswill be conducted to examine the variance of responses of the items: different item function (DIF) to seek the difference between gender;rating scale analysis to test the appropriateness of the items to fit Indonesian students.Findings of this analysis might provide some insights how reasonable are the PISA 2006 assumption that item parameters do not vary among nations.

Rasch Model

A Rasch modeling is widely used to measure invariance and determine equivalence across groups at the items (Schulz &Fraillon, 2009). It is also an effective and accurate way to analyze different dimensions of a survey separately. Rasch models propose that responses to a set of items can be explained by a person's ability along a continuum of the unidimensional construct underlying the items and by characteristics of the items, or item parameters. Several advantages of Rasch measurement have been described (Andrich, 1988; Wright, 1997 & Fischer, 1995). A key characteristic of the models is that Rasch measurement can be considered sample independent as well as instrument-independent. That is, if a Rasch model fits a set of data, item characteristics are not dependent upon a specific sample; item parameters estimated across different groups and contexts will be equivalent (Andrich, 1988). Consequently, the Rasch model can be used to assess the extent to which a set of test items is sample- or context-free (Raczek et al., 1998). Item characteristics should remain relatively fixed so that an invariant construct of physical functioning can be used to compare abilities and discriminate between levels of physical functioning across different samples. Rasch procedures also enable the test developer to examine the equivalence of item calibrations across different samples and contexts, including various cultural-linguistic settings and translations. In this case,Racsh analysis enables a more detailed (item level) examination of the structure and operation of the scales on the survey. While the persons and items are multifaceted in any measurement situation, affective measures on PISA 2006 study survey need to be thought of and behave as if the different facets acts in unison (Green, 1996). Therefore, revalidating the item constructs of affective measures using Rasch model is required as it might provide evidence whether the survey test of PISA 2006 is equivalence with Indonesian context.

Instruments/data

This study examined Indonesian data from an International study PISA 2006. Publicly available datasets student surveys files are taken from OECD's website (www.pisa.oecd.org).

The datasets include information collected from 57 countries. The student file containing survey data will be extracted and used. PISA 2006 surveyis designed to collect information about students. A 30-minute student survey asking for information on student characteristics, perceptions of science and school, and family background was given after the literacy assessment. Over 10.647 of15-year-old Indonesian students, about (50.3%) female and (50.3%) male students answered this study survey. The items on the survey include:

1. Student and family background
2. Views on various issues related to science
3. The environment
4. Career and (broad science)
5. Learning time
6. Teaching and learning science.

As examining all these related items is not feasible, this proposed study will examine some views (attitudes) items variables only, in particular motivation and self-concept variable. The selected variables are described below.

Science motivation items (STQ3501- STQ3505)

How much do you agree with statements below?

a. Making an effort in my <school science> subject (s) is worth it because this will help me in the work I want to do later on
b. What I learn in my <school science> subject (s) is important for me because I need this for what I want to study later on
c. I study my school science because I know it is useful for me
d. Studying my <school science> subject (s) is worthwhile for me because what I learn will improve my career prospect
e. I will learn many things from my <school science> subject (s) that will help me get the job

Science self-concept items (STQ3701-STQ3706)

How much do you agree with statements below?

a. Learning advanced <school science> topics would be easy for me
b. I can usually give good answers to <test question> on <school science> topic
c. I learn <school science> topic quickly
d. <school science> are easy for me
e. When I am being taught <school science>, I can understand the topic very well
f. I can easily understand new idea in <school science>

Item categories for both variables were "strongly agree (1)", "agree (2)", "disagree(3)" and "strongly disagree (4)".

The motivation and self-concept related items are especially appropriate for Raschanalyses, as the items represent the attitude measure with polytomousscales. The data of the response on these scales were transported from SPSS to Text format. To conduct the analysis on the scales, the syntaxes

were created. These are available on the appendixes.

Analysis

As a description of data procedures, information of recoding and rationale of selecting sub group of item are given, investigating items equivalence using Raschmodel was connducted. A series of Rasch analyses were performed to address two major questions: (1) Do the responses on motivation and self-concept items operate differently between students' gender?, and (2) Do the motivation and self-concept items form a unidemensionality, thus appropriate for fitting Indonesian students' context ? The chosen analyses were:

1. Different Item Functioning (DIF)

To test for differences in responses on the proposed items between groups, this study used (DIF) analysis.DIF analyses evaluate whether examinees from different groups (e.g., females and males) who are of comparable ability on the entire survey test have equal probabilities of success on an item. In DIF analyses, conditioning procedures are used to systematically match students of similar ability across groups to distinguish between overall group differences on an item (item impact) and potential item bias. If the probability of a particular response differs significantly across test takers who are equivalent on ability, items are considered to be functioning differentially across groups (Hauger and Sireci, 2008). This is then called Item bias which refers to the situation in which a statistically significant difference in response on an item is observed across two groups (female and male).

2. Rating Scale Model

Rating scale model is used to test the unidimensional of the proposed items for fitting Indonesian students' context. The Rasch rating scale model is based on the assumption that all motivation and self concept related-items have the same underlying structure for the common three-point rating scale. For example, the improvement in moving from "Strongly disagree" to "Strongly agree" is assumed to be approximately the same for all those items. This model provides estimates of item locations (calibrations) that define the hierarchical order of the items along a common measurement continuum (Darmawan, 2005). Item and person calibrations are generally expressed in log-odd units (logits) that are positioned along the hierarchical scale. For items, a logit represents the log-odds of the level of response of an item relative to the response to a total set of items. Logits of greater magnitude represent increasing item difficulty or person ability (Ludlow & Haley, 1995). Lower values on this scale represent easier items or, for persons, lower response functioning ability. High values on the scale represent more difficult items and higher response functioning ability.

To perform these two analyses, the data of science motivation and self-concept variable extracted from Indonesia student data set were

subjected to Rasch analysis using Conquest 2.0 software (Wu, Adam, Wilson&Handale, 2007). Inspecting the infit mean-squares provides evidence about the fit of the data to the model. The Infit mean-squares are used to determine the fit of the item within the construct. In this study, critical values chosen for Infit Mean Square (IMS) fit statistic were 0.72-1.30 (Linacre, Wright, Gustafsson& Martin-Lof, 1994). Item whose infit mean square values fall above 1.30 are generally considered misfitting and do not discriminate well, while below 0.72are overfitting and provide redundant information (Tilahun, 2004).

DIF analysis was used to check the presence of items bias and the significance of differences between different groups of students. Gender (female and male) is person factor used in this study.

1.1 Gender Differences in Motivation Scale

The five items of motivation in learning science was subjected to analysis using DIF model. This was carried out to test whether the items operate differently between female and male. Female and male bias estimate for motivational scale were examined.The results of analysis are shown in the Table 1.1.1.

Results and Discussion

1. Gender Differences

Table 1.1.1 General Gender Differences in Motivation Scale

Gender	Estimate	Error	IMS	CI	t
1 (female)	-0.188	0.03	0.98	(0.96, 1.04)	-1.0
2 (Male)	0.188	0.03	0.94	(0.96, 1.04)	-3.2

Chi-square test of parameter equality = 39.81, df = 1, Sig Level = 0.000

As can be seen from Table 1.1.1, the female group had low estimate value. This indicates that girlshad a high agreement on the items.Lower values on this scale represent easier items or, for persons, lower response functioning ability (in this case exhibits more likely positive attitude) (Ludlow & Haley, 1995). It also shows that the female students score 0.376 lower that male student. The fact that the parameter estimate is more than twice its standard error indicates that this difference is statistically significant(Wu, Adam, Wilson &Handale, 2007). The significant variance within the items are shown in the Table 1.1.2

Table 1.1.2 Gender Differences on Motivation Scale

	Female (N=5326)			Male (N=5291)			
Item	**Est**	**SE**	**IMS**	**Est**	**SE**	**IMS**	**Differences**
ST35Q01	-0.001	0.022	0.96	0.001	0.021	1.03	-0.002
ST35Q02	-0.101	0.021	0.82	0.101	0.021	0.91	-0.202
ST35Q03	0.044	0.021	0.91	-0.044	0.021	0.84	0.088
ST35Q04	0.028	0.021	0.95	-0.028	0.021	0.98	0.056
ST35Q05	0.031	0.021	1.05	-0.031	0.021	1.13	0.062

Chi-square test of parameter equality = 28.85, df = 4, Sig Level = 0.000

It was evident in the Table 1.1.2 that girls were more likely than boys to respond positively to two items (ST35Q01 and ST35Q02), boys were more likely than girls to respond positively to three items (ST35Q03, ST35Q04, and ST35Q05). The significant Chi-square (28.85, df = 4) also show the existence of DIF.However, both Table 1.1.1 and Table1.1.2 do not provide enough information in which step they are different. Therefore, the item characteristics curves were examined. Based on item characteristic curves, generally both female and males students had a high agreement on most items except item ST35Q02 (What I learn in my <school science> subject (s) is important for me because I need this for what I want to study later on). Figure 1.1.1 shows that the differences seem to be larger in highly motivated students, and the difference is not obvious for less motivated students. It also shows the degree of agreement of both groupsis slightly different.

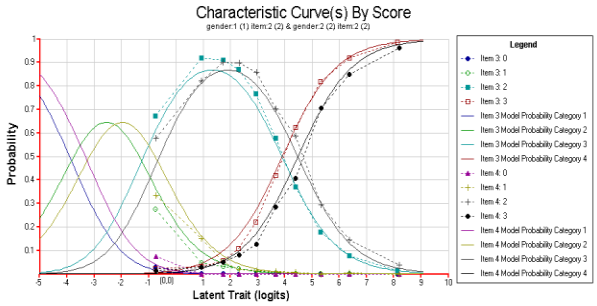

Figure 1.1.1 Plot Characteristic Curves for Item ST35Q02.

As can be seen in the Figure above, given particular ability level, the probability of havinghigh agreement response on this item is slightly higher for girls. This means that although both groups had high motivation, most female students were more optimism than males in perceiving the important of science for their future study. Therefore, there no item on this scale was biased in fovour of females and males.

1.2. Gender Differences in Self-concept Scale

DIF analysis was also conducted to examine self-concept scale. The results of analysis for six items of self-concept scale, using all examinee in each group, are summarized inTable 1.2.1. Generally, boys were more likely

than girls to respond positively (towards agree) to self-concept items, which indicated by the low estimate value of males (-0.124). The difference is statistically significant since the boy students score 0.248 lower that girl students and its parameter estimate is more than twice its standard error. It seems that the self-concept items were biased significantly in favour of girls. However, care was taken in considering this finding. The magnitude of DIF for each item is also important to determine the existence of DIF between the groups(Wu, Adam, Wilson &Handale, 2007). This can be seen in the Table 1.2.2.

Table 2.1 General Gender Differences on Self-concept Scale

Gender	Estimate	Error	IMS	CI	T
1 (female)	0.124	0.024	1.05	(0.96, 1.04)	2.3
2 (Male)	-0.124	0.024	1.03	(0.96, 1.04)	1.6

Chi-square test of parameter equality = 27.15, df = 1, Sig Level = 0.000

Table 1.2.2 Gender Differences in Self-concept Scale

	Female (N=5326)			Male (N=5291)			
Item	Est	SE	IMS	Est	SE	IMS	Differences
ST37Q01	0.029	0.019	1.12	-0.029	0.019	1.13	0.058
ST37Q02	-0.028	0.019	0.93	0.028	0.019	0.94	-0.056
ST37Q03	0.037	0.019	0.88	-0.037	0.019	1.03	0.074
ST37Q04	0.025	0.019	0.90	-0.025	0.019	1.01	0.050
ST37Q05	-0.097	0.019	0.96	0.097	0.019	1.00	-0.194
ST37Q06	0.033	0.019	1.00	-0.033	0.019	1.08	0.066

Chi-square test of parameter equality = 35.33, df = 5, Sig Level = 0.000

It was evident in Table1.2.2 that the girls had low agreement on four items of self-concept (ST37Q01, ST37Q03, ST37Q04 and ST37Q06). This means that estimates of pessimism in girls were slightly higher than the boys. However, they were more likely than boys to respond positively to two items (ST35Q01 and ST35Q02). As both Tables 1.2.1 and Table 1.2.2 do not provide enough information in which step they are different, it is necessary to examine theitem characteristics curves.It shows that the difference on item ST37Q03 and ST37Q04 seems to be larger in students who disagree. This can be seen in the Figure 1.2.1 and Figure 1.2.2.

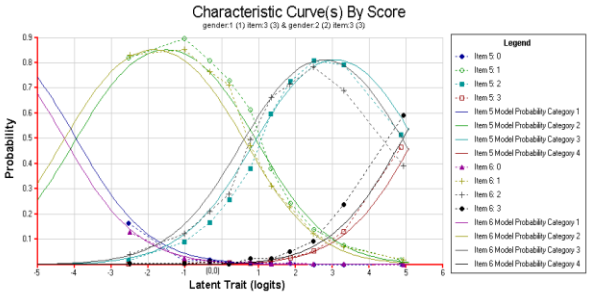

Figure 1.2.1 Plot Characteristic Curves for Item ST37Q03

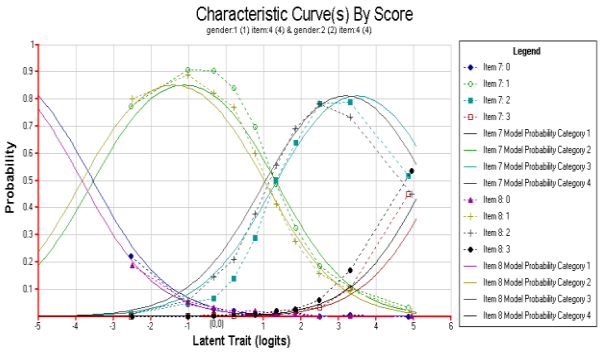

Figure 1.2.2 Plot Characteristic Curves for Item ST37Q04

The both Figure 1.2.1 and Figure 1.2.2,show that given particular ability level, the probability of being negatively response on this item is slightly higher for girls. This may indicate that the girls were more pessimism on learning science topic quickly (ST37Q03) and science is easy for them (ST37Q04).

As the Table 1.2.2 shows the biggest difference estimate between the groups was on their response on item ST37Q05, which is biased significantly in favour of boys, the item characteristic curve exhibits that difference seems to be larger in the agreement level. This illustrates in the Figure 1.2.3.

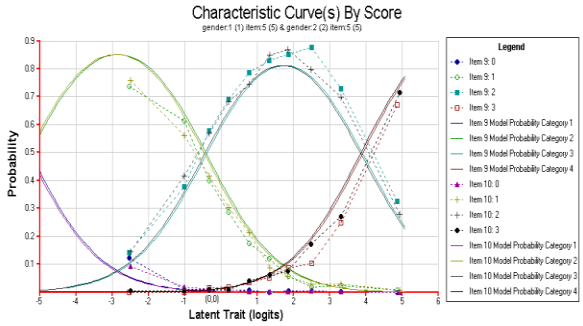

Figure 1.2.3 Plot Characteristic Curves for Item ST37Q05

The Figure 2.3 shows that given particular ability level, the probability of being positively response towards "Agree" on this item is slightly higher for girls. This means that the boys had low agreement that they can understand science well.Based on evidences provided in the Tables and Figures, it can be concluded that two items of self-concept scale (ST37Q03 and ST37Q04) were significantly biased in favour of females.

2. Item analysis with the Rating Scale Model

Rating scale analyses were completed using the total sample of participant (Indonesian students). All proposed items were examined using ConQuest software. This was carried out to test the unidemensionality of the items within Indonesian sample model. This involved examining each item's fit statistics. More specifically, the infit mean square (INFIT MNSQ) statistic was used as a basis for model fitting or non-fitting items. Items whose infit mean square values fall above 1.30 are generally considered misfitting and do not discriminate well, while below 0.72 are overfitting and provide redundant information(Ben, 2010). The

following section provides detailed analysis on each scale.

2.1Motivation items analysis

The 5 items of motivational scale was subjected to item analysis using the rating scale model. Items with goodness-of-fit values fall into the range of critical value (0.72-1.30) are presented in Table 2.1.1

Table 2.1. the response model parameter estimate of motivation scale for the Indonesian student

Variables	Estimates	Error	Unweighted Fit		
			INFT MNSQ	CI	*T*
ST35Q01	-0.660	0.017	0.99	(0.97, 1.03)	-0.9
ST35Q02	-0.107	0.016	0.91	(0.97, 1.03)	-6.9
ST35Q03	-0.147	0.016	0.86	(0.97, 1.03)	-10.4
ST35Q04	0.386	0.016	0.97	(0.97, 1.03)	-2.2
ST35Q05	0.528	0.033	1.05	(0.97, 1.03)	3.4

Separation Reliability = 0.999
Chi-square test of parameter equality = 2247.12, df = 4,
Sig Level = 0.000

As can be seen in Table 2.1.1, all five items related to motivation have a good fit to the measurement model, indicating of the items' infit mean square values fall within the acceptable range (0.72-1.30). All the items delta value are ordered from low to high indicating that the person have answer consistently and logically with the ordered response format used (Ben, 2010). The index of separation reliability for the 5-item scale is 0.999, which means that the proportion of observed variance considered to be true is 99 per cent. This indicated that the discrimination power of the scale is high, indicating that the items discriminate between the high motivated and low motivated students (Alagumalai& Curtis, 2005).Additionally, the good fit measurement model of the motivation scale is also indicated by the chi square probability value which appears to be significant.

There were three items (ST35Q01, ST35Q02 and ST35Q03) that most students probably would find it 'easy', in this case, easy to change their responses from negative to more positive one and two items (ST35Q04 and ST35Q05) that most students probably would find it 'hard' to say that they agree with the statements . The easy items were indicated by low estimate values and hard items were indicated by high estimate values.

In regard to the five motivations of learning science, namely (a) help in the work later, (b) the important for future study, (c) science usefulness, (d) improve career prospect,

and (e) help to get a job, it seemed that most students were high motivated since they perceived that studying science may help them in their work later, be important in their future study and be useful for them. They seemed slightly less motivated when they think that learning science will improve their career prospects and help them to get a job. Learning science to get a job seemed to be the least choice of motivation responded by the students. This is inline with evidence in Table 2.1.1 which shows this item as the hardest one for students to agree with.This can be seen from plot item characteristic curves illustrating in the Figure 2.1.1.

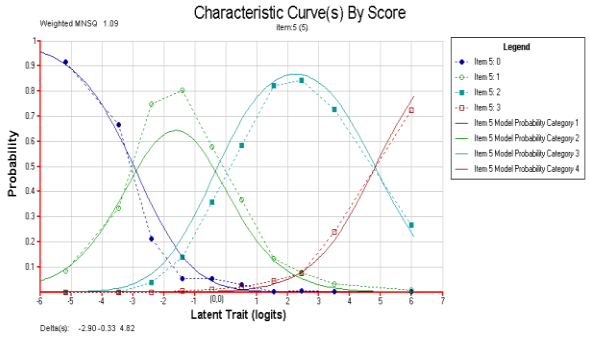

Figure 2.1.1. Response Probability for item ST35Q05 (help to get a job)

Within Indonesia data, the Rasch model generally confirmed the hypothesized structure of motivation and self concept items. It is evident that both scales model were fit well to Rasch Analysis.

Table 2.2.1 The response model parameter estimate of self-concept scale for the Indonesian student

Variables	Estimates	Error	Unweighted Fit		
			INFT MNSQ	CI	*T*
ST37Q01	-0.372	0.015	1.14	(0.97, 1.03)	9.4
ST37Q02	-0.386	0.015	0.94	(0.97, 1.03)	-4.5
ST37Q03	0.584	0.015	0.97	(0.97, 1.03)	-2.1
ST37Q04	1.005	0.015	0.95	(0.97, 1.03)	-3.7
ST37Q05	-0.610	0.015	0.98	(0.97, 1.03)	-1.3
ST37Q06	-0.220	0.033	1.02	(0.97, 1.03)	1.7

Separation Reliability = 1.000
Chi-square test of parameter equality = 9022.58, df = 5
Sig Level = 0.000

As can be seen in Table 2.2.1, all of the items' infit mean square fall within the acceptable range (0.72-1.30). Examination of the item deltas values shows that they are in order of increasing value, which indicates that the response choices on a scale are also in order. The separation reliability index is 1.0 indicating the proportion of observed variance considered to be true is 100 per cent. Additionally, the chi square probability value appears to be significant.

Based on estimate values, four items of self-concept scale (ST37Q01, ST37Q02, ST37Q05 and ST37Q06) seemed to be 'easy' for most of students, while the other two items (ST37Q03 and ST37Q04) were probably 'hard' for most students to say that they agree with the statements. It was expected that there were would be some variation in each person's response to these all items on self-concept scale. As the self-concept scalemeasuredsix factors, it seemed that most students had high confidence on their ability tolearn science, in particular, to learn advanced science, to give good answer, to understand the concept well and to understand new idea easily (see Figure 2.2.1). They seemed slightly less confident on their ability to learn science topic quickly and to understand the science topic easily. However, the least confidence level of most students was of their ability to learning science topic quickly. This isevidence in Table 2.2.1 that this item (ST37Q04) is the hardest statement for students to agree with. The probability of students' response on this item can be seen from plot item characteristic curves illustrating in the Figure 2.2.2.

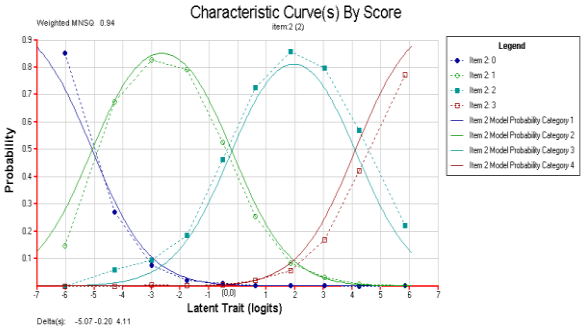

Figure 2.2.1 Response Probability for item ST37Q02 (give good answer)

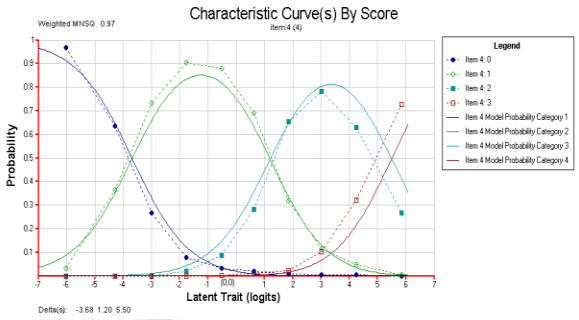

Figure 2.2.2. Response Probability for item ST37Q04 (learn science topic easily)

The response probability of the easy item (Figure 2.2.1) shows that there no big discrepancy in the curve illustrating the expected value scores for four levels of agreement formed very close to the curves, while the response probability of the hardest item (Figure 2.2.2) illustrates slightly big discrepancy on the level of 'disagreement' and 'agreement'. This indicates that most students were probably hard to move from 'disagree' to 'agree'on this item.

Conclusion

In this study, data from Indonesian PISA 2006 survey were used to investigate the equivalence of the survey study. Twoanalyses of Rasch model were employed. The Rasch model was found to be useful in validating a scale of student motivation and self-concept on PISA 2006 study. From DIF analyses, it was found that male and female students had a high agreement on most the items. On motivation scale, the difference only in the agreement level, which indicates no significant DIF. One important finding of DIF analysis using

Indonesia data was that two items of self-concept scale have been shown favoring males. Although these items showing substantial DIF, they were not necessarily deleted from future tests. However, these items were among those that needed to be carefully reviewed prior subsequent use. This finding suggests that examining gender DIF in individual test language groups is necessary in international test.

Regarding rating scale analysis, the 5-items scale of motivation and the 6-items scale of self-concept had desirable measurement properties.Within Indonesia data, the Rasch model generally confirmed the hypothesized structure of motivation and self concept items. It is evident that both scales model were fit well to Rasch Analysis. All of the items' mean square falls within the acceptable range indicating all itemshad good fit. The delta values are ordered from low to high indicating that the students have answered consistently and logically with the order response format used. Item difficulties and person measures are calibrated on the same scale. As theRasch model fits the data set, the item characteristics on PISA 2006 survey are not dependent upon the sample of Indonesian students. Both sample and item are independent. Therefore, it can be concluded that the survey test PISA 2006 is equivalence. The test is fair among different groups and contexts.However, more scales analysis is required as it will provide more comprehensive findings about the equivalence of this survey test.

References

Alagumalai, S. & Curtis, D. D. (2005).Classical Test Theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars* (pp. 1-14). The Netherlands: Springer.

Andrich, D. (1978). A rating formulation for ordered response categories.*Psychometrika, 43*(4), 561-573.

Ben, F (2010), *Students' uptake of Physics: A study of South Australian and Filipino physics students*. Unpublished Doctoral Dissertation.The University of Adelaide.

Byrne, B. M. (2003). Testing for equivalent self-concept measurement across culture. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *Internationaladvances in self-research: Speaking to the future* (pp. 291–314). Greenwich: Information Age Publishing.

Curtis, D. D. (2004). Comparing classical and contemporary analyses and Rasch measurement. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars – Papers in honour of John P. Keeves* (pp. 181-197). The Netherlands: Springer.

Darmawan, I. G. N. (2005). Creating a scale as a general measure of satisfaction for information and communication technology user. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars – Papers in honour of John P. Keeves* (pp. 271-286). The Netherlands: Springer.

Darmawan, I. G. N. (2003).*Implementation of Information Technology in Local Government in Bali, Indonesia.*Adelaide, South Australia: Shannon Research Press.

Green, K. E. (1996). Applications of the Rasch model to evaluation of survey data quality. *New Directions for Evaluation*, 70, 81-92.

Grisay, A., de Jong J., Gebhart, E., Berezner, A., &Halleux-Monseur, B. (2006, April).*Translation equivalence across PISA countries*. Paper presented in the meeting of the American Educational.

Hauger, J. B., &Sirecy, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. International Journal of Testing, 8, 237-250.

Linacre, J. M., Wright, B. D., Gustafsson, J-E & Martin-Lof, P. (1994).Reasonable mean-square fit values.*Rasch Measurement Transactions, 8*(3), 370.

Ludlow, L. H., & Haley, S. M. (1995).Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, 55: 967–975.

OECD.(2006a). PISA 2006. Science Competencies for Tomorrow's World. Volume 1 Analysis.Viewed on 5 Oct 2010.< 0,333,en_32252351_32236191_39718850_1_1_1_1,00.html.>

OECD.(2006b). Assessing Scientific, Reading, and Mathematical literacy.A framework for PISA 2006.Viewed on5 Oct 2010. < http:www.pisa.oecd.org/dataoecd/63/35/37464175.pdf>

OECD (2009).*PISA 2006 Technical Report.*Viewed on 15 October 2010. <http://www.pisa.oecd.org/dataoecd/0/47/42025182.pdf>

Raczec, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., & et al., (1998). Comparison of Rasch and summated rating scale s constructed from SF-36 phisical function items in seven countries: Result from the IQOLA project. *J ClinEpidemiol*, 51(11), 1203-1214.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Chicago: DanmarksPegagogiskeInstitut, University of Chicago Press.

Schibeci, R. A., & Riley, J. P. (1986). Influence of student background and perceiption on science attitude and achievement. *Journal of Research in Science Teaching*, 23 (3), 177-187.

Schulz, W., &Fraillon, J. (2009).The analysis of measurement equivalence in international studies using the Rasch model.Paper presented to the symposium on "Rasch measurement: present, past and future" at the European Conference on Educational Research (ECER) in Vienna, 28-30 September 2009.

Tilahun M. A. (2004). Monitoring mathematics achievement over time: a secondary analysis of FIMS, SIMS and TIMS: a Rasch analysis. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars – Papers in honour of John P. Keeves* (pp. 63-79). The Netherlands: Springer.

Wright, B. D. & Stone, M. H. (1999).*Measurement Essentials* (2nd ed.). Wilmington, Delaware: Wide Range, Inc.

Wright, B. D. (1997). Solving measurement problems with the Rasch model.*Journal Educational Measurement*, 14(2): 97–116.

Wright, B. D., & Linacre, J. M. (1994).Reasonable mean-square fit values.*Rasch Measurement Transactions, 8*(3), 370.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007).ConQuest Version 2.0 [Generalised Item Response Modeling Software]. Camberwell, Victoria: ACER Press.

Yate, S. M. (2005). Rasch and attitude scales: Explanatory style. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars – Papers in honour of John P. Keeves* (pp. 207-226). The Netherlands: Springer.

Yildrim, H.H. (2006). *The differential item functioning (DIF) analysis of mathematic items in the international assessment programs.*Phd. Thesis.Middle East Technical University.