

Psychometric properties of teacher classroom assessment literacy instrument using Rasch model analysis

Nor Hafizi Mohd Khalid¹, Ibtnatul Jalilah Yusof¹, Adibah Abdul Latif^{1,2}, Md Daud Md Jani³

¹School of Education, Faculty of Social Science & Humanities, Universiti Teknologi Malaysia, Skudai, Malaysia

²Centre of Research for Fiqh Science and Technology (CFIRST), Ibnu Sina Institute for Scientific and Industrial Research (ISI-SIR), Universiti Teknologi Malaysia, Skudai, Malaysia

³Department of Education, Institut Pendidikan Guru Temenggong Ibrahim, Johor Bahru, Malaysia

Article Info

Article history:

Received Jan 1, 2022

Revised Dec 24, 2022

Accepted Jan 23, 2023

Keywords:

Assessment literacy

Classroom assessment

Rasch model

Reliability

Validity

ABSTRACT

A teacher classroom assessment literacy (TeCAL) instrument was developed to measure the level of teacher classroom assessment literacy in schools. TeCAL contains 66 multiple choice items with four options based on four constructs namely purpose, measurement, evaluation and use. Thus, this study aims to identify the psychometric properties of TeCAL using Rasch measurement model (RMM) analysis through Winstep software version 3.72.3. The findings show that the compatibility values of mean square (MNSQ) infit and outfit items ranged from 0.64 to 1.46 and 0.40 to 2.23, respectively. The value of MNSQ outfit was outside the set range, but still met the other fit statistics indicator which has a positive point measure correlation (PTMEA) value. In addition, the findings show that the empirical raw variance explained by measures is 38.2%. It was very close to the modeled value of 38.4% with the empirical unexplained variance in 1st contrast being 7.5% less than the maximum controlled 15%. Largest standardized residual correlations identified 10 pairs of dependent items to be less than 0.7. The person and item reliability index values were 0.94 and 0.89 with separation index values of 2.90 and 3.80, respectively. Overall, this psychometric analysis is crucial to ensure that the TeCAL instrument has good quality and meaningful to use.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nor Hafizi Mohd Khalid

School of Education, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia

81310 Skudai, Johor, Malaysia

Email: norhafizie@gmail.com

1. INTRODUCTION

Assessment literacy is defined as teachers' knowledge, abilities and understanding of the concepts, and implementation of basic procedures set out in classroom assessments [1]. It is important because its effectiveness can influence educational decisions [2], such as improving the quality of instructions that drive expected achievement by testing, analyzing, interpreting and using student performance data aimed at identifying learning needs. Thus, a conclusion can be made that assessment literacy is one of the most important connections between assessment quality, instruction, and student performance [3]–[6].

A series of studies to measure teacher assessment literacy have been conducted since 1990. These include the developments of several previous literacy assessment instruments such as teacher assessment literacy questionnaire (TLAQ) [7], classroom assessment literacy inventory (CALI) [8], assessment literacy inventory (ALI) [9], and approaches to classroom assessment inventory (ACAI) [10]. Most of the researchers referred to the standards for teacher competency of educational assessment of students (STCEAS) which was

issued by The National Council on Measurement in Education (NCME) and several world bodies in 1990 and only a few referred to classroom assessment standard (CAS) published in 2015. However, Brookhart [11] stated that STECAS does not consider the current concept of formative assessment and the social problems faced by teachers in the context of constructing and using assessment in standard-based education reforms. Gotch and French [2] support this by pointing out that the STECAS measures no longer satisfy the competency, ability, and expectation requirements for measuring teacher assessment literacy, and that the psychometric evidence used to support these claims is insufficient.

Thus, the rationale of this study is to bridge this gap by focusing on classroom assessment literacy based on the more recent and comprehensive classroom assessment standard (CAS) [12]. It is the process of making judgments or decisions in providing quality values for the learning process of students in the classroom. This includes the ability of teachers to know and clearly understand the purpose, measurement, evaluation and use of assessment to report the level of student achievement on a continuous basis in a particular subject or period of time in order to take follow-up actions to enhance student performance. This is in line with the guidelines for the implementation of effective classroom assessment that have been set by the Malaysian Ministry of Education (MoE) starting with: i) Instructional planning and assessment methods; ii) Performing instructional and assessment; iii) Recording and analyzing student mastery; and iv) Reporting student mastery levels [13].

According to McMillan [14], the first step in an assessment is purpose which explains specific objective to collect particular data either before, during, or after the lesson. This step is necessary to set a clear vision of what needs to be achieved by the implemented assessment [6]. This is aligned with the first step proposed by curriculum development division, MoE [13], while implementing classroom assessments such as instructional planning and identifying assessment methods to be used. When doing instructional planning, teachers need to assess and clearly understand the content of each topic in the curriculum and assessment standards document (CASD) before determining the objectives to be achieved by students and to be interpreted by teachers. Teachers also need to identify assessment methods that are appropriate to assess students' abilities [13]. This is important to ensure that the design, administration and data generated by the assessment activities can be used appropriately [14]. Researchers found that at this stage, teachers need to build instruments as a tool to measure and it depends on the instructional activities that have been planned and assessment methods that have been chosen either based on the components of assessment for learning (AfL), assessment as learning (AaL) or assessment of learning (AoL). This is as stated in the classroom assessment standard (CAS) [12] document, where the type and method of assessment used should enable students to demonstrate their learning. In addition, stakeholders should also be informed of the purpose for which the classroom assessment is conducted and its uses.

Measurement is the systematic process of setting numbers for behaviors and performance or differentiated behaviors using a variety of techniques [14]. It is also used to define or describe how many traits, attributes and characteristics each individual has achieved and needs to improve. This step is aligned with the second step proposed by curriculum development division, MoE [13], while implementing classroom assessment; which involve performing instructional and assessment tasks. The assessment methods in instructional tasks should focus on aspects to be assessed and can be implemented with a variety of designs and approaches such as oral, observation, written or a combination of all these approaches. Teachers also need to improve the assessment framework by emphasizing the characteristics of higher order thinking skills (HOTS) items which assess the skills of analyzing, evaluating, and creating, in addition to involving students directly. This is aligned with CAS which recommends that assessment process should involve students in a meaningful way and evidence of assessments can be used to enhance their learning. Moreover, the teachers should emphasize the features of validity and reliability and should not be influenced by factors unrelated to the intended purpose of assessment [12]. The information obtained from this step can provide a brief overview of the development and progress of students in the learning process.

Evaluation is the procedure of putting a certain value on specific numbers and observations primarily based on a particular reference framework [14]. It involves the process of judging the quality and interpreting the process of measuring the extent of a student's behavioral performance in learning. This step is aligned with the third step proposed by curriculum development division, MoE [13], while implementing classroom assessment i.e. recording and analyzing students' achievement. This recording activity can store information systematically related to the development, ability, progress and achievement of students before, during and after the learning process. According to McMillan [14], this evaluation does not only emphasize quality, error-free and accuracy, but also relates to value. Therefore, the evaluation process conducted by the teacher should go through professional judgment to decide the extent of achievement of students by primarily referring to a set of criteria in the learning process as to whether they have achieved the minimum level of achievement or not. This information needs to be updated by teachers, stored, and maintained properly either in teaching records, teacher notebooks, checklists, and reporting templates to facilitate references on student learning before being analyzed and assessed for following-up and reporting purposes.

Usage is the last stage of implementing classroom assessment which shows how evaluation is applied. The use of scores and other data is highly relevant to the decisions that teachers need to enhance their instructions to assess their students and as a report for parents [14]. This step is parallel to the fourth step proposed by the curriculum development division, MoE [13], while implementing classroom assessment i.e. reporting the level of mastery of students. This reporting involves the process of communicating assessment information about the development and progress of student learning achievement to stakeholders either orally or through writing. It covers the level of mastery, interest, attitude and behavior of students, as well as suggestions for follow-up actions.

In addition, the effectiveness of classroom assessment implementation requires continuous monitoring and reviewing, or reflection. This is proposed in the CAS which mentioned that the implementation of the classroom assessment should be monitored and reviewed in order to improve the overall quality [12]. This is also stated by DeLuca *et al.* [15] whereby these steps are necessary to support implementation and provide opportunities for teachers to improve teacher assessment literacy competency. It also coincides with the statement of the classroom assessment implementation guide document [13], which proposes four approaches to ensure quality in implementing classroom assessment, namely mentoring, coordination, monitoring and detection. Researchers see these steps as necessary to ensure that the implementation of classroom assessment by teachers is aligned with the guidelines.

Psychometric properties of the instrument is necessary to ensure that its measurements are accurate and repeatable [16]. Therefore, there are several related psychometric theories that predict the results of psychological tests. Among them is the item response theory (IRT) which introduced a new approach to analyzing psychological test items [17]. IRT gives response focus to items in a test compared to classical test theory (CTT) which uses test level analysis [18], [19]. There are several statistical models introduced in IRT and they are generally divided into two main branches namely unidimensional and multidimensional. Unidimensional models are for measuring one dimensional trait, while multidimensional models measure multiple dimensions of traits [18]. The complexity that exists in multidimensional models causes most researchers to use unidimensional models, namely the dichotomous and polytomous data response model [20], [21]. There are three models for dichotomous data which are the simplest one-parameter logistics model (1-PL) [19], the two-parameter (2-PL), and the three-parameter (3-PL) models [20].

Rasch measurement model (RMM) refers to an idea, principle, guideline or technique that allows a measurement to be made for a latent trait [21]. RMM is a model of parameters, individual abilities tested, and item difficulty levels, as well as being the simplest response model; the one-parameter logistics model (1-PL) in IRT [22]. This model uses a mathematical formula where the probability of an individual answering or supporting an item correctly depends on the individual's ability or the difficulty of the item [22]. The RMM assumption emphasizes the same discrimination index, where individuals with low abilities do not guess the answers of items in order to answer it correctly [23]. Thus, there is only one parameter measured in the RMM, which is the level of difficulty of the item with the probability of success depending on the differences in individual abilities and the difficulty of an item [23]. RMM uses a combination of algorithms in the form of mathematical equations that express the expected probability of an item as i and the individual's ability as n [24]. Bond and Fox [22] stated the mathematical formula as in (1):

$$P_{ni} \left(x_{ni} = \frac{1}{B}, D \right) = \frac{e^{(Bn - Di)}}{1 + e^{(Bn - Di)}} \quad (1)$$

With

$P_{ni} \left(x_{ni} = \frac{1}{B}, D \right)$ = the probability of an individual n on item i giving the correct response ($x=1$)

Bn = individual ability

Di = item difficulty level

$Bn - Di$ = the probability of the possibility of a success

2. RESEARCH METHOD

2.1. Research design

Cross-sectional survey methods and fully quantitative data analysis were used in the study. The use of surveys has comprehensive features, is a fast way of handling or collecting data, can be used with a large sample, enables information to be obtained directly from the respondents, and enables the findings to be generalized [25]. Meanwhile, cross-sectional studies can estimate the prevalence of outcomes of interest or provide an overview of the population at a point or within a short time (cohort studies) [26]. This feature is seen to be appropriate with the study conducted by the researcher.

2.2. Study sample

The study was administered to 103 respondents consisting of primary school teachers throughout Malaysia. The disproportional stratified and multi stage cluster sampling technique was used to ensure that every respondent within the population had an equal probability of being as a respondent [25]. Respondents were selected based on demographic factors such as gender, position, subjects taught, teaching experience, and training attended.

2.3. Study instrument

The instrument is a part of the measurement tool used to collect data of the study [27]. In this study, only one instrument was used which is a test known as teacher classroom assessment literacy (TeCAL). The instrument is self-developed which is constructed according to the steps of the instrument construction design model [28].

2.4. Research procedure

The research was conducted by distributing the survey through an online platform. The respondents were given one hour to complete the test individually or simultaneously in groups. They must also adhere to the test procedures.

3. RESULTS AND DISCUSSION

In psychological testing, content specifications are usually less explicit but the dimensions of response required by the researcher are broad and clear [29]. Thus, RMM has been chosen to analyze the psychometric characteristics possessed by TeCAL. This is aligned with the statement by Bond and Fox [22] where RMM can be used as an approach in developing an instrument. Researchers need to ensure that the two assumptions of RMM are measured first, which are unidimensionality and local item dependency [30]. The conducted analyses were the fit analysis of item and person - mean square (MNSQ) infit-outfit, Z-Std standardized fit statistics (ZSTD); polarity item - point measure correlation (PTMEA); unidimensional - residual partial component analysis (PCA); standardized residual correlations; reliability values; item and person separation index.

3.1. MNSQ infit, MNSQ outfit, and ZSTD values

The first evaluation procedure that needs to be performed is to identify each incompatible item (misfit) through MNSQ analysis or standardized fit statistics (ZSTD) [31]. This is to ensure that the data or responses match or comply with the RMM [32], [33]. Therefore, this analysis is performed before other analyses, and the findings are as shown in Table 1.

Table 1. Fit indices

Indices	MNSQ		Point measure
	Infit	Outfit	Correlation
Range	0.64 to 1.46	0.40 to 2.23	0.12 to 0.66
Mean	1.02	0.97	-

According to Linacre [34], this MNSQ value must be in the range of 0.5 to 1.5 to give productive implications for measurement. If the MNSQ value exceeds the range, it indicates a homogeneous item, while if the value is less than the range, it indicates that the domain overlaps with other items in one measurement scale. On the other hand, Bond and Fox [22] considered the ZSTD values in the range between -2.0 and 2.0 as acceptable for sample sizes of 30 to 300. Overall range for the infit MNSQ in this study is still within the acceptable logit values. Nevertheless, the overall range for the MNSQ outfit is slightly off the range as suggested by Linacre [34] which involves four unfit items as tabulated in Table 2.

Table 2. Misfit item–outfit MNSQ

Item	Outfit		Point measure
	MNSQ	ZSTD	Correlation
T64	2.23	1.7	0.35
T52	1.58	1.9	0.12
G34	0.45	-2.1	0.54
N28	0.40	-2.2	0.54

Items T64 and T52 have values of 2.23 and 1.58, respectively, above the recommended outfit MNSQ value of 1.50 but still meet the ZSTD range. Meanwhile, for items G34 and N28, the value of Outfit MNSQ is less than the proposed value of 0.5 and the value of ZSTD exceeds the value of 2.0 set. However, an MNSQ misfit value below 0.5 is considered less threatening than a larger MNSQ misfit value [34], [35]. In addition, these four items still meet the other fit statistics indicator that has a positive point measure correlation (PTMEA). This indicates that the items are functioning in the same direction to the measured and expected domains. Thus, the positive value of point measure correlation explains that items can measure what should be measured and should be retained [21].

3.2. Residual partial component analysis (PCA)

PCA is used to ensure the uniformity of the dimensions of an instrument [21]. This feature of unidimensional can determine a one direction measure of instrument [21]. Findings from Table 3 show that the value of empirical raw variance explained by the measures is 38.2%, very close to the modeled value of 38.4%. In addition, the empirical value of unexplained variance in 1st contrast has a value of 7.5% less than the maximum controlled 15% and is acceptable [21]. However, the Eigen value of unexplained variance in 1st contrast is 5.3 which is more than 5 as suggested by Linacre [36], but this value pertains to the entire instrument representing the four constructs used. This value indicates contrasts between opposing factors but not loadings on one factor [36] and can be scaled down if analyzed separately by construct.

Table 3. Residual partial component analysis

Standard residual variance	Eigenvalue	Empirical	Modelled
Total raw variance in observations	71.2	100.0%	100.0%
Raw variance explained by measures	27.2	38.2%	38.4%
Raw variance explained by persons	10.5	14.7%	14.8%
Raw variance explained by items	16.7	23.5%	23.6%
Raw unexplained variance (total)	44.0	61.8%	100.0%
Unexplained variance in 1st contrast	5.3	7.5%	12.1%
Unexplained variance in 2nd contrast	4.1	5.8%	9.4%
Unexplained variance in 3rd contrast	3.1	4.3%	6.9%
Unexplained variance in 4th contrast	2.6	3.7%	6.0%
Unexplained variance in 5th contrast	2.3	3.2%	5.2%

3.3. Standardized residual correlations

Table 4 highlights the largest standard residual correlations used to identify the multicollinearity of the two items in the test. A pair of listed items is in the range of -0.54 to 0.56, with standard balance correlation values. All items have point measure correlation value of not more than 0.7, which indicates that the respondents saw this pair of related items as a matter that can be distinguished and not confusing [21] or known as having local independence [37]. This shows that the items in TeCAL have less effect of item noise on the measurements carried out.

Table 4. Largest standardized residual correlations

Correlation	Paired item	
	Number of items	Number of items
0.56	U20	N28
0.55	G15	N28
0.49	N28	G34
0.47	U57	G63
0.47	T16	N48
0.46	G15	G34
-0.54	N48	G63
-0.50	N28	G31
-0.49	N28	U54
-0.48	N28	G50

3.4. Reliability values and separation index

The person and item reliability values for TeCAL are 0.94 and 0.89, respectively. The reliability value of person fulfils the requirements set by Linacre [36] which is 0.8, with an interpretation of a strong accepted reliability value, while the reliability value of the item fulfills the requirements set by Azrilah *et al.* [21] which is 0.78 with an interpretation of the adequacy of the item which measures what needs to be

measured. The separation index values for persons and items were 2.90 and 3.80, respectively, exceeding 2.0 as suggested by Bond and Fox [22] showing a good and acceptable index. These obtained reliability values and separation index indicate the ability of TeCAL to provide different hierarchies along the measured variables. Table 5 presented the summary statistics of the field data.

Table 5. Summary statistics of field data

Measure	Total	Max Logit	Min Logit	Separation	Reliability
Person	48	6.01	-2.27	2.90	0.94
Item	103	6.43	-6.47	3.80	0.89

3.5. Abortion of items

Findings revealed that of the 66 TeCAL items analyzed, four were eliminated and 20 were modified. There were four items that dropped: N46 and N29 from the evaluation construct and G51 and G49 from the use construct. Meanwhile, the 20 modified items were T64, T37, and T52 from the purpose construct; U56, U21, U42, U23, and U41 from the measurement construct; N45, N61, N47, N44, N14, N26, N60, and N28 from the assessment construct; G32, G35, G6, and G63 from the usage construct. The items were dropped or modified based on the findings of the analysis because they did not meet the MNSQ, ZSTD, or PTMEA correlation values. Table 6 shows a summary of the number of items dropped, modified and retained in the constructs.

Table 6. A summary of the number of items by construct

Construct	Number of original items	Number of items dropped	Number of items modified	Number of items retained
Purpose	9	-	3	9
Measurement	21	-	5	21
Evaluation	21	2	8	19
Use	15	2	4	13
Total	66	4	20	62

3.6. Person-item map

Mapping is a diagram that characterizes an individual's ability and item difficulty as a location on a latent variable (model/linear line) by placing individuals or items with high and low ability or difficulty with positive and negative values on the logits scale or parameter [19], which have uniform intervals, respectively [21]. The position of the person or item at the top of the scale or parameter logits indicates a person with high ability or an item with a high level of difficulty, while the position of the person or item at the bottom of the scale or parameter logits indicates a person with low ability or item with low difficulty level [22].

Based on Table 5 and Figure 1, the distribution of respondents' responses is from a maximum logit of 6.01 to a minimum logit of -2.27, which is equivalent to 8.28, while the item response spread is from a maximum logit of 6.43 to a minimum logit of -6.47 which is equivalent to 12.90. Figure 1 also shows that items N43 and U39 are the items estimated with the highest level of difficulty (maximum measure) and items T1 and T36 are the items estimated with the lowest level of difficulty (minimum measure) in TeCAL. There are three people estimated to have the highest ability which are persons with entry numbers 151, 154, and 155, and two people estimated to have the lowest ability which are persons with the entry numbers 37 and 52. Overall, TeCAL has item representation with various levels of difficulty such as easy, medium and difficult with persons and items involved scattered along the parameter logit depending on their respective difficulty abilities and levels.

In addition, the item map can identify the gaps between students' ability and item difficulty in the study. A noticeable gap should be calculated to define whether the gap is acceptable or not [21]. Figure 1 shows that there were two item gaps between items N43, U49 and item U38, and items T1, T36 and item U22 with difference of 3.32 and 3.46 logit, respectively. These gaps are marked with blue upward-downward arrow. A noticeable person gap identified was marked with a red upward-downward arrow that is between persons with entry numbers 151, 154, 155 and 101,103 with a difference of 3.49 logit. Furthermore, there are four items; G65, U22, T1 and T36 with no representation of respondents in the study conducted marked with green upward-downward arrow.

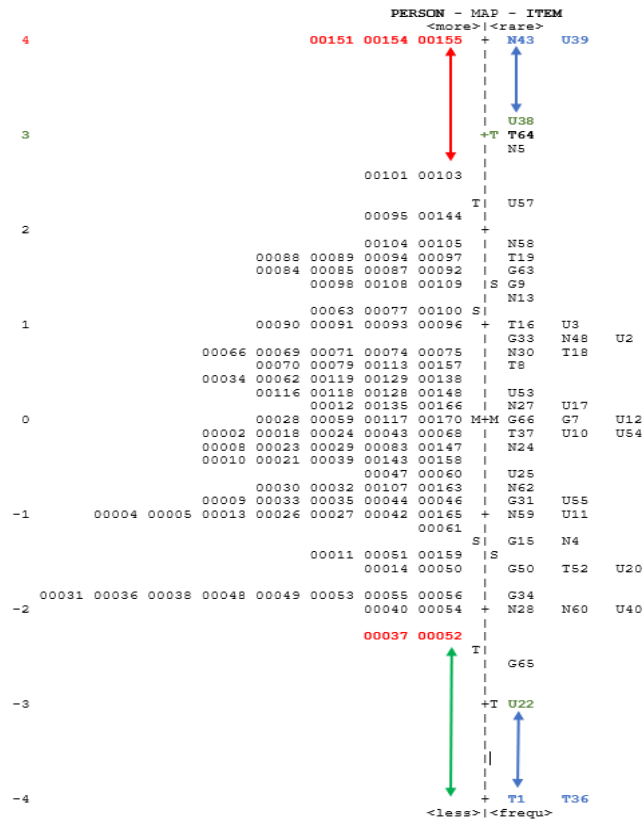


Figure 1. Item-person representativeness

4. CONCLUSION

A newly developed instrument namely TeCAL needs to be emphasized on its psychometric characteristics to ensure that it has a good level of validity and reliability to guarantee its quality. Thus, RMM analysis was conducted on 66 items representing four main constructs, namely purpose, measurement, evaluation, and use. Findings showed that 48 of the original 66 items met the required psychometric characteristics, while 20 items were updated and four items were deleted. This updated TeCAL instrument should be used by researchers to identify teachers’ strengths in assessment literacy while implementing classroom assessment in schools. Stakeholders under the MoE can use these provided items as an option to measure the literacy level of teacher classroom assessment in schools based on the high validity and reliability of TeCAL and tested using RMM.

Indirectly, this study can contribute to the development of teacher classroom assessment through the use of items from the four constructs of assessment literacy in a new context that refers to CAS and is based on the local situation in Malaysia. TeCAL is expected to reduce the dependence of measuring instruments from abroad, which are less suited to the local educational culture and frequently refer to the STECAS, which are less suited to current learning and assessment practices in Malaysia. The process of validation of these items is very important to facilitate stakeholders to identify the literacy level of teacher classroom assessment more accurately. Thus, teacher development programs can be planned more effectively through the analysis of the four constructs of teacher classroom assessment literacy in TeCAL. Teachers themselves as implementers can take more effective initiatives while implementing classroom assessment in schools. In addition, several initiatives using different statistical approaches such as structural equation modelling (SEM) and Rasch multidimensionality analysis can be conducted to add psychometric empirical evidence of items representing constructs in TeCAL.




REFERENCES

[1] H. Xu, “Exploring novice EFL Teachers’ classroom assessment literacy development: A three-year longitudinal study,” *Asia-Pacific Education Researcher*, vol. 26, no. 3–4, pp. 219–226, 2017, doi: 10.1007/s40299-017-0342-5.
 [2] C. M. Gotch and B. F. French, “A systematic review of assessment literacy measures,” *Educational Measurement: Issues and Practice*, vol. 33, no. 2, pp. 14–18, 2014, doi: 10.1111/emip.12030.




- [3] H. Ashraf and S. Zolfaghari, "EFL teachers' assessment literacy and their reflective teaching," *International Journal of Instruction*, vol. 11, no. 1, pp. 425–436, 2018, doi: 10.12973/iji.2018.11129a.
- [4] M. Mellati and M. Khademi, "Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development," *Australian Journal of Teacher Education*, vol. 43, no. 6, pp. 1–18, 2018, doi: 10.14221/ajte.2018v43n6.1.
- [5] F. Zolfaghari and A. Ahmadi, "Assessment literacy components across subject matters," *Cogent Education*, vol. 3, no. 1, pp. 1–16, 2016, doi: 10.1080/2331186X.2016.1252561.
- [6] M. Nurfirdawati, M. A. Normazita, Z. Surianorbaya, Y. Noor Azlin, and I. Nur Nadiyah, "Examining assessment literacy: A study of technical teacher," *European Journal of Molecular & Clinical Medicine*, vol. 07, no. 08, pp. 705–717, 2020.
- [7] B. S. Plake, J. C. Impara, and J. J. Fager, "Assessment competencies of teachers: A national survey," *Educational Measurement: Issues and Practice*, vol. 12, no. 4, pp. 10–12, 1993, doi: 10.1111/j.1745-3992.1993.tb00548.x.
- [8] C. A. Mertler, "Preservice versus inservices teachers' assessment literacy: Does classroom experience make a difference?" in *Annual meeting of the Mid-Western Educational Research Association*, Columbus, OH, 2003, pp. 1–28.
- [9] C. A. Mertler and C. Campbell, "Measuring Teachers' Knowledge & Application of Classroom Assessment Concepts: Development of the 'Assessment Literacy Inventory'," Paper presented at the *Annual Meeting of the American Educational Research Association*, 2005, pp. 1–17.
- [10] C. DeLuca, D. LaPointe-McEwan, and U. Luhanga, "Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy," *Educational Assessment*, vol. 21, no. 4, pp. 248–266, 2016, doi: 10.1080/10627197.2016.1236677.
- [11] S. M. Brookhart, "Educational assessment knowledge and skills for teachers," *Educational Measurement: Issues and Practice*, vol. 30, no. 1, pp. 3–12, 2011, doi: 10.1111/j.1745-3992.2010.00195.x.
- [12] D. Klinger, P. J. McDivitt, B. B. Howard, M. A. Munoz, W. T. Roger, and E. C. Wylie, *The Classroom Assessment Standards for PreK-12 Teachers*. Kindle Dir, 2015.
- [13] Curriculum Development Division, *Panduan Pelaksanaan Pentaksiran Bilik Darjah*, 2nd ed. Putrajaya: Malaysia of Education Ministry, 2019.
- [14] J. H. McMillan, *Classroom assessment: Principles and practice for effective standards-based instruction*, 5th ed. Boston, MA.: Allyn & Bacon, 2011.
- [15] C. DeLuca, D. LaPointe-McEwan, and U. Luhanga, "Teacher assessment literacy: a review of international standards and measures," *Educational Assessment, Evaluation and Accountability*, vol. 28, no. 3, pp. 251–272, 2015, doi: 10.1007/s11092-015-9233-6.
- [16] K. Coaley, *An introduction to psychological assessment and psychometric*. London: SAGE Publications Ltd, 2010.
- [17] P. W. Holland and M. Hoskens, "Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test," *Psychometrika*, vol. 68, no. 1, pp. 123–149, 2003, doi: 10.1007/BF02296657.
- [18] M. L. Nering and R. Ostini, *Handbook of polytomous item response theory models*. New York: Routledge. Taylor & Francis Group, 2010. doi: 10.16309/j.cnki.issn.1007-1776.2003.03.004.
- [19] R. K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing, 1985.
- [20] F. M. Lord, "Small justifies the Rasch Model," in D. J. Weiss, Ed., *New Horizons in Testing. Latent Trait Test Theory and Computerized Adaptive Testing*, New York: Academic Press, 1983, pp. 51–61.
- [21] A. A. Azrilah, M. Mohd Saidudin, and Z. Azami, *Asas pengukuran Rasch. Pembentukan skala dan struktur pengukuran*. Bangi, Selangor: Penerbit UKM, 2017.
- [22] T. G. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*, 3rd ed. New York, London: Routledge Taylor & Francis Group, 2015. doi: 10.1017/CBO9781107415324.004.
- [23] M. E. E. M. Matore and A. Z. Khairani, "Pengujian Ciri Psikometrik Item USMEQ-I Dalam Kalangan Pelajar Politeknik Menggunakan Model Rasch," *Jurnal Teknologi*, vol. 75, no. 1, pp. 251–257, 2015, doi: 10.11113/jt.v75.3901.
- [24] M. M. Mohd Effendi, "Pembinaan Instrumen Kecerdasan Menghadapi Cabaran (IKBAR) bagi pelajar politeknik menggunakan model Rasch," PhD thesis, Universiti Sains Malaysia, 2015.
- [25] Y. P. Chua, *Kaedah dan statistik penyelidikan: Kaedah penyelidikan*. Shah Alam: McGraw Hill Education, 2011.
- [26] K. A. Levin, "Study design III: Cross-sectional studies," *Evidence-Based Dentistry*, vol. 7, no. 1, pp. 24–25, 2006, doi: 10.1038/sj.ebd.6400375.
- [27] J. W. Creswell, *Educational research. Planning, conducting, and evaluating quantitative and qualitative research*, 4th ed. Boston, MA.: Pearson Education Inc, 2012.
- [28] L. A. Miller, R. L. Lovler, and S. A. McIntire, *Foundations of psychological testing. A practical approach*, 13th ed. US: SAGE Publications. Inc, 2013.
- [29] E. Haertel, "Construct validity and criterion-referenced testing," *Review of Educational Research*, vol. 55, no. 1, p. 23, 1985, doi: 10.2307/1170406.
- [30] I. J. Yusof, A. Abdul Latif, and M. D. Mat Jani, "Research literacy level of education postgraduate research students using rasch measurement model," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3S2, pp. 791–796, 2019, doi: 10.35940/ijrte.C1242.1083S219.
- [31] W. P. Fisher, "Survey design recommendations," *Rasch Measurement Transactions*, vol. 20, no. 3, pp. 1072–1074, 2006.
- [32] T. G. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd ed. NJ, USA: Lawrence Erlbaum Associates, Mahwah, 2007.
- [33] A. S. Rahayah, *Inovasi dalam pengukuran dan penilaian pendidikan*. Bangi: Penerbit UKM, 2008.
- [34] J. M. Linacre, "What do Infit and Outfit Mean-Square and Standardized mean?" *Rasch Measurement Transactions*, vol. 16, no. 2, p. 878, 2002.
- [35] H. Y. Chong, "A simple guide to the Item Response Theory (IRT) and Rasch modeling," Oct. 2017. [Online]. Available: <http://www.creative-wisdom.com/computer/sas/IRT.pdf> (accessed: Jun. 23, 2020).
- [36] J. M. Linacre, *A User's Guide to Winsteps Ministep: Rasch-Model Computer Programs*. Chicago: Winsteps.com, 2012. [Online]. Available: <https://www.winsteps.com/winman/copyright.htm>
- [37] M. Balsamo, G. Giampaglia, and A. Saggino, "Building a new Rasch-based self-report inventory of depression," *Neuropsychiatric Disease and Treatment*, vol. 10, pp. 153–165, 2014, doi: 10.2147/NDT.S53425.

BIOGRAPHIES OF AUTHORS






Nor Hafizi Mohd Khalid    is a Ph.D. Candidate, School of Education, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, 81310 Skudai, Johor Baharu Malaysia. He received his first degree in Science from the Universiti Kebangsaan Malaysia and a Master's degree in Measurement and Evaluation in Education from the Universiti Teknologi Malaysia. His research focuses on Classroom Assessment, Assessment Literacy, Bibliometric Analysis, Educational Measurement and Evaluation, Psychometric and Instrument validation, and measurement theory such as Item Response Theory. He can be contacted at email: norhafizie@gmail.com.






Ibnatul Jalilah Yusof    has a master's degree in Measurement and Evaluation from Universiti Teknologi Malaysia (2015) and Ph.D. in Measurement and Evaluation from Universiti Teknologi Malaysia (2019). She's currently serving as senior lecturer at School of Education, Universiti Teknologi Malaysia, working in the department of Educational Foundation and Social Sciences. She's currently teaching educational subjects such as educational assessment, psychological testing, psychometric properties, program evaluation, quantitative data analysis, and research methodology to both undergraduate and postgraduate students. Her research interests include educational assessment, quantitative research methodology, and statistical applications for the behavioral and social sciences. She can be contacted at email: ijalilah@utm.my.



Adibah Abdul Latif    received a Ph.D. degree in Educational Measurement and Evaluation from the Universiti Teknologi Malaysia, 81310 Skudai, Johor Baharu, Malaysia. She has over 20 years of experience as an Academician with the Universiti Teknologi Malaysia (UTM), where she is currently an Associate Professor in the School of Education, Faculty of Social Sciences and Humanities. Her current expertise research includes Educational Measurement and Evaluation; Psychometric and Instrument validation; Holistic Assessment; and Constructive Alignment and Outcome Based Education. She can be contacted at email: p-adibah@utm.my.



Md Daud Md Jani    is a lecturer at the Temenggong Ibrahim Teacher Education Institute, Johor Malaysia, Department of Education. He received a Ph.D. Measurement and Evaluation in Education from the University Teknologi Malaysia (UTM). His research focuses on constructive alignment in higher education, development and validation Instrument, psychometric evaluation, classroom assessment, school-based assessment, and alternative assessment. He can be contacted at email daudmdjani@yahoo.com.