INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 10, Issue 4, (2021) E-ISSN: 2226-6348

Validity and Reliability of Mathematics Intervention Instrument Based on the Learning Style of Students with Learning Disability (IMGAP) Using Rasch Assessment Model

Nafisah Baharom, Norshidah Mohamad Salleh & Mohd Mokhtar Tahar

National University of Malaysia, Malaysia

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v10-i4/11839 DOI:10.6007/IJARPED/v10-i4/11839

Published Online: 16 November 2021

Abstract

Survey questionnaires have been widely used to measure important outcomes in special education. However, the reliability and validity of these questionnaires are often measured using the Classical Test Theory approach. In the meantime, Rasch Analysis based on Item Response Theory provides a better alternative for examining the psychometric quality of rating scales and informing scale improvements. This article outlines a six-step process for using Rasch Analysis to review the psychometric properties of a rating scale, namely item functional inspection, reliability and item-respondent separation, polarity and suitability of items for construct measurement, and the standard residual correlation values. The questionnaires were distributed to 140 special education (learning disabilities) teachers. The final analysis found that no items were dropped from the questionnaire. Thirty-nine items were deemed suitable for evaluating constructs related to learning styles-based mathematical intervention for students with learning disabilities. The findings from the analysis have statistically proved that the items in this instrument have a high degree of validity and reliability as well as suitable to be used on special education teachers for learning disabilities in determining the importance of considering learning styles as one of the important factors when conducting mathematical interventions of students with learning disabilities.

Keywords: Mathematical Intervention Instrument, Learning Style, Student with Learning Disability, Rasch Measurement Model, Mathematical Interventions

Introduction

Mathematics interventions are targeted to mathematics disabilities to cater to their specific academic needs (Taylor-Cox, 2016). Mathematics interventions for these students play an important role in reducing the impact of their disabilities, especially aspects that have been scientifically proven to be challenging and can affect students' readiness to engage in the teaching and learning process (Kementerian Pendidikan Malaysia, 2015). Teaching students with mathematics disabilities pose a unique challenge to the educational environment, as teachers strive to pay individual attention to students with different learning styles.

Challenges these students face in understanding mathematical concepts could be attributed to their learning styles as the intervention processes mostly involve their senses. These senses are responsible for driving mind development as the intervention takes place. Learning style is one of the methods used by individuals to obtain and focus on new information (Sengodan & Iksan, 2012). Teachers who teach based on different student learning styles are oriented towards improving learning processes and outcomes. They will also be more open to changes than those who do not consider learning styles as a pedagogical base (Boström, 2011). This study on instrument development was implemented based on several features of mathematical interventions: screening, progress monitoring, data-based decision, and mathematics.

This ensures that the proposed intervention approach can lead to a greater emphasis on aspects like teaching students with a mathematics learning disability and learning styles of students with learning disabilities. The use of the survey method remains one of the popular and significant research methodologies either in graduate studies or the publication of research articles. Studies have extensively used the survey research methods. This shows the importance of developing a survey instrument that measures what it intends to measure. The quality of the instruments used in the measurement process plays an important role in analysing the data collected. It is important to start at the measurement level and identify weaknesses that could limit the reliability and validity of the measures in a survey instrument (Bond & Fox, 2015). Thus, this study used the Rasch Measurement Model to evaluate the quality and assessment scale structure of the learning styles-based mathematics intervention instrument for students with learning disability (IMGaP).

Problem Statement

The implementation of math interventions in schools could provide continuous support for students with mathematics learning disability. Past studies have shown that mathematical interventions can reduce skills gaps and prevent deficits that may occur in the future (Clements & Sarama, 2007; Fuchs et al., 2002; Sophian, 2004). The success of mathematics interventions for students with learning disabilities has been extensively demonstrated through various means and recommendations(Geary, 2013; Gersten et al., 2005, 2009; Jaspers et al., 2017; Kroesbergen & Luit, 2003; Lemons et al., 2015; Suhaimin & Mohamed, 2017). Nevertheless, there is an evidence gap (Jacobs, 2011; Miles, 2017; Mueller-Bloch & Kranz, 2015) on the correlation between the success of mathematics interventions with the incorporation of learning styles of students with learning disabilities. Accordingly, to fill the existing gap, researchers have conducted an initial assessment of the required characteristics to systematically and effectively implement mathematical interventions on students with learning disabilities. It is important to examine the required characteristics for mathematical interventions to gain a collective understanding of successful interventions. This measure also helps create an integrated model suitable for all states, districts, and schools (Fuchs & Fuchs, 2009). An integrated intervention model can also reduce variability in practice, improve communication, and improve the ability to determine whether it achieves intended goals.

The psychometric quality of a survey is typically assessed based on Classical Test Theory (CTT). However, CTT has several limitations, including scores obtained are sample-dependent and biased toward the central score (Bradley et al., 2015). In this regard, the missing data creates a challenge in calculating the overall score. Furthermore, the reliability of a measure is often presented as Cronbach's alpha, and evidence of validity is based on item content and the correlation of a scale's score with other measures, which subsequently raised doubts about the level of reliability and validity. Finally, it examines the operation of individual items to determine the effectiveness of these items for the target population and their contribution to the overall measurement of the construct. More complicating matters, the measurement problem through survey methods using questionnaires involved respondents' self-reporting of their perceptions which caused many biased responses (Bradley et al., 2015; Zlatkin-Troitschanskaia et al., 2015).

In the meantime, the Rasch analysis based on Item Response Theory by (IRT) Embretson & Reise (2000) provides a highly effective alternative for exploring the psychometric properties of measurements and calculating bias responses (Bradley et al., 2015). The original Rasch model was developed for dichotomously printed items (correct or incorrect items) based on the early work of (Thurstone & Chave, 1929). Unlike in CTT, where standard measurement errors are considered the same across all testers and dependent on the sample, in IRT, measurement errors are considered different between individuals and do not depend on a particular sample of respondents. Estimates of latent properties were measured based on respondents' and items' characteristics, while respondents' ability and item disability were measured on a similar scale (logit). Therefore, the researcher used IRT-based analysis to determine whether the item's disability corresponded to a person's level of ability on the properties of the development construct. By matching item disability with one's ability more accurately, IRT allows researchers to develop measurements with higher score reliability by using fewer test items.

The Study

This study aims to develop a mathematical intervention instrument based on the learning style of students with learning disabilities. This study involved special education teachers in determining the instrument's validity and reliability through the following diagnoses:
i. Reliability and item-respondent separation
ii. Item polarity
iii. The fit between the measurement item and the construct
iv. Determining dependent items based on standardised residual correlation values.
v. Uniformity of dimensions
vi. Validity of the best rating scale

Methodology

A questionnaire instrument was used to survey 140 special education (learning disability) teachers. Respondents were selected using purposive sampling. The study specifically selected special education(learning disability) teachers from the State of Melaka based on the specified characteristics for this study (Noraini, 2013). The Fuzzy Delphi method was implemented when designing and developing the instrument. The process involved 11 special education teachers, officers from the district education office, lecturers from the Institute of Teacher Education Malaysia (IPGM), and public universities. The Delphi method allows validity to be measured based on the Delphi experts' validation of the constructs developed according to the researcher's interpretation and categorisation (Okoli & Pawlowski, 2004; Skinner et al., 2015). In this regard, instruments constructed with the help of individual or group expertise will have a high level of validity and reliability (Gay et al., 2012; Rubin & Babbie, 2005). The Fuzzy Delphi questionnaire instrument was reviewed based on consensus, comments, and views among the panel. These appointed experts can resolve issues identified due to their capability to give an objective opinion on the issue, rather than based on mutual consensus (Hsu & Sandford, 2007; Yousuf, 2007).

Table 1. Construct and Number of Test Items Developed

No	Construct	Item
1	Screening	1-6 (6 item)
2	Progress Monitoring	7-14 (8 item)
3	Data-Based Decision	15-26 (12 item)
4	Learning Styles	27-39 (13 item)

Findings and Discussion

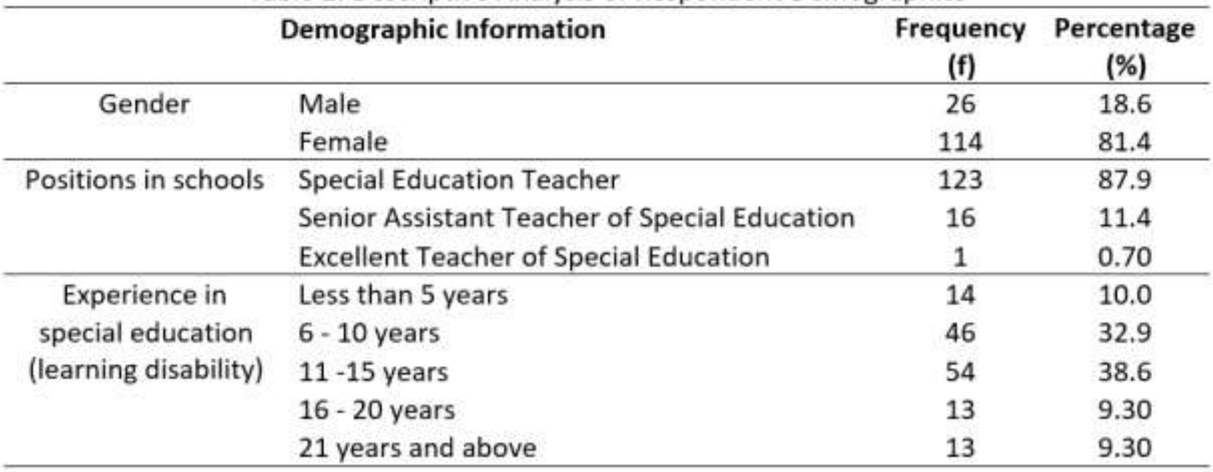
This section discusses the findings on the respondents' profiles. The data were analysed using SPSS V26, and the functional examination of items was conducted using Winstep 4.8.0.0.

Respondents' Profile

The distribution of data presented in Table 2 shows that 26 respondents (18.6%) are male, while 114 respondents (81.4%) are female. In terms of their position in schools, 123 respondents (87.9%) are primary special education (learning disability) teachers, 16 respondents (11.4%) are senior assistant teachers of primary special education schools, and 1 respondent (0.7%) is an excellent teacher for primary-level special education. In terms of their teaching experience, 14 (10.0%) respondents have been teaching special education for primary students with learning disabilities, 46 (32.9%) respondents have been teaching for 6 to 10 years, 54 (38.6%) respondents have been teaching for 11 to 15 years, 13 respondents (9.3%) have been teaching for 16 to 20 years, and 13 people (9.3%) have been teaching for 21 years and more. As shown, 26 respondents (18.6%) are male, while 114 respondents (81.4%) are female.

This study recommends that further studies administer the IMGaP instrument with a large enough sample involving special education (learning disabilities) teachers in various states in Malaysia. Contributions of male special education (learning disabilities) teachers should also be given attention and consideration in future studies. In doing so, it is expected that the respondents' gender distribution could be significantly balanced. Linacre (1994) presented guidelines to choose the appropriate sample size and recommended a minimum of 10 respondents for each scale point to achieve sufficient statistical strength. In this study, at least 60 respondents were required for each of the four conditions or 240 total respondents. This sample size allows item counting accuracy in +/- 1/2 logit (α \ 0.05). As decisions are based on measurement results are more significant, the desired measurement accuracy should be greater. Linacre (1994) suggested that the maximum number of respondents to achieve an accurate decision is 500.

Table 2. Descriptive Analysis of Respondent Demographics



Demographic Information		Frequency (f)	Percentage (%)
Gender	Male	26	18.6
	Female	114	81.4
Positions in schools	Special Education Teacher	123	87.9
	Senior Assistant Teacher of Special Education	16	11.4
	Excellent Teacher of Special Education	1	0.70
Experience in special education (learning disability)	Less than 5 years	14	10.0
	6 - 10 years	46	32.9
	11 -15 years	54	38.6
	16 - 20 years	13	9.30
	21 years and above	13	9.30

Item Functional Inspection

The findings of the survey were analysed using Winsteps software through the Rasch measurement model approach. The researcher performed a functional inspection of the items' (i) Reliability and item-respondent separation, (ii) detecting the polarity of the items measuring the construct based on the PTMEA CORR value, (iii) detecting the suitability of the items (item fit) measuring the construct, (iv) determining the dependent items based on standardised residual correlation values, (v) measuring dimensional uniformity (unidimensionality) using the Residual Principal Component Analysis (PCA) technique, and (vi) the validity of the best rating scale. This diagnosis depends on the needs of the study that is selecting and filtering quality items from the tested items. Items that do not meet the characteristics of the analysis will be repaired or dropped. The explanation for each item's functional inspection is as follows:

a. Reliability

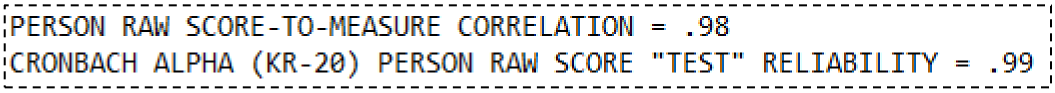
Before an instrument is administered in an actual study, each item's validity and reliability need to be measured to ensure the quality of the instrument and the data obtained. The examination started by removing the item-respondent data with extreme values (outlier) and performing a misfit person removal procedure involving removing respondents with an entry value above the better fitting omitted line or OUTFIT value MS> 2.0. Thus, referring to the results of both tests, after 39 outliers were eliminated, the number of respondents was 101 (N = 101). Then, using the Rasch measurement model approach (Bond & Fox, 2015; Boone, 2016), the functionality of each item was measured based on (1) reliability and item-respondent separation; (2) polarity of items measuring constructs based on PTMEA CORR values; (3) the fit of the construct measuring item; and (4) dependent items based on standardised residual correlation values. According to the Rasch measurement model approach, the acceptable reliability value based on Cronbach's Alpha (α) is between 0.71–0.99 (best level), as shown in Table 3 (Bond & Fox, 2015). This diagnosis depends on the study's requirement in selecting the best item from the tested items. The explanation for each item's functional inspection is shown below.

Table 3. Interpretation of Alpha-Cronbach Scores

Cronbach's Alpha Score	Reliability
0.8-1.0	Very good and effective with a high degree of consistency
0.7-0.8	Good and acceptable
0.6-0.7	Acceptable
<0.6	The item needs to be repaired
<0.5	Items need to be dropped

The analysis found that the reliability value based on the Cronbach's Alpha (α) is 0.99, as illustrated in Table 4. This result clearly shows that the instrument used is in very good condition. Furthermore, it shows high effectiveness with a high level of consistency. Hence, it can be used in actual research.

Table 4. Reliability Value (Cronbach Alpha)

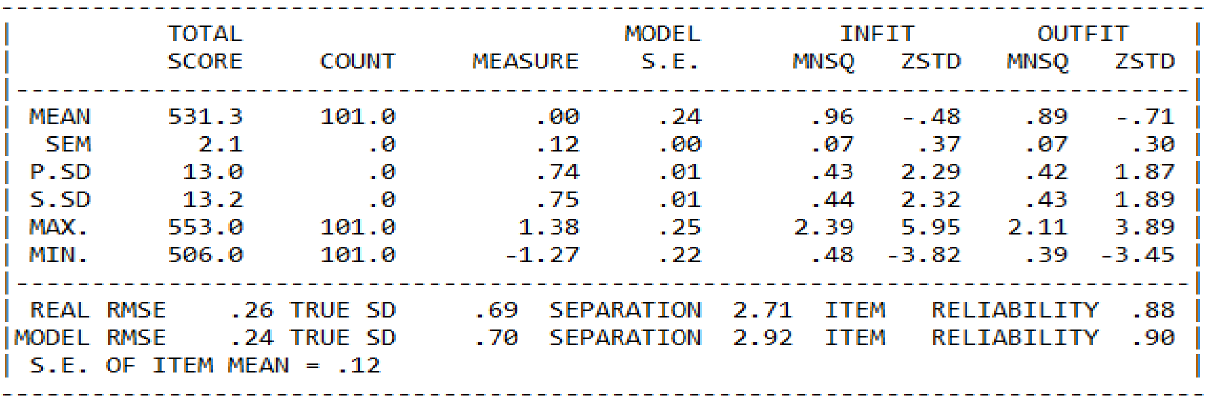


```
PERSON RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .99
```

Next, an overall analysis of the instrument is carried out by looking at the reliability index and the separation of items and respondents. Table 5 shows the item reliability value is 0.88 while the item separation value is 2.71. This result indicates that the item reliability index is very good and effective with a high level of consistency as it approaches the value of 1.0. In this regard, there is a high expectation for the construct to be repeated if administered to another group of respondents with similar abilities (Bond & Fox, 2015). Meanwhile, as the separation index exceeds 2.0 (Bond & Fox, 2015) at 2.71, the items were statistically divided into three strata or levels of agreement.

Table 6 shows that the respondents' reliability value is 0.99, and the respondents' separation value is 9.44. This result indicates a high and good respondents ' reliability value. Bond & Fox (2015) explained that reliability values exceeding 0.8 are good and strongly accepted. In the meantime, the respondents' separation value indicates 9 levels of respondents' ability to agree on items. A good separation value against the item disability level is in line with Linacre (2004), which that explained a separation value greater than 2.0 is a good value.

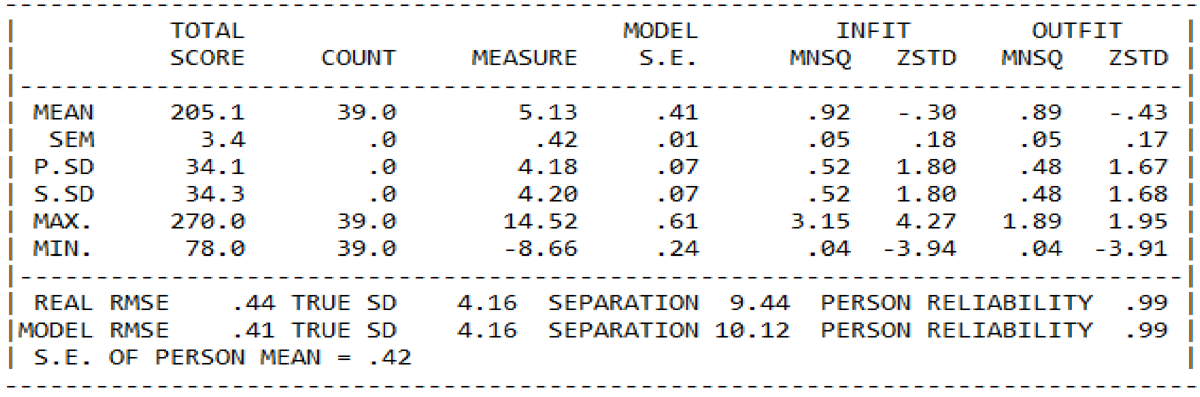
Table 5. Reliability Values and Item Separation for the Instrument Construct



```
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN     531.3      101.0         .00    .24        .96   -.48    .89   -.71 |
|  SEM       2.1         .0         .12    .00        .07    .37    .07    .30 |
| P.SD      13.0         .0         .74    .01        .43   2.29    .42   1.87 |
| S.SD      13.2         .0         .75    .01        .44   2.32    .43   1.89 |
| MAX.     553.0      101.0        1.38    .25       2.39   5.95   2.11   3.89 |
| MIN.     506.0      101.0       -1.27    .22        .48  -3.82    .39  -3.45 |
|------------------------------------------------------------------------------|
| REAL RMSE     .26 TRUE SD     .69  SEPARATION  2.71  ITEM    RELIABILITY  .88 |
|MODEL RMSE     .24 TRUE SD     .70  SEPARATION  2.92  ITEM    RELIABILITY  .90 |
| S.E. OF ITEM MEAN = .12                                                      |
-------------------------------------------------------------------------------
```

Table 6

Reliability Values and Respondent Separation for the Instrument Construct



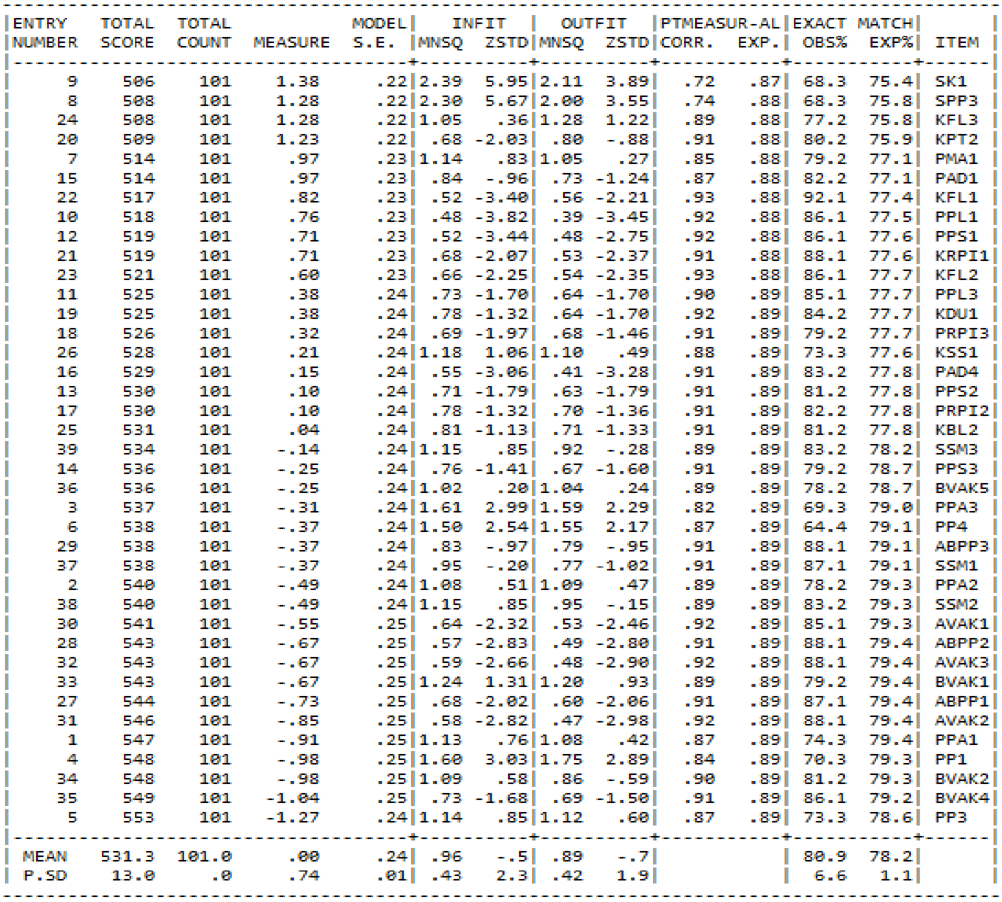
```
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE     S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     205.1      39.0       5.13      .41       .92   -.30    .89   -.43 |
|  SEM       3.4        .0        .42      .01       .05    .18    .05    .17 |
| P.SD      34.1        .0       4.18      .07       .52   1.80    .48   1.67 |
| S.SD      34.3        .0       4.20      .07       .52   1.80    .48   1.68 |
| MAX.     270.0      39.0      14.52      .61      3.15   4.27   1.89   1.95 |
| MIN.      78.0      39.0      -8.66      .24       .04  -3.94    .04  -3.91 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .44 TRUE SD    4.16  SEPARATION  9.44  PERSON RELIABILITY  .99 |
|MODEL RMSE     .41 TRUE SD    4.16  SEPARATION 10.12  PERSON RELIABILITY  .99 |
| S.E. OF PERSON MEAN = .42                                                   |
-------------------------------------------------------------------------------
```

b. Item Polarity Detection

Item polarity analysis (PTMEA CORR) is a very important basic procedure to produce true items in parallel with other items to measure the construct to be measured. In this regard, all items used are functioning in a parallel direction when the measure exhibits a positive index for all items. On the other hand, if a negative index is obtained, the researcher needs to re-examine the data to determine whether it needs to be refined or dropped. Table 7 shows the results of the polarity analysis for each construct. Each construct's polarity or item correlation measurement point is between 0.72 to 0.92, while no item in the PTMEA CORR section shows a negative value, indicating an encouraging result. This means that a positive value indicates the items are functioning in the same direction, in line with the measured construct (Linacre, 2002). Thus, these items statistically indicate that it is moving in a set direction, and out of the total 39 items, no item was dropped.

Table 7

Item Polarity and Item Fit Suitability



```
-----------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL             MODEL|   INFIT   |  OUTFIT   |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE   S.E.  |MNSQ   ZSTD|MNSQ   ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|--------------------------------------+-----------+-----------+-----------+-----------+------|
|     9    506    101    1.38       .22|2.39   5.95|2.11   3.89|  .72   .87| 68.3  75.4| SK1  |
|     8    508    101    1.28       .22|2.30   5.67|2.00   3.55|  .74   .88| 68.3  75.8| SPP3 |
|    24    508    101    1.28       .22|1.05    .36|1.28   1.22|  .89   .88| 77.2  75.8| KFL3 |
|    20    509    101    1.23       .22| .68  -2.03| .80   -.88|  .91   .88| 80.2  75.9| KPT2 |
|     7    514    101     .97       .23|1.14    .83|1.05    .27|  .85   .88| 79.2  77.1| PMA1 |
|    15    514    101     .97       .23| .84   -.96| .73  -1.24|  .87   .88| 82.2  77.1| PAD1 |
|    22    517    101     .82       .23| .52  -3.40| .56  -2.21|  .93   .88| 92.1  77.4| KFL1 |
|    10    518    101     .76       .23| .48  -3.82| .39  -3.45|  .92   .88| 86.1  77.5| PPL1 |
|    12    519    101     .71       .23| .52  -3.44| .48  -2.75|  .92   .88| 86.1  77.6| PPS1 |
|    21    519    101     .71       .23| .68  -2.07| .53  -2.37|  .91   .88| 88.1  77.6| KRPI1|
|    23    521    101     .60       .23| .66  -2.25| .54  -2.35|  .93   .88| 86.1  77.7| KFL2 |
|    11    525    101     .38       .24| .73  -1.70| .64  -1.70|  .90   .89| 85.1  77.7| PPL3 |
|    19    525    101     .38       .24| .78  -1.32| .64  -1.70|  .92   .89| 84.2  77.7| KDU1 |
|    18    526    101     .32       .24| .69  -1.97| .68  -1.46|  .91   .89| 79.2  77.7| PRPI3|
|    26    528    101     .21       .24|1.18   1.06|1.10    .49|  .88   .89| 73.3  77.6| KSS1 |
|    16    529    101     .15       .24| .55  -3.06| .41  -3.28|  .91   .89| 83.2  77.8| PAD4 |
|    13    530    101     .10       .24| .71  -1.79| .63  -1.79|  .91   .89| 81.2  77.8| PPS2 |
|    17    530    101     .10       .24| .78  -1.32| .70  -1.36|  .91   .89| 82.2  77.8| PRPI2|
|    25    531    101     .04       .24| .81  -1.13| .71  -1.33|  .91   .89| 81.2  77.8| KBL2 |
|    39    534    101    -.14       .24|1.15    .85| .92   -.28|  .89   .89| 83.2  78.2| SSM3 |
|    14    536    101    -.25       .24| .76  -1.41| .67  -1.60|  .91   .89| 79.2  78.7| PPS3 |
|    36    536    101    -.25       .24|1.02    .20|1.04    .24|  .89   .89| 78.2  78.7| BVAK5|
|     3    537    101    -.31       .24|1.61   2.99|1.59   2.29|  .82   .89| 69.3  79.0| PPA3 |
|     6    538    101    -.37       .24|1.50   2.54|1.55   2.17|  .87   .89| 64.4  79.1| PP4  |
|    29    538    101    -.37       .24| .83   -.97| .79   -.95|  .91   .89| 88.1  79.1| ABPP3|
|    37    538    101    -.37       .24| .95   -.20| .77  -1.02|  .91   .89| 87.1  79.1| SSM1 |
|     2    540    101    -.49       .24|1.08    .51|1.09    .47|  .89   .89| 78.2  79.3| PPA2 |
|    38    540    101    -.49       .24|1.15    .85| .95   -.15|  .89   .89| 83.2  79.3| SSM2 |
|    30    541    101    -.55       .25| .64  -2.32| .53  -2.46|  .92   .89| 85.1  79.3| AVAK1|
|    28    543    101    -.67       .25| .57  -2.83| .49  -2.80|  .91   .89| 88.1  79.4| ABPP2|
|    32    543    101    -.67       .25| .59  -2.66| .48  -2.90|  .92   .89| 88.1  79.4| AVAK3|
|    33    543    101    -.67       .25|1.24   1.31|1.20    .93|  .89   .89| 79.2  79.4| BVAK1|
|    27    544    101    -.73       .25| .68  -2.02| .60  -2.06|  .91   .89| 87.1  79.4| ABPP1|
|    31    546    101    -.85       .25| .58  -2.82| .47  -2.98|  .92   .89| 88.1  79.4| AVAK2|
|     1    547    101    -.91       .25|1.13    .76|1.08    .42|  .87   .89| 74.3  79.4| PPA1 |
|     4    548    101    -.98       .25|1.60   3.03|1.75   2.89|  .84   .89| 70.3  79.3| PP1  |
|    34    548    101    -.98       .25|1.09    .58| .86   -.59|  .90   .89| 81.2  79.3| BVAK2|
|    35    549    101   -1.04       .25| .73  -1.68| .69  -1.50|  .91   .89| 86.1  79.2| BVAK4|
|     5    553    101   -1.27       .24|1.14    .85|1.12    .60|  .87   .89| 73.3  78.6| PP3  |
|--------------------------------------+-----------+-----------+-----------+-----------+------|
| MEAN    531.3  101.0     .00       .24| .96    -.5| .89    -.7|           | 80.9  78.2|      |
| P.SD     13.0     .0     .74       .01| .43    2.3| .42    1.9|           |  6.6   1.1|      |
-----------------------------------------------------------------------------------------------
```

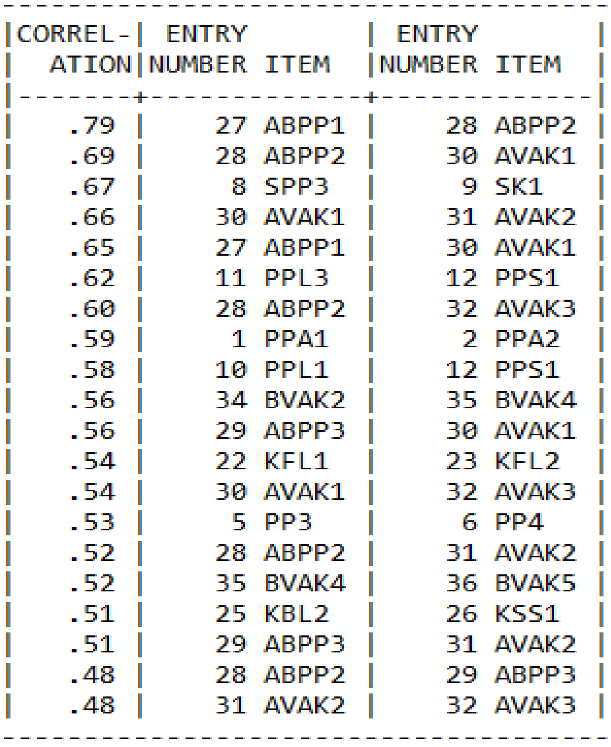
c. Item Suitability (Item Fit)

Each item’s suitability in measuring the developed constructs could be determined through the Mean-Square outfit index (MNSQ). Boone (2016) described that an item’s suitability range or productive MNSQ value should be between 0.5 and 1.5. The values of MNSQ items outside of the MNSQ range normally indicate high Z-STD values surpassing the accepted range of -2.0 <Zstd <+2.0. This step is important to ensure that the items developed are suitable for measuring the study’s constructs. An item with an MNSQ value exceeding 1.5 logits is considered confusing and difficult to answer by respondents. Meanwhile, an item with an MNSQ value less than 0.5 logit indicates that the item is too easy or could be easily guessed by the respondents. If this condition is not met, the item can be refined or dropped. As shown in Table 5, 8 items (SK1 (2.11), SPP3 (2.00), AVAK2 (0.47), AVAK3 (0.48), ABPP2 (0.49), PAD4 (0.41) , PPS1 (0.48) and PPL1 (0.39)) are not within the set range. This means that they need to be refined or dropped from the questionnaire. All items were refined based on the researchers’ requirements and the experts’ views. After examining the eight items in terms of their common features, sentence structure, and language, the experts asserted that these items were deemed relevant under the construct. It was also found that all items have unique strengths and significance to their respective constructs.

d. Detecting Standardised Residual Correlation Values

Measuring the Standardised Residual Correlation can determine local dependence, specifically whether the item is dependent on other items. The residual correlation values should be referred to To identify overlapping items. As shown in Table 8, the high residual correlation for the two items indicates that the items are dependent. This is because they have similar characteristics or because the two combine several other shared dimensions. If the correlation value of two items exceeds 0.7, the correlation value is high, and only one item is required for measurement (Linacre, 2021) as there is a pair of items with a high correlation value at 0.79. This means that these items have the same measurement meaning or are combined with several other dimensions. Therefore, this item was reviewed, and each pair involved would be dropped. However, based on the experts' agreements, this item was refined due to its importance in measuring the construct of this study. Thus, the standardised residual correlation values obtained showed that none of the respondents viewed the item pairs as confusing and combined with other items.

Table 8
Standardized Residual Correlation Values

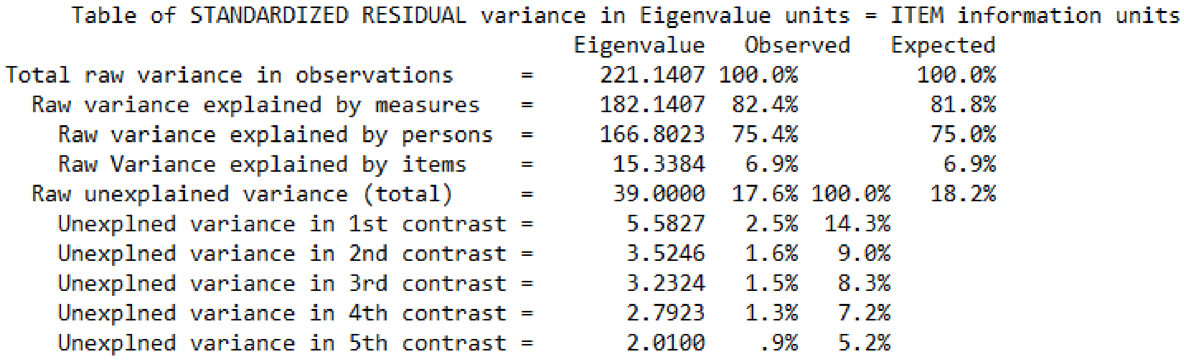


```
------------------------------------
|CORREL-| ENTRY        | ENTRY        |
| ATION|NUMBER ITEM   |NUMBER ITEM   |
|-------+-------------+-------------|
|   .79 |     27 ABPP1 |     28 ABPP2 |
|   .69 |     28 ABPP2 |     30 AVAK1 |
|   .67 |      8 SPP3  |      9 SK1   |
|   .66 |     30 AVAK1 |     31 AVAK2 |
|   .65 |     27 ABPP1 |     30 AVAK1 |
|   .62 |     11 PPL3  |     12 PPS1  |
|   .60 |     28 ABPP2 |     32 AVAK3 |
|   .59 |      1 PPA1  |      2 PPA2  |
|   .58 |     10 PPL1  |     12 PPS1  |
|   .56 |     34 BVAK2 |     35 BVAK4 |
|   .56 |     29 ABPP3 |     30 AVAK1 |
|   .54 |     22 KFL1  |     23 KFL2  |
|   .54 |     30 AVAK1 |     32 AVAK3 |
|   .53 |      5 PP3   |      6 PP4   |
|   .52 |     28 ABPP2 |     31 AVAK2 |
|   .52 |     35 BVAK4 |     36 BVAK5 |
|   .51 |     25 KBL2  |     26 KSS1  |
|   .51 |     29 ABPP3 |     31 AVAK2 |
|   .48 |     28 ABPP2 |     29 ABPP3 |
|   .48 |     31 AVAK2 |     32 AVAK3 |
------------------------------------
```

e. Measuring Unidimentionality

Dimensional uniformity is critical in determining whether an instrument can measure in one direction and form (Aziz et al. 2013). Ambiguous and confusing items need to be reviewed and refined to ensure the instrument provides a robust and achievable measurement. In this light, Rasch analysis using the Principal Component Analysis of Residuals (PCA) technique helps determine the dimensions of a data set. It can detect an instrument's capability in one dimension align with the acceptable level of item interference (Bond & Fox, 2015). In the meantime, unidimensionality assumptions need to be proven so that the data collected are consistent or unidirectional to form a pattern. Conditions proving unidimensionality assumption (Wright & Masters, 1982) include 1) variance explained by measures 40% or more and (2) Unexplained variance in 1st, 2nd, 3rd and 4th factors less than 5 or 10%. In this regard, without the implicit assumption of dimensional uniformity, testing of the relationship between aggregate test scores cannot be performed. Linacre (2002) emphasised that the optimal value of variance is> 60%. However, each construct shown in the raw variance has fulfilled the instrument uniformity requirement of almost 7%. Table 9 shows that it has reached 82.4%, exceeding the limit of 40%. In the meantime, the variance value not explained in contrast 1 is as high as 2.5%, which is well controlled and far from the ceiling value of 15%.

Table 9
Unidimentionality (Standardised Residual Variance)



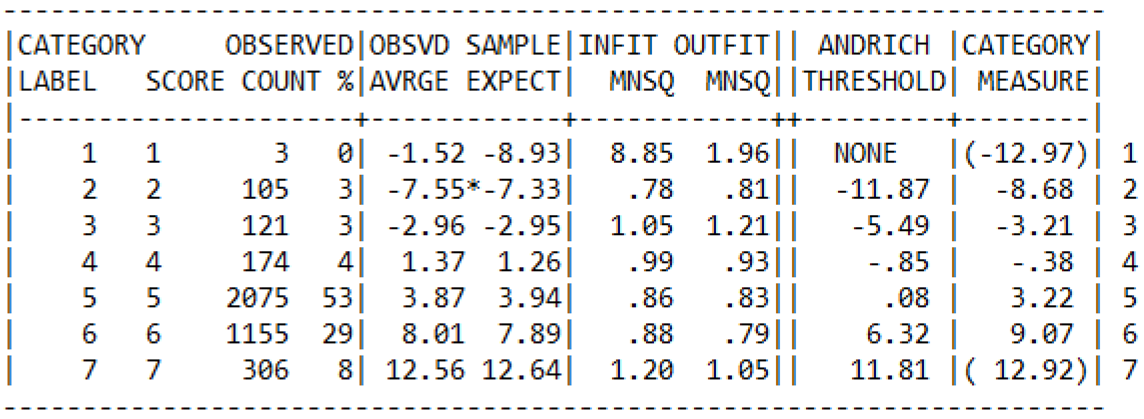
```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                          Eigenvalue   Observed   Expected
Total raw variance in observations     =     221.1407 100.0%         100.0%
  Raw variance explained by measures   =     182.1407  82.4%          81.8%
    Raw variance explained by persons  =     166.8023  75.4%          75.0%
    Raw Variance explained by items    =      15.3384   6.9%           6.9%
  Raw unexplained variance (total)     =      39.0000  17.6% 100.0%   18.2%
    Unexplned variance in 1st contrast =       5.5827   2.5%  14.3%
    Unexplned variance in 2nd contrast =       3.5246   1.6%   9.0%
    Unexplned variance in 3rd contrast =       3.2324   1.5%   8.3%
    Unexplned variance in 4th contrast =       2.7923   1.3%   7.2%
    Unexplned variance in 5th contrast =       2.0100    .9%   5.2%
```

f. Measuring the Best Rating Scale Validity

The process of isolating the mismatched raw data was carried out through match analysis. This step ensures that calibrated data-based analysis is carried out so that the resulting scale works well in forming a response in line with the expected scale increase. A nearby scale should be included if the mean measure of the category is not significantly different (<1.4) and does not show any improvement. The scale needs to be separated if the scale value is (> 5.0). In this study, a 7-point Likert scale was used ranging from (1) Very strongly disagree; (2) Strongly disagree; (3) Disagree; (4) Moderately agree, (5) Agree, (6) Strongly agree, and (7) Very strongly agree. This type of ranking scale provides an opportunity for respondents to mark the level based on their perceptions (Najib, 2003). Some scholars have assumed that respondents accurately perceived the construct, evaluated items according to reproducible criteria, and recorded their evaluations accurately within a uniform scale range. Yet, respondents' perceptions in a survey study are usually based on personal variable criteria. Often, they are not interpreted as intended or properly recorded (Bradley et al., 2015).

The Andrich Rating Scale Model allows a systematic diagnosis of fit items for each response option to indicate each item is functioning optimally to accurately measure the construct. The findings presented in Table 10 show the difference between the 1.4 to 5.0 on the Andrich Threshold section. Thus, the scales expected and used for each construct were deemed appropriate, and the scales do not need to be separated or summarised. Furthermore, the Observed Average section proves that the response pattern is normal. This is evident through the regular increase from negative to positive values for all four constructs. Thus, the validity of this scale is verified and confirms that the scale selection for each construct is appropriate and answers could be spread equally between the scale set.

Table 10. Rating Scale Calibration Structure for Constructs



```
-------------------------------------------------------------------
|CATEGORY     OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL    SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|----------------------+------------+------------++---------+--------|
|    1   1       3   0| -1.52 -8.93|  8.85  1.96||  NONE   |(-12.97)| 1
|    2   2     105   3| -7.55*-7.33|   .78   .81|| -11.87  |  -8.68 | 2
|    3   3     121   3| -2.96 -2.95|  1.05  1.21||  -5.49  |  -3.21 | 3
|    4   4     174   4|  1.37  1.26|   .99   .93||   -.85  |   -.38 | 4
|    5   5    2075  53|  3.87  3.94|   .86   .83||    .08  |   3.22 | 5
|    6   6    1155  29|  8.01  7.89|   .88   .79||   6.32  |   9.07 | 6
|    7   7     306   8| 12.56 12.64|  1.20  1.05||  11.81  |( 12.92)| 7
-------------------------------------------------------------------
```

Conclusion

This article described how the Rasch Measurement Model could be used to empirically examine the psychometric properties and quality of assessment scales. A systematic process was carried out by collecting and analyzing data and comparing the results with specific criteria that have been determined so that the researcher can conclude the quality of the evaluation scales and items of the IMGaP instrument. The functional inspection of items should be done to increase the validity and reliability of a measuring instrument by implementing procedures such as (1) Inspection of item and respondent separation index so that items in the measuring instrument have item levels and respondent abilities to widen the distribution, (2) Items developed should have unidimensional properties. Moreover, it indicates that a good item is an item that does not measure other dimensions, and (3) A combination of scale measurement categories should be done to create a meaningful functionality of the measurement category. Researchers can repeatedly use information from these analytical steps to review, revise, and refine measurements until they reach the level of measurement accuracy required so that the right decision can be made. In the end, the empirical results of the Rasch analysis of this study are combined with the evaluation of an appointed expert to determine the best path to take. Thus, based on the IMGaP instrument evaluation, all items are valid and fair to measure the construct of mathematical intervention based on the learning style of students with learning disabilities. Findings from the analysis have successfully proved statistically that these items have a high level of validity and reliability and should be used on special education teachers of learning disabilities in determining the importance of considering learning style as one of the important factors when conducting mathematics interventions.

Acknowledgment

We thank PP-FPEND-2021 for the support. The first author’s grateful thank is extended to the Ministry of Education for the scholarship awarded for pursing doctoral degree and sincere gratitude is hereby extended to Dr. Norshidah Mohamad Salleh and Dr. Mohd Mokhtar Tahar, as supervisors, both of whom have unceasingly helped in the completion of this paper.

Corresponding Author

Dr. Norshidah Mohamad Salleh
National University of Malaysia, Malaysia, Centre Of Education and Community Wellbeing,
Faculty of Education, National University of Malaysia, 43600 Bangi, Selangor.
Email: nshidah@ukm.edu.my

References

Bond, T., & Fox, C. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd Ed). Routledge.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education*, *15*(4), rm4.

Boström, L. (2011). Students' Learning Styles Compared with Their Teachers' Learning Styles in Upper Secondary School – a Mismatched Combination. *Education Inquiry*, *2*.

Bradley, K., Peabody, M., Akers, K., & Knutson, N. (2015). Rating Scales in Survey Research: Using the Rasch Model to Illustrate the Middle Category Measurement Flaw. *Survey Practice*, *8*(1), 1–12.

Clements, D. H., & Sarama, J. (2007). *Effects of a Preschool Mathematics Curriculum: Summative Research on the Building Blocks Project*. *38*(2), 136–163.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. L. Erlbaum Associates.

Fuchs, L. S., & Fuchs, D. (2009). On the Importance of a Unified Model of Responsiveness-To-Intervention. *Child Development Perspectives*, *3*(1), 41–43.

Fuchs, L. S., Fuchs, D., Yazdian, L., & Powell, S. R. (2002). Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies. *School Psychology Review*, *31*(4), 569–583.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2012). *Educational Research: Competencies for Analysis and Applications* (10th ed). Pearson Education.

Geary, D. C. (2013). Early Foundations for Mathematics Learning and Their Relations to Learning Disabilities. *Current Directions in Psychological Science*, *22*(1), 23–27.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics Instruction for Students With Learning Disabilities: A Meta-Analysis of Instructional Components. *Review of Educational Research*, *79*(3), 1202–1242.

Gersten, R., Jordan, N., & Flojo, J. (2005). Early Identification and Interventions for Students With Mathematics Difficulties. *Journal of Learning Disabilities*, *38*, 293–304.

Hsu, C.-C., & Sandford, B. (2007). The Delphi Technique: Making Sense Of Consensus. *Practical Assessment, Research and Evaluation*, *12*(10), 1–8.

Idris, N. (2013). *Penyelidikan dalam Pendidikan*. McGraw-Hill Education.

Jacobs, R. L. (2011). Developing a Research Problem and Purpose Statement. In *Dlm The Handbook of Scholarly Writing and Publishing, T. S. Rocco and T. Hatcher, San Francisco: Jossey-Bass* (pp. 125–141).

Jaspers, K. E., McCleary, D. F., McCleary, L. N., & Skinner, C. H. (2017). Evidence-based Interventions for Math Disabilities in Children and Adolescents. In *L. A. Theodore. Handbook of Evidence-based Interventions for Children and Adolescents.* (pp. 99–110). Springer Publishing Company.

Kementerian Pendidikan Malaysia. (2015). *Buku Panduan Pengoperasian Program Pendidikan Khas Integrasi* (p. hlm 44). Bahagian Pendidikan Khas, KPM.

Kroesbergen, E., & Luit, J. E. H. (2003). Mathematics Interventions for Children with Special Educational Needs: A Meta-Analysis. In *Remedial and Special Education* (Vol. 24).

Lemons, C., Powell, S., King, S., & Davidson, K. (2015). Mathematics interventions for children and adolescents with Down syndrome: A research synthesis. *Journal of Intellectual Disability Research*, *59*(8), 767–783.

Linacre, J. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*, *7*, 328.

Linacre, J. M. (2002). What Do Infit and Outfit, Mean-Square and Standardized Mean? *Rasch Meas Trans*, *16*(2), 878.

Linacre, J. M. (2004). Test Validity and Rasch Measurement: Construct, content, etc. *Rasch Measurement Transactions*, *18*, 970–971.

Linacre, J. M. (2021). *Winsteps® Rasch Measurement Computer Program*. Winsteps.com.

Miles, D. (2017). A Taxonomy of Research Gaps: Identifying and Defining the Seven Research Gaps. *Journal of Research Methods and Strategies*, 1–10.

Mueller-Bloch, C., & Kranz, J. (2015). A Framework for Rigorously Identifying Research Gaps in Qualitative Literature Reviews. *International Conference on Information Systems*, 1–19.

Najib, A. G. M. (2003). *Reka Bentuk Tinjauan Soal Selidik Pendidikan*. Universiti Teknologi Malaysia

Okoli, C., & Pawlowski, S. D. (2004). The Delphi Method as a Research Tool: An Example, Design Considerations and Applications. *Information & Management*, *42*(1), 15–29.

Rubin, A., & Babbie, E. R. (2005). Research Methods for Social Work. In *Research Methods for Social Workers*. Thomson/Brooks/Cole.

Sengodan, V., & Iksan, Z. H. (2012). Students' Learning Styles and Intrinsic Motivation in Learning Mathematics. *Asian Social Science*, *8*(16), 17–23.

Skinner, D., Nelson, R., Chin, W., & Land, L. (2015). The Delphi Method Research Strategy in Studies of Information Systems. *Communications of the Association for Information Systems*, *37*, 31–63.

Sophian, C. (2004). Mathematics for the Future: Developing a Head Start Curriculum to Support Mathematics Learning. *Early Childhood Research Quarterly*, *19*(1), 59–81.

Suhaimin, S. H., & Mohamed, M. (2017). Intervention for Children with Specific Learning Disabilities (SpLD) in Mathematics Disorders: A Framework. *Sains Humanika*, *9*(3–2), 83–90.

Taylor-Cox, J. (2016). *Math Intervention 3–5: Building Number Power with Formative Assessments, Differentiation, and Games, Grades 3–5*. Routledge.

Thurstone, L. L., & Chave, E. J. (1929). *The Measurement of Attitude: A Psychophysical Method and Some Experiments with a Scale for Measuring Attitude Toward the Church*. The University of Chicago Press.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. MESA Press.

Yousuf, M. (2007). Using Experts' Opinions Through Delphi Technique. *Practical Assessment, Research & Evaluation*, *12*(4), 1–8.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The International State of Research on Measurement of Competency in Higher Education. *Studies in Higher Education*, *40*(3), 393–411.