MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2021, 17(12), em2056
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/11426

Application of Rasch Measurement Model in Developing Calibrated Item Pool for the Topic of Rational Numbers

Ahmad Zamri Khairani [1*], Hasni Shamsuddin [2]

[1] School of Educational Studies, 11800 Universiti Sains Malaysia, Penang, MALAYSIA
[2] Sekolah Menengah Sains Kepala Batas, 13200 Kepala Batas, Penang, MALAYSIA

Received 5 June 2021 ▪ Accepted 29 November 2021

Abstract
Rational Numbers is an essential topic in mathematics since it necessitates the learning progression of more advanced topics. Nevertheless, previous literature shows that students are having difficulties in understanding the topic for numerous reasons. The inability of teachers in providing good examples during teaching is identified as one of the major causes. Thus, this study aims to develop a calibrated pool of items to facilitate teachers in giving appropriate examples for the topic of Rational Numbers. We employed a descriptive design to provide descriptions of the item statistics for the calibrated pool of items. Samples of the study consisted of 1,292 secondary school students. We used the Rasch measurement model framework via a quantitative approach to analyse the data. The results showed that all items demonstrated an acceptable quality of measuring students' ability in rational numbers while at the same time demonstrated high evidence of validity and reliability as well. Ultimately, we also provided suggestions on how teachers can use the pool of items in delivering appropriate examples in the classroom.

Keywords: rational numbers, calibrated item pool, Rasch model, item difficulty measure, learning standards

INTRODUCTION

Success in school and beyond is greatly influenced by mathematics proficiency (Ritchie & Bates, 2013). And, according to Tian and Siegler (2018), one of the prime factors that contribute to mathematics proficiency is knowledge about rational number. Rational number is defined as any number that can be expressed as a ratio of two integers with the denominator ≠ 0 (Blinder, 2013). For example, 2 is a rational number since it is a product of 2/1 or 4/2 etc. Decimals such as 0.125 is also a rational number since it can be expressed in terms of 1/8. In general, since integers take the values of positive, zero, and negative numbers, rational numbers also have similar properties. Rational numbers and the concepts connected to them are essential for learning mathematics since the understanding of these concepts helps students to progress better in more advanced topics (Mozacco et al., 2013; Siegler et al., 2012). For example, since probabilities are widely expressed as fractions, decimal, and percentages, it requires an understanding of the magnitudes of these rational numbers to understand the concept of probabilities and therefore the decision-making contexts.

The following examples might give a better insight into the importance of understanding the concept of rational numbers that can be further applied beyond the classroom. In inferential statistics, to be able to recognize the different meaning of p = .01, p = .10, p = .05, or p = .001 requires some level of understanding of decimals. Also, in engineering, fractions and decimals are always being used in the conversion of units, while ratios and proportions are also used in medical practice for calculating the right amount of dosage of medication. More than that, rational numbers also play an important part in our daily life. For instance, knowledge of fractions helps us to understand discounts for items on sale, while understanding decimals will surely encourage precisions.

✉ ahmadzamri@usm.my **(*Correspondence)** ✉ emelhasni@gmail.com

Contribution to the literature

- This study emphasizes the use of Rasch model in developing calibrated item pool in the topic of Rational Number.
- The study shares new findings about alignment of learning standards with the actual test items.
- The study contributes literature about how teachers can use the calibrated item pool for giving examples in classroom instructions.

LITERATURE REVIEW

Despite its importance, previous literature shows that the topic of Rational Numbers is very challenging for students. Notably, there are different level of difficulties associated with the topic. One of the most resounding difficulty relates to the "whole number bias" phenomenon, in which the inappropriate application of natural number rules was used (Ni & Zhou, 2005). For example, Sun (2019) listed some difficulties in the operations of addition, subtraction, multiplication and division of fraction which resulted from the algorithm that was not supported by the whole number rule. Further, according to van Hoof et al. (2015), whole numbers differ from rational numbers in four distinct aspects, namely, (1) density, (2) representation, (3) number size, and (4) arithmetic operations. Applying whole number rules in these aspects may lead to systematic errors. Other than that, research by Yetim and Alkan (2013) identified basic mistakes such as failure to convert rational numbers into decimal numbers and vice versa and stating that -8/5 is equal to -8.5. Besides, there is the Longer-is-Larger rule, where numbers with more digits are commonly considered bigger (Liu et al., 2014). To illustrate, students who adopt the rule believe that 4.9 is smaller than 4.34 since the latter has more digits.

Apart from the "whole number bias" phenomenon, Sigler and Lortie-Forgues (2017) identified two other sources of difficulties encountered by the students, which they termed as inherent and culturally difficulties. Inherent sources of difficulty include difficulty in understanding individual rational number (such as why $\frac{1}{2}$ is bigger than $\frac{1}{3}$ when 3 is bigger than 2?), the relationship between rational and whole number (such as why $\frac{1}{5}+\frac{2}{5}=\frac{3}{5}$ and not $\frac{3}{10}$?), as well as the relation among rational number (for example, why $\frac{3}{7}+\frac{2}{7}=\frac{5}{7}$ but why $\frac{3}{7}\times\frac{2}{5}=\frac{6}{35}$?).

On the other hand, as the name suggests, a culturally contingent source of difficulty involves the culture within which the learners originated from. It is well acknowledged that teachers' knowledge differs based on their countries of origin. For example, while Canadian pre-service teachers find it difficult in explaining the concept of multiplication using two fractions (Siegler & Lortie-Forgues, 2015), a large majority of their Chinese counterparts reported otherwise (Lin et al., 2013).

One of the possible explanations for the disparity was the conception of teacher professional development (TPD). While TPD in the Western countries often takes place in the form of workshops that are considered remote, inconsistent, and sometimes contradictory (Guskey, 2003), the culture of professional development in East Asian countries such as China can happen at any point of time in the teachers' daily routine (Huang, 2006). Plus, there is also an Asian culture of learning from more experienced teachers that in turns improve the younger teachers' knowledge and skills in teaching (Li et al., 2006).

Another identified source of difficulty is textbook content. Lack of coverage in the textbook may influence students' exertion in understanding a particular mathematical concept. For example, Son and Senk (2010) found that the US textbook contained fewer examples of fraction division problems for the fifth and sixth graders compared to fraction multiplication problems. This may explain the poorer results among American students. Apart from that, language is also decidedly an important culturally contingent source of difficulties. To exemplify, numerical terms used in East Asian countries seem to facilitate students' better achievement in mathematics (Dowker et al., 2008).

Like other topics in mathematics, one of the important approaches to teaching rational numbers is by providing examples. The primary purpose of providing examples is to assist retention by repetition of the procedure so that students develop proficiency. Likewise, it is hoped that while working on examples, students can construct new awareness and understanding with regards to both procedure and concept. There are many ways that teachers can give examples. Among the common approaches is by introducing an idea or explaining a concept. Several researchers have conducted studies to explore strategies used in providing examples. For instance, Bills and Bills (2005) suggest that teachers should use simple examples first, such as using small numbers and minimum operations and use examples that build on students' prior knowledge to scaffold students' learning. Teachers should also use examples that allow them to attend to common errors and misconceptions (Zodik & Zaslavsky, 2008).

Yet, studies also show that teachers struggle to do just that since they depended heavily on the examples and exercises from the textbook. Compounding this issue is

the fact that the difficulty of the items in the textbook was not empirically tested. Also, there is an abundance of items in the textbook to choose from, making it a laborious task for the teachers to choose the best possible items to be used as classroom examples. Moreover, literature shows that teachers have been known as having poor ability to estimate the difficulty of a particular item (Impara & Plake, 1998; van de Watering & van der Rijt, 2006). Thus, teachers, especially the less-experienced, might find it challenging to find items from the textbooks that tailor to their teachings at a particular learning standard.

Calibrated Item Pool

One of the possible solutions to this is by having a pool of calibrated items for teachers to choose as examples in the classroom. Calibrated item pool is defined as a group of items that have been arranged according to their difficulty intensity. Thus, teachers might use easy items from the pool to introduce new concepts as well as reducing misconceptions. Gradually, more difficult items can be added to increase students' understanding of a particular topic. According to the literature, a calibrated item pool can be used for a variety of purposes. One of the most important benefits is that it aids in constructing tests that are relevant to the testing objectives. To give an example, Aung and Lin (2020) established a calibrated pool of 164 mathematics items for Grade 6 children, and based on the statistics of the items in the bank; they were able to develop a psychometrically sound new 60-item test to evaluate the average ability students' mathematical ability.

The Classical Test Theory (CTT) and the Item Response Theory (IRT) are two widely used measurement theories for developing calibrated item pools (IRT). The IRT, on the other hand, is more commonly employed for item calibration. Many researchers choose the Rasch measuring framework within the IRT family because it requires less parameter estimate and is thus easier to deal with. For example, Bjorner et al. (2017) used the Rasch model to create a pool of high-quality items, with only five items from the pool having a very high concordance with the score based on all items. Kallinger et al. (2019) used the same model to calibrate an item bank of anxiety-related questions for orthopedic patients. The item bank serves as the foundation for a computer-adaptive exam that can be used to assess a wide variety of anxiety in orthopedic rehabilitation patients. Meanwhile, Nieto et al. (2017) used the adaptive power of a calibrated item pool to demonstrate that only one-third of the pool questions are sufficient to assess the Five-Factor Model personality facets accurately.

Despite its potential, however, research on calibrated item pools in education is minimal. Hence, the purpose of the present study is to develop a calibrated item pool in the topic of rational numbers so that teachers can use the items effectively as examples during classroom instructions. As a result, teachers are no longer required to estimate the difficulty of the items as classroom examples. Instead, teachers can continue to identify such items to use as examples based on their difficulty statistics.

METHODOLOGY

Participants

Participants in this study consist of 1,292 secondary school students with an average age of 13 years old. The gender distribution is 590 males (45.7%) and 702 females (54.3%) from schools in the states of Kedah, Penang, and Perak in the northern parts of Malaysia. The selection of the schools was based on purposive sampling, in which the researchers identified schools with various degrees of achievement in mathematics.

Instrument

This study employed ten mathematics tests that were administered to ten schools. The tests were conducted separately but were linked together by several common items using the common item non-equivalent group design (Kolen & Brennan, 2014). Altogether, we employed 81 common items to link the ten tests and 362 unique items measuring 13 topics specified in the curriculum specifications (Ministry of Education, 2016). However, only results involving the topic of Rational Numbers will be presented in this article. The tests were developed both by the researchers as well as by the practising teachers. Content validity of the test was observed by the head of the mathematics panel of each school. The tests included both multiple-choice and partial credit items. In the multiple-choice format, participants chose one correct answer from a list of four possible choices. One mark was given to the correct answer and no mark for the incorrect answer. In the partial credit format, the scoring was based on the completion of the steps in solving the problem. The marks for each item ranged from 1 to 4 marks, and the total marks for each test were 100. Correspondingly, items that shared the same stem in the partial credit format were treated as different items. Examples of a multiple-choice item and a 2-marks partial credit item are given in Table 1.

Data Analysis

The quality of each item in the item pool was examined by using Rasch model software WINSTEPS 3.74. Apart from its simplicity, the model is favored to others in the IRT family, such as the 2-parameter model, since each item must have the same discriminatory power, allowing students to be estimated solely by item difficulty and not by how well they know the content being tested. Meanwhile, all forms of data are accepted

Table 1. Example of test items and its scoring

	Multiple-choice	Partial Credit Item
Item	Which of the following is correct? A –2 < –5 B –3 > 0 C –6 > –2 D –5 > –9	Solve the following: $3\frac{1}{3} \times \left(\frac{2}{5} - \frac{3}{4}\right)$ (2 marks)
Scoring	D …… 1 mark A, B, or C …… 0 mark	$3\frac{1}{3} \times \left(-\frac{7}{20}\right)$ or equivalent…1 mark $-1\frac{1}{6}$ or equivalent ……1 mark

when utilizing the 3-parameter model because the model adjusts for any disparities in the data. Nevertheless, we believe that erratic data that did not fit the model's expectations for achievement tests will not be accepted for analysis. Similarly, guessing is also not accepted and is considered as reflecting the unreliability of the respondents.

The plan of analysis started with assessing the assumptions of the Rasch model, specifically, (1) the model-data fit and (2) the unidimensionality assumptions. This is a crucial step since the Rasch model is considered as a model with a strict assumption that must be met to create the equal-interval scale (Bond & Fox, 2015). The first assumption was that the data must fit the model's expectation. Model-data fit refers to the extent to which the data collected matches expectations from the model. This assumption was examined using the infit and outfit mean-square (MNSQ) values generated from WINSTEPS 3.74. While both statistics are sensitive towards unexpected responses, the infit MNSQ deals with responses by the respondents that are targeted towards them while the outfit MNSQ explains far from the targeted respondents (Linacre, 2002). According to Bond and Fox (2015), the assumption is met when the values of the infit and outfit MNSQs were in the range from 0.6 to 1.4. Meanwhile, the unidimensionality assumes that items in a test measure a single construct (Wright & Masters, 1982). The assumption was examined from the principal component analysis of the residuals procedure in the software. The assumption is met when the variance explained by the measurement dimension from the procedure is more than 40% (Linacre, 2006).

In this study, apart from examining the assumptions, we also reported statistics at the item level, specifically, the item reliability and item separation indices. Item reliability statistics refer to the ratio between true to observed item variance (Linacre, 2006). This provides information on the consistency of the ordering of item difficulty if an instrument is administered to a comparable sample of participants. High item reliability statistic indicates the consistent ordering of the items' difficulty and vice versa. Meanwhile, the item separation index is an indication of the adequacy of the measurement to distinguish between participants. For example, if the separation index is 2, then it is possible to distinguish the participant into two ability groups. It should be noted that a proper measurement should be able to distinguish clearly the ability of the participants. For a proper measurement, the item reliability index should be more than 0.94 (Fisher, 2007), while the separation index should not be less than 2.0 (Bond & Fox, 2015).

At the same time, statistics for each item were also reported. Apart from the item difficulty and the fit statistics, the point measure correlation (PTMEA) statistic was also included. The positive values of this statistic indicate that the particular item is working together with other items in the same direction to measure the intended construct (Bond & Fox, 2015).

Apart from the abovementioned analysis, the present study also provided information regarding the learning standards for the topic of Rational Numbers. In the curriculum, learning standards are indicators of the quality of learning and achievement that can be measured (Ministry of Education, 2016). The analysis is essential to identify the most difficult-to-master learning standards so that teachers can benefit from the information when providing examples during the classroom. The topic of Rational Numbers consists of 21 learning standards which are the most for any topic in the curriculum (Ministry of Education, 2016). The list of learning standards for this topic is presented in Table 2. Note that learning standards 1.2.4 (Describe the laws of arithmetic operations, which are Identity Law, Communicative Law, Associative Law, and Distributive Law) was not targeted by any item since it is supposed to be measured orally and not through test items. Meanwhile, learning standards 1.3.1 (Represent positive and negative fractions on number lines) was also not targeted by any items. This is because the knowledge and skills for the learning standards are similar to learning standards 1.4.1.

FINDINGS

In terms of a model-data fit, results from the calibration of all 447 items showed that the software dropped three items due to a lack of responses by the students. Meanwhile, 21 items that exhibited the infit and outfit MNSQ values outside the acceptable 0.6 -1.4 guideline was manually deleted (see Table 3).

Table 2. Learning standards

Learning Standards		No. of Item
1.1.1	Recognize positive and negative numbers based on real-life situations.	2
1.1.2	Recognize and describe integers.	7
1.1.3	Represent integers on number lines and make connections between the values and positions of the integers with respect to other integers on the number line.	4
1.1.4	Compare and arrange integers in order.	8
1.2.1	Add and subtract integers using number lines or other appropriate methods. Hence, make a generalization about the addition and subtraction of integers	3
1.2.2	Multiply and divide integers using various methods. Hence make a generalization about the multiplication and division of integers.	1
1.2.3	Perform computations involving combined basic arithmetic operations of integers by following the order of operations.	9
1.2.4	Describe the laws of arithmetic operations, which are Identity Law, Communicative Law, Associative Law, and Distributive Law.	0
1.2.6	Solve problems involving integers.	9
1.3.1	Represent positive and negative fractions on number lines.	0
1.3.2	Compare and arrange positive and negative fractions in order.	3
1.3.3	Perform computations involving combined basic arithmetic operations of positive and negative fractions by following the order of operations	4
1.3.4	Solve problems involving positive and negative fractions.	1
1.4.1	Represent positive and negative decimals on number lines.	2
1.4.2	Compare and arrange positive and negative decimals in order.	3
1.4.3	Perform computations involving combined basic arithmetic operations of positive and negative decimals by following the order of operations.	5
1.4.4	Solve problems involving positive and negative decimals.	2
1.5.1	Recognize and describe rational numbers.	2
1.5.2	Perform computations involving combined basic arithmetic operations of rational numbers by following the order of operations.	4
1.5.3	Solve problems involving rational numbers.	2

Table 3. Descriptive statistics

Statistics	Purpose	Guidelines	Empirical
Infit and outfit MNSQ	To ensure the empirical data matches the model's specifications	0.6-1.4 (Bond & Fox, 2015)	0.62-1.38
Percentage of variance explained in the 1st contrast	To examine whether the scale is measuring a unidimensional construct.	> 40% (Linacre, 2006)	54%
Item reliability	Consistency of the ordering of item difficulty if an instrument	> 0.94 (Fisher, 2007)	0.97
Item separation	The adequacy of the measurement to distinguish between participants	> 2.0 (Bond & Fox, 2015)	6.10

Conversely, results from the PCA of residuals showed that raw variance explained by both the students and the items measures was 54%, which was more than the intended value of 40% (Linacre, 2006). As such, we provided ample evidence that the unidimensionality assumption was also fulfilled. Besides, both item reliability and item separation indices exceed the intended values.

Table 4 showed statistics for all 71 items measuring 17 learning standards. Two items were measuring the

Table 4. Item statistics

Learning Standard	No of Item	Format	Item	Difficulty (in logits)	SE	Infit MNSQ	Outfit MNSQ	PTMEA
1.1.1	2	MCQ	HA1	-3.18	0.59	1.03	1.15	0.30
1.1.1		MCQ	KA4	-1.29	0.22	1.04	0.99	0.34
1.1.2	7	PC	KB1	-1.48	0.11	1.13	1.27	0.56
1.1.2		PC	LB4	-1.24	0.13	1.35	1.30	0.48
1.1.2		MCQ	L11R82	-0.69	0.17	0.98	0.95	0.24
1.1.2		MCQ	TA2	-0.42	0.25	1.04	1.12	0.17
1.1.2		PC	L1R7	-0.39	0.10	1.02	0.89	0.36
1.1.2		PC	SB1	-0.29	0.09	1.29	1.32	0.26
1.1.2		PC	TB3	1.63	0.18	1.03	1.01	0.33

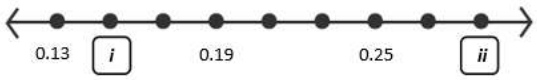
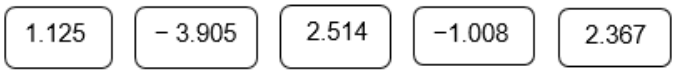
*MCQ= multiple-choice question, PC= partial credit question

Table 4 (Continued). Item statistics

Learning Standard	No of Item	Format	Item	Difficulty (in logits)	SE	Infit MNSQ	Outfit MNSQ	PTMEA
1.1.3	4	PC	LB7	-1.15	0.18	0.68	0.62	0.72
1.1.3		PC	LC14	-1.06	0.13	0.83	0.73	0.70
1.1.3		PC	L9R76	-0.90	0.15	1.04	0.90	0.21
1.1.3		MCQ	LA2	0.45	0.31	1.08	1.08	0.22
1.1.4	8	MCQ	HA2	-2.46	0.42	1.01	0.83	0.13
1.1.4		MCQ	KA1	-2.46	0.29	0.80	0.51	0.50
1.1.4		MCQ	QA1	-1.88	0.20	0.90	0.83	0.41
1.1.4		PC	L4R30	-1.75	0.19	0.97	0.94	0.37
1.1.4		MCQ	RA2	-1.60	0.46	0.98	0.75	0.20
1.1.4		MCQ	LA1	-1.41	0.28	0.98	0.96	0.38
1.1.4		MCQ	LA7	-1.03	0.27	1.00	1.08	0.35
1.1.4		PC	QB1	0.07	0.16	0.87	0.85	0.48
1.2.1	3	MCQ	MA2	-1.93	0.18	1.04	1.03	0.39
1.2.1		PC	LC2	-0.56	0.17	1.22	1.19	0.38
1.2.1		PC	HC5	0.37	0.07	1.00	1.05	0.42
1.2.2	1	PC	QC1	-1.19	0.09	0.72	0.68	0.65
1.2.3	9	MCQ	MA3	-3.67	0.26	1.00	0.73	0.34
1.2.3		MCQ	KA2	-2.84	0.33	1.02	1.05	0.22
1.2.3		MCQ	NA3	-1.22	0.26	1.02	1.01	0.12
1.2.3		MCQ	L1R1	-0.90	0.17	1.00	1.06	0.16
1.2.3		PC	HB1	-0.57	0.08	0.91	0.92	0.46
1.2.3		PC	SC17	0.12	0.12	1.06	1.03	0.28
1.2.3		PC	KC2	0.46	0.14	1.34	1.27	0.38
1.2.3		PC	LB8	0.53	0.21	0.95	1.00	0.40
1.2.3		PC	QC2	1.10	0.09	0.90	0.87	0.55
1.2.6	9	MCQ	SA2	-1.57	0.28	1.04	1.22	0.05
1.2.6		MCQ	HA4	-1.54	0.28	1.01	0.97	0.16
1.2.6		MCQ	LA4	-1.41	0.28	0.95	1.00	0.40
1.2.6		PC	QC4	-0.46	0.08	1.22	1.37	0.43
1.2.6		MCQ	LA3	-0.44	0.27	0.86	0.81	0.52
1.2.6		PC	KC4	0.67	0.11	0.89	0.85	0.70
1.2.6		PC	TC1	0.87	0.09	1.12	1.12	0.46
1.2.6		PC	NC6	1.37	0.08	1.13	1.20	0.36
1.2.6		MCQ	NA2	1.60	0.18	0.99	0.99	0.24
1.3.2	3	MCQ	TA19	-1.74	0.40	1.00	0.80	0.19
1.3.2		PC	LC5	-0.72	0.15	0.90	0.87	0.61
1.3.2		MCQ	L11R81	0.83	0.13	0.96	0.93	0.34
1.3.3	4	MCQ	TA1	-1.88	0.43	1.00	0.88	0.17
1.3.3		PC	MC2	-1.25	0.09	1.15	1.38	0.68
1.3.3		PC	HC2	0.54	0.10	0.98	0.95	0.38
1.3.3		PC	QC3	1.36	0.10	1.03	1.00	0.44
1.3.4	1	PC	L7R64	0.94	0.06	1.02	1.11	0.52
1.4.1	2	PC	L1R6	-1.20	0.11	1.02	0.91	0.31
1.4.1		PC	L7R63	-0.06	0.10	1.07	1.09	0.27
1.4.3	5	MCQ	MA4	-2.00	0.18	0.99	0.97	0.44
1.4.3		MCQ	NA1	-0.65	0.22	1.03	1.07	0.11
1.4.3		MCQ	L3R20	-0.63	0.16	0.94	0.86	0.49
1.4.3		PC	KC3	0.10	0.13	1.14	1.03	0.51
1.4.3		PC	L11R87	0.44	0.10	0.99	0.95	0.37
1.4.4	2	MCQ	NA4	-0.61	0.21	0.95	0.90	0.30
1.4.4		PC	KC23	-0.08	0.10	1.32	1.33	0.57
1.5.1	2	PC	MB1	-1.81	0.08	0.80	0.64	0.80
1.5.1		PC	NC1	0.23	0.09	0.95	0.99	0.46
1.5.2	4	PC	LC15	-0.13	0.11	1.11	0.81	0.55
1.5.2		PC	NC2	0.22	0.08	1.18	1.13	0.37
1.5.2		PC	HB5	0.74	0.06	0.98	0.91	0.53
1.5.2		PC	KC12	0.92	0.12	1.23	1.23	0.51
1.5.3	3	PC	L11R86	-0.57	0.13	1.10	1.15	0.28
1.5.3		PC	MC3	-0.23	0.11	1.08	0.86	0.62
Total	71							

*MCQ= multiple-choice question, PC= partial credit question

Table 5. Example of items according to difficulty

Item label	Difficulty	Level	Item
MA3	-3.67	Easy	$468 \div (6 \div 3) \times 2$ A 26 C 104 B 84 D 162
HA1	-3.18	Easy	Sarah is at 4 m below sea level. What is the appropriate integer to represent Sarah's position? A 4 C × 4 B -4 D ÷ 4
KA2	-2.84	Easy	$5\ (6 - 20) - (-9) =$ A – 76 B 76 C 61 D -61
QB1	0.07	Moderate	Arrange the following in descending order -4, -9, -7, 0, -5, 2
KC3	0.10	Moderate	Solve $1.32 - \left(\frac{-2.8}{0.7}\right) + (-6.4)$
L7R63	-0.10	Moderate	Complete the following number line. 
TB3	1.63	Difficult	The following shows a few numbers. State the prime numbers. 
NA2	1.60	Difficult	A submarine is 190 m below sea level. The submarine descends at 12 m per minute for the first 5 minutes. Then for the following 10 minutes, the submarine ascends 15 m per minute. Find the final position of the submarine. A 193 m C 280 m B -100 m D -187 m
L5R45	1.42	Difficult	The following diagram shows cards with decimal numbers.  Arrange the cards in ascending order. [2 marks]

first learning standards 1.1.1, with both were in the form of multiple-choice questions (MCQ). The difficulty of items was estimated from the Winsteps software. Since the mean difficulty was set at 0, then the negative sign showed that respondents have more than a 50% chance of getting HA1 and KA4 correctly, with the latter was considered as more difficult based on its larger values. The SE indicated the standard error of the estimation. The infit and outfit MNSQ values of 1.03 and 1.15 signify that there were only 3% and 15% variation from the model's expectations for the on-target and off-target participants. Finally, the positive value of 0.30 of the point-measure correlation (PTMEA) yielded evidence that Item HA1 was working together with other items in measuring the participants' ability in Rational Numbers. In general, it seems that the teachers developed relatively easy items for this topic since the respondents have more than 50% of getting correct answer for 44 (61.97%) items. Table 5 shows examples of easy, moderate, and difficult items.

DISCUSSION

From Table 5, it can be observed that the results for the easiest items were duly expected. Previous studies in Malaysia have shown that students have a high mastery level when answering items that measure procedural understanding, such as items MA3 and KA2. One possible explanation was that the ability to perform a series of computational tasks has always been exposed to the students since primary school (Rittle-Johnson et al., 2001). Therefore, the students were quite familiar with the types of items and had no problem solving them.

Item HA1 was endorsed as one of the easiest-to-score since the item was very similar to the examples in the textbook. It is plausible that the teachers had gone through similar items with the students in the classroom. Materials from textbooks are always used as primary sources for teaching and learning activities, as demonstrated by Lepik et al. (2015). As a result, when asked again in these tests, students need to recollect the solution steps taught in class instead of engaging in high-level cognitive tasks like interpreting or evaluating. While there was a possible explanation for the easy-to-score items, the same could not be generalized to the difficult items. This is because, based on the explanation given by the teachers, these items were considered easy items since they are measuring low-level learning standards such as recognizing integers (item TB3) or arrange positive and negative decimals (item L5R45). Even though item NA2 requires students to solve a problem, it is considered a routine problem, and the students may have discussed it with their teachers during lessons. Since the teachers themselves did not have the explanations why these items were perceived as difficult, perhaps there is a need to retrace the

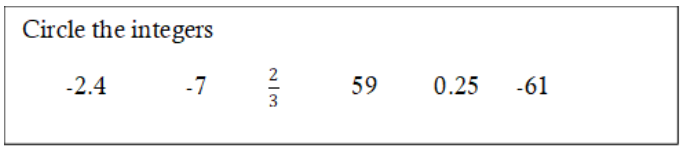


Figure 1. Item KB1 (Difficulty measure = -1.48 logits)

Table 6. Statistics for Items measuring Learning Standards 1.1.2 (Recognize and describe integers)

Label	Item	Difficulty Measure
LB4	Mark √ for prime numbers 27 ( ) 41 ( ) 55 ( ) 33 ( ) 43 ( ) 61 ( ) 37 ( ) 51 ( ) 70 ( )	-1.24 logits
L11R82	The following list showed prime numbers in ascending order. State the values of X and Y. 17, 19, X, 29, 31, Y, 41 A 21, 33 B 21, 37 C 23, 33 D 23, 37	-0.69 logits
TA2	The first prime numbers that are more than 40 are A 43, 45, 53 B 43, 47, 53 C 41, 45, 47 D 41, 43, 47	-0.42 logits
LIR7	Mark √ for prime numbers 19 ( ) 25 ( ) 31 ( ) 39 ( ) 45 ( ) 49 ( ) 53 ( ) 69 ( )	-0.39 logits
SB1	Circle the integers -2.4 0.4 -7 $\frac{2}{5}$ 0 59 0.25 -.51	-0.29 logits
TB3	The following shows several numbers. State the prime numbers (1) (2) (9) (17) (33)	1.63 logits

students' responses and identify if there was a problem during the teaching and learning of these items.

The calibrated pool of items developed from this study may help teachers in multiple ways. Firstly, it can be used to provide appropriate examples in the classroom. It is widely accepted that teachers should begin by providing easy examples to help students understand a concept before going progressively with more challenging ones (Bills & Bills, 2005; Rowland, 2008). We believe that the effecting of instructional scaffolding like this can be best implemented using the calibrated pool of items. To illustrate, to teach learning standards 1.1.2 (Recognize and describe integers), teachers may use the item KB1 as the first example to convey the concept of integers (see Figure 1). This is because the item is the easiest, and it is conceivable that the students would be able to understand how they can come up with the answer. Teachers may start by explaining the definition of integers and ask the students whether -2.4 is an integer or not. Note that the answer is 'no' because it is a decimal number and not a whole number. Teachers then should ask the students why it is not an integer. Next, the teachers may proceed with the subsequent number, i.e., -7 and ask the students the same question again. After that, teachers can ask the students to identify whether $\frac{2}{3}$, 59, 0.25, and -61 are integers or not. We would expect that the students will circle 59 and -61 and provide justifications. If somehow the students have difficulty in recognizing the integers, teachers may provide remedial activities at this early stage before students get more difficulties at the more advanced stage.

After that, teachers may use more difficult items from the pool, such as item LB4 as examples to strengthen the students' ability in recognizing and describing integers, particularly with regards to the prime numbers (see Table 6). Then teachers may ask the students to try answering the more difficult items L11R82 and TA2 on their own since not only both items involve prime numbers, but also possess a similar degree of difficulty. Finally, teachers might want to give items L1R7, SB1, and TB3 as part of homework together with other items from the textbook. Note that since TB3 was the most difficult items within this learning standard, then perhaps the teacher might want to revisit this item in the next class to see whether the students have difficulties in answering it.

Even though the development of the pool of items was able to help teachers with instructional scaffolding for students by starting with easy examples and followed by more difficult ones, we believe that there is still a need to include more items. For example, we can see that only KB1 involves integers, decimals, and fraction, while other items in this learning standard

measured students' knowledge in prime numbers. Therefore, we believe there is a need to add more items that measure decimals and integer since these two concepts were often considered as challenging for the students (Barnett-Clark et al., 2010; Chval et al., 2013; Idris & Narayanan, 2011; Morge, 2011; Razak et al., 2011).

Apart from facilitating teachers in providing examples, the pool of items developed in this study also has several other potentials. For example, at the end of the topic, teachers may use the pool of item for diagnostic purposes. That is, teachers can assemble some of the items to form a diagnostic test and then chart students' performance for each learning standards. Using this approach, teachers can diagnose each students' strengths and weaknesses with regards to the specified learning standards. Teachers can then plan a more focused intervention based on the diagnostic information.

For students who demonstrated a high level of proficiency, the teachers can engage the students in enrichment activities such as disseminating more challenging items from the item pool. Also, teachers can develop different forms of test from the pool so that even though the students may not need to sit for the same test, their performance can still be compared. For instance, a teacher might choose ten items from the pool to create a short test on Rational Numbers to be administered to one particular class. The teacher then can select different items from similar difficulties to create another set of test to be administered to another class. Since all items were calibrated on a common scale, the performance of students in both classes can still be compared despite them answering different sets of test items (Holland & Dorans, 2006). This practice effectively helps maintain the security of the test items.

CONCLUSION

The current study described the process of developing a calibrated pool of items in the topic of Rational Numbers to facilitate teachers in giving examples in classroom learning. We were able to pool 71 high-quality test items with varying degree of difficulties that were calibrated on a common scale. We were also able to provide evidence of validity and reliability of all items to measure the students' ability in rational numbers. Subsequently, we presented indications that the difficulty measure of each item is highly reproducible when subsets of the pool items were administered to other groups of students. We also furnished some guidelines on how teachers could use the pool of items in giving examples as well as for other assessment purposes.

Whilst several encouraging outcomes were demonstrated, the present study is bounded by few limitations. Firstly, this study delivers a strong assumption that the test items were administered to a relatively homogenous sample of students. We would have little knowledge about the results of the calibration if the samples were drawn from a heterogeneous population. Secondly, even with the large sample of items tested in this study, we believe that the items still represent a limited sample of stimuli for measuring students' ability in rational numbers. As such, there is still a need to develop more items before the pool of items can maximize its capability to be used for formative purposes especially with regards to learning standards 1.1.1, 1.2.2, 1.3.4, 1.4.4, 1.5.1 since we believe each of the standards should be measured by at least three items.

Author contributions: All authors have sufficiently contributed to the study, and agreed with the results and conclusions.

Funding: The research was funded by the Ministry of Education through the Fundamental Research Grant Scheme (FRGS 203.PGURU/6711857).

Acknowledgements: We would like to thank Universiti Sains Malaysia for providing the opportunity to conduct this study as well as all students for their time and willingness to participate.

Declaration of interest: No conflict of interest is declared by authors.

REFERENCES

Aung, A. A., & Lin, S. (2020). Development of grade six mathematics item bank by applying item response theory. *Learning Science and Mathematics, 15*, 40-55.

Barnett-Clark, C., Fisher, W., Marks, R., & Ross, S. (2010). *Developing essential understandings of rational numbers for teaching mathematics in grades 3-5*. NCTM.

Bills, C., & Bills, L. (2005). Experienced and novice teachers' choice of examples. In P. Clarkson, A. Downton, D. Gronn, M. Horne, A. McDonough, R. Pierce, & A. Roche (Eds.), *Proceedings of the 28th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 146-153). MERGA.

Bjorner, J. B., Kosinski, M., & Ware Jr, J. E. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT™). *Quality of Life Research, 12*, 913-933. https://doi.org/10.1023/A:1026163113446

Blinder, S. M. (2013). *Guide to essential math: A review for physics, chemistry and engineering students (2nd ed.)*. Elsevier. https://doi.org/10.1016/B978-0-12-407163-6.00002-3

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

Chval, K., Lannin J., & Jones D. (2013). *Putting essential understanding of fractions into practice in grades 3-5*. NCTM.

Dowker, A., Bala, S., & Lloyd, D. (2008). Linguistic influences on mathematical development: How

important is the transparency of the counting system? *Philosophical Psychology, 21*, 523-528. https://doi.org/10.1080/09515080802285511

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

Guskey, T. R. (2003). What makes professional development effective? *Phi Delta Kappan, 84*, 748-750. https://doi.org/10.1177/003172170308401007

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Praeger.

Huang, R. (2006). Tension and alternative of in-service secondary mathematics teacher profession development in China. In *Proceeding of the Second International Forum on Teacher Education* (pp. 162-179).

Idris, N., & Narayanan, L. (2011). Error patterns in addition and subtraction of fractions among form two students. *Journal of Mathematics Education, 4*(2), 35-54.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69-81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x

Kallinger, S., Scharm, H., Boecker, M., Forkmann, T., & Baumeister, H. (2019). Calibration of an item bank in 474 orthopaedic patients using Rasch analysis for computer-adaptive assessment of anxiety. *Clinical Rehabilitation*, *33*(9), 1468-1478. https://doi.org/10.1177/0269215519846225

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Lepik, M., Grevholm, B., & Viholainen, A. (2015). Using textbooks in the mathematics classroom – the teachers' view. *Nordic Studies in Mathematics Education, 20*(3-4), 129-156.

Li, Y., Kulm, G., & Smith, D. (2006). *Facilitating mathematics teachers' professional development in knowledge and skills for teaching in China and the United States* [Invited presentation]. Second International Forum on Teacher Education. October 25-27, 2006. Shanghai, China.

Lin, C. Y., Becker, J., Byun, M. R., Yang, D. C., & Huang, T. W. (2013). Preservice teachers' conceptual and procedural knowledge of fraction operations: A comparative study of the United States and Taiwan. *School Science and Mathematics, 113*, 41-51. https://doi.org/10.1111/j.1949-8594.2012.00173.x

Linacre, J. M. (2002). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878. http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2006). *User's guide to WINSTEPS computer program*. Winsteps.com.

Liu, R. D., Ding, Y., Zong, M., & D. Zhang, D. (2014). Concept development of decimals in Chinese elementary students: A conceptual change approach. *School Science and Mathematics, 114*(7) 326-338. https://doi.org/10.1111/ssm.12085

Mazzocco, M. M., Myers, G. F., Lewis, K. E., Hanich, L. B., & Murphy, M. M. (2013). Limited knowledge of fraction representations differentiates middle school students with mathematics learning disability (dyscalculia) versus low mathematics achievement. *Journal of Experimental Child Psychology, 115*(2), 371-387. https://doi.org/10.1016/j.jecp.2013.01.005

Ministry of Education (2016). *Dokumen standard kurikulum dan pentaksiran matematik Tingkatan 1* [Form 1 mathematics curriculum and assessment standard documents]. Kementerian Pendidikan Malaysia.

Morge, S. P. (2011). Family Connections: Helping children understand fraction concepts using various contexts and interpretations. *Childhood Education, 87*(4), 282–282. https://doi.org/10.1080/00094056.2011.10523193

Ni, Y., & Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40*(1), 27-52. https://doi.org/10.1207/s15326985ep4001_3

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema*, *29*(3), 390-395. https://doi.org/10.7334/psicothema2016.391

Razak, F., Noordin, N., Dollah, R., & Alias, R. (2011). How do 13-year-olds in Malaysia compare proper fractions? *Journal of ASIAN Behavioral Studies, 1*(3), 31-40. https://doi.org/10.21834/jabs.v2i2.177

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. Psychological Science, 24(7), 1301-1308. https://doi.org/10.1177/0956797612466268

Rittle, J. B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346-362. https://doi.org/10.1037/0022-0663.93.2.346

Rowland, T. (2008). The purpose, design and use of examples in the teaching of elementary mathematics. *Educational Studies in Mathematics, 69*(2), 149-163. https://doi.org/10.1007/s10649-008-9148-y

Siegler, R. S., & Lortie-Forgues, H. (2015). Conceptual knowledge of fraction arithmetic. *Journal of Educational Psychology, 107*(3), 909-918. https://doi.org/10.1037/edu0000025

Siegler, R. S., & Lortie-Forgues, H. (2017). Hard lessons: Why rational number arithmetic is so difficult for so many people. *Current Directions in Psychological Science, 26*(4), 346-351. https://doi.org/10.1177/0963721417700129

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., et al. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691-697. https://doi.org/10.1177/0956797612440101

Son, J. W., & Senk, S. (2010). How reform curricula in the USA and Korea present multiplication and division of fractions. *Educational Studies in Mathematics, 74*(2), 117-142. https://doi.org/10.1007/s10649-010-9229-6

Sun, X. H. (2019). Bridging whole numbers and fractions: Problem variations in Chinese mathematics textbook examples. *ZDM Mathematics Education, 51*, 109-123. https://doi.org/10.1007/s11858-018-01013-9

Tian, J., & Siegler, R. S. (2018). Which type of rational numbers should students learn first? *Educational Psychology Review, 30*(2), 351-372. https://doi.org/10.1007/s10648-017-9417-3

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review, 1*(2), 133-147. https://doi.org/10.1016/j.edurev.2006.05.001

van Hoof, J., Verschaffel, L. & van Dooren, W. (2015). Inappropriately applying natural number properties in rational number tasks: Characterizing the development of the natural number bias through primary and secondary education. *Educational Studies in Mathematics, 90*, 39-56. https://doi.org/10.1007/s10649-015-9613-3

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

Yetim, S. & Alkan, R. (2013). How middle school students deal with rational numbers? A mixed methods research study. *Eurasia Journal of Mathematics, Science & Technology Education, 9*(2), 213-221. https://doi.org/10.12973/eurasia.2013.9211a

Zodik, I., & Zaslavsky, O. (2008). Characteristics of teachers' choice of examples in and for the mathematics classroom. *Educational Studies in Mathematics, 69*(2), 165-182. https://doi.org/10.1007/s10649-008-9140-6