# THE DEVELOPMENT OF AN ITEM BANK OF WORD PROBLEMS TO MONITOR PROGRESS OF STUDENTS WITH LEARNING DIFFICULTIES IN MATHEMATICS

[1*]Lim Luck Siew

*National Institute of Education, Singapore*

[*]*lucksiew.lim@nie.edu.sg*

## ABSTRACT

As children progress from preschool to primary school, the content of mathematics learning gains more depth and an emphasis on problem solving is one of the core processes in the mathematics curriculum. Word problems are accorded such significance as the problem solving skills involve the development of other life skills such as communication and reasoning. Students with difficulties in mathematics often encounter difficulties solving word problems. Prior to intervention in supporting students, an item bank of word problems with different levels of difficulty should be developed. This allows for word problems of similar difficulties to be administered to monitor the progress. In addition, the difficulty level of the word problems in the item bank could also inform teachers of the potential difficulty level of the word problems that could affect their students' performance. In order to develop the item bank of word problems, the word problem type of combine, change and compare were administered to 191 students in Primary 3. Using their performance in the word problems, the Rasch model of measurement was used to analyze the data to determine the difficulty level of the word problems. There were 75 word problems administered and the difficulty level of the word problems ranged from most difficult word problem at 5.58 logit while the easiest was -4.92 logit. Among the three types of word problems, the Change Word Problem had the widest range of difficult and less difficult word problems. The difficulty level of the word problems enables word problems of varying difficulties to be included at baseline and also during intervention to monitor students' progress during intervention.

**Keywords:** word problems, mathematics difficulties, item bank, Rasch analysis

## 1. INTRODUCTION

The joy of thinking mathematically is present in young children (Clements & Sarama, 2009). Such an assumption encourages preschool teachers to create an environment that provides the experience of mathematics learning through play and exploration. As children progress from preschool to primary school, the content of mathematics learning gains more depth (e.g. number and operations, algebra, geometry, measurement). There is also an emphasis on problem solving as one of the core processes in the mathematics curriculum.

DOI: https://zenodo.org/record/8336676
Published by: https://publication.seameosen.edu.my/index.php/icse/issue/view/5

According to Fuchs et al. (2016), the importance of word problems can be observed from their presence being an integral component of the mathematics curriculum. Word problems are accorded such significance as the problem solving skills involves the development of other life skills such as communication and reasoning (Niss et al., 2016).

Therefore, schools place emphasis on students learning to become proficient problem solvers. The teachers need to apply alternative strategies and teach specific skills in word problem solving as soon as they have identified the students who struggle with word problems. In the implementation of an intervention to improve students' word problem solving skills, teachers would be required to monitor students' progress to check if the intervention is effective in improving their word problem solving.

The objective of this study is to develop an item bank of word problems (pertaining to combine, change and compare word problems) to provide a set of word problems that could be used to create tests to monitor the progress of students during intervention. The tests to be created for monitoring progress can be of similar difficulty so that students' performance can be measured objectively. The difficulty level of the word problems in the item bank can also inform teachers of the potential difficulty level of the word problems that can affect their students' performance.

## 2. LITERATURE REVIEW

The literature review focused on the types of word problems and the use of Rasch analysis in determining difficult level of word problems.

### Types of Word Problems

According to Verschaffel, Schukajlow, Star and Van Doreen (2020), word problems are one of the most challenging aspects of mathematics learning. Riley et al. (1983) had used the classification of the word problems into change, combine, compare and equalizer. The equalizer word problems were not included in the present study, as most of the research that involved word problem solving (e.g. schema-based instruction) had mainly used the three types of word problems, combine, change and compare. Riley et al. (1983) reiterated that the semantic structure and the identity of the unknown quantity in the word problems are challenges, which make word problem solving difficult. They defined semantic relationship as, "conceptual knowledge about increases, decreases, combination and comparisons involving sets of objects" (p.159). Similarly, Peltier, VanDerHeyden and Hott (2022) also recognised the importance of identifying the structure of the word problem to support students in devising an appropriate plan to solve word problems.

Riley and Greeno (1988) reported that *Combine Word Problems* that required finding the subset of the combination were more difficult than the unknown combination. In *Change Word Problems*, the word problems in which students had to find the unknown result were the easiest, and this was followed by the change unknown which was less difficult than when the start quantity was unknown. In the *Compare Word Problems*, the most difficult subtype was word problem that needed students to calculate the referent's amount. The difference unknown and the compared quantity unknown were less difficult to solve than the unknown referent word problems. Riley et al. (1983) highlighted that in the *Change Word Problems*, the unknown starting amount was difficult for students from kindergarten to Grade 3. They also noted that the *Compare Word Problems* in which the referent was unknown were more difficult than the other *Compare Word*

*Problems*. The difficulties of the subtypes within each word problem type were used as a reference in the development of the word problems for the item bank.

## Use of Rasch Analysis

The Rasch model of measurement is based on a stochastic model used to describe the probable outcome when a person takes an assessment or a test (Bond & Fox, 2007). According to Bond and Fox (2007), the Rasch model of measurement assumes that the outcome is governed by two parameters – the ability of the participant and the difficulty of the item. Therefore, when the ability of the participant is greater than the difficulty of the word problem, the participant is likely to succeed on that word problem. Similarly, when the difficulty of the word problem is greater than the ability of the participant, it is less likely that the participant will get the correct answer for the word problem.

The raw test scores obtained from assessments indicate the gaps between scores, but it does not inform equal interval of measurement. Rasch model of measurement is known to achieve the “ideal of interval scale measurement of the latent trait to the same degree” (Baylor et al., 2011 p. 246). The Rasch model transforms raw scores to their natural logarithm or logit, which provides for a more useful equal interval scale of measurement (Bond & Fox, 2007). The logit is sometimes compared to a “ruler” in which the logit is the measurement that places word difficulties and students’ ability on a standardized scale. According to Bond and Fox (2007), the logit range is usually within three logits and minus three logits, with zero logit as the mean logit representing the test sample. The higher logits are associated with higher ability of obtaining a correct answer, representing higher ability. At the same time, it would also mean that the higher value of the logit would also mean that the question is more difficult. The converse is also true. The illustration showing the difficulty level of items and the ability of the participants can be seen in Figure 1. In the illustration, there were seven participants in zero logit and 1, indicating that most of the participants were in the average range to above average in terms of ability. For the questions, question 12 was the most difficult. The questions (8, 2, 3, 4) were considered to be of average difficulty as they were at the level of zero logit.

***Figure 1:*** An Example Showing Participant Ability and Question Difficulty

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | | | | x | 12 | | | |
| 2 | | | | x | | 5 | | |
| 1 | | x | x | x | 1 | 6 | 7 | |
| 0 | x | x | x | x | 8 | 2 | 3 | 4 |
| -1 | | | x | x | 9 | | | |
| -2 | | | | x | | | | |
| -3 | | | | | | | | |
| | | | | | | | | |
| Logits | | Participants | | | | Questions | | |

One of Rasch models of analysis that can evaluate the fit of the items to the underlying construct is the item fit analysis. According to Bond and Fox (2007), the “item fit” is able to ascertain if the items being scrutinized are either diverged or closely unacceptable from the expected ability/difficulty pattern. The item fit in Rasch model analysis provides the construct validity of the tests. It is considered the “quality control mechanism” (Bond & Fox, 2007, p. 35) as the analysis keeps in check if the items measure only one construct, thus holding the assumption of unidimensionality. Using the item fit analysis, items that do not fit the underlying construct of the study, have to be reconsidered for inclusion in the study.

The mean square residual (MNSQ) weighted infit and unweighted outfit is commonly used in the item fit statistics (Bond & Fox, 2007). The MNSQ weighted infit statistics and the unweighted outfit statistics were used to determine if the items were "behaving" as expected of the Rasch model (Bond & Fox, 2007). According to Bond and Fox (2007), the MNSQ weighted infit is more sensitive to unexpected patterns than the MNSQ unweighted outfit statistics. The infit statistics place greater weight for the person ability, which is nearer to the item difficulty. Wright and Stone (1999) reiterated that infit statistic is useful as it is "robust with respect to idiosyncratic outliers" (p. 112). The unweighted outfit statistic is not weighted and thus, is sensitive to outlying scores.

In addition, the z-standard scores or the t-standard infit (infit *t*) scores would be used to determine the suitability of the items to be included in the research. The infit *t* is one of the alternative measures that would inform the deviation or the adherence to the Rasch model. Bond and Fox (2007) recommended that the values of the z-score or the infit *t* scores fall between -2.0 and +2.0. This was the range of values used in this study. The item reliability index would also be used in the word problem analysis.

In summary, the Rasch model of measurement was used to determine the difficulty level of the word problems. It used the MNSQ infit and outfit statistics and infit and outfit *t* scores for the item fit analysis, item reliability and item difficulty in the analysis of the word problems.

## 3. METHODOLOGY

The objective of constructing an item bank of word problems was to provide a set of word problems that can be used to monitor the progress of the students. Wright and Stone (1999) described the item bank as one that consists of items that had been jointly calibrated which provide an "operational definition of one variable" (p. 107). They also emphasized that a well-constructed item bank would allow for "the best possible test for any possible assessment" (Wright & Stone, 1999, p. 10). Therefore, the item bank is likely to have items of varying difficulties. The application of the Rasch model analysis to empirical data allows for the word problems to be arranged from the least to the most difficult according to logit.

### Design of The Item Bank Construction

The design of the item bank construction was similar to the steps recommended by Wright and Stone (1999) in forming a group of items into a calibrated bank. Wright and Stone (1999) outlined the following steps. First, the bank plan focused on the items to be banked and the creation of the items. This was followed by the test administration which comprised the assembling of forms, giving out of tests and filing of responses. Finally, the calibration of the items as well as the item and student reports, were generated. In this study, the first stage of bank plan focused on the principles of the word problems and the processes in item bank development. These processes include the writing of word problems, teacher validation and student validation. At the second stage, the tests were administered after the recruitment of participants. The final stage was the analysis of the word problems using Rasch analysis. Following this, the item bank of word problems continued with the categorisation of the word problems into word problem type and level of difficulty.

## Development of Word Problems

The construction of the item bank involved the development of word problems for the three types of word problems. In the design of the item bank, two important qualities, relevance and different levels of difficulties, were emphasized. In order to achieve these two qualities, the 2015 MOE primary mathematics curriculum, local mathematics resources that were used in the lower primary mathematics (e.g. text books and school examination papers) were referenced. The word problems in the item bank had the following characteristics: the use of either addition or subtraction in the solution to the word problem, the solutions requiring one or two equations to be formulated and the numerical value of the quantity (when summed) should not exceed 1,000.

The numerical values in the word problems were deliberately kept small to maintain the motivation of the students in attempting and learning the strategies for word problems (Riley et al., 1983). The numerical values were also kept within the Primary 3 mathematics syllabus, where students are required to learn numbers up to 10, 000 (MOE, 2012). The adherence to the lower primary mathematics syllabus addressed the internal validity, since the characteristics of word problems did not deviate from the national mathematics syllabus.

With the given characteristics of the word problems as a guide, the researcher adapted the word problems from current Primary 3 mathematics textbooks, workbooks (e.g. My Pals Mathematics and My Pals Mathematics workbook) and one-step word problems from Riley et al. 1983. The characteristics of word problems that were often found to be difficult for students were cross-checked against studies such as Riley et. (1983) and Ostad (1998). The differentiating characteristics of the word problems were the different contexts and the numerical values used. The numbers were randomly generated such that the sum of the two numbers would not exceed 1,000.

After adapting the word problems from various sources such as textbooks and journals, the researcher also consulted a senior teacher in mathematics regarding the suitability of the word problems. The researcher had 10 years of experience in teaching primary school mathematics, while the senior teacher had more than 20 years of teaching experience in mathematics.

It was decided that 75 word problems would be developed for the item bank based. It was planned that there would be 20 tests for each word problem type to monitor students' progress. The 75 word problems for the item bank were distributed across six pen and paper tests. The tests were labelled Test A – Test F with 20 word problems within each test. In each test, the 20 word problems consisted of a combination of *Combine*, *Change* and *Compare Word Problems*. Among these 20 word problems, there were at least six from each of the word problem categories. There were nine anchoring word problems that were common to all the tests. The purpose of these anchoring word problems was to have a network of common items in which the tests are connected.

Following the development of the word problems, it was planned that the word problems would be drawn randomly for a pilot test to check the clarity of the word problems. After the pilot test, modifications were made to word problems that were found to be ambiguous. The student's workings were also analysed to find out if student was able to understand the wording in the word problem. The next step after developing the word problems was to recruit participants to attempt the word problems. The participants had to be Primary 3 students from local schools.

## Participants

Before administrating the word problem tests to the students, the research proposal was sent to the NTU Institutional Review Board (NTU-IRB) for approval and permission to administer the word problem tests to students in schools was also sought from the Ministry of Education (MOE), Singapore. After the approval was given, an email to the primary schools inviting participation was sent to the principals of the schools. Subsequent to approval being given by the principals to collect data in the school, parents' consent was obtained before students were able to participate in the study. This was followed by the students' consent to participate in the study.

The participants were Primary 3 students recruited from four primary schools and one afterschool care center. This included students of different academic abilities and represented by different ethnic groups and genders. There were 191 students who participated in this phase. Among these students, there were 64 boys and 127 girls. There were more girls among the participants as one of the participating schools was an all-girls school.

## Procedure

The tests were administered to the students who consented to participate in the study. The tests were stratified so that each school had the right selected tests based on participants who had given consent. After the students completed the tests, the tests were scored by the researcher. An answer key was created to provide an objective method of accepting solutions to the word problem.

The researcher scored all the word problems in the tests using the prepared answer key. To ensure that there was interrater reliability, a primary school senior teacher with 24 years of mathematics teaching experience scored 25% of the completed tests using the same prepared answer key. Interrater reliability was computed using point by point agreement ratio. This meant the researcher would score each word problem according to the given criteria of a correct equation in the first time. For the second round of scoring, the senior teacher would score the tests. The researcher and the senior teacher would check if they agreed on the criteria for each word problem. Thus, in scoring each word problem, both the researcher and senior teacher had to agree that either all the criteria were present or a particular criterion was missing. The interrater reliability was calculated using the following formula –

$$\frac{number\ of\ agreements\ on\ the\ number\ of\ word\ problem}{number\ of\ agreements\ on\ the\ number\ of\ word\ problems\ and\ the\ disagreements\ on\ the\ number\ of\ word\ problems}$$

The interrater reliability for the tests was 99.3%. When there was no agreement, the researcher and the senior teacher discussed on the point of disagreement and came to a consensus with respect to awarding the mark or not awarding it.

## ANALYSIS

The final step in the construction of the item bank was the analysis and calibration to determine the difficulty level of the word problems. In this study, Winstep Version 3.68.0 was used to analyse the data. Word problem answered with the correct equation or correct result of the equation were scored 1 while those with incorrect equation was scored 0. When the data are scored and entered as 0 or 1 in Winstep, the Rasch dichotomous analysis is used to obtain the statistical information such as item reliability, item fit, and the difficulty level of the word problems. The steps in Figure 2 provide the sequence in which the analysis of the word problems was conducted to obtain the difficulty level.

After the word problems were analysed using Rasch analysis, they were categorised into their word problem types – combine, change and compare. The difficulty level of the word problems was calibrated again after grouping the word problems. This was to determine the difficulty of the word problems when those belonging to the same category had been grouped together. Within each category, the word problems were subdivided according to their level of difficulty. These subdivisions were obtained from the standard deviation in the analysis. The levels of difficulty were determined by the standard deviations from the mean. The levels of difficulty are as follows: Level 1 (one to two standard deviation below mean), Level 2 (half to one standard deviation below mean), Level 3 (half-standard deviation below mean to mean), and Level 4 (mean to one standard deviation above mean).

***Figure 2**: Steps in the Analysis of Word Problems Using Winstep Program*

| Steps in the analysis of word problems |
|---|
| 1. Analyse the item fit of the word problems |
| -Outfit statistics then infit statistics |
| -MNSQ statistics before *t* statistics |
| -High values before low or negative values |
| 2. Check the Reliability index |
| -Person reliability index |
| -Item reliability index |
| 3. Remove word problems which do not fit |
| 4. Check the reliability index |
| 5. Check the variance |
| -Person standard error of person mean |
| -Person standard deviation |
| 6. Collate the difficulty level of the word problems |

# RESULTS AND FINDING

### DIFFICULTY LEVELS OF WORD PROBLEMS

Following the analysis of the statistics to confirm the reliability of the word problems for the item bank, the difficulty level of the word problem was collated. The most difficult word problem was 5.58 logit while the easiest item was -4.92 logit. The mean difficulty of all the word problems was -0.7 logit. This means that a majority of the word problems were not difficult for most of the Primary 3 students. As there was a wide range in the difficulty level, it was decided that word problems which were more than 2 standard deviation below the mean and word problems more which were than 1 standard deviation above the mean would not be included.

The range of difficulty level, the standard deviation and item reliability are summarized in Table 1. Among the three types of word problems, the *Change Word Problem* had the widest range of difficult and less difficult word problems. The standard deviation was the largest for *Compare Word Problems*. The item reliability is interpreted in the same way as Cronbach's alpha statistics.

***Table 1**: The Range of Level Difficulty, Standard Deviation and Item Reliability*

| Type of word problem | Range of difficulty level | Standard deviation | Item reliability |
|---|---|---|---|
| Combine | 4.43 to -4.92 | 1.99 | 0.88 |
| Change | 7.28 to -4.16 | 2.21 | 0.92 |
| Compare | 5.28 to -3.35 | 3.23 | 0.94 |

After the word problems were categorised, they were placed into their various levels of difficulty, starting from Level 1 to Level 4. The number of word problems distributed across Level 1 to Level 4 of difficulty level is seen in Table 2.

***Table 2**: The Distribution of Word Problems According to Difficulty Levels*

| Levels of difficulty | No. of combine word problem and percentage | No. of change word problem and percentage | No. of compare word problem and percentage |
|---|---|---|---|
| Level 1<br>(1 to 2 standard deviation below mean) | 3 (15%) | 2 (10.5%) | 4 (19%) |
| Level 2<br>(Between 1 to ½ deviation below mean) | 5 (25%) | 7 (36.8%) | 4 (19%) |
| Level 3<br>(Between ½ deviation below mean to mean) | 4 (20%) | 5 (26.3%) | 2 (9.5%) |
| Level 4<br>(Between mean and 1 deviation above mean) | 8 (40%) | 5 (26.3%) | 11 (52.4%) |

Within the *Combine Word Problems*, the word problems that required two or more equations to solve were more difficult than those that required one equation. These word problems were either in Level 3 or Level 4. A total of 60% of *Combine Word Problems* were between Level 1 and Level 3. This suggests that the *Combine Word Problems* were generally less difficult in their level of difficulty.

Similar to the *Combine Word Problem*, *Change Word Problems* requiring two or more equations to solve were in Level 4 difficulty. Riley et al. (1983) identified three subtypes of *Change Word Problems* – result unknown, change unknown and start unknown. From the analysis, word problems at Level 3 and Level 4 are mostly those which involved solving the start unknown or change unknown value.

The *Compare Word Problems* involve the comparison of the two quantities. Riley et al. (1983) grouped comparison word problems to finding the difference between the quantities, the quantity compared unknown and the referent unknown. In *Compare Word Problem*, word problems that had the referent unknown were generally more difficult than those involving finding the difference between quantities and the quantity compared unknown.

## 4. DISCUSSIONS, RECOMMENDATIONS AND CONCLUSIONS

In order to have a range of less difficult and more difficult word problems, the word problems were developed with a variation of the unknown position in the equation (e.g., the start of the equation being an unknown). The following characteristics of word problems were found to be categorised in the Level 4 word problem difficulty: two-step word problems, and unknown starting quantity for *Compare Word Problems*, unknown referent quantity (e.g*., "John has 10 marbles. He has 4 marbles fewer that Tom. How many marbles does Tom have?"*).

Most of the students were familiar with one-step word problem. The difficulty of the two-step word problems arose as the word problems required the students to understand the semantic relationships in a more complex word problem. In the *Change Word Problem*, the difficulties could arise due to weak understanding of the part-whole relationship. When the start value is unknown, the part-whole concept is important as it enables the students to understand that whole is formed from the change quantity and the quantity that results from the change. The *Compare Word Problems* can be identified by the comparison of the two quantities. In *Compare Word Problem*, the structure of where the referent is unknown requires students to have a strong conceptual understanding of the concepts of "more than" and "less than". Many students have difficulties with this subtype of *Compare Word Problem*. The word problems which were more difficult were consistent with the subtypes of those which were found in the literature. Their difficulty level was also linked to the subtypes of the word problems which were often associated with the semantic structure and the position of the unknown quantity in the equation. Peltier et al. (2022) in their research of word problem solving had also emphasized on the importance of categorising word problems according to their schema structure. They suggested adhering to the sequence of combine, change and compare in order for students to experience success in their word problem solving as they build their skill level.

While the literature has compared the subtypes within each type of word problems and described their difficulty in the process of solving them, the Rasch analysis was able to provide a "standardized measurement" for comparison. The item bank of word problems with their level of difficulty allows for teachers to select less difficult word problems to be used to test the students' basic understanding of the word problem and more difficult ones to be used to check if students apply the strategies taught at intervention.

The limitation of this study is that the item bank of word problems consists only of the word problem types – combine, change and compare. It could be expanded to include other word problem types such as equaliser word problem type.

### Conclusion

In summary, the use of the Rasch model in the analysis of word problem difficulty allowed the word problems to be categorised into different levels of difficulty in an item bank. The item bank would be beneficial for teachers to select word problems of varying difficulties to form tests or assessments which can monitor the progress of their students' progress objectively.

As this research had focused on word problems with whole numbers, future research could also focus on word problems involving fractions, money with the same word problem category of combine, change and compare. We could find out if students who recognise the word problem schema would be able to apply to other types of word problems involving fractions or money.

## REFERENCE

Baylor, C., Hula, W., Donovan, N. J., Doyle, P.J., Kendall, D., & Yorston, K. (2011). An introduction to item response theory and Rasch model for speech-language pathologists. *American Journal of Speech-Language Pathology, 20*, 243-259. doi: 10.1044/1058-0360(2011/10-0079)

Bond, T.G., & Fox, C.M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). Erlbaum.

Clements, D.H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach.* Routledge.

Fuchs ,L. S., Gilbert, J. K., Powell, R.S., Cirino, P.T., Fuchs, D., Hamlet, C.L., Seethaler, P.M., & Tolar, T.D. (2016). The role of cognitive processes, foundational math skill, and calculation accuracy and fluency in word-problem solving versus prealgebraic knowledge. *Developmental Psychology,52, 2085-2098.* doi:http://dx.doi.org/10.1037/dev0000227

Ministry of Education (2012). *Primary mathematics teaching and learning syllabus.* Singapore: Curriculum Planning and Development Division.

Niss, M., Bruder, R., Planas, N., Turner, R., & Villa-Ochoa, J. A. (2016). Survey team on: Conceptualisation of the role of competencies, knowing and knowledge in mathematics education research. *ZDM Mathematics Education, 48*, 611-632. doi: 10.1007/s11858-016-0799-3

Ostad, S. A. (1998). Developmental differences in solving simple arithmetic word problems and simple number-fact problems: A comparison of mathematically normal and mathematically disabled children. *Mathematical Cognition, 4*(1), 1-19.

Peltier, C. VanDerHeyden, A.M., & Hott, B. (2022). Strategies to help students solve addition and subtraction word problems. Beyond Behaviour, 31, 29-41. doi: 10.1177/10742956211072260

Riley, M. S. & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*(1), 49-101.

Riley, M. S., Greeno, J. G, & Heller, J. I. (1983). The development of children's problem-solving ability in arithmetic. In H.P. Ginsburg (Ed.), *The development of mathematical thinking*, (pp. 153-190). Academic Press.

Verschaffel L., Schukajlow S., Star, J., & Van Doreen, W. (2020). Word problems in mathematics education: a survey. ZDM, 52, 1-16. doi: https://doi.org/10.1007/s11858-020-01130-4

Wright, B.D., & Stone, M. (1999). *Measurement essentials* (*2nd Ed*.). Wide Range.