International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 13, Issue 7, 2020

Development of an Assessment Literacy Super-item Test for Assessing Preservice Teachers' Assessment Literacy

Lim Hooi Lian* **and Wun Thiam Yew**

[1]School of Educational Studies, Universiti Sains Malaysia,11800 Pulau Pinang, Malaysia. *Corresponding author email: hllim@usm.my

This study aimed to develop and determine the psychometric properties of an assessment literacy superitem test that assesses preservice teachers' assessment literacy. This was achieved through investigating their competence in selecting of assessment method and constructing assessment task. This study used a survey approach to assess 397 preservice teachers' level of assessment literacy. The test consisted of eight superitems with the total of 24 items. Data collected was analysed by using the Rasch model. The result of unidimensionality， reliability, and the item hierarchy revealed that the test pointing toward one dimension adhered to Rasch model's expectation. The majority of the samples did not perform well in both constructs assessed as they were stuck at lower levels. This result provides the vital information to the authority for planning the proper training and identifying actions to improve the quality of educational assessment systems. The assessment framework can become a useful reference for developing other superitem tests that assess different constructs.

Key words: *Assessment literacy super-item test, Preservice teachers, SOLO model, Rasch analysis, Psychometric properties*

INTRODUCTION

Over the last two decades has witnessed marked and obvious changes in educational assessment. As part of the movement towards 21st century teaching and learning goals, the focus and direction of educational assessment has been shifted significantly from conventional classroom assessment to school-based assessment system. School-based assessment emphasises the holistic assessment of students' development in cognitive, affective and psychomotor domains. Thus, various assessment methods need to be applied by teachers in order to assess accurately the different domains of learning outcomes (Norazilawati, Noorzeliana, Mohd Sahandri Gani, & Saniah, 2015; Salmiah, Ramlah, & Abdullah, 2013). As a result, the assessment literacy of teachers becomes the main catalyst and the most imperative criterion to the success of school-based assessment (Boudett, City, & Murnane, 2013; Lee, 2017; Rohaya, 2014).

Assessment literate teachers basically understand 'what to assess' (learning outcomes), 'how to assess professionally' (procedures), 'why they assess' (purposes) and 'how to use the assessment data' (decision making) (Khadijeh & Amir, 2015; Vahid & Nasree, 2019). Thus, the quality of school-based assessment is determined by teachers who play the role of instrument constructor, administer, examiner and interpreter. Many previous studies have investigated school teachers' assessment literacy at either the primary or secondary school level (Kahl, Hofman, & Bryant, 2012; Kanjee & Mthembu, 2015; Rohaya & Mohd Najid, 2008; Suah, 2012; Webb, 2002). The findings revealed that the inability of teachers to select appropriate assessment method and construct assessment task are the main factors affecting the effectiveness of implementing school-based assessment (Sewornoo, 2016; Suah, 2012). They are unable to effectively apply various assessment method and tend to use paper-and–pencil tests as they can adapt or adopt the assessment tasks directly from reference books or internet sources (Suah, 2012).

The process of selecting appropriate assessment methods and constructing assessment tasks requires the application of some systematic procedures and fundamental principles of assessment to ensure the high degree of assessment validity (Nitko & Brookhart, 2014). Thus, the combination of knowledge and skills in this aspect need to be assessed in a comprehensive and detailed manner in order to identify precisely the weaknesses and difficulties faced by teachers. In previous studies, teachers' assessment literacy in selecting assessment method and constructing assessment task was assessed far too generally and across various constructs as well. Likert-scale questionnaire and multiple-choice questions test were commonly used (Metler, 1999; Kanjee & Mthembu, 2015; Norazlina, 2014; Rohaya & Mohd Najid, 2008; Suah, 2012; Yamtim & Wongwanich, 2014). Hence, detailed information about the difficulties faced by teachers is in dire need of further research.

Therefore, in this study, the researchers aimed to develop and determine the quality (in terms of psychometric properties) of an assessment tool for assessing preservice teachers' assessment literacy in these both constructs, namely selecting appropriate assessment methods and

constructing assessment tasks. A comprehensive understanding of preservice teachers' ability to select an assessment method and construct an assessment task serves the dual purpose of informing the nature of teacher education reforms and the future direction of professional development training. Furthermore, there is still relatively little research devoted to the understanding of assessment literacy among preservice teachers. In fact, the investigation on assessment literacy should begin at the faculty or school of education in higher education institutions as they play the vital role in equipping the pre-service teachers with educational assessment skill and knowledge. Moreover, some researchers (DeLuca & Linger, 2010; Kim, 2014; Plake, 1993) found that teachers often claimed that they had a lack of test preparation skill is largely due to inadequate preservice professional training in educational evaluation and assessment. This potentially implies that preservice teachers claimed the assessment training in their undergraduate courses did not prepare them to be confident in developing school-based assessment.

To be a valid assessment tool, it must be a standard which does not change, just like the measure of the height of a building or wall. One of the fundamental principles of Item Response Theory is that it highlights that a measurement must be independent of the observer and is not dependent on the samples selected for measurement (Bond & Fox, 2015). Once calibrated, the scale should measure assessment literacy ability independent of the samples selected. Thus, in this study, one parameter of Item Response Theory, namely Rasch analysis, has been applied to perform data analysis in evaluating the appropriate degree of construct validity and reliability of the 24-item assessment literacy superitem test. The Rasch model is one of the parameters from the Item Response Theory and has been applied in many validation studies in various fields of study, such as sport education (Hecimovich1 & Marais, 2017), mental health (Chang, Ailey, Heller, & Chen, 2013) medical health (Bagraith, Strong, Meredith, & McPhai, 2017) counselling (Mallinckrodt, Miles, & Recabarren, 2016), psychology (Sartori & Pasini, 2006), and social science (Carpita & Golia, 2012). Previous studies have shown that Rasch analysis offers a powerful and useful examination of psychometric properties for new and adapted instruments.

Purpose of the Study

This study aimed to develop and validate an assessment literacy superitem test for assessing preservice teachers' competency in selecting assessment method and constructing assessment task. The quality of the test in terms of psychometric properties was examined using the Rasch model. The tables and figures in this article are annotated to summarise and highlight the main points of psychometrics analysis. A superitem is a format of item that provides more user friendly and effective way to determine the learners' ability level and detects their strengths and weakness if they are not progressing past a certain level.

Theoretical framework

The SOLO (Structure of the Observed Learning Outcome) model, which was developed by Biggs and Collis (1982), is a cognitive psychology model that concerns more on the structure of the response, analysing 'how' a task is responded to rather than whether the response is correct or not. In this study, a combination of this model and a superitem format had been applied to assess preservice teachers' assessment literacy pertaining to the competence in selecting an assessment method and constructing ~~an~~ assessment task. In this combination, each task consists of a problem situation, followed by three different complexity levels of questions related to it. The problem situation is represented by text or diagram while the questions represent the levels of cognitive reasoning defined by the SOLO model which include unistructural, multistructural, relational, and extended abstract. Thus, a correct response to a question within any superitem would indicate the cognitive ability at a certain level reflected in the SOLO model. This format of item provides more user friendly and effective way to determine the learners' ability level and detects their strengths and weakness if they are not progressing past a certain level.

The researchers hypothesised that preservice teachers involved in this study should exhibit three basic levels of assessment literacy. Therefore, the theoretical framework had been developed along with the expected preservice teachers' assessment literacy pertaining to the competence in selecting assessment methods and developing assessment tasks based on the three levels of SOLO model, namely unistructural, multistructural, and relational. Table 1 shows an example of the superitem task which was used in this study. There are two main stages of skills that need to be equipped in order to achieve the competency of selecting the appropriate assessment method: identifying the intended learning outcomes to be assessed and locating the domains involved. This type of item format allows the assessor to detect easily the weaknesses of preservice teachers if they perform poorly in this competency.

Table 1 *The Framework on the Characteristics of Assessment Literacy Pertaining to the Competence in Selecting Assessment Method*

Mr Jeffry is a mathematics teacher of the Form Two Waja class. The class has 20 students who have different levels of ability. He wants to give his students the opportunity to communicate mathematically about the concepts of linear patterns, identify the students' ability in generalising the linear pattern and applying the concept of algebraic expressions and linear equations.

Level of task	Task	Description
Unistructural	What is one of the intended learning outcomes to be assessed by Mr Jeffry?	This level requires the response to directly refer to a piece of concrete information in the task. This task requires the understanding of the intended learning outcome. The task can be responded to based on the concrete information given; that is, identify an intended learning outcome from the given information.
Multistructural	Classify the intended learning outcomes into domains.	This task requires the given information to be applied in order. That is, identify all the intended outcomes and do the classification into domains. The information given in the task is still used directly.
Relational	Which assessment method is the most appropriate that can be proposed to Mr. Jeffry? Give reason to support your response.	The task requires the integration of all given information to make a decision. The learner has to consider all the given information in order to decide on the most appropriate assessment method.

METHODOLOGY

This study used a quantitative approach to assess 397 preservice teachers' level of assessment literacy. The preservice teachers were final year undergraduate students. They did their first education degree at a local university. They had completed their educational measurement and evaluation course and waited to be posted to secondary schools for their teaching practicum.

In this study, the instrument of data collection consisted of eight superitems to assess preservice teachers' level of assessment literacy in selecting assessment method and constructing assessment tasks. All the three items in each superitem are in an open-ended format. Open-ended item format might require the preservice teachers to respond with a word, a phrase, or they may require a long and complex response. Superitem 1 to superitem 4 with the total of 12 items (items 1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4a, 4b, and 4c), assessed the preservice teachers' ability in detecting the weaknesses of the test item and then revised it. Superitem 5 to superitem

8 with the total of 12 items (items 5a, 5b, 5c, 6a, 6b, 6c, 7a, 7b, 7c, 8a, 8b, and 8c), assessed the preservice teachers' ability in identifying learning outcomes based on the information given and suggested appropriate assessment method to assess the particular learning outcomes (refer to example of superitem 7 in Table 1). The time duration of completing the 24-item test was approximately an hour.

The test paper results were analysed by using a rating scale analysis of the Rasch model. Rating scale analysis (Wright & Masters, 1982) is a statistical model that specifically incorporates the possibility of having the same number of steps or levels for the items in a test (Bond & Fox, 2001). For example, the ordered values of 0, 1, and 2 might be applied to each item in the superitems which has three ordered performance category levels as follows: 0 = totally wrong, 1 = partially correct, and 2 = completely correct. WINSTEP software program was used to run the analysis. It estimated the psychometric properties in terms of reliability and construct validity.

The Statistical Analysis

a. The unidimensionality was determined by examining the principal component analysis (PCA), fit statistics and point-measure (PTMEA) correlations. These allow determining whether all the tasks developed represent the single construct.
b. The reliability was examined by estimating the item and person reliability as well as the item and person separation. The reliability ranges from 0 to 1, with higher values indicating higher degrees of reliability. Meanwhile the item and person separations estimate the degree of assessment literacy superitem test in discriminating items or person into levels.
c. Item hierarchy was examined to provide the additional indication of construct validity that is whether all the items are ordered in terms of endorsability. The person-item map estimates the difficulty levels of the item and whether the person-item targeting is adequate.

RESULTS

1. Dimensionality

Evidence of dimensionality was derived from the (a) item fit mean square (MNSQ), (b) point measure (PTMEA) correlation, and (c) Rasch residual based principal components analysis (PCA).

Item fit MNSQ in this study included both infit and outfit statistics, used to measure item fit on preservice teachers' sample scale. Table 2 contained the item fit statistics for samples. All items had acceptable infit and outfit statistics, either above 0.60 or below 1.4 threshold values (Bond & Fox, 2015), except items 2A and 3A which the MNSQ values were slightly above 1.4. However, the misfit superitems were retained because of the consideration that the mean for the overall infit and outfit lied within the acceptable range of 0.6 and 1.4 for the mean of the

mean square scores, that is 1.06 for infit and outfit respectively, suggesting there existed no redundancy and heterogeneity of items for the samples.

Table 2

Item Statistics

Item	Measure	Infit		Outfit		PT-MEASURE
		MNSQ	ZSTD	MNSQ	ZSTD	CORR.
2A*	-1.78	1.50	5.3	1.56	4.8	.26
3A*	-.43	1.46	7.4	1.47	6.9	.32
4A	-.88	1.38	5.7	1.40	5.4	.32
2C	1.43	1.17	1.8	1.40	3.1	.25
6C	.91	1.34	4.5	1.29	3.0	.37
4C	1.51	1.34	3.4	1.33	2.5	.33
5C	1.96	1.30	2.4	1.17	1.1	.34
7C	1.54	1.28	2.8	1.16	1.3	.31
8C	.79	1.24	3.4	1.26	2.9	.37
7A	-2.49	1.24	2.0	.89	-.8	.52
1A	-3.50	1.10	.6	1.22	.9	.23
8A	-2.77	1.21	1.6	1.14	.9	.39
5A	-1.57	1.17	2.2	1.12	1.3	.51
3C	2.01	1.09	.8	.94	-.4	.29
3B	1.29	1.04	.5	1.08	.7	.32
6A	-1.66	1.03	.4	.99	-.1	.48
4B	1.26	.78	-2.9	.90	-.9	.35
1C	1.41	.79	-2.5	.84	-1.4	.38
6B	.35	.77	-4.4	.84	-2.5	.46
2B	.24	.63	-7.6	.77	-3.8	.39
1B	.73	.69	-5.3	.76	-3.2	.44
5B	-.03	.73	-5.5	.73	-4.8	.52
8B	-.30	.55	-9.9	.60	-7.9	.39
7B	-.03	.56	-9.9	.56	-8.6	.50
Mean	.00	1.06	-.1	1.06	.0	
S. D	1.55	.29	4.7	.27	3.7	

Remark: * misfit

In the aspect of point measure (PTMEA) correlation, all 24 items exhibited positive and moderate to strong PTMEAs, ranged from 0.23 to 0.52 (Refer to Table 2). According to Bond and Fox (2015), items with the value of point measure (PTMEA) correlation above 0.20 are acceptable. All items were acting as expected with regard to the underlying construct but not multidimensional.

The analysis results of the principal components analysis (PCA) for the samples were presented in Table 3 and Figure 1. In total, 54.3% of the variance was accounted for by the unidimensional model, closely matched the modelled value of 54.8%. Additionally, the unexplained variance in first contrast had an eigenvalue of 3.1 and accounted for 6.0% for the unmodeled data.

Table 3

Standardised Residual Variance (in Eigenvalue units)

		Empirical		Modelled
Total raw variance in observations	52.5	100.0%		100.0%
Raw variance explained by measures	28.5	54.3%		54.8%
Raw variance explained by persons	6.9	13.2%		13.4%
Raw Variance explained by items	21.5	41.0%		41.4%
Raw unexplained variance (total)	24.0	45.7%	100.0%	45.2%
Unexplained variance in 1st contrast	3.1	6.0%	13.1%	

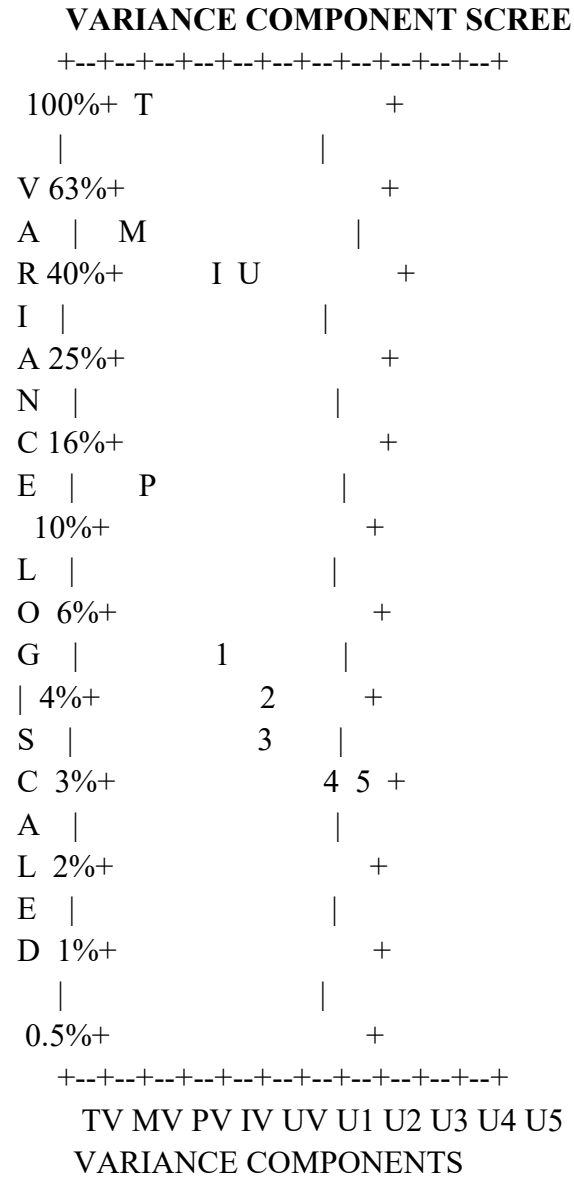
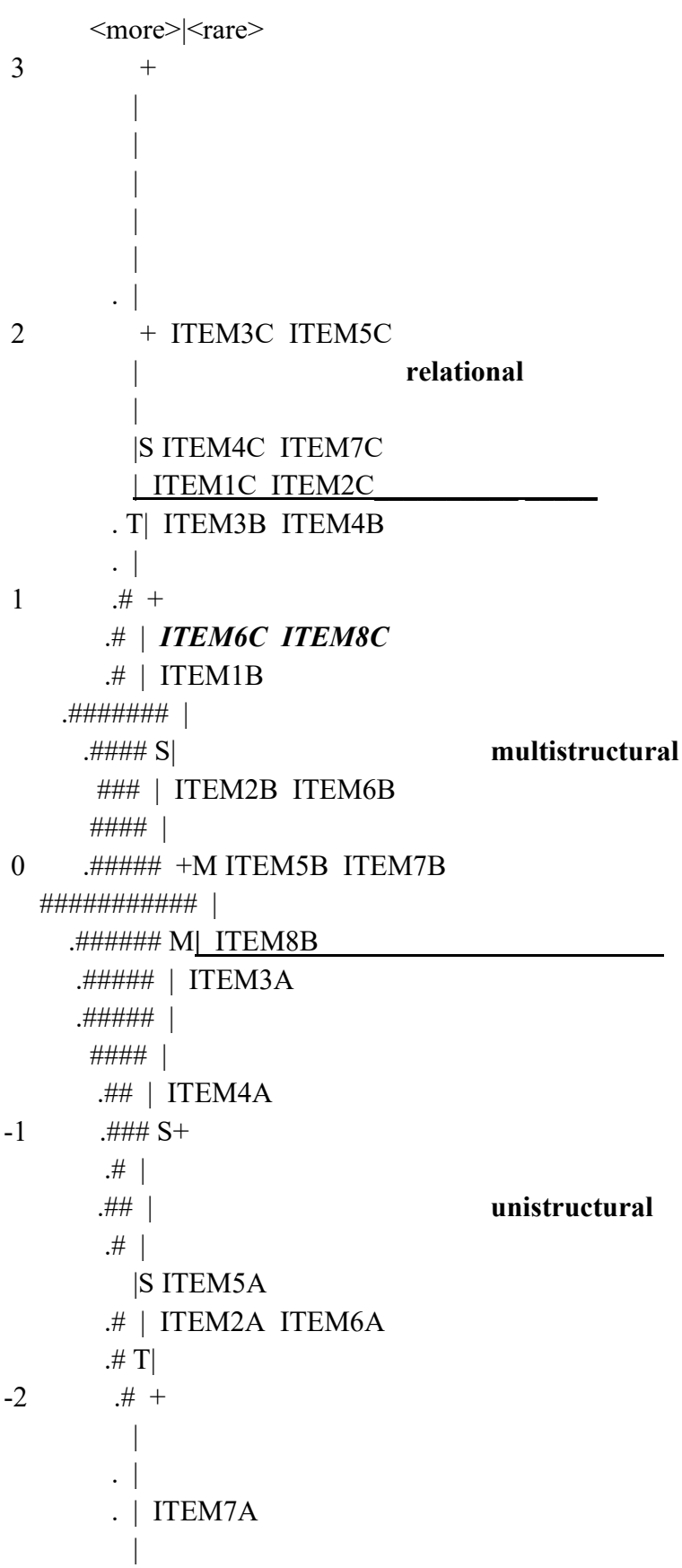


Figure 1. **Variance component scree plot**

In general, each and every evidence of item statistics [item fit mean square (MNSQ)], point measure (PTMEA) correlation as well as Rasch residual based principal components analysis (PCA) presented above support the fact that the assessment literacy items appeared to be unidimensional for the samples in this study.


Persons - MAP - Items
<more>|<rare>
3
2
1
0
-1
-2
ITEM3C ITEM5C
relational
ITEM4C ITEM7C
ITEM1C ITEM2C
ITEM3B ITEM4B
ITEM6C ITEM8C
ITEM1B
multistructural
ITEM2B ITEM6B
ITEM5B ITEM7B
ITEM8B
ITEM3A
ITEM4A
unistructural
ITEM5A
ITEM2A ITEM6A
ITEM7A

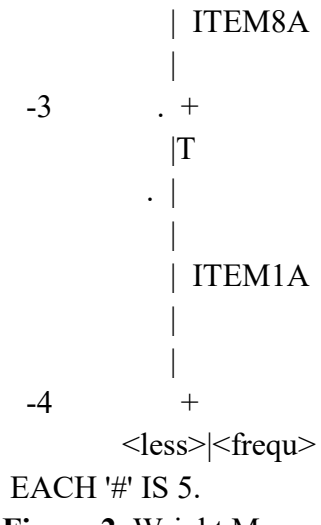


Figure 2. Wright Map

Person-Item Reliability and Separation Indices

Table 4 shows the person-item reliability and separation indices. From the table, item reliability was 1.00. This indicated that all the 24 items in the assessment literacy superitem test were highly reliable. Meanwhile, the separation value for item was 14.63. This suggested that the items can be grouped into 14 levels of difficulty.

On the other hand, person reliability was .72. This indicated that the samples involved in this study were reliable. In addition, person separation value (1.63) indicated that approximately two distinct groups can be identified in the data.

Table 4 *Person-Item Reliability and Separation Indices*

	Separation	Reliability
Item	14.63	1.00
Person	1.62	.72

2. ***Appropriateness of the Item Difficulty Level for Samples***

Figure 2 illustrated the map of persons and 24 items of assessment literacy superitem test for the samples recruited in this study. Person latent trait (ability) and item difficulty (known as item measure) were arranged following the sequence of highest to lowest. Hence, samples with higher level of assessment literacy and items gauging more severe degree of difficulty, were located at the top of the Wright map. Item difficulty distribution covered the person ability distribution. Item covered a range of approximately – 3.50 to + 2.01 logits (coverage of more than 2 standard deviations) while person covered a range of approximately - 3.28 to + 2.12 logits (within 2 standard deviations). It means that all the items can cover the range of traits measured. The mean of the person was lower than the mean of the items, indicating that the samples' assessment literacy was lower than the difficulty of items.

The level of item for each superitem corresponded to the alphabet of the item. For example, item 1A means the item in the lowest level (unistructural) for superitem 1, 2B means the item in multistructural level for superitem 2 and 5C means the item in relational level for superitem 5. Generally, within each superitem, the expected order of the items was unistructural, multistructural and relational. This expected order was found to hold for all superitems. Item 1A until 8A were generally placed on the lowest position (the easiest). Item 1B until 8B were generally placed on the middle position. The highest position was occupied by the most difficult items, namely items 1C, 2C, 3C, 4C, 5C and 7C; except items 6C and 8C which were easier and fall into the multistructural level zone.

Based on the clusters of items (horizontal lines) and samples as shown in the previous paragraph, three different levels could be distinguished, namely unistructural, multistructural and relational. These levels are the hierarchy of SOLO model applied in this study. Samples at unistructural level could understand the task and able to identify a learning outcome or types of item format correctly based on the information given. Meanwhile students at multistructural level were able to identify all the learning outcomes correctly or identify all the weaknesses of items based on the information given. At the third level, namely relational level, samples had shown their abilities in suggesting one appropriate assessment method or revise the problematic item based on the weaknesses detected.

DISCUSSION

Prior to the detailed analysis of the results obtained by using Rasch analysis, it was vital to examine the aspect of unidimensionality of the instrument in order to ensure the data collected fit to the Rasch model reasonably (Green & Denver, 2002). If it did not fit the Rasch model, another model would need to be utilised. In line with this, fit statistics result was used to determine how well the raw data fit the Rasch model (Bond & Fox, 2015; Boone, Staver, & Yale, 2014; Chang & Wu, 2008). Infit and outfit MNSQ for person and item are expected to be 1.00 (Green & Frantom, 2002). However, Boone, Staver, and Yale (2014) stated that generally, a range between 0.5 and 1.5 suggests a reasonable fit of the data to the model. Based on the results of this study, infit MNSQ for person was 1.03 and infit MNSQ for item was 1.06. Meanwhile, the outfit MNSQ for both person and item was 1.06. Both fit statistics values showed that the data obtained fit to the Rasch model expectations. In other words, most of the samples had shown that their responses are within the expectations of the model. The results suggested that the items of the developed instrument were able to discriminate the samples with different assessment literacy levels.

Moreover, the fit statistics were also used to further examine the item-level model fit. Based on the results, there are only two items out of 24 items that are not in the reasonable range, which are items 4 and 7. However, the deviations are small (item 4: infit MNSQ = 1.50; outfit MNSQ = 1.56; item 7: infit MNSQ =1.46; outfit MNSQ =1.47). According to Wu and Adams (2007), and Staver and Yale (2014), the misfit item in one test may fit well with items in another test.

Thus, instead of setting rules for accepting or rejecting the item, fit statistics result should be served as an indicator for identifying problematic items and then revise the items.

On top of that, point measure (PTMEA) correlation analysed for this study revealed that the values ranged from 0.23 to 0.52. Each item was found to have positive and moderate strength of PTMEA values. In general, item with point-biserial correlation, r_{ph}> .20 is acceptable whereas r_{ph}< .15 should be examined for further action (McCormack, Masse, Bulsara, Pikora, & Giles-Corti, 2006). According to Linacre (2006), the value of point measure more than 0.3 indicates that all the items correlate positively toward measuring the same construct. The results of this study showed that 22 out of 24 items had positive values of PTMEA results which ranged from 0.3 and 0.52. This indicated that the instrument acting as expected with regard to the underlying construct.

Principal component analysis of residuals was referred to evaluate whether a substantial factor existed in the residuals after the primary measurement dimension had been estimated (Smith, 2002). This analysis allowed the determination of whether the items developed represented a single construct. In terms of Rasch residual based PCA comparison, this study found that 54.3% of the variance was accounted for the model, closely matched the modelled value of 54.8%. According to Linacre (2011), and Teh and Lim (2016), the minimum value of recommended variance is 40. However, Bagraith, Strong, Meredith, and McPhail (2017) stated that at least 50% of total variance should be determined to support the unidemensionality. Obviously, the principal component analysis showed the acceptable unidimensionality and also indicated the assessment literacy superitem test is appropriate to be used for assessing assessment literacy.

The PCA analysis was used to test the assumption of Rasch model. The eigen value of the unexplained variance in first contrast obtained in this study was 6 percent, as stated by Fisher (2007) and far from the ceiling value, that is 15 percent. The result from the PCA seems to support a unidimensional construct of the instrument, thus suggesting that the application of the level of SOLO model in assessing assessment literacy is warranted.

The Rasch analysis showed the acceptable value of person reliability (0.72) and good value of item reliability (1.0) (Masran, Rahim, Faizal, & Marian 2017). These estimations indicate the high replicability of result across both person and item. Meanwhile, the person separation and item separation indices represent the very important additional information to the evaluation of the developed instrument's function (Boone, Staver, & Yale, 2014). Based on the results of this study, the value of item separation and person separation were 14.63 and 1.63 respectively. According to Linacre (2012), there is no ceiling for these indexes. Thus, it can be ranged from 0 to infinity. However, for the purpose of introductory analysis, the higher value of separation will be better. Tennant and Conghan (2007) stated that if the items are analysed at an individual level, the item separation value of 1.5 is required. If the items are analysed at the group level, the minimum of 2.5 for the item separation is required. Meanwhile, Duncan, Bode, Lai, and Perera (2003), and Garzón Umerenkova, de la Fuente Arias, Martínez-Vicente, Zapata

Sevillano, Pichardo, and García-Berbén (2017) revealed that an acceptable value of person separation is 1.50, a good level is 2.00 and 3.00 represents an excellent level of separation. Based on the criteria suggested by experts, it can be concluded that the value of person separation in this study is acceptable and item separation value showed the appropriateness of items to be analysed at group level.

Next, person-item distribution was analysed by investigating the Wright map (see Figure 1). It is very useful to determine and ensure that the developed instrument is able to detect the full variability of population. As claimed by Alquraan, Alshraideh, and Bsharah (2010), and Garzón Umerenkova, de la Fuente Arias, Martínez-Vicente, Zapata Sevillano, Pichardo, and García-Berbén (2017), an instrument should be able to assess individual at both high levels and low levels of wisdom. The results of the study showed that all the 24 items could cover the range of traits measured. In other words, the results give suggestions that the items can discriminate students with different assessment literacy levels. Generally, the ordering of items on the Wright map matched the hierarchy level of SOLO model. Since the measures are in interval scale, one important observation is that the most difficult item of the test, namely items 1c, 2c, 3c, 4c, 5c and 7c were high in difficulty. In particular, the item difficulty measures showed that this test consists of items in which the difficulty level did not correspond to the level of assessment literacy of the samples. It revealed that majority of preservice teachers were unable to perform well in both constructs assessed due to the fact that they were 'stuck' at the unistructural and multistructural levels.

Summary

The quality of the test in terms of psychometric properties was determined using the Rasch model. The result of unidimensionality, reliability, and the item hierarchy revealed that the developed test pointing toward one dimension adhered to Rasch model's expectation. Moreover, the findings also revealed that majority of the samples did not perform well in both constructs assessed as they were stuck at lower levels, namely unistructural and multistructural.

CONCLUSION

These results provide vital information to the authority for planning proper training and also to identify actions to improve the quality and efficiency of educational assessment systems. In addition, the development of the assessment literacy superitem test highlights the need for school teachers to practice self-assessment continuously. Through the newly developed instrument, teachers can easily diagnose the strengths and weaknesses of their assessment literacy. They will be clearly informed of the reason behind their inability to achieve the highest level. In short, the results are potentially to be used to analyse assessment literacy in a detailed manner; either collectively or individually.

Although the developed test had been revised and determined to be valid and reliable, future studies can be carried out to improve the test, such as including more items. Moreover, it can become a useful reference for developing other superitem tests that assess different constructs as well.

The framework of the study is expected to be able to contribute meaningfully and significantly to the development of assessment literacy among teachers, both preservice and inservice teachers. Thus, the application of this superitem test is not only limited to preservice teachers but can also be appropriately adopted or adapted to evaluate inservice teachers' assessment literacy in different countries.

ACKNOWLEDGEMENT

This paper was made possible with funding from the Short-Term Grant of University Sains Malaysia, Penang, Malaysia.

REFERENCES

Alkharusi, H. (2011). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia - Social and Behavioral Sciences*, *29*, 1614 – 1624.

Alias, I. (1993).*Tanggapan terhadap penilaian formatif, pembinaan item penilaian dan penggunaanya oleh guru sains dan matematik* [*Perception towards formative assessment, construction of assessment items and its usage by science and mathematics te*achers]. Unpublished master dissertation, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Asri, S. (2007).*Amalan pentaksiran pengajaran dan pembelajaran di sekolah-sekolah menengah Bestari Negeri Johor Darul Takzim* [*Assessment practice of teaching and learning at Bestari high schools of Johor Darul Takzim*].Unpublished doctoral thesis, Universiti Teknologi Malaysia, Skudai, Malaysia.

Alquraan, M., Alshraideh, M., & Bsharah, M. (2010). Psychometric properties and differential item functioning (DIF) analyses of Jordanian version of Self-assessed Wisdom Scale (SAWS-Jo).*International Journal of Applied Educational Studies, 9*(1)*,* 52-66.

American Federation of Teachers, National Council on Measurement in Education, National Education Association. (1990). *The standards for competence in the educational assessment of students*. Retrieved July 22, 2003, from http://www.unl.edu/buros/article3.html

Bagraith, K. S., Strong, J., Meredith, P. J., & McPhail, S. M. (2017). Rasch analysis supported the construct validity of self-report measures of activity and participation derived from patient ratings of the ICFlow back paincore set. *Journal of Clinical Epidemiology*, *84*, 161-172.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101-118.

Beziat, T. L. R., & Coleman, B. K. (2015). Classroom assessment literacy: Evaluating pre-service teachers. *The Researcher*, *27*(1), 25-30.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* New York: Academic Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd. Ed)*. New York: Taylor & Francis Group.

Boone, W. J,, Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human science.* New York: Springer.

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: A step by step guide to using assessment results to improve teaching and learning*. Cambridge: Harvard Education Press.

Campbell, C., Murphy, J.A., & Holt, J. K. (2002). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at the Annual Meeting of the Mid-Western Educational Research Association. October, 2006. Columbus, OH.

Carpita, M., & Golia, S. (2012). Measuring the quality of work: The case of the Italian social cooperatives. *Quality & Quantity*, *46*, 1659–1685.

Chang, S. F. (1988). *Teachers' assessment practices: Assessing phase ii pupils' progress in KBSR English.* Unpublished master's thesis, Universiti Malaya, Petaling Jaya.

Chang, H. L., & Wu, S. C. (2008). A multi-facet analysis on rating the academic scientific papers. *Psychological Testing, 55*(1), 105-128.

Chang, Y. C., Ailey, S. H., Heller, T,, & Chen, M. D. (2013). Rasch analysis of the Mental Health Recovery Measure. *American Journal of Occupational Therapy, 67*(4), 469-477.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17* (4), 419-438.

Duncan, P. W., Bode, R., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specify outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84* (7), 953.

Fisher, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Garzón Umerenkova, A., de la Fuente Arias, J., Martínez-Vicente, J.M., Zapata Sevillano, L., Pichardo, M. C., & García-Berbén, A. B. (2017). Validation of the Spanish short self-regulation questionnaire (SSSRQ) through rasch analysis. *Front. Psychol*, *8*, 276.

Green, K. E., & Frantom, C. G. (2002). S*urvey development and validation with the Rasch model.* A paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, November 14-17, 1-42.

Hecimovich, M., & Marais, I. (2017). Examining the psychometric properties of a sport-related concussion survey: A Rasch measurement approach. *PMC Journal*, *10*, 1-11.

Khadijeh, B., & Amir, R. (2015). Importance of teachers' assessment literacy. *International Journal of English Language Education*, 139-145.

Kahl, S. R., Hofman, P., & Bryant, S. (2012). *Assessment literacy standards and performance measures for teacher candidates and practicing teachers***.** Retrieved on August 30, 2017, from caepnet.org/~/media/Files/caep/standards/assessment-literacy-in-teacher-preparati.pdf

Kanjee, A., & Mthembu, J. (2015). Assessment literacy of foundation phase teachers: An exploratory study. *South African Journal of Childhood Education 5*(1), 142-168.

Kim, K. (2014). Developing preservice teachers' assessment literacy: A problem-based learning approach. In P. Preciado Babb (Ed.), *Proceedings of the IDEAS: Rising to Challenge Conference* (pp. 113-120). Calgary, Canada: Werklund School Education, University of Calgary.

Lee, I. (2017). *Classroom assessment literacy for l2 writing teachers: In classroom writing assessment and feedback in l2 school contexts.* Singapore: Springer.

Lerdal, A., Johnson, S., Kottorp, A., & Koch, L. V. (2010). Psychometric properties of the fatigue severity scale: Rasch analyses of responses in a Norwegian and a Swedish MS cohort. *Multiple Sclerosis*, *16*(6), 733-741.

Linacre, J. M. (2012). *Winsteps (Version 3.74)* [Software]. Available from http://www.winsteps.com/index.html

Mallinckrodt, B., Miles, J. R., & Recabarren, D. A. (2016). Using focus groups and Rasch Item Response Theory to improve instrument development. *The Counselling Psychologist*, *44*(2), 146–194.

Mappiasse, S. (2006). Developing and validating instruments for measuring democratic climate of the civic education classroom and students' engagement in North Sulawesi, Indonesia. *International Education Journal, 7*(4), 580-597.

Masran, S. H., Rahim, M. B., Faizal A. N. Y., & Marian, M. F. (2017). Validity and reliability of an e-portfolio indicators instrument for Malaysian skills certification (MSC). *Pertanika Journal Social Science and Humanities,* 25 (S), 47 – 56.

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*(2), 285-296.

Mertler, C. A. (2003). *Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference*? Paper presented at the Annual Meeting of the Mid-Western Educational Research Association. Oct 15-18, 2003. Columbus, OH.

Mertler, C. A. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts:Development of the assessment literacy superitem test.* Paper presented at the Annual Meeting of the American Educational Research Association. Apr 11-15, 2005. Montreal, Quebec, Canada.

Mohamad, A. M. A. (2006). *Amalan pentaksiran di sekolah menengah* [*Assessment practice in secondary schools*].Unpublished doctoral thesis, Universiti Malaya, Petaling Jaya.

Muhamad, S. N. (2001). Pengujian selaku pemangkin perubahan pendidikan: Satu peluang atau retorik [Testing as a catalyst for educational change: An opportunity or rhetoric]? *Jurnal PengurusanPendidikan Institut Aminuddin Baki, 11*(2), 25-36.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of student*. Harlow: Pearson Education Limited.

Norazlina, H. T. (2014). *Literasi pentaksiran dalam kalangan calon guru teknikal di IPTA zon selatan* [*Assessment literacy among pre-service technical teachers at south zone of public higher institution*]. Retrieved on August 30, 2017, fromhttp://eprints.uthm.edu.my/5353/

Norazilawati, A., Noorzeliana, I., Mohd Sahandri Gani, H., & Saniah, S. (2015). Planning and implementation of school-based assessment (SBA) among teachers. *Procedia - Social and Behavioral Sciences 211*, 247-254.

Perry, M. L. (2013). *Teacher and principal assessment literacy*. Retrieved on August 14, 2017, from http://scholarworks.umt.edu/etd/1391/

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher*, *6*(1), 21-27.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12.

Rasidayanty, S. (2014). *Persepsi guru-guru kemahiran hidup bersepadu terhadap literasi pentaksiran dalam pelaksanaan pentaksiran berasaskan sekolah [Perceptions of integrated life skills teachers on assessment literacy in the implementation of school-*

based assessments]. Retrieved on August 14, 2015, from http://eprints.uthm.edu.my/5360/1/RASIDAYANTY_BINTI_SAIO N.pdf

Rohaya, T., & Mohd Najid, A. G. (2008). *Pembinaan dan pengesahan instrument bagi mengukur tahap literasi pentaksiran guru sekolah menengah di Malaysia [Construction and validation of instrument to measure the level of assessment literacy of secondary school teachers in Malaysia]*. Retrieved on May 10, 2016, from http://eprints.utm.my/7906/1/EDUPRES_%28F2%29_10.pdf

Rohaya, T. (2014). *Assessment Literacy: A catalyst to the success of school-based assessment in Malaysian schools*. Proceedings of the International Conference on Science, Technology and Social Sciences (ICSTSS) 2012, pp. 197-202.

Salmiah, J., Ramlah, H. A. R. B., & Abdullah, M. R. (2013). Acceptance towards school based assessment among agricultural integrated living skills teachers: Challenges in implementing a holistic assessment. *Journal of Technical Education and Training, 5* (1), 44-50.

Sartori, R., & Pasini, M. (2006). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity*, *41,* 359–374.

Sewornoo, S. (2016). ***A****ssessment literacy of mathematics teachers and challenges in the implementation of the school-based assessment in senior high schools of Ghana*. Retrieved on May 10, 2017, from http://ir.uew.edu.gh:8080/xmlui/handle/123456789/636

Smith, E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205–231.

Stiggins, R. J. (1995). Assessment literacy for the 21st century.*Phi Delta Kappan, 77*(3), 238-245.

Suah, S. L. (2012). *Analisis model literasi dan amalan pentaksiran guru sekolah serta kajiantentang jurang antara keduanya* [*Analysis of literacy model and assessment practices of school teacher and study of the gap between them*]. Unpublished doctoral thesis, Universiti Sains Malaysia, Penang, Malaysia.

Teh, J., & Lim, H. L. (2016). Examining psychometric properties of malay version children depression inventory (CDI) and prevalence of depression among secondary school students. *Pertanika Journal Social Science and Humanities, 24* (4), 1349 – 1379.

Tennant, A., & Conghan, P. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research, 5* (8), 1358–1362.

Title, C.K. (1994). Toward an educational psychology of assessment for teaching and learning: Theories, contexts and validation arguments. *Educational Psychologist*, *29*, 149-162.

Vahid, N., & Nasreen, B. (2019). A review of literature on language assessment literacy in last two decades (1999-2018). *International Journal of Innovation, Creativity and Change. 8* (11), 44-59.

Waugh, R. F. (2002). Creating a scale to measure motivation to achieve academically: Linking attitudes and behaviours using Rasch measurement. *British Journal of Educational Psychology, 72*(1), 65-86.

Webb, N. L. (2002). *Assessment Literacy in a standards-based urban education setting*. Retrieved on May 5, 2017, from http://facstaff.wcer.wisc.edu/normw/AERA%202002/Assessment%2 0litera cy%20NLW%20Final%2032602.pdf

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Retrieved on May 5, 2016, from http://works.bepress.com/geoff_masters/102/

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

Wu, P.C., & Chang, L. (2008). Psychometric properties of the chinese version of the beck depression Inventory-II using Rasch model. *Measurement and Evaluation in Counselling and Development, 41*, 13-31.

Yamtim, V., & Wongwanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia - Social and Behavioral Sciences*, *116*, 2998 – 3004.