Journal of Engineering and Applied Sciences 14 (9): 2783-2798, 2019
ISSN: 1816-949X


# Validation of the Malaysian-Based Assessment Practice Inventory for Teacher Educators (MAPITE) using Rasch Model

[1]Siti Eshah Mokshein, [1]Hishamuddin Ahmad,[1]Othman Lebar, [1]Mohd Uzi Dollah, [2]Jamal Yunus, [3]Azali Rahmat and [1]Hamsa Hameed Ahmed
[1]Faculty of Human Development,
[2]Faculty of Management and Economy,
[3]Faculty of Sport Science and Coaching, Universiti Pendidikan Sultan Idris, 35900 Tg. Malim, Perak, Malaysia

**Abstract:** A psychometrically sound instrument to measure teacher educator's assessment practice suitable for the Malaysian context needs to be developed. This study aimed to validate an Assessment Practice Inventory for Teacher Educators in Malaysia (MAPITE) to be used in the project 'NRGS-developing and validating an assessment and accountability framework for preparing quality teachers for the future' using Rasch Model. An Assessment Practice Inventory for Teacher Educators (MAPITE) consisting 70 items has been developed in the NRGS-assessment sub-project 'developing and validating an assessment and accountability framework for preparing quality teachers' 2014-2019. Analysis of construct validity by Exploratory Factor Analysis (EFA) and internal consistency on the data of pilot study showed that the instrument seem to be sound and can be used to measure assessment practice related to assessment literacy standards assessment principles and frequency of carrying out described items. However, 21 items needed to be removed, leaving 49 items in the new version of the instrument. A more detailed analysis with a larger sample using Item Response Theory (IRT) or Rasch Model was suggested before the instrument could be finalised. Data about teacher educator assessment practice were collected from 763 teacher educators in the Teacher Education Institutions (TEI's) and an Education University in Malaysia. The data gathered were re-analysed in the light of the Rasch Model of measurement using Winsteps software. The MAPITE was found to be a reasonably sound instrument for measuring assessment practice among teacher educators in Malaysia covering all important aspects of assessment literacy standards as demonstrated by the fit and Z-standardized statistics, item polarity and PTMEA correlation. Six items were removed from the early version of the instrument due to misfit, leaving 64 items in the new version. This instrument can be used to measure the assessment practice among teacher educators in Malaysia, so that, appropriate follow-up actions can be implemented towards the betterment of teacher education quality. The MAPITE adds to the limited collection of locally developed instruments in the field of educational assessment and evaluation.

**Key words:** Assessment assessment practice inventory, teacher educators, validation, Rasch Model, instruments, MAPITE

## INTRODUCTION

Research in many parts of the world has shown that assessment literacy of pre-service and in-service teachers were at low to medium level (Plake *et al*., 1993; Campbell *et al*., 2002; Talib, 2009; Hamzah and Sinnasamy, 2009; Faizah, 2011). Quite similar findings was observed in Malaysia where the School-Based Assessment (SBA) was not implemented according to guidelines and objectives provided by the Malaysian examination syndicate (Talib, 2009). SBA demands new ways of conducting assessments and multiple assessment tasks to assess a student and these pose a new challenge to teachers. The role of teachers has become more prominent than before. Teachers are given empowerment in assessing their students and these assessments carry certain weight in the overall assessment in the national examinations or at the end of educational level completion. That is because of the purpose of assessment is to track overall growth and development of students.

**Corresponding Author:** Aos A. Z. Ansaef Al-Juboori, Faculty of Computer and Creative Industry, University Pendidikan Sultan Idris, 35900 Tg. Malim, Perak, Malaysia

One of the factors that contribute to lack of assessment competency among teachers is limited assessment education that is potentially misaligned to assessment standards and classroom practices (DeLuca and Bellara, 2013). This in turn has raised questions on whether or not teacher training institutions has prepared pre-service teachers adequately in terms of necessary values, skills and knowledge to carry out assessment for 21st century education. How much assessment education has been provided by teacher training institutions in the country and what kind of education do our institutions provide to our pre-service teachers is determined partly by the assessment literacy among teacher educators. The measurement of teacher educator's assessment practice in the country requires a psychometrically sound instrument suitable for the Malaysian context. Thus, the purpose of this study is to validate an assessment practice inventory for teacher educators in Malaysia to be used in the NRGS assessment sub-project 'developing and validating an assessment and accountability framework for preparing quality teachers for the future' using Rasch Model.

The NRGS-assessment sub-project 'developing and validating an assessment and accountability framework for preparing quality teachers for the future 2014-2019' explores teacher educator's knowledge about assessment, their beliefs about assessment, their assessment practice and competency in conducting assessment, so that, necessary features can be incorporated in new framework for the future. For these purposes, an instrument called A Malaysian Assessment Practice Inventory for Teacher Educators (MAPITE) was developed to be used in the study. The instrument was developed based on the assessment standards, 9 principles of best practices in educational assessment, classroom assessment practices and the teacher's self-perceived assessment skills (DeLuca and Bellara, 2013) and assessment of student's learning: practice among malaysian teachers.

Siti Eshah *et al.* cited that Zhang and burry-stock had developed an instrument containing 67 items to investigate teacher's assessment practices across education levels and content areas as well as teacher's self-perceived assessment skills as a function of teaching experience and measurement training. Data from 297 teachers on the assessment practices inventory were analyzed in a MANOVA design. They found that as grade level increases, teachers rely more on objective tests in classroom assessment and show an increased concern for assessment quality ($p<0.001$). Across content areas, teacher's involvement in assessment activities reflects the nature and importance of the subjects they teach ($p<0.001$). Regardless of their teaching experiences, teachers with measurement training report a higher level of self-perceived assessment skills in using performance measures; standardized testing, test revision and instructional improvement, as well as in communicating assessment results ($p<0.05$) than those without measurement training. In the recent development, teacher educators in Malaysia have started to use e-portfolio to assess the development of soft skills among student-teachers.

Suah *et al.* in obtaining information on school teacher's assessment practice in Malaysia developed an instrument adapted from Zhang and Burry-Stock. The inventory consists of three sub-sections namely information about teacher's background, training and knowledge on assessment and assessment practices implemented by school teachers. The sample for the study was 602 teachers from the Northern region of Malaysia. Data from the study were analyzed by calculating the mean values of the responses and percentage of respondent's practices. The results showed that the form of assessment frequently used by school teachers was multiple-choice objective test. There were significant differences among teachers from different school levels for aspects like developing marking scheme, giving feedbacks of evaluation results and the use of written test and the use of other strategies. Comparison among teachers teaching different subject areas showed significant difference only in written test. However, these two studies focused on assessment practice among teachers and not teacher educators.

Literature from best practices in educational assessment highlighted nine important principles of which seven that were suitable with the local context were chosen and further developed in the NRGS-assessment project to be matched with the essential values, skills and knowledge. The 7 principles and related attributes to be developed in the NRGS project are as follow:

- The assessment of student learning begins with educational values
- Assessment is most effective when it reflects an understanding of learning as multidimensional integrated and revealed in performance over time
- Assessment works best when the programs it seeks to improve have clear, explicitly stated purposes
- Assessment requires attention to outcomes but also equally to the experiences that lead to those outcomes

- Assessment works best when it is ongoing not episodic
- Assessment fosters wider improvement when representatives from across the educational community are involved
- Through assessment, educators meet responsibilities to students and to the public

The MAPITE in the NRGS-assessment study consists of 78 items (excluding educator's perception of competence in conducting described items-DS) that were organized in five sections as below:

- A: Demography (8 items)
- B: Assessment literacy standards (10 items)
- C: Beliefs about assessment (9 items)
- DU: Frequency in conducting described items (51 items)
- DS: Competence in conducting described items (51 items)

Items on section DS were not analyzed in this study due to the low interpretability of the results. For instance, if a set of items was grouped together in factor analysis based on respondent's perception of their competence in conducting the described item what can we say about the validity? We cannot say that those items measure the same construct when they obviously are not related to one another. For that reason, items on section DS were excluded from the validity and reliability analysis. The assessment literacy standards covered in this instrument were based on the assessment literacy standards by the American Federation of Teachers (Ling *et al*., 2009) as:

**Standard 1:** Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

**Standard 2:** Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

**Standard 3:** The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.

**Standard 4:** Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum and school improvement.

**Standard 5:** Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

**Standard 6:** Teachers should be skilled in communicating assessment results to students, parents, other lay audiences and other educators.

**Standard 7:** Teachers should be skilled in recognizing unethical, illegal and otherwise inappropriate assessment methods and uses of assessment information.

The initial instrument contained 70 items (excluding background information) that were divided into three important sections practice related to assessment literacy standards (section B), belief about assessment principles (section C) and frequency of carrying out described items (section D). The instrument was administered to 254 teacher educators from a teacher education university and a teacher training institute. Exploratory Factor Analyses (EFA) and reliability tests were performed on the data. Results showed that the instrument developed yielded high values of internal consistency as reflected by the Cronbach alpha values. Results of EFA suggested that 21 items need to be removed due to their non-dimensionality as they have more or less equal loadings on several factors. Thus, the final draft of the instrument contained 49 items. Even though the reliability and validity of the instrument are within the acceptable range, more data need to be gathered using bigger sample size, so that, further analysis using item response theory can be used to explore deeper into the psychometric characteristics of the items before the instrument can be finalized. Some of the items suggested to be removed in the pilot study could be important to measure teacher educator's assessment practice for the assessment project (DeLuca and Bellara, 2013).

**Item response theory and Rasch Model:** Item Response Theory (IRT) and Rasch Model which is commonly associated with 1-parameter IRT have been widely used in education and psychological testings because of their advantages over Classical Test Theory (CTT). Even though IRT and Rasch are similar to each other in terms of computation, their philosophical foundations are vastly different from each other (Yu, 2013). In research modelling, there is an ongoing tension between fitness and parsimony. IRT is concerned with a model that reflects or fits "reality" while the Rasch inclines to simplicity. To be more specific, IRT modelers might use up to three parameters but Rasch stays with one parameter only. In other words IRT is said to be descriptive in nature

because it aims to fit the model to the data. In contrast, Rasch is prescriptive for it emphasizes fitting the data into the model. The item characteristic curve known as ICC item response function (Tucker, 1946) for Rasch Model is represented by the following Eq. 1 (Bond and Fox, 2007):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \quad i = 1, 2, ..., n \tag{1}$$

Where:

$P_i(\theta)$ = The probability of a person with ability theta ($\theta$) obtaining a correct answer on a particular item

(i) and $b_i$ = The difficulty of the item or the location of item i

In short, $P_i(\theta)$ is a function of person's ability and item difficulty and the relationship is shown in Eq. 1. The IRT and Rasch Model are also widely used in instrument development. Unlike the CTT, both the item and respondent's ability are located at the same continuum. Even though the items can be placed at different location, the discrimination parameter that differentiates individuals is always constant across items. The Rasch Model sets the discrimination parameter $\alpha$ at 1.0 while $\alpha$ in 1PL Model is set to constant. The ICC for 1PL Model is as in Eq. 2:

$$P_i(\theta) = \frac{e^{\alpha(\theta - b_i)}}{1+e^{\alpha(\theta - b_i)}} \quad i = 1, 2, ..., n \tag{2}$$

**Reliability and validity in Rasch measurement's perspective:** Validity and reliability are the important attributes for the quality of an assessment. Kelley (1927) state that the problem of validity concerns with whether a test really measures what it purports to measure while reliability is the question on how accurately a test measures the thing which it does measure. Validity is also referred to which theory or evidence of the test score interpretations entailed by proposed uses of the test (AERA, 1999). Each assessment is said capable to meet its intended purposes from logical and/or empirical evidence provided Educational Testing Service (2002). Educational Testing Service (2002) also described reliability as the extent to which results obtained can be generalized to scores obtained in different forms of the assessment which administered at particular times and possibly scored by some particular raters. In other words, reliability is the consistency of such measurements on the repeated testing procedures to the population whether individuals or groups Educational Testing Service (2002). Furthermore, the reliability of a behavioral measure is the stability of that measure to produce the same results when measuring a construct.

Rasch measurement model has been used widely in developing and validating an instrument among researchers. Rasch analysis is helpful in providing guidance to assess the reliability and validity of an instrument (Sabah *et al*., 2009). Furthermore, Rasch analysis is described as a powerful tool for evaluating construct validity (Baghaei, 2008). In general, the Rasch Model is a simple logistic latent trait Item Response Theory (IRT) Model concerned with the quality of outcome measures. Rasch analysis is a method of statistical approach for obtaining objective, fundamental, additive measures (qualified by standard errors and quality-control fit statistics) from ordinal observations (Linacre, 2005). It was formulated by Georg Rasch, a Danish mathematician in 1953 to analyze responses to a series of reading tests (Rasch, 1992).

In Rasch Model, FIT statistics of items and persons describe how well each item and person fit the model. Two mean square fit statistics are used to assess the extent to which unpredicted responses to an item are given by students whose position in the hierarchy as determined by their measure is either close to the item's position infit (MNSQ) or far from the item's position outfit (MNSQ) in the hierarchy of items (Tan and Yates, 2007). Rasch measurement model also produces item-person map which presents the estimation of person ability and item difficulties on the same continuum (Sabah *et al*., 2009). Item-person map can be used to improve the instrument, since, Rasch analysis produces the order and spacing of items on the hierarchical scale. Boone and Scantlebury as cited by Sabah *et al.* (2009) stated that items need to be added to fill the gaps on the hierarchical scale to improve the instrument.

Rasch measurement also provides reliability (separation indices) which is meant "reproducibility of relative measure location" (Linacre, 2005). Bond and Fox (2007) stated that these indices determine whether the items are spread enough along the continuum (item separation) and the ability is spread enough among persons (person separation). According to Linacre (2005), "person reliability" in Winsteps is equivalent to the traditional "test" reliability. Low values indicate a narrow range of person measures or a small number of items. The instrument should test persons with more extreme abilities (high and low) or lengthen the test to increase person reliability. While "item reliability" in Winsteps has no traditional equivalent, low values indicate a narrow range of item measures or a small sample. Bigger sample is needed to increase "item reliability". In general, low item reliability means that the sample size is too small for stable item estimates based on the current data. There are also differences between reliability indices between KR 20, Cronbach alpha and in Winsteps. KR-20 estimates reliability by summarizing item point-biserials, Cronbach

alpha estimates it with an analysis of variance while Winsteps estimates it using the standard error measures. In this study, Rasch analysis was used to explore the validity of the MAPITE.

## MATERIALS AND METHODS

The validity and reliability analyses of the MAPITE using actual data involving 763 respondents were conducted in 2015, a few months after the pilot test. The instrument was administered to teacher educators from several teacher education institutes (IPG's) and an education university in the country. Respondents from eight teacher education programs were selected through stratified random sampling technique. Since, this instrument is not an achievement test or speeded test, time is not a critical element in the administration of the questionnaire. Respondents were given sufficient time to complete the questionnaire until they satisfied with the responses provided. The questionnaires were then gathered and data were entered into spreadsheets using Statistical Package for Social Science (SPSS). Data analysis involved computation of descriptive statistics and examination of psychometric characteristics of the items using and internal consistency measure Cronbach alpha and exploratory factor analysis.

Analysis of scree plots by section of the instrument shows a big jump indicating that possibly there is only one dominant factor present in the test (Hambleton *et al*., 1991). Thus, the unidimensionality and local independence are assumed and IRT or Rasch analyses can be performed on MAPITE data. The scree plot obtained for the instrument is shown in Fig. 1.

The contribution of the first factor on the variance explained for sections B, C and DU were 53.1, 54.9 and 39.6%. These percentages are adequate as according to Reckase cited by Siti Eshah *et al*., more than 20% variance explained is needed for accurate estimation. McGill (2009) found that if the ratio of Eigenvalues of first to second dimension is greater than four, it could be considered as a "good" evidence of unidimensionality

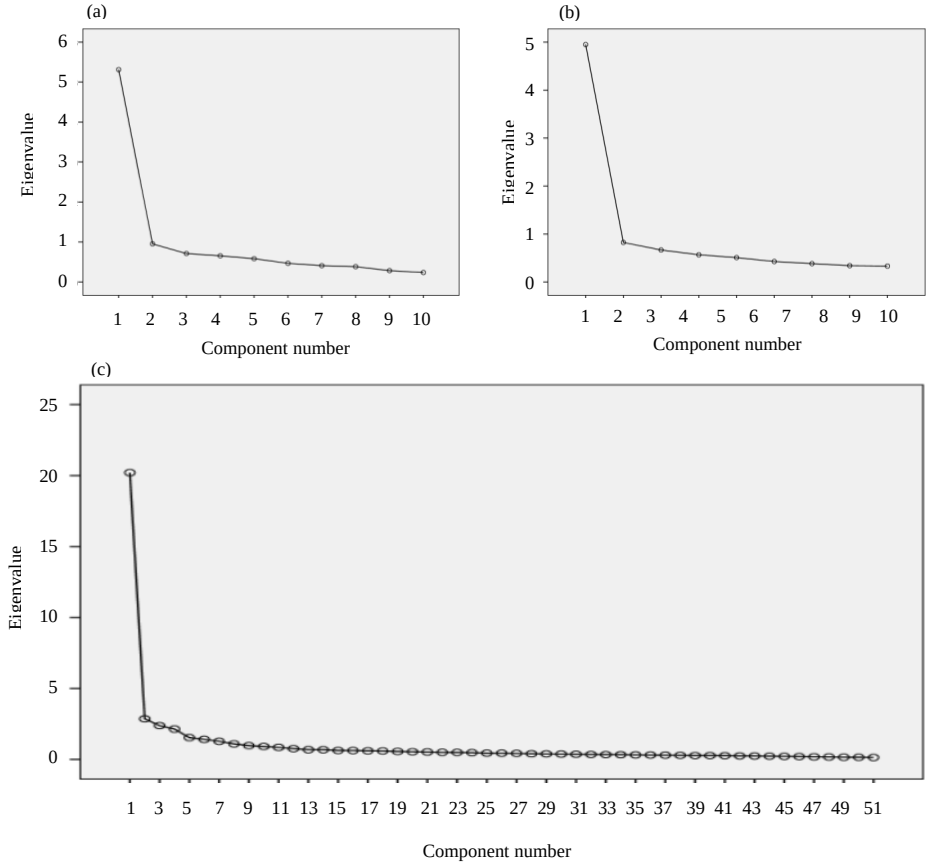


Fig. 1: a-c)Scree plot of EFA for Section B, C and DU of MAPITE: Section B:1 factor extracted, variance explained 53.1%; Section C:1 factor extracted, variance explained 54.9%; Section DU:8 factors extracted, variance explained for first factor 39.6%

In this study, the ratio was 4.6:1.06. Therefore, unidimensionality assumption was assumed to hold and Rasch analysis can be performed on the data.

Winsteps control files were prepared prior to the analysis. Since, the Winsteps Software reads text file, data from the SPSS format were converted into the text format and saved as Winsteps file (Linacre, 2005). The item codes were checked to ensure they matched with the data. Data were analyzed using polytomous rating scale model. In this study, Rasch analysis was used to provide validity and reliability evidence of the scales of the constructs in the MAPITE.

## RESULTS AND DISCUSSION

**Validity of the instrument:** The Rasch Model analysis using Winsteps Software produces several important outputs that could be used to answer research questions in this study. One of the main purposes of this study is to determine the appropriateness of the scale measured by the instrument. The fit statistics of persons and items in testing the unidimensionality of the model, the Z-Standardized Statistics (ZSTD) and the person-item maps for scale confirmation analysis are important outputs used to check on the validity of the items in the instrument. Infit and outfit statistics can provide an evidence of validity of the scale (Sabah *et al*., 2009).

Rasch Model analyses were performed on sections B (Practice related to assessment literacy standards), C (beliefs about assessment) and DU (frequency carrying out described items) of the instrument. Results of the first level of analysis of section B showed that all the 10 items fit the model well except for Item 8 which fits and Z-standardized statistics were outside the acceptable range (MNSQ between 0.6 and 1.4). The final analysis of section B that excluded Item 8 was then performed (Table 1).

Another useful statistics in the winsteps output is point-measure correlation (PTMEA CORR) which is the point-measure correlation between scored responses and ability measures. The Point-Measure Correlations (PTMEA CORR) report the extent to which this is true for each item. Good items would show strong positive PTMEA values. Small positive PTMEA values indicate the need of further investigation while negative values indicate that the responses to the item contradict the direction of the latent variable. The PTMEA values for all the nine items show strong positive correlations. Result of Principal Component Analysis (PCA) showed that the variance explained for the nine items on this section was 46.4%. Details of the results are shown in Table 1.

Results of the first level of analysis of section C showed that all the 9 items fit the model perfectly with the fit and Z-standardized statistics fall within the acceptable range. PTMEA correlations for all the 9 items were found to be high (between 0.69-0.76). Result of Principal Component Analysis (PCA) showed that the variance explained by the measures was 45.4%. Thus, no item was removed from this study. Details of the results are shown in Table 2.

Section DU (frequency of carrying out described items) consists of 51 items. Results of the first level of analysis of section DU showed that all the items fit the model well except for Item 2 and 48 which infit and outfit values were beyond the acceptable range (Table 3). Similarly, the ZSTD and PTMEA correlations for these two items were also beyond the acceptable range. Thus, the second level of analysis was carried out without the two items above mentioned.

In the second level of analysis, items 4, 44 and 49 were found to have fit statistics (infit and outfit) and ZSTD outside the acceptable range. Thus, the next analysis was carried out with exclusion of these items. PTMEA showed strong positive values, between 0.47 and 0.67. Results of the final analysis of section DU suggested

Table 1: Final analysis-misfit order of items in section B final analysis

| Entry number | Total score | Count | Measure | Model SE | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | PTMEA Corr. | Exact OBS (%) | Match Exp (%) | Items | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 3282 | 763 | 0.55 | 0.09 | 0.77 | -3.8 | 0.70 | -4.1 | 0.78 | 82.1 | 75.9 | B4 | 0 |
| 10 | 3208 | 763 | 0.41 | 0.08 | 1.41 | 4.9 | 1.36 | 4.2 | 0.67 | 72.7 | 74.0 | B10 | 0 |
| 6 | 3267 | 763 | 0.01 | 0.09 | 0.93 | -1.0 | 0.84 | -2.4 | 0.75 | 74.7 | 73.5 | B6 | 0 |
| 7 | 3284 | 763 | 0.00 | 0.08 | 1.26 | 3.6 | 1.34 | 4.4 | 0.66 | 63.7 | 72.3 | B7 | 0 |
| 3 | 3244 | 763 | -0.03 | 0.09 | 1.01 | 0.2 | 0.97 | -0.4 | 0.73 | 74.1 | 73.6 | B3 | 0 |
| 1 | 3370 | 763 | -0.07 | 0.09 | 0.91 | -1.4 | 0.81 | -2.5 | 0.74 | 77.8 | 76.1 | B1 | 0 |
| 5 | 3286 | 763 | -0.18 | 0.09 | 0.80 | -3.0 | 0.75 | -3.5 | 0.77 | 79.7 | 75.0 | B5 | 0 |
| 2 | 3327 | 763 | -0.30 | 0.10 | 0.85 | -2.6 | 0.75 | -3.4 | 0.76 | 81.6 | 76.9 | B2 | 0 |
| 9 | 3354 | 763 | -0.39 | 0.09 | 0.98 | -0.4 | 0.96 | -0.6 | 0.73 | 75.6 | 75.9 | B9 | 0 |
| Mean | 2771.3 | 659 | 0.00 | 0.09 | 0.99 | -0.4 | 0.94 | -0.9 | | 75.8 | 74.8 | | |
| SD | 48.7 | 0 | 0.29 | 0.00 | 0.20 | 2.8 | 0.24 | 3.0 | | 5.3 | 1.4 | | |

that 46 items were retained in the instrument. Principal Component Analysis (PCA) results showed that the variance explained by the measures was 58.2%. Thus, 5 items would be removed from this section in the final version of the instrument. Details of the results are shown in Table 3.

Table 2: First level analysis-misfit order of items in section C

| Entry number | Total score | Count | Measure | Model SE | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | PTMEA Corr. | Exact Obs. (%) | Match Exp. (%) | Items | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 3168 | 763 | 0.41 | 0.09 | 0.96 | -0.6 | 0.93 | -1.0 | 0.76 | 78.3 | 74.9 | C8 | 0 |
| 7 | 3178 | 763 | 0.36 | 0.09 | 1.02 | 0.3 | 0.99 | -0.1 | 0.74 | 76.3 | 75.9 | C7 | 0 |
| 6 | 3204 | 763 | 0.09 | 0.08 | 1.16 | 2.4 | 1.15 | 2.1 | 0.70 | 71.9 | 72.8 | C6 | 0 |
| 9 | 3217 | 763 | 0.06 | 0.08 | 1.24 | 3.6 | 1.24 | 3.3 | 0.69 | 69.0 | 70.7 | C9 | 0 |
| 3 | 3350 | 763 | -0.01 | 0.09 | 0.88 | -2.1 | 0.85 | -2.2 | 0.73 | 78.3 | 74.6 | C3 | 0 |
| 1 | 3341 | 763 | -0.21 | 0.09 | 1.06 | 0.8 | 1.04 | 0.6 | 0.69 | 74.4 | 75.9 | C1 | 0 |
| 2 | 3292 | 763 | -0.23 | 0.09 | 0.83 | -2.5 | 0.76 | -3.2 | 0.76 | 78.0 | 75.0 | C2 | 0 |
| 4 | 3291 | 763 | -0.23 | 0.09 | 0.87 | -2.0 | 0.84 | -2.2 | 0.75 | 77.1 | 74.3 | C4 | 0 |
| 5 | 3293 | 763 | -0.23 | 0.09 | 0.93 | -1.0 | 0.89 | -1.5 | 0.74 | 72.7 | 73.7 | C5 | 0 |
| Mean | 2739.3 | 659 | 0.00 | 0.09 | 1.00 | -0.1 | 0.97 | -0.5 | | 75.1 | 74.2 | | |
| SD | 64.9 | 0 | 0.24 | 0.00 | 0.13 | 2.0 | 0.15 | 2.0 | | 3.1 | 1.6 | | |

Table 3: Final analysis - Misfit order of items in Section DU

| Entry number | Total score | Count | Measure | Model SE | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure Corr. | Exp. | Exact Obs. (%) | Match Exp. (%) | Items | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 2590 | 763 | 0.55 | 0.05 | 1.29 | 5.2 | 1.39 | 6.3 | A.59 | 0.66 | 44.4 | 47.0 | D9U | 0 |
| 41 | 2740 | 763 | 0.24 | 0.05 | 1.18 | 3.3 | 1.33 | 5.3 | B.57 | 0.64 | 51.6 | 50.6 | D41U | 0 |
| 50 | 2930 | 763 | -0.22 | 0.05 | 1.13 | 2.1 | 1.22 | 3.4 | C.56 | 0.60 | 58.2 | 56.2 | D50U | 0 |
| 1 | 2913 | 763 | -0.57 | 0.06 | 1.17 | 2.8 | 1.21 | 3.4 | D.47 | 0.57 | 61.8 | 63.8 | D1U | 0 |
| 11 | 2640 | 763 | 0.47 | 0.05 | 1.15 | 2.8 | 1.18 | 3.1 | E.61 | 0.65 | 52.7 | 50.1 | D11U | 0 |
| 8 | 2752 | 763 | 0.06 | 0.05 | 1.15 | 2.6 | 1.17 | 2.9 | F.56 | 0.61 | 56.9 | 55.7 | D8U | 0 |
| 25 | 2380 | 763 | 1.06 | 0.05 | 1.09 | 1.8 | 1.12 | 2.2 | G.64 | 0.67 | 50.3 | 48.5 | D25U | 0 |
| 3 | 2947 | 763 | -0.79 | 0.06 | 1.11 | 2.0 | 1.12 | 2.0 | H.52 | 0.58 | 62.2 | 61.1 | D3U | 0 |
| 26 | 2535 | 763 | 0.74 | 0.05 | 1.06 | 1.2 | 1.11 | 1.9 | I.63 | 0.65 | 56.1 | 51.2 | D26U | 0 |
| 51 | 2930 | 763 | -0.21 | 0.05 | 1.09 | 1.6 | 1.07 | 1.2 | J.58 | 0.61 | 58.2 | 54.8 | D51U | 0 |
| 43 | 2867 | 763 | -0.05 | 0.05 | 1.00 | 0.0 | 1.08 | 1.3 | K.60 | 0.61 | 59.0 | 56.4 | D43U | 0 |
| 24 | 2426 | 763 | 0.90 | 0.05 | 1.04 | 0.9 | 1.08 | 1.5 | L.67 | 0.67 | 51.4 | 47.2 | D24U | 0 |
| 5 | 2758 | 763 | 0.05 | 0.05 | 1.05 | 0.9 | 1.07 | 1.2 | M.61 | 0.62 | 56.8 | 54.3 | D5U | 0 |
| 46 | 2749 | 763 | 0.18 | 0.05 | 1.05 | 0.9 | 1.06 | 1.1 | N.62 | 0.63 | 53.6 | 51.9 | D46U | 0 |
| 38 | 2835 | 763 | -0.01 | 0.05 | 0.97 | -0.5 | 1.06 | 1.0 | O.63 | 0.63 | 55.7 | 52.6 | D38U | 0 |
| 30 | 2664 | 763 | 0.46 | 0.05 | 0.96 | -0.7 | 1.05 | 1.0 | P.63 | 0.63 | 56.2 | 52.7 | D30U | 0 |
| 19 | 2975 | 763 | -0.52 | 0.06 | 1.02 | 0.4 | 1.05 | 0.8 | Q.58 | 0.59 | 64.0 | 58.5 | D19U | 0 |
| 20 | 2963 | 763 | -0.84 | 0.06 | 1.04 | 0.9 | 1.04 | 0.8 | R.57 | 0.59 | 60.1 | 58.8 | D20U | 0 |
| 21 | 2858 | 763 | -0.24 | 0.06 | 1.01 | 0.3 | 1.03 | 0.5 | S.60 | 0.60 | 60.8 | 57.4 | D21U | 0 |
| 42 | 2870 | 763 | -0.10 | 0.05 | 0.99 | -0.2 | 1.01 | 0.2 | T.62 | 0.61 | 56.8 | 54.9 | D42U | 0 |
| 18 | 3001 | 763 | -0.89 | 0.06 | 1.01 | 0.2 | 1.01 | 0.1 | U.57 | 0.58 | 61.6 | 61.1 | D18U | 0 |
| 27 | 2699 | 763 | 0.28 | 0.05 | 0.94 | -1.0 | 1.01 | 0.2 | V.63 | 0.62 | 60.1 | 55.3 | D27U | 0 |
| 13 | 2962 | 763 | -0.64 | 0.06 | 0.98 | -0.4 | 1.01 | 0.1 | W.60 | 0.59 | 58.5 | 59.0 | D13U | 0 |
| 45 | 3003 | 763 | -0.66 | 0.06 | 1.00 | 0.0 | 0.98 | -0.4 | w.59 | 0.59 | 62.0 | 58.7 | D45U | 0 |
| 22 | 2613 | 763 | 0.35 | 0.05 | 0.98 | -0.4 | 1.00 | -0.1 | v.66 | 0.65 | 53.0 | 50.8 | D22U | 0 |
| 17 | 2557 | 763 | 0.57 | 0.05 | 0.97 | -0.5 | 0.99 | -0.2 | u.66 | 0.65 | 57.4 | 51.3 | D17U | 0 |
| 47 | 2915 | 763 | -0.30 | 0.05 | 0.98 | -0.3 | 0.99 | -0.2 | t.61 | 0.60 | 59.5 | 56.9 | D47U | 0 |
| 28 | 2740 | 763 | 0.20 | 0.05 | 0.95 | -0.9 | 0.99 | -0.2 | s.63 | 0.61 | 62.2 | 56.8 | D28U | 0 |
| 31 | 2848 | 763 | -0.18 | 0.06 | 0.99 | -0.2 | 0.99 | -0.2 | r.60 | 0.60 | 62.9 | 58.4 | D31U | 0 |
| 32 | 2893 | 763 | -0.32 | 0.06 | 0.98 | -0.4 | 0.99 | -0.2 | q.57 | 0.58 | 66.0 | 61.9 | D32U | 0 |
| 35 | 2558 | 763 | 0.65 | 0.05 | 0.98 | -0.4 | 0.98 | -0.3 | p.67 | 0.65 | 54.9 | 49.7 | D35U | 0 |
| 7 | 3032 | 763 | -0.82 | 0.06 | 0.98 | -0.4 | 0.95 | -0.9 | o.58 | 0.57 | 63.9 | 63.7 | D7U | 0 |
| 33 | 2722 | 763 | 0.28 | 0.05 | 0.97 | -0.5 | 0.98 | -0.4 | n.64 | 0.62 | 60.2 | 53.9 | D33U | 0 |
| 36 | 2572 | 763 | 0.62 | 0.05 | 0.95 | -0.9 | 0.96 | -0.7 | m.68 | 0.65 | 54.3 | 49.4 | D36U | 0 |
| 16 | 2891 | 763 | -0.34 | 0.06 | 0.92 | -1.4 | 0.95 | -0.8 | l.64 | 0.60 | 58.6 | 57.2 | D16U | 0 |
| 10 | 2695 | 763 | 0.07 | 0.05 | 0.92 | -1.5 | 0.92 | -1.6 | k.67 | 0.63 | 55.7 | 53.1 | D10U | 0 |
| 15 | 2782 | 763 | -0.01 | 0.05 | 0.91 | -1.7 | 0.91 | -1.6 | j.65 | 0.62 | 61.8 | 55.2 | D15U | 0 |
| 29 | 2862 | 763 | -0.28 | 0.06 | 0.91 | -1.7 | 0.91 | -1.6 | I.63 | 0.59 | 65.8 | 60.0 | D29U | 0 |
| 40 | 2788 | 763 | 0.10 | 0.05 | 0.89 | -2.1 | 0.90 | -1.7 | h.66 | 0.63 | 58.1 | 52.6 | D40U | 0 |
| 6 | 2975 | 763 | -0.73 | 0.06 | 0.90 | -1.9 | 0.89 | -2.1 | g.63 | 0.58 | 65.2 | 61.7 | D6U | 0 |
| 39 | 2842 | 763 | 0.02 | 0.05 | 0.88 | -2.2 | 0.89 | -2.0 | f.65 | 0.62 | 55.7 | 53.6 | D39U | 0 |

Table 3: Continue

| Entry number | Total score | Count | Measure | Model SE | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure Corr. | Exp. | Exact Obs. (%) | Match Exp. (%) | Items | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 2692 | 763 | 0.19 | 0.05 | 0.83 | -3.5 | 0.87 | -2.4 | e.69 | 0.63 | 58.1 | 54.1 | D14U | 0 |
| 34 | 2694 | 763 | 0.30 | 0.05 | 0.86 | -2.6 | 0.87 | -2.4 | d.68 | 0.63 | 60.3 | 53.7 | D34U | 0 |
| 12 | 2845 | 763 | -0.15 | 0.06 | 0.87 | -2.4 | 0.87 | -2.5 | c.67 | 0.61 | 61.4 | 56.5 | D12U | 0 |
| 23 | 2795 | 763 | 0.04 | 0.05 | 0.85 | -2.7 | 0.82 | -3.4 | b.68 | 0.62 | 60.8 | 54.9 | D23U | 0 |
| 37 | 2594 | 763 | 0.49 | 0.05 | 0.84 | -3.2 | 0.84 | -3.1 | a.69 | 0.64 | 59.4 | 52.8 | D37U | 0 |
| Mean | 2780.3 | 763 | 0.00 | 0.05 | 1.00 | -0.1 | 1.02 | 0.3 | | 58.4 | 55.1 | | | |
| SD | 157.4 | 0 | 0.48 | 0.00 | 0.10 | 1.8 | 0.12 | 2.0 | | 4.3 | 4.2 | | | |

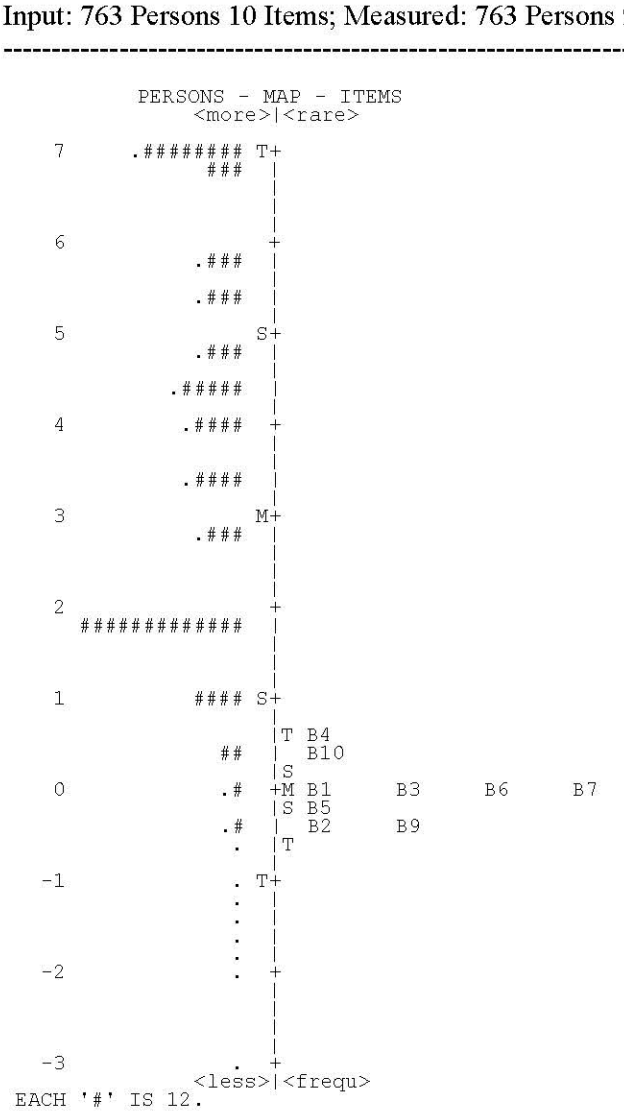


Fig. 2: Item map for section B

**Item-person map** One of the most important outputs obtained from Rasch analysis is the item-person map. This map locates the position of person's ability (or in this study is attitude) and item difficulty on the same continuum. Item map of Section B shows that the average item difficulty in general is lower than person's ability (attitude). This indicates that the items are not able to provide any information about persons of high ability (or positive attitude), suggesting 'harder' items to be included (Fig. 2). However, the

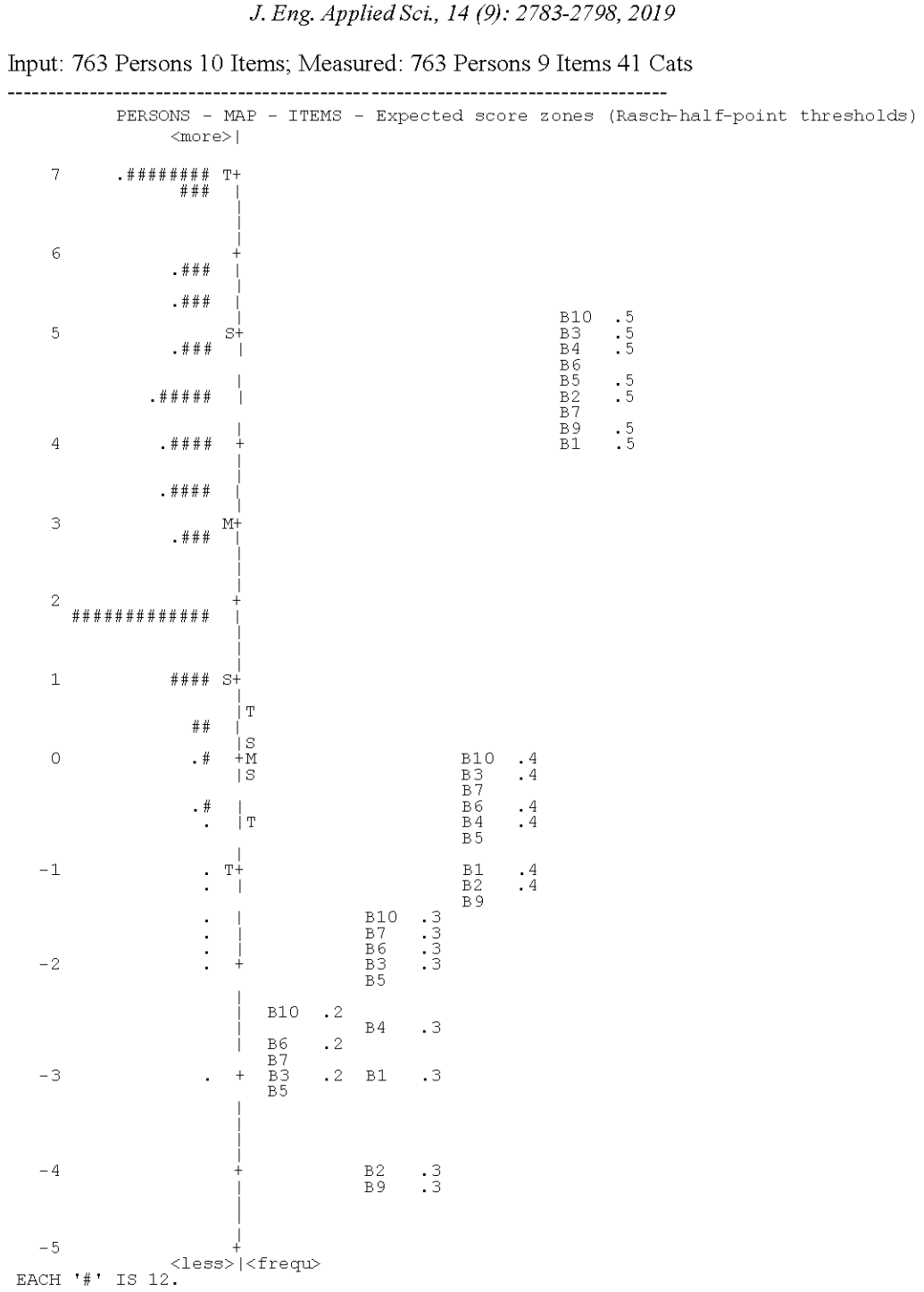


Fig. 3: Item map for section B by response category

person separation value (>2) in Table 1 suggests that the number of items in the test is sufficient. This can be explained by the item map by response category where by the responses actually spread along the continuum (Fig. 3).

Further, examination of the item map shows that for a person to respond a '5' from '4', it requires a big jump in ability (attitude). However, the gaps between responses '1-2' and '2-3' are almost similar while the gap between '3-4' is slightly bigger than the former. This observation is true for almost all items. It is also shown that even a person with lowest ability (attitude) is most unlikely to choose a '2' or '1' response. Besides, response category '2' for several items (B3, B5, B6, B7, B10)

Input: 763 Persons 51 Items; Measured: 763 Persons 46 Items

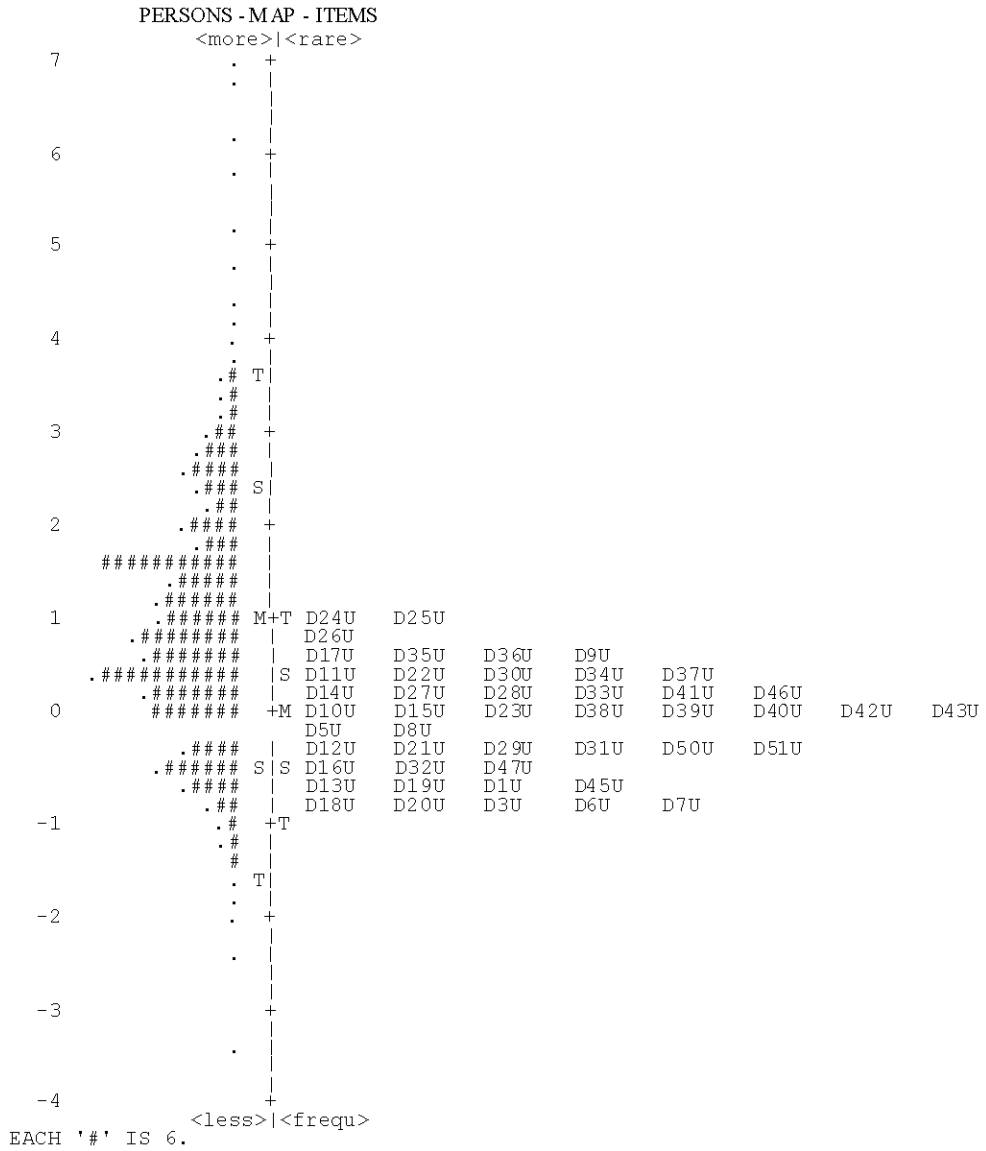


Fig. 4: Item map for section DU

overlaps with response category '3' for the rest of the items. Therefore, even though the mean difficulty for the individual items generally found to be low but there are persons choosing high (a '4' or '5') for the items. The item-person map is shown in Fig. 3.

The item maps for sections C and DU show quite similar pattern with section B. The average mean difficulty of the items is lower than the average mean ability of the persons. However, for each item, there are people high scores. The distribution of average item difficulty however approaches normal for section DU (Fig. 4).

**Item category measure:** The item category measure output shows the distance between response categories on the vertical logit scale for each item. The output shows that the intervals between response categories are not the same across item. This confirms that is the responses in the instrument are indeed in ordinal scale, not interval. Item category measures for sections B, C and DU are illustrated in Fig. 5 and 6.

Across section and item, the interval between responses 4-5 is the greatest whereas intervals between 1-2 and 2-3 are almost similar. The intervals

between 3-4 are somewhat different for different sections of the instrument. For sections B and C, the interval is slightly bigger compared to interval between 1-2 but smaller than 4-5. For sections DU however, the interval between 3-4 is almost similar with 1-2 interval. Overall, it requires a lower ability (attitude) for a person to respond a '4' and '5' in section DU as compared to sections B and C. Overall, category function output shows that even though the items have 5 categories of responses, only 3 to 4 categories or options are functioning.

**Reliability of the instrument:** Rasch Model analysis using winsteps produces item and person reliability and separation that are useful in determining the reliability of the instrument. Person separation is used to classify people. Low person separation (separation <2, person reliability <0.8) with a relevant person sample implies that the instrument may not be not sensitive enough to distinguish between high and low performers. More items may be needed to measure the construct. Acceptable value of person separation should greater than 2.0 with reliability >0.8. Item separation is used to verify the item hierarchy. Low item separation implies that the person sample is not large enough to confirm the item difficulty hierarchy (construct validity) of the instrument. Acceptable values of item separation should be >3.0 and reliability >0.9.

In this study, reliability of the instrument was found to be high as demonstrated by the person and item separation and reliability for each section of the instrument. For section B, person separation was 2.26 with reliability 0.84 while item separation was 3.11 with

Input: 763 Persons; Measured: 763 Persons

---

Section B

```
-3   -2    -1     0     1     2     3     4     5     6     7
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM  ITEM
|             2       3            4                  5     |    4  B4
|                                                           |
|          1    2      3            4                    5   |   10  B10
|                                                           |
|        1            3            4                  5     |    6  B6
|         1    2       3            4                 5      |    7  B7
|        1     2       3             4                  5    |    3  B3
|           2       3            4                  5        |    1  B1
|                                                           |
|       1     2       3            4                  5     |    5  B5
|                                                           |
|          2        3            4                  5      |    2  B2
|           2        3           4                  5       |    9  B9
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM  ITEM
-3   -2    -1     0     1     2     3     4     5     6     7
```

Section C

```
EXPECTED AVERAGE MEASURES FOR PERSONS (scored) (ILLUSTRATED BY AN OBSERVED CATEGORY)
-3   -2    -1     0     1     2     3     4     5     6     7
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM  ITEM
|           1     2      3           4                    5 |    8  C8
|           1    2       3           4                    5 |    7  C7
|                                                           |
|           1    2       3           4                   5   |    6  C6
|           1    2       3           4                  5    |    9  C9
|              2      3          4                  5        |    3  C3
|                                                           |
|          1    2     3           4                   5     |    1  C1
|          1    2      3           4                   5    |    2  C2
|          1    2      3           4                   5    |    4  C4
|          1    2      3           4                   5    |    5  C5
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
-3   -2    -1     0     1     2     3     4     5     6     7
```

Fig. 5: Continue

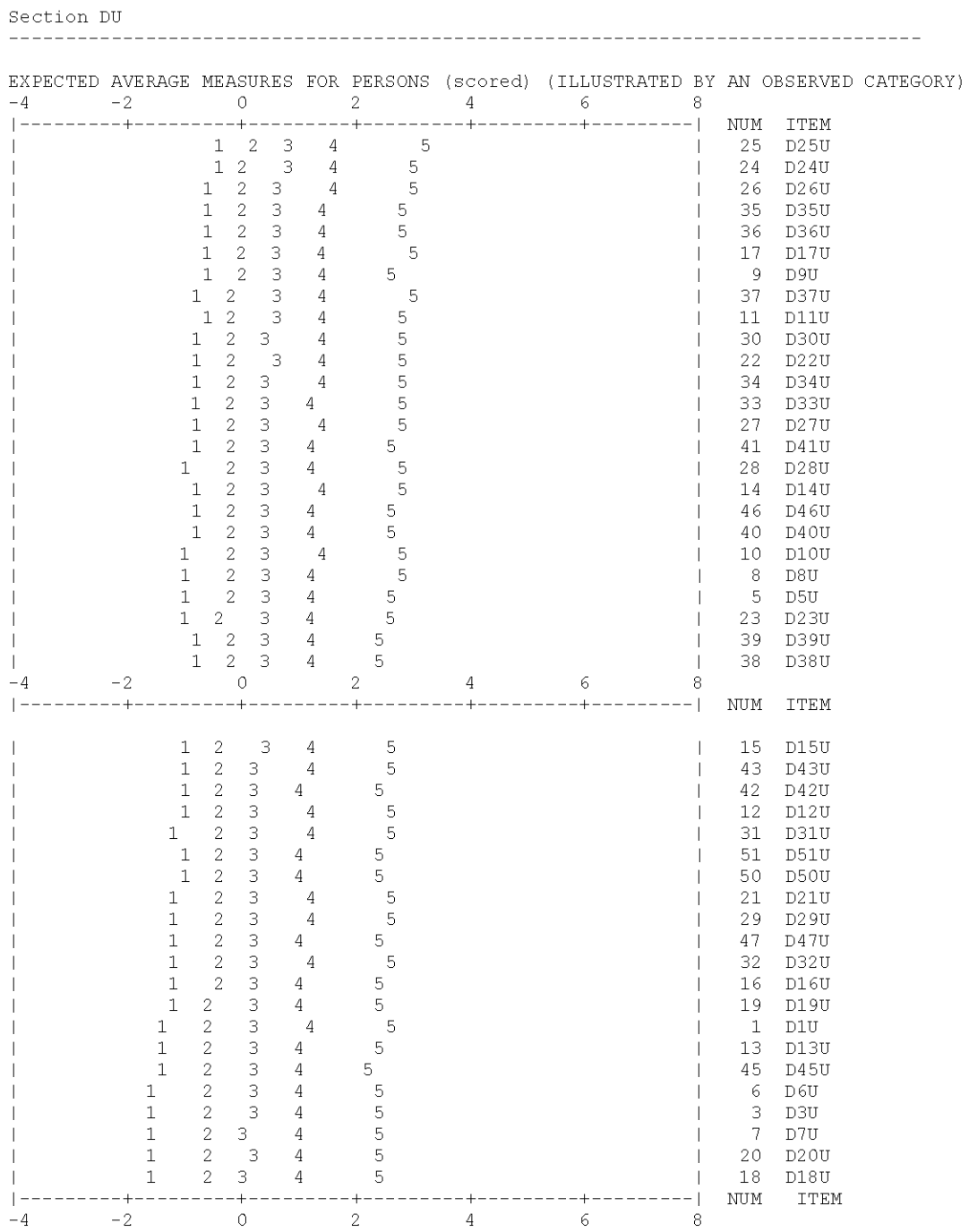


Fig. 5: Item category measure for sections B, C and DU

reliability 0.91. For section C, person separation was 2.30 with reliability 0.84 and item separation was 2.59 with reliability 0.87. This indicates that a bigger sample size is needed to respond to this section. However, researchers decided to accept this value as it is very close to 0.9 and the internal consistency of the instrument of this section as reported by the Cronbach alpha value was 0.9.

For section DU, person separation was 5.31 with reliability 0.97. Item separation was 8.88 with reliability 0.99. These values are well within the acceptable ranges. Details of person and item measures are shown in Table 4-9.

This study is similar to the earlier study in Malaysia by Ling *et al*. (2009) in the sense that both try to measure

Table 4: Summary of 763 measured (extreme and non-extreme) persons for sections B

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 38.8 | 9.0 | 3.73 | .94 | | | | |
| S.D. | 4.1 | .0 | 2.51 | .39 | | | | |
| MAX. | 45.0 | 9.0 | 8.02 | 1.86 | | | | |
| MIN. | 18.0 | 9.0 | -2.91 | .38 | | | | |

REAL RMSE 1.08 ADJ.SD 2.27 SEPARATION 2.10 PERSON RELIABILITY .81
MODEL RMSE 1.02 ADJ.SD 2.30 SEPARATION 2.26 PERSON RELIABILITY .84
S.E. OF PERSON MEAN = .09
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .90

Table 5: Summary of 9 measured (non-extreme) items for sections B

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 2771.3 | 659.0 | .00 | .09 | .99 | -.4 | .94 | -.9 |
| S.D. | 48.7 | .0 | .29 | .00 | .20 | 2.8 | .24 | 3.0 |
| MAX. | 2850.0 | 659.0 | .55 | .10 | 1.41 | 4.9 | 1.36 | 4.4 |
| MIN. | 2688.0 | 659.0 | -.39 | .08 | .77 | -3.8 | .70 | -4.1 |

REAL RMSE .09 ADJ.SD .27 SEPARATION 3.00 ITEM RELIABILITY .90
MODEL RMSE .09 ADJ.SD .28 SEPARATION 3.11 ITEM RELIABILITY .91
S.E. OF ITEM MEAN = .10

Table 6: Summary of 763 measured (extreme and non-extreme) persons for sections C

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 38.4 | 9.0 | 3.74 | .92 | | | | |
| S.D. | 4.2 | .0 | 2.50 | .39 | | | | |
| MAX. | 45.0 | 9.0 | 8.22 | 1.86 | | | | |
| MIN. | 18.0 | 9.0 | -2.46 | .39 | | | | |

REAL RMSE 1.07 ADJ.SD 2.26 SEPARATION 2.12 PERSON RELIABILITY .82
MODEL RMSE 1.00 ADJ.SD 2.29 SEPARATION 2.30 PERSON RELIABILITY .84
S.E. OF PERSON MEAN = .09
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .90

Table 7: Summary of 9 measured (non-extreme) items for sections C

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 2739.3 | 659.0 | .00 | .09 | 1.00 | -.1 | .97 | -.5 |
| S.D. | 64.9 | .0 | .24 | .00 | .13 | 2.0 | .15 | 2.0 |
| MAX. | 2830.0 | 659.0 | .41 | .09 | 1.24 | 3.6 | 1.24 | 3.3 |
| MIN. | 2648.0 | 659.0 | -.23 | .08 | .83 | -2.5 | .76 | -3.2 |

REAL RMSE .09 ADJ.SD .22 SEPARATION 2.52 ITEM RELIABILITY .86
MODEL RMSE .09 ADJ.SD .22 SEPARATION 2.59 ITEM RELIABILITY .87
S.E. OF ITEM MEAN = .08
UMEAN=.000 USCALE=1.000

Table 8: Summary of 763 measured (extreme and non-extreme) persons for sections DU

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 167.6 | 46.0 | 1.02 | .23 | | | | |
| S.D. | 26.7 | .0 | 1.36 | .10 | | | | |
| MAX. | 230.0 | 46.0 | 8.07 | 1.83 | | | | |
| MIN. | 58.0 | 46.0 | -3.37 | .17 | | | | |
| REAL RMSE .28 ADJ.SD 1.33 SEPARATION 4.76 PERSON RELIABILITY .96 | | | | | | | | |
| MODEL RMSE .25 ADJ.SD 1.34 SEPARATION 5.31 PERSON RELIABILITY .97 | | | | | | | | |
| S.E. OF PERSON MEAN = .05 | | | | | | | | |

PERSON RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .97

Table 9: Summary of 46 measured (non-extreme) items for sections DU

| | RAW | | | MODEL | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | SCORE | COUNT | MEASURE | ERROR | MNSQ | ZSTD | MNSQ | ZSTD |
| MEAN | 2780.3 | 763.0 | .00 | .05 | 1.00 | -.1 | 1.02 | .3 |
| S.D. | 157.4 | .0 | .48 | .00 | .10 | 1.8 | .12 | 2.0 |
| MAX. | 3032.0 | 763.0 | 1.06 | .06 | 1.29 | 5.2 | 1.39 | 6.3 |
| MIN. | 2380.0 | 763.0 | -.89 | .05 | .83 | -3.5 | .82 | -3.4 |
| REAL RMSE .05 ADJ.SD .48 SEPARATION 8.73 ITEM RELIABILITY .99 | | | | | | | | |
| MODEL RMSE .05 ADJ.SD .48 SEPARATION 8.88 ITEM RELIABILITY .99 | | | | | | | | |
| S.E. OF ITEM MEAN = .07 | | | | | | | | |

assessment practice among educators. However, both studies differ in terms of the instrument used, the level of educators assessed and the methods used to validate the instruments. Ling *et al.* (2009) used her own instrument to measure assessment practice among teachers in schools and instrument was validated through content validity. In this study, the MAPITE was developed to measure assessment practice among teacher educators in teacher training institutions and the instrument was validated through construct validity using Rasch Model.

The results of the MAPITE analysis showed that none of the items were too easy or too difficult for the respondents. The item difficulties ranged between -1.0-1.06. Most of the items that have been dropped due to their poor characteristic were from the DU section (frequency in conducting described items). Only one item was dropped from other sections. The discarded items are as follow:

- B8-Using assessment result to make decision.
- D2U-Selecting test book-provided test item for classroom assessment
- D4U-Assessing students through observation
- D44U-Incorporating attendance in the calculation of grades
- D48U-Communicating classroom assessment result to other educators
- D49U-Protecting students confidentiality with regards to test scores

Most of the items discarded are related to practice that is rarely done by normal classroom teachers that might cause respondents to guess or provide ingenuine responses. Researchers have suggested a variety of problems that may contribute to inaccurate estimates of an individual's responses on the construct assessed including ingenuine response, distorting factors can act to either artificially inflate construct estimates, (e.g., intentional distortion on responses to personality items) and deflate construct estimates (e.g., answer sheet miscues on cognitive ability test items) (Schmitt *et al.*, 1999). In exploring the quality of items using the Item Response Theory (IRT), Hishamuddin and Siti (2016) also suggested that items with high values of guessing parameter should be discarded (Ayala, 2009; Rasch, 1980; Reckase, 1979; Zhang and Burry-Stock, 2003).

## CONCLUSION

The Rasch Model analyses using Winsteps Software suggested that 6 items [B8, DU2, DU4, DU44, DU48 and D49] be removed from the earlier version of the Malaysian Assessment Practice Inventory for Teacher Educators (MAPITE) due to their fit and Z-standardized statistics

that were beyond the acceptable ranges. Thus, a new version of MAPITE consists of 64 items covering the important assessment literacy standards (9 items), beliefs about assessment (9 items) and frequency in conducting described items (46 items) is suggested. This new version of the instrument was found to possess sound item characteristics and can be used to measure assessment practice among teacher educators in Malaysia, so that, appropriate follow-up actions can be implemented towards the betterment of teacher education quality. This new instrument adds to the limited collection of locally developed instruments in the field of educational assessment and evaluation.

## RECOMMENDATION

However, it is recommended that items be recoded to 3 or 4 response categories accordingly only before analyzing the data using Rasch Model.

## ACKNOWLEDGEMENTS

The researchers gratefully acknowledge the financial supports provided by the Ministry of Education Malaysia through the NRGS Developing Teacher Education Model for the 21st Century (Project Code: 2014-0001-107-82-3) the NRGS Team Leader, Research Management and Innovation Centre Sultan Idris Education University (UPSI) and educational institutions involved in this study including UPSI, Teacher Education Institutions (TEI's) all over Malaysia and all teacher educators who participated in this study.

## REFERENCES

AERA., 1999. Standards for educational and psychological testing. American Educational Research Association, Washington, USA.

AFT., 2009. Assessment literacy standard. American Federation of Teachers, Washington, USA.

Ayala, R.J.D., 2009. The Theory and Practice of Item Response Theory. The Guilford Press, New York, USA., ISBN:978-1-59385-869-8, Pages: 448.

Baghaei, P., 2008. The Rasch model as a construct validation tool. Rasch. Measure. Trans., 22: 1145-1146.

Bond, T.G. and C.M. Fox, 2007. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2nd Edn., Lawrence Erlbaum Inc., New Jersey, USA.,.

Campbell, C., J.A. Murphy and J.K. Holt, 2002. Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers. Educational Research Association, Columbus, Ohio.

De Ayala, R.J. and M.A. Hertzog, 1991. The assessment of dimensionality for use in item response theory. Multivariate Behav. Res., 26: 765-792.

DeLuca, C. and A. Bellara, 2013. The current state of assessment education: Aligning policy, standards and teacher education curriculum. J. Teacher Educ., 64: 356-372.

Educational Testing Services, 2002. ETS Standards for Quality and Fairness. Educational Testing Service, Princeton, New Jersey, Pages: 81.

Faizah, A.M., 2011. School-based assessment in Malaysia schools: The concerns of the English teachers. US. China Educ. Rev., B3: 393-402.

Hambleton, R.K., H. Swaminathan and H.J. Rogers, 1991. Fundamentals of Item Response Theory. Sage Publications Inc., Thousand Oaks, California, USA., Pages: 173.

Hamzah, M. and P. Sinnasamy, 2009. Between the idea and reality: Teachers' perception of the implementation of school-based oral English assessment. English Teach. J., 38: 13-29.

Kelley, T.L., 1927. Interpretation of Educational Measurements. World Book Company, New York.

Linacre, J.M., 2005. Rasch dichotomous model vs. one-parameter logistic model (1PL 1-PL). Rasch Meas. Trans., 19: 1032-1032.

Ling, S.S., O.S. Lan and S. Osman, 2009. [Student learning assessment: The practice of teachers in Malaysia (In Malay)]. Majlis Dekan Pendidikan Malaysia, 5: 1-22.

McGill, M.T., 2009. An investigation of unidimensional testing procedures under latent trait theory using principal component analysis. Ph.D. Thesis, Virginia Polytechnic Institute, State University, USA.

Plake, B.S., J.C. Impara and J.J. Fager, 1993. Assessment competencies of teachers: A national survey. Educ. Meas.: Issues Pract., 12: 10-12.

Rasch, G., 1980. Probabilistic Models for Some Intelligence and Attainment Tests. University of Chicago Press, Chicago, Illinois, ISBN:9780226705538, Pages: 199.

Rasch, G., 1992. Probabilistic Models for Some Intelligence and Attainment Tests. MESA Press, Chicago, IL., USA.

Reckase, M.D., 1979. Unifactor latent trait models applied to multifactor tests: Results and implications. J. Educ. Stat., 4: 207-230.

Sabah, S., H. Hammouri and M. Akour, 2009. Validation of a scale of attitudes toward science across countries using Rasch model: Findings from TIMSS. J. Baltic Sci. Educ., 12: 692-702.

Schmitt, N., D. Chan, J.M. Sacco, L.A. McFarland and D. Jennings, 1999. Correlates of person fit and effect of person fit on test validity. Applied Psychol. Meas., 23: 41-53.

Talib, R., 2009. [Instrument construction and verification to measure the literacy level of secondary school teachers in Malaysia]. Ph.D Thesis, Universiti Teknologi Malaysia, Johor Bahru, Malaysia. (In Malay)

Tan, J.B.Y. and S.M. Yates, 2007. A Rasch analysis of the academic self-concept questionnaire. Int. Educ. J., 8: 470-484.

Tucker, L.R., 1946. Maximum validity of a test with equivalent items. Psychometrika, 11: 1-13.

Yu, C.H., 2013. A simple guide to the item response theory (IRT) and Rasch modeling. http://www.creative-wisdom.com/computer/sas/IRT.pdf.

Zhang, Z. and J.A. Burry-Stock, 2003. Classroom assessment practices and teacher;s self-perceived assessment skills. Appl. Meas. Educ., 16: 323-342.