ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 244
1st International Conference on Early Childhood and Primary Education (ECPE 2018)

Evaluating Person and Item Fit in Science Achievement Test of TIMSS 2015 for Australian Grade 4 Students Using Rasch Measurement

Yulia Linguistika
Elementary School Teacher Education
Universitas Negeri Malang
Malang, Indonesia
yulia.linguistika.fip@um.ac.id

***Abstract*—The purpose of the study was to evaluate the person and item fit of science achievement test for Australian grade 4 students (N = 421) using Rasch measurement. This study was conducted in two stages. First, the software evaluated the person fit then the misfitting persons are removed based on the criteria. Second, the software examined the item fit of the reduced data set. The result indicated that there were ten underfitting persons with the range of person fit index from 1.512 to 1.984. The item difficulty, infit, and outfit MNSQs range of the initial analysis and those of after misfitting person removal are slightly different. The Rasch item and person reliability slightly change. For the final analysis, the infit MNSQs and the outfit MNSQs are within the reasonable range for high stakes MCQs which implies that the items are good and indicate the sufficient fit to Rasch model for practical measurement purposes.**

Keywords—Rasch model, students' science achievement, person fit, item fit

I. INTRODUCTION

Science is one of the core subjects which are essential for students. By mastering science, they can get succeed in their life and work in the 21st century [1]. The importance of science seizes the attention from countries around the world either in developing or developed countries. Australia has focused on developing science in the education system which is reflected in the Melbourne Declaration on Educational Goals for Young Australians [2]. The implementation of the Melbourne Declaration can be found in the curriculum and assessment design that strongly focuses on literacy and numeracy in Science, Technology, Engineering, Arts, and Mathematics which can enable students to develop knowledge, skill, and value in the discipline of each subject. The intention of mastering science is also to provide a high quality of life which will influence the ability to compete in the global economy [2].

International Association for the Evaluation of Educational Achievement (IEA) has conducted regular international comparative assessments of student achievement in mathematics and science named the Trends in International Mathematics and Science Study (TIMSS) in more than 60 countries. [3] reported that in the year of 2015 with an average score of 524 score points on the TIMSS Year 4 science scale, Australian students significantly outperformed students in 17 other countries, such as Portugal, New Zealand and France. However, Australian Year 4 students were outperformed by students in 17 other countries, including the United States and England, as well as the participating East Asian countries such as Singapore, Korea, Japan, Hong Kong and Chinese Taipei.

Education policy makers in Australia have tried to improve the system based on the research which has been conducted by the experts to enhance the students' science achievement. The TIMSS 2015 results will also be a consideration for the policymakers to create a new policy or revise the existing policy. This situation will remain some questions about the validity and the reliability of the instruments used in TIMSS assessments. Unfortunately, not all countries evaluate the validity and the reliability of the instruments.

II. LITERATURE REVIEW

A. Australian Students Grade 4 in TIMSS 2015

A stratified random sample of 287 primary schools in Australia participated in the data collection for TIMSS 2015. The stratification of the sample ensured that the TIMSS sample was representative of the Australian Year 4 populations (according to jurisdiction, school sector, geographic location of the school and socioeconomic category for the area of the school). There was 6057 Year 4 students participated in the study [3]. The results revealed that in the science achievement test, Australia's average score of 524 score points was significantly higher than the scores for 17 other countries, such as New Zealand, Portugal and France, and places average achievement at the higher end of the Intermediate benchmark. However, Australia's average score was significantly lower than the average scores for 17 other countries, including the United States and England, as well as the participating East Asian countries Singapore, Korea, Japan, Hong Kong and Chinese Taipei.

Based on the comparison amongst jurisdictions [4], Australian Capital Territory was the highest-performing jurisdiction. The spread of average scores across the jurisdictions was quite large, being 69 score points (almost three-quarters of a standard deviation) between the average scores of students in the Australian Capital Territory and those in the Northern Territory. The performance of students in the Australian Capital Territory was significantly higher than that of students in all other jurisdictions. Students in the Northern Territory performed significantly below students in all other jurisdictions. The jurisdiction with the highest percentage of students achieving the Advanced international benchmark was the Australian Capital Territory, in which 14 per cent of students achieved the highest level. In Western

Australia, nine per cent of students and in New South Wales and Victoria eight per cent of students achieved this benchmark.

The Northern Territory had the lowest proportion of students at this level, with just three per cent achieving the Advanced benchmark. Forty-two per cent of students in the Northern Territory did not reach the Intermediate international benchmark, which is the proficient standard for Australia. In the other jurisdictions, this proportion ranged from 15 per cent in the Australian Capital Territory to 30 per cent in Western Australia. Queensland and South Australia both significantly improved their scores in Year 4 science since TIMSS 2011, with increases of 23 score points and 18 score points, respectively. Regarding trends over 20 years, only Queensland showed a significant difference in performance between 1995 and 2015, an increase of 20 score points.

B. Science Assessment Domains in TIMSS 2015

In TIMSS 2015, the science assessment test was organised around two dimensions: the content dimension, specifying the subject matter to be assessed; and the cognitive dimension, specifying the thinking processes to be assessed.

Three major content domains define the science content for the TIMSS Science-Fourth Grade assessment: life science, physical science, and earth science [5]. Life science is represented by five topic areas, i.e. characteristics and life processes of organisms; life cycles, reproduction, and heredity; organisms, environment, and their interactions; ecosystems; and human health. The topic areas for the physical science include the following: classification and properties of matter and changes in matter; forms of energy and energy transfer; and forces and motion. Earth science is represented by some topics such as earth's structure, physical characteristics, and resources; earth's processes and history; and earth in the solar system.

The cognitive dimension is divided into three domains that describe the thinking processes students are expected to use when encountering the science items developed for TIMSS 2015. The first domain, knowing, addresses the student's ability to recall, recognise, and describe facts, concepts, and procedures that are necessary for a solid foundation in science. The second domain, applying, focuses on using this knowledge to generate explanations and solve practical problems. The third domain, reasoning, includes using evidence and science understanding to analyse, synthesise, and generalise, often in unfamiliar situations and complex contexts.

C. Students' Science Achievement

Reynolds & Walberg [6] showed that cognitive, behavioural, and attitudinal outcomes are influenced by three sets of factors such as aptitude and students attribute (student ability or prior achievement, motivation, and age or developmental level); instruction (quantity of time and quality); and psychological environment (classroom climate, stimulating qualities of the home environment, peer environment, and leisure time).

Students' science achievement was described as content-based science performance which can be measured by some tests, for example, students' comprehension of a science passage, science course grade, and state science test scores. That definition was stated by O'Reilly & McNamara [7] who conducted students' science performance by assessing their cognitive variables such as knowledge, reading skill, and reading strategy knowledge. The result showed the cognitive variables reliably influence all three science achievement tests.

D. The Rasch Model for Measurement

Rasch model is one-parameter Item Response Theory (IRT) model which can overcome the classical test theory problem by producing items and person statistics independent [8]. The analysis of Rasch is a psychometric technique that was established to improve the accuracy with which researchers create instruments, monitor instrument quality, and compute respondents' performances [9].

Rasch analysis enables researchers to construct alternative forms of measurement instruments. Rasch analysis also helps researchers think in more complex ways concerning the constructs (variables) they wish to measure. Bond & Fox [10] stated that the Rasch model focuses on a person who has ability encountering an item which has the level of difficulty and the likelihood of the person gets the items correct. The probability of success depends on the difference between the ability of the person and the difficulty of the item.

III. Research Gaps, Purpose, and Questions

A. Research Gaps

In 2015, there had been conducted a survey at Australian Primary schools which studied about students' science achievement for Grade 4. Nonetheless, there is no advanced research which examines the person fit and item fit of students' attitude towards science questionnaire for Australian students in Grade 4.

B. Purpose

The purpose of the study was to evaluate the person and item fit of science achievement test for Australian grade 4 students using Rasch measurement.

C. Research Questions

1) How is the person fit analysis of science achievement test?
2) How is the comparison of mean-square between initial and final analysis of person fit?
3) How is the item fit analysis of science achievement test after removing misfitting person?
4) How is the Wright map of the science achievement test for Australian grade 4 students?

D. Limitation of the Study

Although the study used a sufficient amount of large sample size, the sample only represented a particular school in Australia. Therefore, the results of the study could not be generalised to the other school in the target population.

E. Significance of the study

The results of the study could be the considerations for the TIMSS researcher to create a new instrument or revise the existed instrument regarding students' science achievement. The study would also give empirical evidence

to support the theoretical framework which has been established in the previous study.

IV. METHOD

A. Research Design

This study is an ex-post facto study based on the cross-sectional survey design in TIMSS 2015 which examines the Australian Grade 4 students' science achievement. The data was collected by a test to investigate students' science achievement. There are 14 booklets of test items used in the TIMSS 2015. However, this study only examines the test booklet 3.

B. Study Participants and Data Under Examination

The participants of this study were the students grade 4 at primary school level in Australia. The number of total participants in this study was 421 (N = 421). The science achievement test consists of 21 items with the topics such as life science, physical science, and earth science. The data underestimation are the students' science achievement test and the data were dichotomous.

C. Techniques of Analysis Used

This study used Conquest software to validate the instruments and the analysis is conducted in two stages. First, the software evaluates the person fit, then the misfitting persons are removed based on the criteria. Curtis [11] stated that a person is regarded as underfitting if the person fit index is greater than 1.5 or as overfitting if the person fit index is less than 0.6. These values were based on recommendations provided by Bond & Fox [10]. Second, the software examines the item fit of the reduced data set. For a multiple-choice test with high stakes, Bond & Fox [10] recommended an acceptable Infit Mean Square range from 0.8 to 1.2.

V. RESULT AND DISCUSSION

A. Research Question 1: How is the person fit analysis of science achievement test?

The initial analysis involved 421 students. The result indicated that there were ten underfitting persons with the range of person fit index from 1.512 to 1.984. The underfitting persons were removed from the data set because they would make the pattern unpredictable and distort the psychometric properties of the measurements. Another person has also been removed from the data set since that person did not answer the test at all. The analysis also showed there were 31 overfitting persons range from 0.023 to 0.495. Considering that the sample size would be reduced, the overfitting persons were not removed from the data set. This decision might result the redundancy, however, the pattern would still be able to predict.

B. Research Question 2: How is the comparison of mean-square between initial and final analysis of person fit?

The item properties of the science achievement test are presented in table 1 with the comparison of initial and final analysis of the misfitting person. The initial difficulty levels for the items range from -1.588 to 1.282 and the difficulty levels after misfitting person removal are slightly different ranging from -1.683 to 1.340. The initial and final infit MNSQs have the same range, from 0.910 to 1.100 and they are within the range of high stakes MCQs [10]. The initial outfit MNSQs range from 0.760 to 1.160 and the final outfit MNSQs range from 0.770 to 1.140.

The Rasch item reliability slightly changes from 0.971 to 0.972 which means that they are very high. However, although they are not quite high, the Rasch person reliability also changes from 0.676 to 0.679 and both are acceptable.

TABLE I. SCALE PROPERTIES OF THE SCIENCE ACHIEVEMENT TEST IN THE INITIAL AND FINAL ANALYSIS

No	Item	Initial Analysis			Final Analysis		
		Measure (Logits)	*Outfit MNSQ*	*Infit MNSQ*	*Measure (Logits)*	*Outfit MNSQ*	*Infit MNSQ*
1	S051041	0.582	1.030	1.030	0.596	1.030	1.040
2	S051004	-0.123	0.880	0.910	-0.099	0.890	0.910
3	S051026A	-0.326	1.080	1.060	-0.290	1.090	1.050
4	S051026B	0.390	1.080	1.080	0.409	1.110	1.090
5	S051026C	-1.097	0.830	0.960	-1.236	0.770	0.960
6	S051026D	-1.588	0.760	0.950	-1.683	0.780	0.960
7	S051114	0.801	0.960	0.970	0.845	0.930	0.950
8	S051121A	-0.687	1.020	0.970	-0.679	0.990	0.970
9	S051121B	-0.719	1.130	1.060	-0.755	1.040	1.050
10	S051121C	-0.293	1.010	1.040	-0.312	1.000	1.030
11	S051121D	-0.092	1.110	1.060	-0.089	1.120	1.060
12	S051121E	-0.377	0.990	1.000	-0.399	0.990	1.000
13	S051105	0.103	1.020	1.010	0.119	1.010	1.000
14	S051110	0.518	0.940	0.960	0.506	0.920	0.950
15	S051111	0.098	1.020	1.030	0.128	1.000	1.010
16	S061135	-0.757	0.890	0.970	-0.810	0.910	0.970
17	S061134	0.411	0.990	0.980	0.446	1.010	0.990
18	S061140	1.282	1.160	1.100	1.340	1.140	1.100
19	S061022	0.108	0.930	0.980	0.138	0.960	0.990
20	S061118	1.219	1.030	1.020	1.253	1.020	1.000
21	S061097	0.549	0.990	1.000	0.576	0.980	1.000
Person Separation Reliability			0.676		0.679		
Item Separation Reliability			0.971		0.972		

C. Research Question 3: How is the item fit analysis of science achievement test after removing misfitting person?

The item properties of the science achievement test in the final analysis of misfitting person are presented in table 2. The difficulty levels for the items range from -1.683 to 1.340 logits, associated with standard errors of 0.1 or 0.2. These standard errors are not quite small which means that the difficulty estimations are not perfectly precise but they are still acceptable.

The infit MNSQs range from 0.910 to 1.100 and the outfit MNSQs range from 0.770 to 1.140. These values are within the reasonable range for high stakes MCQs [10] which implies that the items are good and indicate the sufficient fit to Rasch model for practical measurement purposes. There are t-values of infit obtained from the analysis and there is an item (S061140) which has t-value more than +2.00. However, this situation can be ignored since the t-values are sensitive to the sample size and the rule of acceptable range from -2.00 to +2.00 is only applicable for the sample size of 30 to 300 [10].

The Rasch item reliability is 0.972 which is very high and it has meaning that if the items are administered to the other sample, they tend to give the similar result. However, the Rasch person reliability is 0.679 which means a little bit low but it is still acceptable. It means that if the person is given the other set of items with the same indicators, the results tend to be similar.

TABLE II. SCALE PROPERTIES OF THE SCIENCE ACHIEVEMENT TEST IN FINAL ANALYSIS

No	Item	Measure (Logits)	S.E.	Outfit		Infit	
				MNSQ	*t*	*MNSQ*	*t*
1	S051041	0.596	0.110	1.030	0.400	1.040	0.800
2	S051004	-0.099	0.122	0.890	-1.600	0.910	-1.400
3	S051026A	-0.290	0.127	1.090	1.200	1.050	0.700
4	S051026B	0.409	0.113	1.110	1.500	1.090	1.700
5	S051026C	-1.236	0.168	0.770	-3.600	0.960	-0.300
6	S051026D	-1.683	0.200	0.780	-3.300	0.960	-0.200
7	S051114	0.845	0.107	0.930	-1.000	0.950	-1.300
8	S051121A	-0.679	0.142	0.990	-0.200	0.970	-0.400
9	S051121B	-0.755	0.145	1.040	0.600	1.050	0.600
10	S051121C	-0.312	0.128	1.000	0.100	1.030	0.400
11	S051121D	-0.089	0.124	1.120	1.700	1.060	0.900
12	S051121E	-0.399	0.132	0.990	-0.200	1.000	0.100
13	S051105	0.119	0.116	1.010	0.200	1.000	0.100
14	S051110	0.506	0.110	0.920	-1.100	0.950	-1.000
15	S051111	0.128	0.116	1.000	0.000	1.010	0.300
16	S061135	-0.810	0.146	0.910	-1.200	0.970	-0.300
17	S061134	0.446	0.111	1.010	0.200	0.990	-0.100
18	S061140	1.340	0.107	1.140	2.000	1.100	2.900
19	S061022	0.138	0.118	0.960	-0.500	0.990	-0.200
20	S061118	1.253	0.109	1.020	0.300	1.000	0.100
21	S061097	0.576	0.112	0.980	-0.300	1.000	0.000
Person Separation Reliability				0.679			
Item Separation Reliability				0.972			

D. Research Question 4: How is the Wright map of the science achievement test for Australian grade 4 students?

The Wright map of the science achievement test is presented in figure 1. It can be seen from the figure that most of the students have high person ability indicated by many "X" positioned above 0 logits. The most difficult items are items number 18 and 20 (S061140 and S061118), but almost half of the students can answer all of the questions correctly. There are two items which are considered as the very easy item, i.e. item number 5 and 6 (S051026C and S051026D), and the figure indicates that almost all of the students could answer those questions correctly.

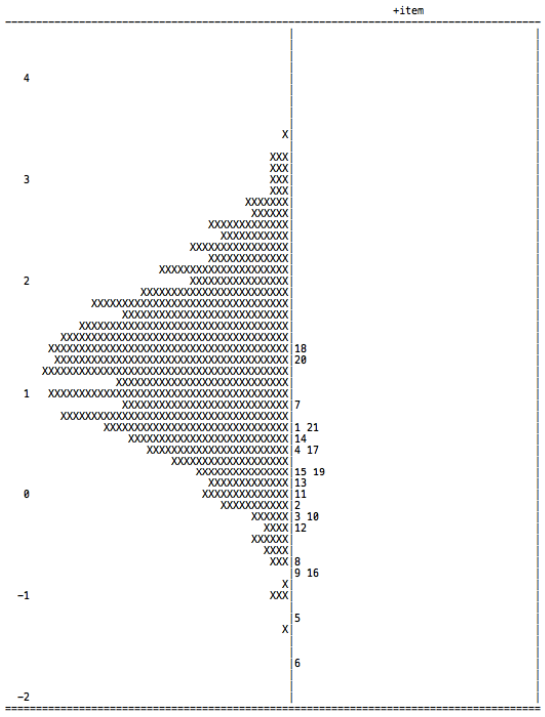

Fig. 1. Wright map of Science Achievement Test for Australian Grade 4 Students. Each X represents 0.6 cases.

VI. CONCLUSION

A. Summary of findings

The result indicated that there were ten underfitting persons with the range of person fit index from 1.512 to 1.984. The item difficulty range of the initial analysis and that of after misfitting person removal is slightly different. The initial and final infit MNSQs have the same range, and they are within the range of high stakes MCQs. The initial and final outfit MNSQs are somewhat distinct. The Rasch item reliability slightly changes and they are very high. However, the Rasch person reliability also changes and both are acceptable.

In the final analysis, the difficulty levels have not quite small standard errors which mean that the difficulty estimations are not perfectly precise but they are still acceptable. The infit MNSQs and the outfit MNSQs are within the reasonable range for high stakes MCQs which implies that the items are good and indicate the sufficient fit to Rasch model for practical measurement purposes. The Rasch item reliability is very high and the Rasch person reliability is a little bit low but it is still acceptable. The Wright map of the science achievement test shows that most of the students have high person ability and almost half of the students can answer all of the questions correctly. There are two items which are considered as the very easy item and it indicates that almost all of the students could answer those questions correctly.

B. Implications of the study

1) Theoretical; This study gave empirical evidence to support the theories which have been established before.
2) Practical; The findings can be used as the identification of the fit to the Rasch model for practical measurement purposes, then it would help the decision to keep the item or remove it from the instruments.

C. Recommendation for future research

This study only evaluates the fit of the instrument to the Rasch model and it is only applied to Australian Grade 4 students. Meanwhile, there are 50 more countries which the instruments have not been evaluated yet. In addition, the results showed by TIMSS only reveal the proportion of students' science achievement score. There is no study which investigates the relationship among variables obtained in TIMSS 2015. Therefore, other countries' instrument evaluation and the relationship-causation analysis are possible to be conducted.

REFERENCES

[1] The Partnership for 21st Century Skills, 'P21 Framework Definitions'. he Partnership for 21st Century Skills, 2009.

[2] Australian Education Ministers, 'Melbourne Declaration on Educational Goals for Young Australians'. Ministerial Council on Education, Employment, Training, and Youth Affairs, 2008.

[3] M. O. Martin, I. V. S. Mullis, P. Foy, and M. Hooper, 'TIMSS 2015 International Results in Science', Boston College, TIMSS & PIRLS International Study Center, 2016.

[4] S. Thomson, N. Wernert, E. O'Grady, and S. Rodrigues, *TIMSS 2015; A First Look at Australia's Result*. Victoria: Australian Council for Educational Research Ltd, 2016.

[5] I. V. S. Mullis and M. O. Martin, 'TIMSS 2015 Assessment Frameworks', 2013. [Online]. Available: http://timssandpirls.bc.edu/timss2015/frameworks.html.

[6] A. J. Reynolds and H. J. Walberg, 'A Structural Model of Science Achievement and Attitude: An Extension to High School.', *J. Educ. Psychol.*, vol. 84(3), p. p 371–382, Sep. 1992.

[7] T. O'Reilly and D. S. McNamara, 'The Impact of Science Knowledge, Reading Skill, and Reading Strategy Knowledge on More Traditional "High-Stakes" Measures of High School Students' Science Achievement', *Am. Educ. Res. J.*, vol. 44, no. 1, pp. 161–196, Mar. 2007.

[8] X. Fan, 'Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics', *Educ. Psychol. Meas.*, vol. 58, no. 3, pp. 357–381, Jun. 1998.

[9] W. J. Boone, 'Rasch Analysis for Instrument Development: Why, When, and How?', *CBE Life Sci. Educ.*, vol. 15, no. 4, 2016.

[10] T. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition*. Routledge, 2015.

[11] D. D. Curtis, 'The influence of person misfit on measurement in attitude surveys', *Unpubl. EdD Diss. Flinders Univ. Adel.*, 2003.