International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 376–394
https://doi.org/10.21449/ijate.1155679
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Classroom assessment that tailor instruction and direct learning: A validation study

Wai Kei Chan [1], **Li Zhang** [2], **Emily Pey-Tee Oon** [3*]

[1]University of Macau, Faculty of Education, Macau, China
[2]University of Macau, Faculty of Education, Macau, China
[3]University of Macau, Faculty of Education, Macau, China

ARTICLE HISTORY

Received: Aug. 04, 2022
Revised: Apr. 26, 2023
Accepted: June. 21, 2023

Keywords:
Classroom assessment,
Constructing measures,
Tailor instruction,
Direct learning.

Abstract: We report the validity of a test instrument that assesses the arithmetic ability of primary students by (a) describing the theoretical model of arithmetic ability assessment using Wilson's (2004) four building blocks of constructing measures and (b) providing empirical evidence for the validation study. The instrument consists of 21 multiple-choice questions that hierarchically evaluate arithmetic intended learning outcomes (ILOs) on arithmetic ability, hierarchically, based on Bloom's cognitive taxonomy for 138 primary three grade students. The theoretical model describes students' arithmetic ability on three distinct levels: solid, developing, and basic. At each level, the model describes the characteristics of the tasks that the students can answer correctly. The analysis shows that the difficulty of the items followed the expected order in the theoretical construct map, where the difficulty of each designed item aligned with the cognitive level of the student, the item difficulty distribution aligned with the structure of the person construct map, and word problems required higher cognitive abilities than the calculation problems did. The findings, however, pointed out that more difficult items can be added to better differentiate students with different ability levels, and an item should be revised to enhance the reliability and validity of the research. We conclude that the conceptualizations of such formative assessments provide meaningful information for teachers to support learning and tailoring instruction.

1. INTRODUCTION

The central purpose of classroom assessment is to provide feedback to improve student learning and teachers' pedagogies (Black & Wiliam, 2010; Dixson & Worrell, 2016; Shepard, 2006; Stiggins, 1994; Wiggins, 1998). However, grading continues to dominate pedagogical practices, even in the context of formative assessment, diluting its effectiveness. Consequently, crucial questions concerning the essential features for collecting, formatting, and acting on evidence of learning through formative assessment remain unanswered.

The situation has not changed much today, as classroom formative assessment practices continue to be counterproductive and incoherently disconnected from each other and from high-stake accountability assessments (Wilson, 2004; NRC, 2006; Gorin & Mislevy, 2013). Contrary

* CONTACT: Emily Pey-Tee Oon peyteeoon@um.edu.mo University of Macau, Faculty of Education, Macau, China

e-ISSN: 2148-7456 /© **IJATE** 2023

to expectations based on the central role of feedback in effective classroom practices (Hattie, 2008; Duckor & Holmberg, 2017), teachers often do not review student responses to formative assessment questions or reflect upon the content measured, such as whether items are ambiguously worded or a scoring criterion is unclear (Black & Wiliam, 1998; Guskey, 2003; Popham, 2009, 2010).

When student scores and grades are the primary focus of assessment and the full value of formative feedback is not obtained, outcome quality is inevitably undercut, even though formative practices have repeatedly been shown to be the most effective means to improve student achievement (Black & Wiliam, 2003; Hattie, 2008). Instructionally-relevant feedback is essential to providing the leverage needed for advancing toward the desired learning outcome (Brookhart et al., 2010).

The purpose of formative assessment is to help teachers identify difficulties obscuring students' conceptual understanding, charting a path forward along a learning progression (Bell & Cowie, 2001; Black et al., 2011; Cardinet, 1989). Formative practices encompass a broad range of qualitative and quantitative assessments of as and learning in the classroom, and are not limited to feedback from formally scored assessments (Baird, et al., 2017; Duckor & Holmberg, 2017; Fisher, 2013). However, without an essential feedback mechanism, formative practices of any kind will fail to produce the desired effect. Properly conceived, designed, and implemented, formative assessment is integrated with instruction, and should be a key tool for monitoring learning progressions and supporting the attainment of learning outcomes (Clark, 2012; Gorin & Mislevy, 2013; Popham, 2009, 2010).

Wilson (2004) proposed constructing measurement instruments using four building blocks: construct map, item design, outcome space, and measurement model, to properly conceive, design, and implement formative practices in classrooms. The blocks ascertained the validity of the formative assessment information produced from the test items. Building blocks are a reference that aid in the assessment design cycle. Each building block focuses on one of the four principles: developmental theory perspective (construct map), a match between instruction and assessment (item design), management by teachers (outcome space), and evidence of high quality (measurement model). When this cycle is reiterative, block coherence is improved because each block's information can optimize other blocks. This model tests construct consistency for objective proof of knowledge, skills, and attitude measurements.

The Rasch measurement model is applied for shortcomings that plagued the classical test theory that it is sample dependent and item dependent (Embretson & Reise, 2000; French, 2001; Hambleton et al., 1991; Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985) that limit the generalizability of research findings (Wright & Master, 1982). Furthermore, Rasch measurement model is applied in the current study to ascertain validity of the test. A number of related studies have reported on scale validity based on content validity, as judged by experts in the relevant domain in science education (e.g. Adillah et al., 2022; Beck, 2020; Hidayati et al., 2019; Luque-Vara et al., 2020; Nasir et al., 2022; Wole et al., 2021). Content validity based solely on professional judgment is insufficient to establish validity (Messick, 1981; 1989). At the most, testing validity merely on content validity is insufficient (Lee & Fisher, 2005) because validity refers to "*the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"* (Messick, 1989, *p*. 13). As quoted by Fisher (1997), "*The conventional focus on content validity has misled us about what is important in educational measurement"*.

Arithmetic skills that involve basic operations of addition, subtraction, multiplication, and division are the foundation for advanced mathematical concepts (Björklund, 2021; Hong Kong Education Bureau, 2000, 2018; Parviainen, 2019; Sievert et al., 2021; Vlassis et al., 2022) such as algebra and geometry (Vlassis et al., 2022). It helps students develop logical thinking and

problem-solving abilities (Björklund, 2021). Thus, arithmetic is an essential component of the primary mathematics curriculum (Engvall et al., 2020). Baroody and Dowker (2003), Dowker (2005), Geary (1993), Goldman and Hasselbring (1997), Hiebert and Lefevre (1986), and Kilpatrick et al. (2001) carried out research on how students attain arithmetic proficiency. However, they did not address the validity issue of the instrument used to measure students' arithmetic abilities.

In the present study, we aimed to provide evidence of the validity of a classroom assessment evaluating the arithmetic ability of primary students based on Wilson 's (2004) four building blocks. We then provide empirical evidence for the validation of the hypothesized model based on Rasch's (1960) measurement model.

1.1. Theoretical Framework

1.1.1. *Construct mapping*

The first building block was a constructed map. This is a diagrammatic representation of the construct specifications. It operationalizes constructs in the successive stages of understanding or abilities. It shows how students' understanding evolves, and how their responses to items might change or develop. The construct assessed in this research was arithmetic ability, defined as accurately solving addition, subtraction, multiplication, and division problems mentally. This includes choosing the correct arithmetic operation and calculating the solution (Millians, 2011).

Students' arithmetic ability can be represented using a person-construct map (Wilson, 2004). Theoretically, it is suggested that students' arithmetic abilities are developed in stages where they start from the basic skills of solving addition and subtraction problems, and then move on to more complex operations of multiplication and division (Hong Kong Education Bureau, 2018). According to the Cognitive Development Theory, students initially use concrete objects and counting strategies to solve arithmetic problems. As they progress through these stages, they develop more abstract and efficient strategies for solving problems. The construct map developed in this study was considered within these contexts.

The person construct map assumes that arithmetic ability is a unidimensional latent variable that extends from low to high (Table 1). Students were categorized into three groups according to their levels of arithmetic ability. At the basic level, students with low ability can remember and understand algorithms (four operations) to solve simple operation problems. At a developing level, mid-range students can analyse information and apply the underlying algorithms to solve less challenging or less complex operational problems. At a solid level, high-ability students can evaluate the information given in words, work out the best operation to solve complex problems in words, and justify their solutions (Table 1).

Table 1. *Person construct map of predicted arithmetic ability levels of primary school students.*

Level	Respondents	Description
3	Solid	➢ Students can evaluate the information given in words and work out addition and subtraction operations to solve mixed operation problems. They can correctly identify a mixed operation to solve two-step problems when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can evaluate the information given in words and work out the best multiplication operation to solve complex problems and justify their solutions. They can accurately identify multiplication to solve and explain two-step problems by multiplying a one-digit number by a one-digit number. (Multiplication construct) ➢ Students can evaluate the information given in words and work out the best division operation to solve complex problems and justify their solutions. They can correctly identify division to solve one-step and two-step problems of the quotient of a one-digit number and explain the reasons. (Division construct)
2	Developing	➢ Students can analyse the question and apply the best strategy to solve addition and subtraction. They can precisely calculate addition with carrying (carrying once and carrying twice) and subtraction with borrowing (borrowing once, borrowing twice and borrowing twice for ones, when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can analyse the information given in numbers and apply the necessary addition and subtraction to solve conceptually less challenging mixed operation problems. Individually, they can calculate mixed operation (addition and subtraction) with and without carrying and borrowing (when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can analyse the information given in numbers and apply the best strategy to solve conceptually less challenging multiplication operation problems. They understand multiplication and can correctly solve problems by multiplying a one-digit number by a one-digit number. (Multiplication construct) ➢ Students can analyse the information given in numbers and apply the best strategy to solve conceptually less challenging division operation problems. They can correctly calculate division with a remainder and the quotient of a one-digit number. (Division construct)
1	Basic	➢ Students can remember and understand addition and subtraction algorithms to solve simple operation problems. They can accurately calculate addition and subtraction with no carrying and borrowing (when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can memorise and understand the multiplication table and use what they recall to solve simple operation problems. Precisely, they can calculate the multiplication of a one-digit number by a one-digit number and use it to solve one-step problems without explanation in words. (Multiplication construct) ➢ Students can memorise and understand division algorithms to solve simple operation problems. They can calculate division without a remainder and the quotient of a one-digit number. (Division construct)

1.1.2. *Item design*

The item design encapsulates the types of items used to provide evidence of students' knowledge and understanding embodied in the theoretical construct map. It guides how the learning outcomes will be measured and aligns the curriculum and assessment using standard conditions. This enabled assessment of each level defined in the construct map (Table 1). A total of 21 multiple choice question (MCQ) items were designed based on the three cognitive knowledge levels (Basic, Developing, and Solid) from the person construct map (Table 2). MCQ tests can save time and reduce grading costs (Alderson, 1990; Liu et al., 2008), can test multiple knowledge domains within the same test (Çataloğlu & Robinett, 2002), enabling more objective grading to ensure the fairness of the test and facilitate item and test analysis, which

can improve teaching and students' learning (Gurel et al., 2015). Therefore, we used MCQs to measure students' attainment of learning outcomes.

Table 2. *Item design.*

Level	Respondents Level	Items	Operation	Problem type	Option type	Bloom's cognitive level
3	Solid	Q13, Q15	Division	Word	Algorithm	Evaluate
		Q4, Q9	Multiplication	Word	Algorithm	Evaluate
		Q21	Mixed	Word	Algorithm	Evaluate
2	Developing	Q5, Q10	Division	Word	Number	Analyse
		Q6, Q12	Multiplication	Word	Number	Analyse
		Q2	Division	Calculation	Number	Apply
		Q19	Mixed	Calculation	Number	Apply
		Q7, Q11, Q16	Subtraction	Calculation	Number	Apply
		Q3, Q17, Q20	Addition	Calculation	Number	Apply
1	Basic	Q1	Division	Calculation	Number	Understand
		Q14Q8	Subtraction	Calculation	Number	Understand
		Q18	Addition	Calculation	Number	Understand
			Multiplication	Calculation	Number	Remember

1.1.3. *Outcome space*

The outcome space encapsulates different student response levels for items correlated with the construct level. It guides the assessment of students' responses to items relative to the construct map. Specifically, it can be used as a scoring guide to ensure that student answers align with the constructed map. Teachers assigned scores to an item designed for a particular knowledge level based on the construct map in the outcome space. When the item design is completed, teachers then decide which factors may affect the item response, and classify and score these factors to ensure meaningful student responses. In this study, MCQs were used to design items that were scored dichotomously (Incorrect = 0 and Correct = 1; Wilson, 2004).

1.1.4. *Measurement model*

The measurement model is the framework by which assessors equate student scores on items from specific construct levels and apply the scored responses to the constructs. The assumption is that student scores on individual items align with the knowledge construct map. The resulting model is a measurement or interpretation model (Wilson, 2004). This helps teachers understand and evaluate student responses to the items. The Rasch measurement model was used in the current study. The model transforms the scores into the locations of items in the construct map. It is an objective measurement suitable for various random, hierarchical, and classified data analyses (Linacre, 2000). It was thus used in this study to relate data to assessment targets and construct maps. The output is a Wright map that displays student performance on elements of the construct map and enables comparisons between students. In addition, it places students and items on the same scale (with an arbitrary scale representing a student's chances of a positive response at that position). This, in turn, documents the measurement system and assesses the construct validity (Wilson, 2004, p. 156-157).

1.1.5. *Validation study for the hypothesized model*

We sought to evaluate whether the theoretical person constructs a map of arithmetic ability following the four building blocks (Wilson, 2004) aligns with the statistical results of the Rasch measurement model. Specifically, we examined whether the following were true.

1. The difficulty of each designed item aligns with the cognitive ability of students in the following order: Level 1 ability is less than Level 2 ability, which is less than Level 3 ability (Table 1).
2. Basic students can solve Level 1 items, developing students can solve Level 1 and 2 items, and solid students can solve items from all levels.
3. Arithmetic calculation problems are more accessible to students' ability levels than arithmetic word problems (at Level 3).

2. METHOD

2.1. Participants

A sample of 138 Primary 3 students from a single-gender school in Macao participated in this study. The students were from four different classes taught by three teachers. One teacher taught two classes and the other two teachers taught one class each. The students had completed their Primary 2 mathematics course and had just entered Primary 3.

2.2. Test Instrument

One month before the school year, researchers (first and second authors) reviewed Primary 2 and Primary 3 mathematics curricula, textbooks, and workbooks used by Macau schools. A test instrument was constructed in consultation with the head of the Elementary Mathematics Department.

It consisted of 21 MCQs encompassing four operations to measure the arithmetic ability of elementary students. The elementary mathematics department head with 16 years of teaching experience and three mathematics teachers with teaching experience ranging from 1 to 3 years validated the test instrument. Four MCQs assessed addition, four assessed subtraction, six assessed division, five assessed multiplication, and two assessed mixed addition and subtraction (Table 2). In addition to the classification by operation, the 21 items were also categorized according to Bloom's cognitive levels: remembering, understanding, applying, analysing, and evaluating (Table 2). Furthermore, the 21 items were divided into arithmetic calculation problems (requiring students to apply one or more operations) and arithmetic word problems. There were 12 calculation problems and nine-word problems (Table 2).

Each item has four options, with three distractors and one correct option. The 21 items were grouped into two groups based on the option types. While 16 items required students to pick the correct numerical figure, five required students to select the correct algorithm (Table 2). All 21 MCQs were ordered randomly during the test. Prior to data collection, item difficulty was validated by 16 preservice science and mathematics teachers.

2.3. Data Collection Procedure

The test was administered at the beginning of the school year. It covered the arithmetic that students should have already learned based on The Curriculum Framework for Formal Education of Local Education and The Macao Requirements of Basic Academic Attainments of Local Education System (BAA). In the first week of September 2019, the teachers informed the students of the assessment date and coverage of this assessment. The teachers distributed a test instrument containing 21 items and a scantron answer sheet on the assessment date. The students recorded their answers by shading the box of the chosen option for each MCQ with a pencil. The students completed the test for 20 minutes.

The teachers collected scantron sheets at the end of the designated time. The sheets were provided to the first author, who used an optical mark reader to scan the answer sheets for data input. Each correct answer was recorded as 1, and each incorrect answer was recorded as 0. Thus, the maximum raw score was 21 and the minimum was 0. All student data (class number, item number, and student's answers) and test scores for each item were entered into a spreadsheet.

2.4. Calibration of Item Difficulties and Person Abilities

The Rasch model uses logarithmic transformation to calibrate a person and items on the same single-dimensional ruler (Wright & Masters, 1982). Based on their respective positions on this single-dimensional continuum, comparisons can be made between person and person, item and item, and person and item, yielding objective and linear measures of person's ability and item difficulty. Data were analyzed using Winsteps 4.4.5 and the dichotomous Rasch model (Rasch, 1960) as each MCQ only had a correct or incorrect answer.

$$P_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

The input of this logistic function is the difference between the person's ability θ_n and item difficulty δ_i. The output is the probability that person n correctly answers item i, P_ni1. Therefore, item difficulty and person ability were measured on a standard interval scale. The items had a mean difficulty of 100.00, a minimum of 81.61 (item 8), and a maximum of 119.50 (item 4). The students had a mean ability of 109.27, minimum of 89.99, and maximum of 130.33.

3. RESULTS

3.1. Quality of the Data

Data quality was evaluated by assessing data fit, reliability, and fairness using the Rasch Model of Winsteps (Lincare, 2012) prior to reporting the validation results.

3.1.1. *Fit diagnosis*

In the Rasch model, the expectation is that students with higher ability will have a higher probability of answering more difficult items correctly than students with lower ability who will have a higher probability of correctly answering easier items than difficult items (Wright & Stone, 1999, p. 48). Infit and Outfit evaluate how well the data fit the structure of this expectation.

An item with Zstd greater than 2.00 distorts the model fit because it is a poor fit due to unexpected, unrelated irregularities (Linacre, 2012). Of the 21 items, only Q9, the third most challenging item, had an Infit Zstd >2.00 and an Outfit Zstd >2.00. All other item Infit and Outfit values lie between -1.75 and +1.40, signifying a good fit to the Rasch model.

Items with MnSq from 0.50 to 1.50 are considered productive for measurement (Linacre, 2012). The item with the lowest Outfit MnSq was Q14 (0.65), and the item with the highest Outfit MnSq was Q1 (1.40). The item with the lowest Infit MnSq was Q20 (0.88) and the item with the highest Infit MnSq was Q9 (1.22). Therefore, all items were within the expected range of 0.50 to 1.50 and they usefully fit the Rasch model (Linacre, 2012). In other words, all items were retained for the subsequent analysis.

Q9 had the highest Infit Zstd and Outfit Zstd of the 21 items. This may imply that its wording is ambiguous, its options are misleading, or both.

3.1.2. *Reliability of data*

Item reliability and Pearson's reliability were used to indicate reliability. From easy to difficult, item locations operationally define the latent variable (Wright & Stone, 1999, p. 151). Thus, the items should be appropriately located and separated along a line to assess relative item difficulties and item redundancy gaps.

Item reliability shows how this sample of students separated the 21 items in the assessment. A value closes to 1.00 indicates a higher precision (Wright & Stone, 1999, p. 151). The item

reliability in this assessment was 0.96, indicating that the sample size was sufficient to support the instrument's construct validity (Lincare, 2012).

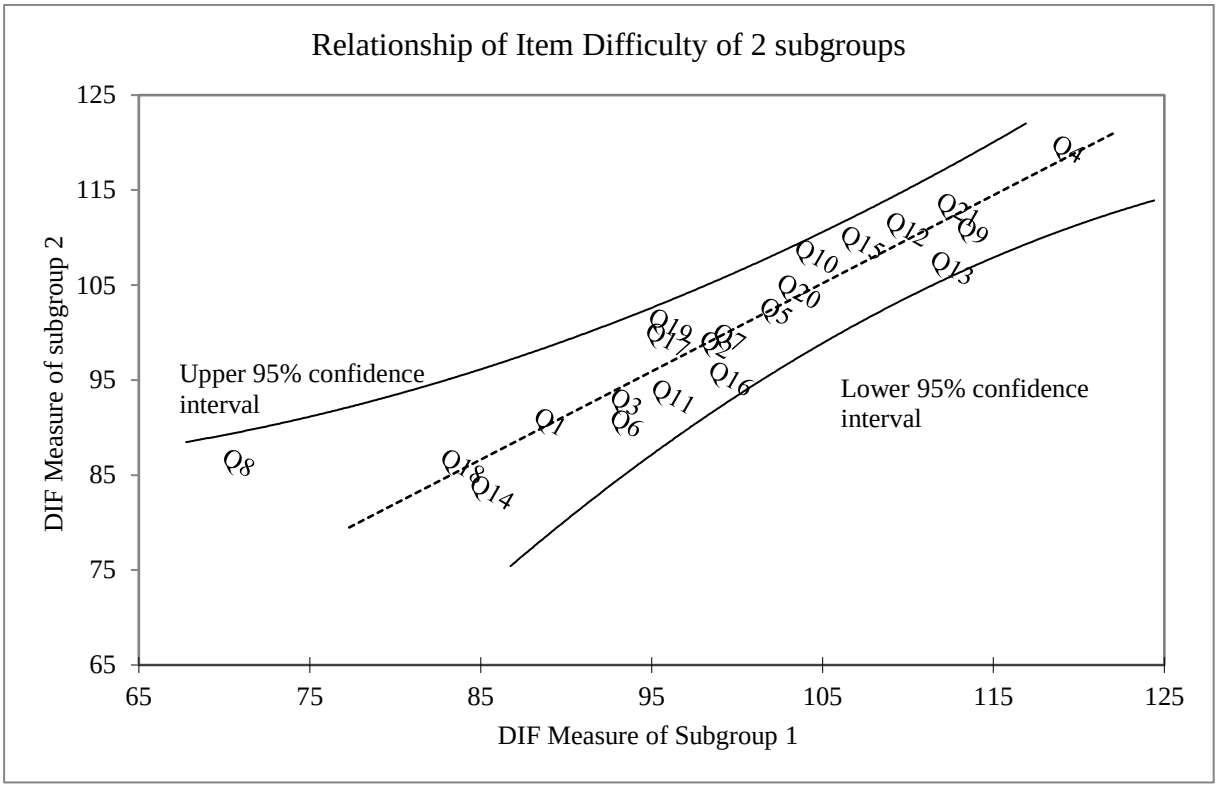
The Pearson reliability is the same as traditional test reliability (Lincare, 2012). This shows how this set of 21 items separated the sample of 138 students. It ranges from 0.00 to 1.00; a value close to 1.00 indicates higher precision (Wright & Stone, 1999, p. 151). The Pearson's reliability for this assessment was 0.60, indicating that the items were not sufficient to classify students into different ability levels. This means that the sample of students had similar ability levels or too few items (Lincare, 2012) to evaluate the latent variable.

3.1.3. *Fairness of data: Differential item functioning*

All items should behave similarly to students with the same abilities. If an item functions differently across different subgroups of students, the validity of this instrument may be questioned (Wilson, 2004). The sample students were randomly assigned to two subgroups and were expected to function the same across these subgroups. In other words, differential item functioning (DIF) should not occur (Bond & Fox, 2015, p. 281-282).

If an item had a p-value greater than 0.05, DIF did not occur significantly across the two subgroups (Linacre, 2009). As the USCALE of this assessment was 8.52, one logit equals 8.52 units (J.M. Linacre, personal communication, 22nd September 2020). Therefore, a DIF may exist if the DIF contrast is greater than 5.43. Q8 and Q13 had DIF contrasts of -15 and 6, respectively. However, because they are still within the 95% confidence interval (Figure 1), their DIF is not considered significant, which means that they behave similarly for different subgroups. The results indicated that the items were fair; therefore, item validity was upheld.

Figure 1. *DIF measures for the two subgroups under 95% confidence interval.*



3.2. Is Item Difficulty for Each Level Aligned with Student Ability in Order?

The Wright map represents item difficulty versus person ability (Figure 2a). The student ability has a range of 90.00 logits to 130.33 logits. The item difficulty ranges from 81.61 logits to 119.50 logits (Figure 2a). Items Q8, Q18, and Q14, which all students could answer, were more

than one standard deviation below the mean item difficulty. Item Q4 was the most difficult and lay almost two standard deviations above the mean item difficulty. Although it is the most difficult item, ten students (7.24% of the sample size) were able to solve it and all the other items. In other words, the 21 designed items were not sufficient to assess the upper limit of the arithmetic ability of the ten students. This indicates that items that are more difficult than Q4 should be included in the instrument. In addition, there is a significant gap between Q4 and Q21. To better assess what students with an ability between 113 and 120 logits can do, more items of intermediate difficulty should be added to fill this gap.

In alignment with Bloom's taxonomy, the four easiest items (Q1, Q14, Q18, and Q8) mainly require students to remember and understand. In contrast, the most difficult items (Q4, Q21, Q9, Q13, and Q15) mainly required students to evaluate. Items with a medium difficulty level mainly require students to analyse and apply (Table 2 and Figure 2b). Observed item difficulty based on Rasch analysis and predicted item difficulty from the person construct map were compared (Table 3) to examine item difficulty alignment. The findings show that item difficulty hierarchy generally aligns with Bloom's taxonomy.

Table 3. *Comparison of alignment between observed items based on Rasch analysis and predicted construct map based on item development.*

Level	Observed item cognitive levels based on Rasch analysis	Predicted construct map based on item development in person construct map	Bloom's cognitive level
3 Solid	Q4, Q9, Q13, Q15, Q21	Q4, Q9, Q13, Q15, Q21	Evaluate
2 Developing	Q2, Q3, Q5, Q6, Q7, Q10, Q11, Q12, Q16, Q17, Q19, Q20	Q2, Q3, Q5, Q6, Q7, Q10, Q11, Q12, Q16, Q17, Q19, Q20	Apply and Analyse
1 Basic	Q1, Q8, Q14, Q18	Q1, Q8, Q14, Q18	Remember and understand

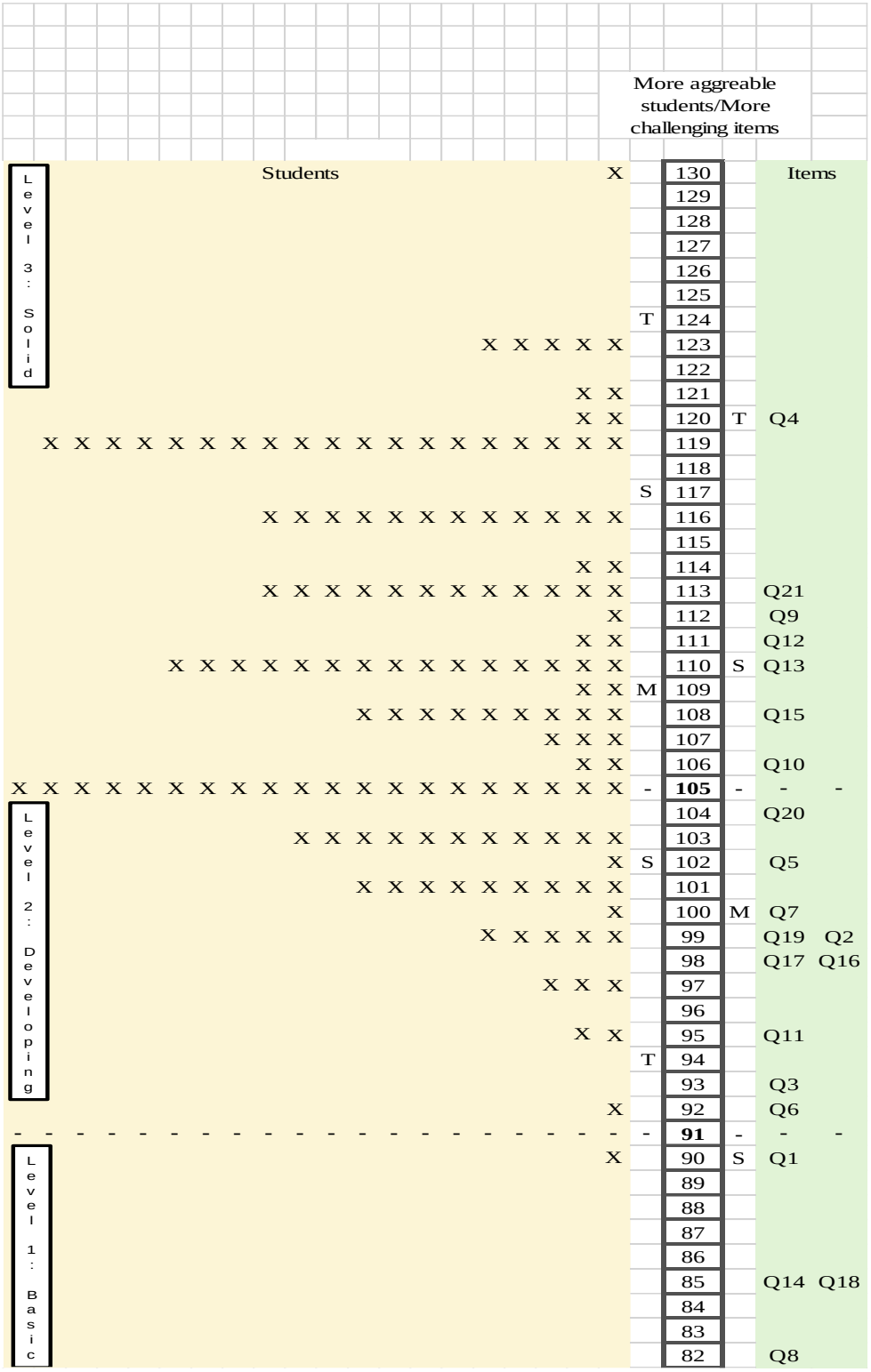
Note: Predicted cognitive items are from the person construct map (Table 2) and the observed items are based on Rasch analysis (Figure 2a).

To lay out the findings more precisely, items were further grouped according to Bloom's cognitive abilities (i.e., remember, understand, apply, analyse, and evaluate), as shown in Figure 2b. In general, the structure of items follows the expected order of the person construct map; that is, Level 1 (Basic) items are the easiest group of items among the three levels, Level 2 (Developing) items are more difficult than Level 1 but are easier than Level 3, and Level 3 (Solid) items are the most difficult among the three levels. A similar pattern is also observed in Figure 2c.

All nine word problems (mean difficulty = 119.50) were in the top half of the scale, and the remaining 12 arithmetic calculation problems (mean difficulty = 104.07) were in the bottom half. Apparently, word problems were more difficult than word problems.

However, the two-word problems (Q5 and Q6) shared the same difficulty level as the arithmetic calculation problems (Figure 2c). Hence, further investigation is required for Q5 and Q6.

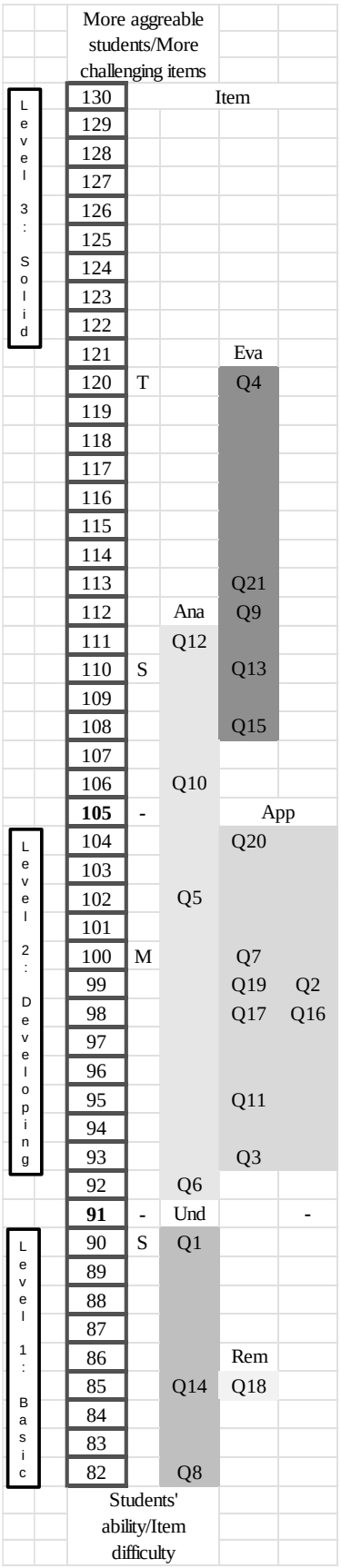
Figure 2a. *Wright map distribution of students' ability and item difficulty*.



Note:1. The observations on the left (in yellow) show the distribution of measured student abilities. The students showed the lowest ability at the bottom and highest ability at the top. The observations on the right in green show the item difficulty distribution, with the least challenging items at the bottom and the most challenging items at the top. M, on the left, indicates the mean student ability, S one standard deviation point, and T two standard deviation points of student ability, respectively. Similarly, on the right, M shows the mean item difficulty, S is one standard deviation point, and T is two standard deviation points of item difficulty.

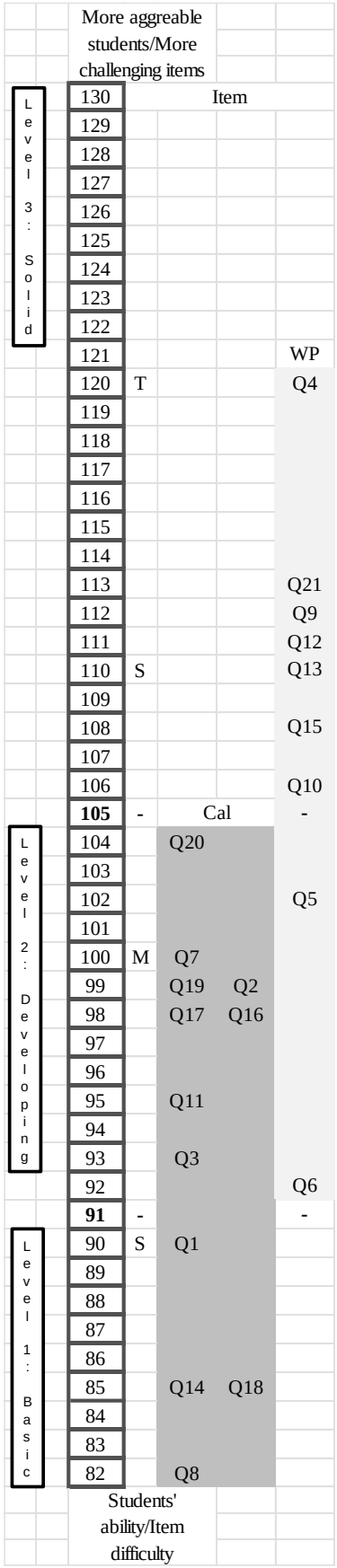
Note: 2. "X" = 1 student; Mean person ability = 109.27 (Standard deviation = 7.44); Mean item difficulty = 100 (Standard deviation = 10.00)

Figure 2b. *Wright map – Cognitive ability.*



Note: "Rem"=Remember; "Und"=Understand; "App"=Apply; "Ana"=Analyse, "Eva"=Evaluate.

Figure 2c. *Wright map – Arithmetic calculation and word problem.*



Note: "Cal" = arithmetic calculation problem (dark gray shading); "WP" = arithmetic word problem (light gray shading).

3.3. Are Respondent Categories (Basic, Developing, Solid) Aligned with the Person Construct Map?

Guided by the item design (Table 2), students were separated into three levels: Level 1 (basic students), Level 2 (developing students), and Level 3 (solid students). The students were assumed to be able to solve Level 1 items. Developing students were assumed to be able to solve Level 1 and 2 items. Lastly, the Solid students were assumed to be able to solve Levels 1, 2, and 3 items.

As the items were designed according to what the primary three students had learned in the first two, and the assessment was conducted in the first month of this school year, most of the students could solve all Level 1 items. Hence, only one student was at Level 1 (Figure 2a). Therefore, even students with the lowest ability could solve Level 1 items, given that their ability is higher than the difficulty of Level 1 items. In other words, the appropriate difficulty level for Level 1 items was overestimated based on the person construct map (Table 1) and item design of the building blocks (Table 2).

Developing students should be able to solve levels 2 and 1. There were 53 students (38.41% of the sample) in the developing student category, and the developing students were aligned with Level 2 items and above Level 1 items (Figure 2a).

Solid students should be able to solve Level 3, 2, and 1 items. There were 84 students (60.87%) in the solid student category. We noted that some solid students' abilities were higher than the difficulty of all items (Figure 2a). This suggests that the Level 3 item difficulty was overestimated.

This non-normal distribution of students for the Basic, Developing, and Solid levels (1, 53, and 84 students, respectively) suggests that more difficult items should be added to better assess arithmetic ability. Overall, the distribution of students (Figure 2a) followed the structure of the predicted construct map based on item development. Students at the Basic level (corresponding to Level 1 in Table 1) could solve the least difficult problems with the lowest cognitive level in Table 2. Students at the Developing level (Level 2) solved more difficult problems. Finally, students at the solid level (Level 3) solved the most difficult problems with a higher Bloom's cognitive level.

3.4. Are Arithmetic Calculation Problems (Levels 1 and 2) More Accessible than Arithmetic Word Problems (Level 3)?

Figure 2c shows the operational grouping of arithmetic word problems and calculations. Items Q5 ("64 books) were packed in boxes of 7. How many books are left?') and Q6 ("There are 14 students. None of the patients had coins. How many coins do they have altogether?") combined arithmetic word problems with an arithmetic calculation problem operation (Figure 2c) because of the 2-step operation. Theoretically, they are expected to cluster with other problems. However, Rasch analysis of student responses placed them as difficult as arithmetic calculation items (Figure 2c).

When we compared these two items with other word problems, we noted the following: the mean number of words in the Level 3 word problems was approximately 24 words. In contrast, the mean numbers of words in Q5 and Q6 were 16. Therefore, one explanation for their position in Figure 2c is that students find these questions easier to read, comprehend, and solve because of the lower word number and complexity than other Level 3, more complex items. Furthermore, the operation that students require for solving the arithmetic word problems is directly given in the question stems of Q5 and Q6 (e.g., they are asked to find out "How many … altogether" and "How many … are left"). In other words, students immediately signalled that Q5 and Q6 were addition and subtraction problems, and they only needed to determine the

correct numerical answer from the four numerical options (Table 2). As such, Q5 and Q6 require a lower level of language proficiency, one-step calculation, and lower thinking skills and may explain why they fall below other arithmetic word problems and into the Level 2 category.

In contrast, all other Level 3 items demanded higher-order thinking. For example, when students read these questions (e.g., "Does she need extra money?' If yes, how much?"; "Mr Lee spent three times as much as Mr Chan"; "At least how many …"; "At most …"), they cannot simply compute the required answer; instead, these require a higher level of language ability, analysis and at least two operation steps to arrive at an answer. In other Level 3 arithmetic word problems, students were required to select the correct algorithm from the four algorithms (Table 2). In other words, all other Level 3 word problems require higher-order thinking skills.

However, aside from items Q5 and Q6, the distribution of the other items fits the structure of the predicted construct map based on item development (Figure 2c). Thus, basic-level students can solve Level 1 items, developing-level students can solve Level 1 and Level 2 items, and solid-level students can solve Level 1, Level 2, and Level 3 items. It is worth noting that Levels 1 and 2 are primarily arithmetic calculation problems. At the same time, Level 3 items were solely arithmetic word problems. This implies that word problems can distinguish solid students from Developing and Basic students, which is consistent with the contention that Level 3 questions require a higher cognitive level. In comparison, Level 1 and 2 questions were accessible to students with lower cognitive levels.

However, to test the premise that word problems are more difficult and require a higher cognitive level than problems involving only calculation, Rasch's Principal Components Analysis (PCA) of residuals was performed on all questions to test unexpectedness (Linacre, 2012). PCA's standardized residual (loadings) was analysed after extracting the primary Rasch dimension. Higher factor loadings indicate substantial unexplained variance (Bond & Fox, 2015). In other words, the residuals of these items are not the result of random noise. This analysis tests whether the common factor can explain variance (Linacre, 1998, p. 636).

The items were clustered into two groups: the items that have a positive loading (from +0.01 to + 0.57) make up cluster 1, and items with a negative loading (from -0.42 to -0.10) make up cluster 2. Thus, items represent two strands of the same latent variable. One strand comprises arithmetic calculation problems, and the other comprises arithmetic word problems. Items with higher positive loadings are primarily single-step calculation problems. By comparison, those with higher negative loadings were primarily word problems. Thus, these findings support the contention of two strands: arithmetic calculation ability and the ability to solve arithmetic word problems.

4. DISCUSSION and CONCLUSION

This work aims to validate a formative assessment based on Wilson's (2004) four building blocks, which can be used to meaningfully measure students' understanding. The current study illustrates a measure of primary student arithmetic ability. The data for the hypothesized model show evidence of reliability and validity based on the Rasch framework, with the exception of item Q9 (*Mr Chan spent 3 dollars. Mr Lee spent 3 times as much as Mr Chan. Ms Fong spent 3 times as much as Mr Lee. How much did Ms Fong spend?*) where it reports an Infit Zstd of 2.95 and an Outfit Zstd of 2.90. We noted that the structure of each sentence in Q9 was simple, but the relationship was complex. This suggests that students might comprehend individual sentences, but not the overall context. Thus, they must select the best algorithm instead of the correct numerical answer, demanding higher-order thinking, metacognition, and language proficiency versus the one-step mathematical operations needed for other items. We retained this information in the assessment analysis of the current study; however, further investigation is required for future research. It also points out the need to systematically analyze individual

test items that students may perceive differently than the teacher's original intention. Low person reliability indicates that there are insufficient items to evaluate the latent variable (Lincare, 2012). Hence, the research findings should be viewed within this limitation. Future research should consider adding more items to the test to enhance its reliability.

The hypothesized model, however, shows sufficient evidence of validity, as follows:

a) The difficulty distribution of the items follows the expected order for the construct map, where the difficulty for each designed item aligns with the student's cognitive level in that order.
b) The students' performance distribution followed the structure of the person-construct map. Students at the Solid level could answer items that require students to evaluate, students at the developing level could answer items that require them to apply and analyse, and students at the basic level could answer items that require them to memorize and remember. The results indicate that internal validity is upheld when the construct is in order, as expected (Wilson, 2004, p. 157–158). However, as only one student fell into the basic level, more challenging Level 1 items may need to be added to the instrument to improve internal validity. Future research may need to redefine the three-level categorization of students. However, the performance of this cohort of students was unexpectedly higher than anticipated.
c) Solving arithmetic calculation problems is the foundation of solving arithmetic word problems. However, the latter requires additional skills such as reading comprehension and analysis, pattern recognition, semantic relations, and problem-model strategies (Chiang & Chen, 2019; Cummins, 1991; Prakitipong & Nakamura, 2006; Riley et al., 1983; Simon, 1978; Weitheimer, 1959). Pertinent literature supports this contention. Given the above discussions, we found that word problems can be categorized into simpler and more difficult items. Therefore, future research is required to study the factors affecting the difficulty of word problems.

The credibility and interpretation of the assessment information are not dependent only on the item content. To be instructionally useful, the items must define a meaningful hierarchy of increasing difficulty in which easier items assess the conceptual understanding needed in the solution of more difficult items (Alonzo & Steedle, 2009; Black et al., 2011; Fisher, 2013). In addition, item content must be aligned with local learning objectives if the goal of coherence is to be realized (Baird et al., 2017). The formative assessment in the current study provides this type of assessment information. Specifically, the conceptualizations of understanding about the topic following the four building blocks (Wilson, 2004) in the current study provide information on where a student stands relative to intended learning outcomes in a person construct map. Classroom assessment of this kind sets up information sources for teachers to formulate valuable and timely feedback for students about '*what might be useful to do next* (Black & William, 1998; Mislevy et al., 2003). This improves coherence in documenting learning, enhances classroom feedback, and shifts focus away from grades to more authentically serve classroom assessment purposes in facilitating learning at the individual level.

Acknowledgments

We want to thank the mathematics teachers in this study. This research would not have been possible without their willingness to share their experiences.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: University of Macau, SSHRE19-APP065-FED.

Authorship Contribution Statement

Wai Kei Chan: Design of the research project, design of assessment, data collection and analysis, and drafting the manuscript. **Li Zhang**: Design of the research project, design of assessment, and drafting the manuscript. **Emily Pey-Tee Oon**: Design of the research project, critical feedback on the manuscript, and writing and editing of the manuscript.

Orcid

Wai Kei Chan https://orcid.org/0000-0003-2023-5141
Li Zhang https://orcid.org/0000-0003-1091-5979
Emily Pey-Tee Oon https://orcid.org/0000-0002-1732-7953

REFERENCES

Adillah, G., Ridwan, A., & Rahayu, W. (2022). Content validation through expert judgement of an instrument on the self-assessment of mathematics education student competency. *International Journal of Multicultural and Multireligious Understanding*, *9*(3), 780-790. http://dx.doi.org/10.18415/ijmmu.v9i3.3738

Alderson, J.C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language*, *6*(2), 425-438.

Alonzo, A.C., & Steedle, J.T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421.

Baird, J.-A., Andrich, D., Hopfenbeck, T.N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317-350.

Baroody, A.J., & Dowker, A. (Eds.). (2003). *The development of arithmetic concepts and skills: Constructing adaptive expertise.* Lawrence Erlbaum Associates Publishers.

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *Frontline Learning Research*, *8*(6), 1-37. https://doi.org/10.14786/flr.v8i6.517

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536-553.

Björklund, C., Marton, F., & Kullberyg, A. (2021). What is to be learnt? Critical aspects of elementary skills. *Educational Studies in Mathematics*, *107*, 261-284. https://doi.org/10.1007/s10649-021-10045-0

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Black, P., & Wiliam, D. (2003). In praise of educational research': Formative assessment. *British Educational Research Journal*, *29*(5), 623-637.

Black, P., & Wiliam, D. (2010). *Inside the black box: Raising standards through classroom assessment*. *Phi Delta Kappan*, *92*(1), 81-90. https://doi.org/10.1177/003172171009200119

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspectives*, *9*, 1-52.

Bloom, B.S. (1956). *Taxonomy of educational objectives. The classification of educational goals*. Handbook 1: Cognitive domain. David McKay.

Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Brookhart, S.M., Moss, C.M., & Long, B.A. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 41-58.

Cardinet, J. (1989). Evaluer sans juger. *Revue Française de Pédagogie*, *88*, 41-52.

Cataloglu, E., & Robinett, R.W. (2002). Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career. American Journal of Physics, *70*(3), 238-251. https://doi.org/10.1119/1.1405509

Chiang, T., & Chen, Y. (2019). Semantically-aligned equation generation for solving and reasoning math word problems. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1272

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205-249.

Cummins, J. (1991). *Interdependence of first- and second-language proficiency in bilingual children.* Language Processing in Bilingual Children, pp. 70-89. Cambridge University Press. https://doi.org/10.1017/cbo9780511620652.006

Dixson, D.D., & Worrell, F.C. (2016). Formative and summative assessment in the classroom. *Theory into Practice*, *55*(2), 153-159. https://doi.org/10.1080/00405841.2016.1148989

Dowker, A. (2005) Early Identification and Intervention for Students with Mathematics Difficulties. *Journal of Learning Disabilities*, *38*(4), 324. http://dx.doi.org/10.1177/00222194050380040801

Duckor, B., & Holmberg, C. (2017). *Mastering formative assessment moves: 7 high-leverage practices to advance student learning*. ASCD Press.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Engvall, M., Samuelsson, J., & Östergren, R. (2020). The effect on students' arithmetic skills of teaching two differently structured calculation methods. *Problems of Education in the 21st Century*, *78*(2), 167-195. https://doi.org/10.33225/pec/20.78.167

Fisher, W.P., Jr. (1997). Is content validity valid? *Rasch Measurement Transactions*, *11*, 548.

Fisher, W.P., Jr. (2013). Imagining education tailored to assessment as, for, and of learning: Theory, standards, and quality improvement. *Assessment and Learning*, *2*, 6-22.

Geary, D.C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, *114*(2), 345–362. https://doi.org/10.1037/0033-2909.114.2.345

Goldman, S.R., & Hasselbring, T.S. (1997). Achieving meaningful mathematics literacy for students with learning disabilities. *Journal of Learning Disabilities*, *30*(2), 198–208.

Gorin, J.S., & Mislevy, R.J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment (K-12 Center at Educational Testing Service No. Invitational Research Symposium on Science Assessment)*. ETS.

Gurel, D.K., Eryilmaz, A., & McDermott, L.C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *EURASIA Journal of Mathematics, Science and Technology Education*, *11*(5). https://doi.org/10.12973/eurasia.2015.1369a

Guskey, T.R. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*(5), 6-11.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer.Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hidayati, K., Budiyono, & Sugiman. (2019). Using alignment index and Polytomous item response theory on statistics essay test. *Eurasian Journal of Educational Research*, *79*, 115-132.

Hiebert, J., & Lefevre, P. (1986). *Conceptual and procedural knowledge in mathematics: An introductory analysis*. In J. Hiebert (Ed.), Conceptual and procedural knowledge: The case of mathematics (pp. 1–27). Lawrence Erlbaum Associates, Inc.

Hong Kong Education Bureau (2018). *Explanatory Notes to Primary Mathematics Curriculum (Key Stage 1)*.

Hong Kong Education Bureau (2000). *Mathematics education key learning area.*

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.

Lee, N.P., & Fisher, W.P., Jr. (2005). Evaluation of the diabetes self-care scale. *Journal of Applied Measurement*, *6*(4), 366-381.

Linacre, J.M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*(2), 636.

Linacre, J.M. (2000). Computer-adaptive testing: A methodology whose time has come. *MESA Memorandum*, *69*, 1991-2000.

Linacre, J.M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement, 10*(2), 157-169.

Linacre, J.M. (2012). *A user's guide to Winsteps. Ministeps. Rasch-model computer programs. Program manual 3.74.0*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Liu, O.L., Lee, H.S., Hofstetter, C., & Linn, M.C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*(1), 33-35.

Luque-Vara, T., Linares-Manrique, M., Fernández-Gómez, E., Martín-Salvador, A., Sánchez-Ojeda, M.A., & Enrique-Mirón, C. (2020). Content validation of an instrument for the assessment of school teachers' levels of knowledge of diabetes through expert judgment. *International Journal of Environmental Research and Public Health*, *17*(22), 8605. https://doi.org/10.3390/ijerph17228605

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9-20.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Macmillan Publishing.

Millians, M. (2011). Computational skills. In S. Goldsteing & J. A. Naglieri (Eds.), *Encyclopedia of child behavior and development*. Springer. https://doi.org/10.1007/978-0-387-79061-9_645

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3-62. https://doi.org/10.1207/s15366359mea0101_02

National Research Council (NRC), (2006). *Systems for state science assessment. Committee on Test Design for K-12 Science Achievement*. M. R. Wilson and M. W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. The National Academies Press.

Nasir, N.A.M., Singh, P., Narayanan, G., Mohd Habali, A.H., & Rasid, N.S. (2022). Development of mathematical thinking test: Content validity process. *ESTEEM Journal of Social Sciences and Humanities*, *6*(2), 18-29.

Parviainen, P. (2019). The development of early mathematical skills - A theoretical framework for a holistic model. *Journal of Early Childhood Education Research*, *8*(1), 162-191.

Popham, W.J. (2009). Our failure to use formative assessment: *Immoral Omission. Leadership, 1*(2), 1-6.

Popham, W. (2010). Wanted: A formative assessment starter kit. *Assessment Matters*, *2*, 182.

Prakitipong, N., & Nakamura, S. (2006). Analysis of mathematics performance of grade five students in Thailand using Newman procedure. *Journal of International Cooperation in Education, 9*(1), 111-122.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Mesa Press.

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H. Ginsburg (Ed.), The development of mathematical thinking (pp. 153-196). Academic Press.

Shepard, L.A. (2006). *Classroom assessment*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 623-646). Rowman & Littlefield.

Sievert, H., van den Ham, A.-K., & Heinze, A. (2021). Are first graders' arithmetic skills related to the quality of mathematics textbooks? A study on students' use of arithmetic principles. *Learning and Instruction, 73*. https://doi.org/10.1016/j.learninstruc.2020.101401

Simon, H.A. (1978). *Information-processing theory of human problem solving*. In W.K. Estes (Ed.), *Handbook of learning and cognitive processes (Volume 5): Human information processing* (pp. 271-295). Psychology Press.

Stiggins, R.J. (1994). *Student-centered classroom assessment*. Merrill.

Vlassis, J., Baye, A., Auqui ère, A., de Chambrier, A.-F., Dierendonck, C., Giauque, N., Kerger, S., Luxembourger, C., Poncelet, D., Tinnes-Vigne, M., Tazouti, Y. & Fagnant, A. (2022). Developing arithmetic skills in kindergarten through a game-based approach: a major issue for learners and a challenge for teachers. *International Journal of Early Years Education*. https://doi.org/10.1080/09669760.2022.2138740

Wertheimer, M. (1959). *Productive thinking*. Enlarged Edition. Harper and Brothers.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* Jossey-Bass.

Wilson, M. (2004). *Constructing measures: An item response modeling approach.* Routledge.

Wole, G.A., Fufa, S., & Seyoum, Y. (2021). Evaluating the Content Validity of Grade 10 Mathematics Model Examinations in Oromia National Regional State, Ethiopia. *Mathematics Education Research Journal*. https://doi.org/10.1155/2021/5837931

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Mesa Press.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials*. Wide Range.