Implementing Assessment *for* Learning: An Application of the Rasch Model for the Construction of a Mathematics Assessment to Inform Learning

Jingjing Yao[1, 2] and Magdalena Mo Ching Mok[1]

[1]Assessment Research Centre, The Hong Kong Institute of Education
[2]Department of Psychology, Zhejiang Normal University

Acknowledgement

This study was supported by a grant from the General Research Grant sponsored by the Research Grants Council of Hong Kong (Project Number 844011).

1. Introduction

Mathematical literacy is a key competence for individuals and organisations in society today (OECD, 2010; National Mathematics Advisory Panel, 2008). In order to prepare students' capacity for understanding and applying mathematics, there has been an increase of attention in recent years on the stronger integration among mathematics curricula, instruction, and assessment of mathematics, especially in the foundation stage of primary school (Abakhani, 2011; Arslan & Ozinar, 2010; Bulut, 2007).

The subject of mathematics has had a long history of using quizzes, tests and examinations for the assessment *of* learning. That is, the identification of standards reached by students, particularly at the end of key learning stages, used to be the sole purpose of assessment. Needless to say, assessment *of* learning is important for accountability purposes; and in this regard, schools and teachers are accountable to taxpayers, parents, and school sponsoring bodies. Assessment *of* learning is also necessary to maintain academic standards of the education system and for articulation between education systems. Nevertheless, assessment *of* learning alone is inadequate to prepare our students in facing

challenges unique to the twenty-first century workplace. Furthermore, we now know a lot more about the power of assessment in transforming teaching and learning. Notably, large scale reviews undertaken by Black and Wiliam (1998), Hattie and Timperley (2007), Kluger and Denisi (1996), Mory (2004), Narciss and Huth (2004), and Shute (2008), and independent classroom-based research conducted by Berry (2008), Carless (2007), Lee (2012), Mok (2010), Salvage (2011), and others have repeatedly shown that through providing diagnostic feedback, assessment can inform and support further learning of students. These studies provided empirical evidence that quality feedback helps students to diagnose their learning progress, how well they have learned, identifies the gaps and the nature of misconceptions. As a result, students are supported with information to improve their learning; a development that would not occur without such feedback information. This means that whereas validity, reliability, and discrimination are essential features of assessment of learning, diagnostic feedback is a critical component of learning assessment.

Despite consistent research on the importance of feedback to learning, the implementation of assessment *for* learning is often hindered by the lack of tools required for the generation of diagnostic feedback. If we step back and reflect on the question: When teachers ask their students to take a test, what feedback information do they want from this test? The answer is two-fold. First, teachers should want to know more about their students; namely, how well each individual student is doing, the level of target knowledge mastery and skills of the whole class, and how to help each and every one of them to further their learning from where they are, based on the diagnostic information derived from students' misconceptions and non-mastery. Second, teachers may also want to know about the quality of the test; namely, the overall test difficulty and the difficulty of individual items. Some teachers may also want to address the questions of reliability, fairness, and validity of the test and its individual items.

A certain amount of the feedback information above can be generated by experienced teachers themselves through inspecting the distribution of raw scores for the whole class, or by inspecting individual item- and student-responses. But other information (e.g., the issue of "where to go from

here?" for individual students) can hardly be obtained without the help of analytic tools. This is the time when such analytic tools as the Rasch model (Bond & Fox, 2007) could be helpful.

Moreover, the literature of Embretson (1996) warns against traditional methods of using raw scores in the analysis of assessment data. Research has shown that raw scores cannot be assumed to be interval-level data (Stevens, 1946), and treating them as such will lead to the misinterpretation of test quality and of student achievement. For instance, to improve from a score of 98 to 99 in a test, with a maximum score of 100, is much harder than to improve from a score of 71 to 72 in the same test. This means the distances between the two pairs of raw scores might be both one unit mathematically, but they are not of the same distance. It is therefore erroneous to compare students directly based on raw scores (Embretson, 1996).

In view of the need for diagnostic tools in support of assessment for learning, this study aims to illustrate, through an example of Rasch analysis on students' responses to a 35-item mathematics assessment designed for Primary 5 students in Hong Kong, how the Rasch model (Bond & Fox, 2007) could be used to optimize the effectiveness of assessment for learning in school-based assessments. In this example, the easily accessible Winsteps software (Linacre, 2011) was used and the main steps of Rasch analysis for extracting diagnostic information in support of teaching and learning were introduced didactically.

2. Method

2.1 Participants

This is part of a larger longitudinal study on assessment feedback, self-directed learning, and mathematics achievement of primary students. Participants for the current study comprised a sample of 1368 Primary 5 students from 16 Hong Kong schools. There were 648 males (47.4%), 716 females (52.3%), and 4 students (0.3%) did not report their gender (Table 1). This study observed all ethical compliances set by the university where the

authors worked, and informed consents from parents and schools were obtained before the commencement of the study.

Table 1. Sample Distribution by Gender

	N	%
Male	648	47.4%
Female	716	52.3%
Missing	4	0.3%
Total	1368	100%

2.2 Instrument

A 35-item mathematics test was developed after careful analysis of the Hong Kong mathematics curriculum, and in consultation with teachers on the suitability of the test for Primary 5 students by the end of Semester One. All the items in the test were multiple choice questions with four options and only one of the options was the correct answer. The test consisted of 25 items in the Number domain, three items in the Shape and Space domain, and seven items in Measures domain. Within the Number domain, nine items involved the understanding of the basic concepts of whole numbers and fractions, 12 items involved performing addition, subtraction, multiplication operations, as well as mixed operations on whole numbers and fractions, and another four items involved solving application problems. In the Shape and Space domain, the three items were on direction and location. In the Measurement domain, the seven items were on the calculation of perimeters and areas (Table 2).

Table 2. Item Domain Contents and Numbers

Domain	**Contents**	**Items**
Number	basic concepts of whole numbers and fractions	9
	addition, subtraction, multiplication and their mixed operations	12
	application problem solving	4
Shape & Space	direction and location	3
Measures	calculation of perimeters and areas	7

2.3 Procedure and Analysis

First, we analysed the local curriculum carefully, selected the representative contents and typical items from textbooks, exercises and other related materials. On the basis of this analysis, we designed items for the assessment. The purpose of the study and drafts of the assessment were presented to participating primary mathematics teachers in order to consult with them on suitability and practicability of the assessment for the targeted cohort of students. After several rounds of consultation and revisions, the final version of the assessment was administered at the end of Semester One to 1368 Primary 5 students from 16 Hong Kong schools under the supervision of mathematics teachers during normal school time.

Rasch analysis was conducted using the Winsteps software (version 3.72.3) (Linacre, 2011) to validate the mathematics assessment. As responses to the items in the assessment were scored, either right or wrong, a dichotomous Rasch model (Rasch, 1960) represented in equation (1) was used to estimate difficulties of the item, or the item measures, and mathematics ability of the student, or the person measure, on a common interval scale of mathematics ability.

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

where P_{ni1} is the probability of person n making a correct response to item i. Correspondingly, P_{ni0} (= 1 − P_{ni1}) is the probability of this person n making a wrong response to the same item i. In the Rasch model expressed in equation (1), the probability of a correct response is a logistic function of the difference between the person ability θ_n and item difficulty δ_i. Thus, we can place item difficulty and student ability on the same measurement scale for interpretation. It will be discussed in later sections of this manuscript that other diagnostic information about the students and the assessment, for example the extent to which items making up the assessment fall into a single dimension, can be generated using the Rasch model.

3. Results

The Rasch analysis conducted below was divided into two parts: (a) the validation of the mathematics assessment, and (b) the data analysis for diagnostic information. In order to validate the assessment, the unidimensionality of the mathematics assessment according to the Rasch model was first tested to ascertain the extent to which the assessment was underpinned by a single Rasch measurement. Next, other indices of validity were generated, including person and item reliabilities, item difficulty, item fit, and gender differential item functioning (DIF) (Wang, 2008). This validation process aimed to guarantee the quality of the mathematics test as an appropriate and valid instrument for assessing the students. It also informed the teachers of the characteristics of the assessment, so that the teachers could decide whether or not the items could be included in an item bank for future assessment purposes.

The diagnosis aspect of the data analysis included the generation of estimated mathematics ability of students for the whole group as well as for individual students, the individualised diagnostic map (called the Person-Kid-Map, abbreviated as PKMAP in Winsteps), which provided information on the Zone of Proximal Development (Vygotsky, 1978) of each student, and the person Keyforms for each student, from which observed and highly unexpected responses could be identified and diagnosed. Collectively, this diagnostic information gathered is made possible through the powerful functions of Winsteps software (Linacre, 2011). The results reported below illustrate how teachers could make effective use of diagnostic information for formative purposes of assessment.

In terms of validation of the mathematics assessment, the results showed that: (1) unidimensionality of the assessment was supported by the data; (2) Rasch person and item reliabilities were acceptable; (3) with one exception, all items had item-fit between 0.5 and 1.5; (4) item difficulties ranged from -2.30 to 2.83; (5) there was good alignment between item difficulty and student ability; and (6) there was no gender DIF detected among the items. In terms of diagnostic information, the analysis showed that (7) student ability ranged between -2.58 and 4.19; (8) person diagnostic PKMAPs provided information on: the Zone of Proximal Development, the items being mastered comfortably, and future learning goals for each of the students; and (9) the person Keyforms provided information on: the extent to which responses to each item for each

student were within or out of expectation. Details of the information are reported in the sections that follow.

3.1 Unidimensionality of the Assessment

Test of unidimensionality of the assessment was conducted through a Principal Components Analysis of Rasch residuals subroutine in the Winsteps software (version 3.72.3) (Linacre, 2011). The Principal Components Analysis of Rasch residuals is used to detect if there is more than one factor that can explain the response structure (i.e., unidimensionality) by comparing differences between the observed and the expected responses (Raîche, 2005; Linacre, 2011). Simulation studies by Raîche (2005) found eigenvalues of the first contrast in the Principal Components Analysis of test from 20 to 60 items to be in the range of 1.4 to 2.1. The results were subsequently replicated by Linacre and Tennant (2009). The literature of Linacre (2011) recommends researchers to use eigenvalue of the first contrast being less than 2.0 as an acceptable criterion for establishing unidimensionality. In this study, the Principal Component eigenvalues in the first contrast was 1.8 (below 2.0), and 31.2% of raw variance were explained by the Rasch measures. The result indicated that the mathematics assessment was likely to be underpinned by a single dimension, which was consistent with the assessment design intent.

3.2 Reliability of Item and Person Measures

Internal consistency of assessment items in terms of Cronbach's alpha was 0.82 for the Primary 5 mathematics assessment, which was an acceptable reliability in accordance with classical test theory. The Rasch analysis also found that the assessment had a Rasch item reliability of 1.00, an item separation index of 17.17, a Rasch person reliability of 0.80, and a person separation index of 1.98. These results mean that the assessment had excellent item reliability, and the items could be separated into nearly 17 groups according to responses by students. On the other hand, the person reliability was just acceptable, and the students could be separated into almost two groups by the items in the assessment. If we take the different number of items and students into consideration, the differences in reliabilities and separation indices between items and persons could be interpreted easily. It is easy to separate 35 items by 1368 students, but it is comparatively more difficult to separate 1368 students by only 35 items.

3.3 Item Fit, Item Difficulty, and Alignment between Item and Person

Item statistics are presented in Table 3, including estimates of item difficulties, their standard errors, item goodness of fit (both Infit and Outfit), and point-measure correlation for each item. These statistics support the validity and reliability of the assessment. More details are given in the sections below.

Table 3. Item Difficulty, Standard Error, Fit, and Point-Measure Correlation

			Infit		Outfit		PTME Corr.
Item	Difficulty	*SE*	MNSQ	ZSTD	MNSQ	ZSTD	
Q16	2.83	0.08	1.16	2.85	2.01	8.28	0.07
Q12	2.23	0.07	1.01	0.33	1.19	2.73	0.32
Q34	1.48	0.06	1.11	4.00	1.23	4.99	0.27
Q22	1.37	0.06	1.03	1.11	1.07	1.73	0.37
Q31	1.17	0.06	1.16	6.21	1.29	7.37	0.24
Q15	1.14	0.06	1.02	0.95	1.10	2.74	0.37
Q25	1.12	0.06	0.96	-1.58	0.98	-0.59	0.44
Q19	0.98	0.06	0.89	-5.12	0.87	-4.02	0.51
Q6	0.96	0.06	1.04	1.87	1.05	1.63	0.37
Q27	0.85	0.06	1.01	0.28	1.02	0.76	0.40
Q13	0.73	0.06	1.04	1.75	1.07	2.36	0.37
Q35	0.73	0.06	1.10	4.59	1.13	4.17	0.31
Q18	0.66	0.06	0.94	-2.94	0.93	-2.31	0.47
Q33	0.50	0.06	1.02	0.87	1.03	0.93	0.39
Q7	0.43	0.06	1.18	7.69	1.24	7.02	0.24
Q20	0.38	0.06	0.99	-0.46	1.00	-0.05	0.42
Q29	0.36	0.06	0.94	-2.68	0.93	-2.21	0.46
Q26	0.26	0.06	0.92	-3.73	0.90	-2.99	0.48
Q28	0.03	0.06	1.19	7.26	1.36	8.55	0.20
Q2	-0.13	0.06	1.01	0.41	1.03	0.66	0.38
Q21	-0.17	0.06	0.89	-4.32	0.83	-4.11	0.50
Q32	-0.21	0.06	0.90	-3.78	0.83	-4.01	0.49
Q24	-0.47	0.07	0.98	-0.81	0.89	-2.18	0.41
Q23	-0.83	0.07	0.93	-2.01	0.91	-1.43	0.42
Q10	-0.91	0.07	0.82	-4.96	0.66	-5.79	0.53
Q11	-0.93	0.07	0.95	-1.20	0.99	-0.11	0.38
Q5	-1.11	0.07	0.95	-1.06	0.95	-0.64	0.37
Q30	-1.21	0.08	0.92	-1.80	0.80	-2.68	0.41
Q8	-1.30	0.08	0.94	-1.20	0.92	-0.95	0.36
Q17	-1.36	0.08	0.88	-2.47	0.76	-2.94	0.43

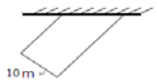
Q1	-1.40	0.08	0.99	-0.25	1.01	0.16	0.32
Q3	-1.62	0.09	1.04	0.65	1.22	2.05	0.22
Q9	-2.07	0.10	0.94	-0.82	1.08	0.68	0.29
Q4	-2.18	0.10	0.93	-0.92	0.80	-1.49	0.31
Q14	-2.30	0.11	0.86	-1.68	0.56	-3.54	0.38
Mean	0.00	0.07	0.99	-0.10	1.02	0.40	0.37
SD	1.24	0.01	0.09	3.10	0.24	3.60	0.10

3.3.1 Item difficulty and Wright Map

In Table 3, the item difficulty estimated values are listed in ascending order of difficulty and they range from -2.30 to 2.83. Their standard errors are all small and in the order of 0.1. The three most difficult items are items Q16 (item difficulty 2.83), Q12 (2.23), and Q34 (1.48), while the three easiest items are Q14 (-2.30), Q4 (-2.18), and Q9 (-2.07). These items are presented in Table 4.

Q16 and Q12 are questions dealing with fractions (number domain) and Q34 is a measures problem dealing with the perimeter and the area of a trapezoid. Although these three items are different in domain types, they share a common feature that the problem expression is a little complicated and needs to be well understood before accurate computation. For instance in Q16, multiple choice option A attracted most students who just added 3 grams of sugar to the existing 2 grams of sugar, and divided the sum by the 100 grams of water. In so doing, they did not take into account that the term sweet soup meant a mixture of water and sugar, and so the 3 grams of sugar needed to be included both in the numerator and in the denominator. Since sweet soup is a common diet in Hong Kong, it is unlikely that students committed the error for cultural reasons. Rather, the error was more likely to have arisen because students were confused by the complicated language expression of sugar, water, and sweet soup. The result suggests possible interference to students' mathematical abilities by way of their language abilities. On the other hand, we found these students were good at calculation of whole numbers (Q14), understanding concepts of numbers (Q4) and identifying positions (Q9). Each of these items had a success rate of over 90%. Overall, the Primary 5 mathematics test is an appropriate and valid instrument to detect the mathematics performance for Primary 5 students.

Table 4. The Three Most Difficult Items and the Three Easiest Items

The 3 most difficult items and the 3 easiest items		Item difficulty	Key
Q16	There are 2g of sugar and 100g of water in a bowl of sweet soup. If 3g of sugar is added, then sugar becomes（　　）of the sweet soup. A.　　B. $\frac{1}{21}$　　C.　　D. $\frac{5}{102}$	2.83	B
	Percentages choosing the Options for Q16 are: A: 48%　　**B: 14%**　　C: 17%　　D: 22%		
Q12	3 is added to the numerator of $\frac{1}{5}$. In order to make the fraction unchanged, which of the following should be done to the denominator? A. + 3　　B. - 3　　C. x 3　　D. x 4	2.23	D
	Percentages choosing the Options for Q12 are: A: 31%　　B: 11%　　C: 37%　　**D: 21%**		
Q34	Referring to the picture below, if we string a rope 58m long from the left side of the figure to the wall, what is the area cordoned off by the rope?  A. 480 m^2　　B. 240 m^2　　C. 580 m^2　　D. 290 m^2	1.48	B
	Percentages choosing the Options for Q34 are: A: 14%　　**B: 33%**　　C: 36%　　D: 17%		
Q9	Referring to the figures in Q8, Tom walks (…) to go from his home to the school and then walks (…) to go to the hospital. A. east, south　　B. east, north C. west, south　　D. west, north	-2.07	A
	Percentages choosing the Options for Q9 are: **A: 91%**　　B: 2%　　C: 5%　　D: 1%		
Q4	In which of the following numbers does “5” have the largest value? A. 15706439　　B. 16905347 C. 79654310　　D. 96574130	-2.18	A
	Percentages choosing the Options for Q4 are: **A: 92%**　　B: 1%　　C: 1%　　D: 6%		
Q14	14 children are playing together. They have spent $840 in total. How much has each child spent on average? A. 840 ÷ 14 = $60　　B. 14 × 14 = $196 C. 840 + 14 = $854　　D. 840 × 14 = $11760	-2.30	A
	Percentages choosing the Options for Q14 are: **A: 93%**　　B: 2%　　C: 2%　　D: 3%		

The item difficulty ranging from -2.30 to 2.83 indicated an appropriate difficulty level span. This is further supported by the Wright Map (Figure 1), which shows that the Primary 5 mathematics test was well matched against the sample. In the Rasch approach, a Wright Map is a visual representation of the distribution of the respondents' abilities in relation to the distribution of the item difficulties. Each # on the left panel represents seven students in this study and the numbers (e.g. Q16) on the right represent the items.

Items were plotted on the Wright Map according to their difficulties along the vertical straight line, which represented the mathematics ability scale, in the middle of the figure. Items at the top of the scale are more difficult items than those at the bottom of the scale. Students were plotted into the map according to their estimated mathematics abilities. More able students were at the top and less able students were at the bottom of the scale.

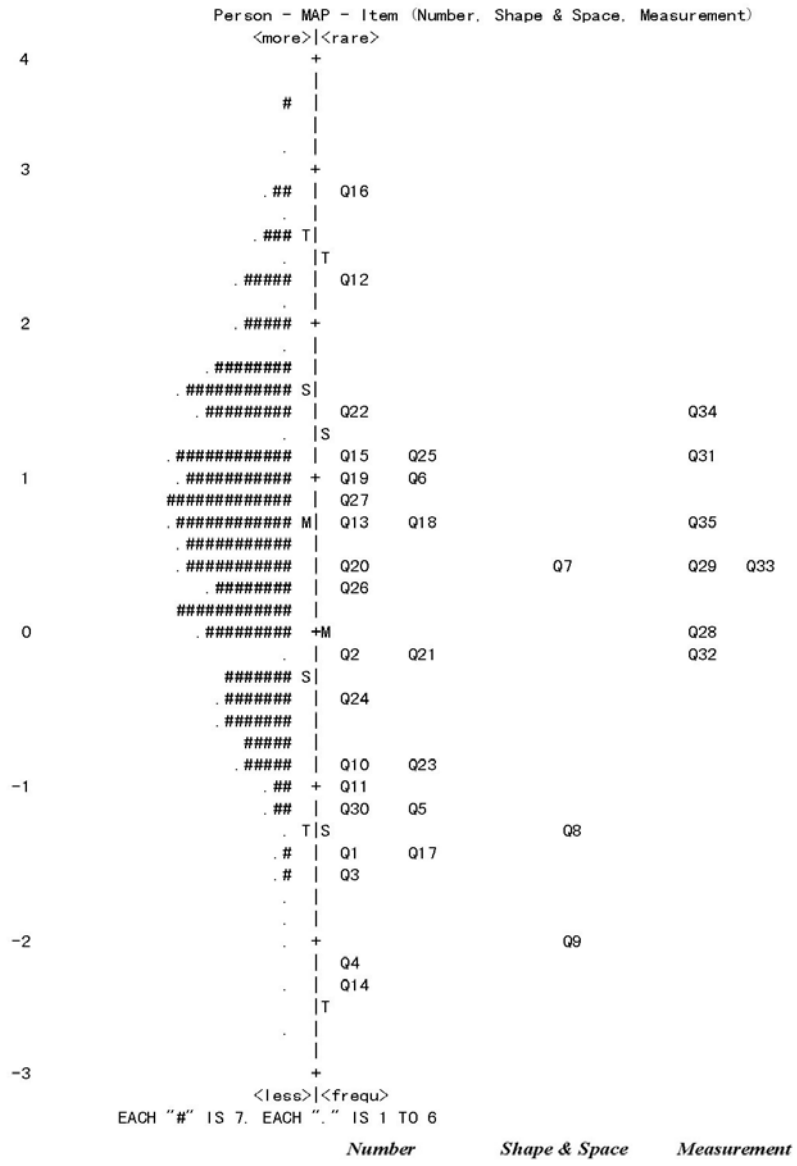


Figure 1. Wright Map of Items and Persons

In Figure 1, the mean (indicated by M on the right panel in Fig. 1) of item difficulty was close to the mean (indicated by M on the left panel in Fig. 1) of student ability within the first standard error (indicated by S). This means there was good alignment between student ability and item difficulty.

In Figure 1, items are listed as clusters according to different domains. The three items of the shape and space domain are listed in the middle of the right panel, while items of the number domain are in the left panel and items of measurement domain are in the right panel. On average, the three domains do not have substantially different mean difficulties. That is, no single domain presented more difficulty than another domain for the group of students.

Most of the items have difficulties near the mid-range of the vertical scale, which is around the mean item difficulty and within one standard deviation of the mean, while the two most difficult items Q16 and Q12 are located two standard deviations above the mean item difficulty. The three easiest items (Q9, Q4, Q14) are located outside two standard deviations toward the bottom of the scale. The majority of students on the left panel have ability levels above these three easiest items. It can be seen from this analysis that because of the good alignment between item difficulty and student ability, teachers can obtain a picture of the performances of the Primary 5 students as a whole, as well as their individual profiles.

3.3.2 Item fit

Fit statistics (Table 3) show the difference, or the residual, between the observed data and the estimated measure according to the Rasch model. Outfit mean square (MNSQ) is a mathematical function based on the mean of squared residuals. The computation of Infit MNSQ is similar to that of outfit except that each observation is weighted by its statistical information or the model variance. Statistical information is higher around the middle of the measurement scale where the observations tend to concentrate, and is lower towards the top and bottom tails of the scale where there are fewer observations. Infit ZSTD and outfit ZSTD are the standardized forms of Infit MNSQ and Outfit MNSQ

respectively (Bond & Fox, 2007; Linacre, 2011). Infit and Outfit statistics provide evidence of construct validity in Rasch measurement. Linacre (2011) recommended that Infit and outfit MNSQ values falling within the range of 0.5 to 1.5 can be taken as indication of good concordance between the data and the Rasch model. Items with goodness of fit values less than 0.5 or more than 1.5 are considered as having poor fit to the Rasch model. Table 3 shows that all items in the Primary 5 mathematics assessment, with the exception of item Q16 (which has an outfit MNSQ value of 2.01), have infit and outfit MNSQ values ranged from 0.56 to 1.36, which is well within the range of good fit.

When we referred to other statistics indexes, we found Q16 had a low point-measure correlation 0.07, while other items were all from 0.2 and 0.53. This result means that except for Q16, items in the assessment are internally coherent. As discussed in an earlier section, students might have failed Q16 because of a deficiency in common sense, which is not directly connected with knowledge and skills about fractions. However, mathematics thinking needs strictness and rationality. Item Q16 revealed deficiencies in the daily learning and training of Primary 5 students, and highlighted possible areas for enhancement in future instruction.

3.4 Differential Item Functioning

Differential Item Functioning (DIF) occurs when test-takers with same abilities in some measured latent trait have different probabilities of achieving a correct response to an item, which is considered an important issue in establishing test fairness (Wang, 2008). The magnitude of DIF signifies the extent to which the item parameter differs between different groups, such as gender, location, or social-economic status, even though the groups under comparison are of equal ability (Wang, 2008). In Rasch models, a DIF value of 0.5 logits or larger could be considered a substantial DIF (Wang, 2008). Recent research (Paek & Wilson, 2011) showed that for short tests with a small sample size, the Rasch model approach to DIF is more effective than the traditional approach of using Mantel-Haenszel probability. This analysis found that all items, with the exception of items Q9 and Q14, had very low DIF (less

than 0.5) for gender. The two exception items Q9 and Q14 had DIF values of 0.52 and 0.74 respectively, although both had low Mantel-Haenszel probability values (0.02 and 0.01 respectively). On this basis, it is concluded that items in the Primary 5 mathematics test assessment revealed no substantial DIF in gender, which means the mathematics assessment is fair to both boys and girls who take the test.

3.5 Person Diagnosis Information from Winsteps

From the teachers' perspective, diagnostic information on students' ability is perhaps the most precious. Discussions in earlier sections have already shown that the Rasch analysis using Winsteps (Linacre, 2011) can generate information on the mathematics ability of individual students. The Wright Map presents students' abilities alongside the items in the assessment and provides two frames of reference for the interpretation of each student: against the other candidates taking the same assessment, and against the ability requirement of the assessment items. Indeed, the Rasch analysis can generate at least two additional pieces of diagnostic information invaluable to teachers; namely, the Person-Kid-Map, and the Person Keyforms. These are discussed in the following sections.

3.5.1 Person-Kid-Map (PKMAP)

The Person-Kid-Map (PKMAP) is a graphical display of the zone of proximal development and response pattern of each individual student. An example taken from this study is the PKMAP of student number five presented in Figure 2. In the PKMAP, the estimated ability level of the student is represented by "xxx." Using this estimate as a focal point, the PKMAP divides the figure vertically into two panels, and horizontally into three panels, resulting in six regions of the graph. Located in the left panel are those items which the student answered correctly and items in the right panel are those which the student answered incorrectly. Items in the top panel were difficult for the student because their difficulty levels are at least 0.5 logits more than the ability level of the student. Items in the top panel are easy for the student because their

difficulty levels are at least 0.5 logits less than the ability level of the student. The middle panel contains items with difficulty levels within ±0.5 logits of the student's ability level. This is the Zone of Proximal Development of the student. Utilising these categorisations by the panels, and counting clockwise from the top-right region, the six regions in the PKMAP are:

1. **Non-mastery Future Goal Region**: Items that are difficult for the student, who answered them incorrectly. The items involved in this region (Q12 in this example; Fig. 2) give direction for future learning goals of student.
2. **Zone of Proximal Development Need-Scaffolding Region**: The student answered items in this region incorrectly (Q22, Q31, Q25, Q6 and Q35 in this example). This is the region where the student has not yet mastered the necessary knowledge and skills required to answer the items correctly, but if given support, the student will be able to achieve mastery.
3. **Carelessness/Special Learning Needs Region**: Items (Q33, Q28, Q23, Q10, and Q11) are easy for student, but the student still answered them incorrectly. The teacher should check items in this region carefully to see if there is something wrong with the items themselves. If the answer is negative, the teacher should find out if the student has made careless mistakes, lacks examination skills, has special learning difficulties (e.g., dyslexia), or misconceptions, and then seek to provide appropriate remediation.
4. **Mastery Region**: This is the region of mastery. Items (Q7, Q20, Q29, Q26, Q2, Q21, Q32, Q24, Q5, Q30, Q8, Q1, Q17, Q3, Q9, Q4, and Q14) in this region are easy for the student, who answered them correctly.
5. **Zone of Proximal Development Need-Consolidation Region**: Although the student answered items (Q15, Q19, Q27, Q13, and Q18) in this region correctly, learning is shaky and needs consolidation.
6. **Pleasant Surprise Region:** Items (Q16 and Q34) in this region are beyond the student's ability level. Nonetheless, the student answered them correctly. The teacher has to check to see if there are elements of luck, dishonesty, or that the student has learned the topics involved at other settings such as a tutorial school.

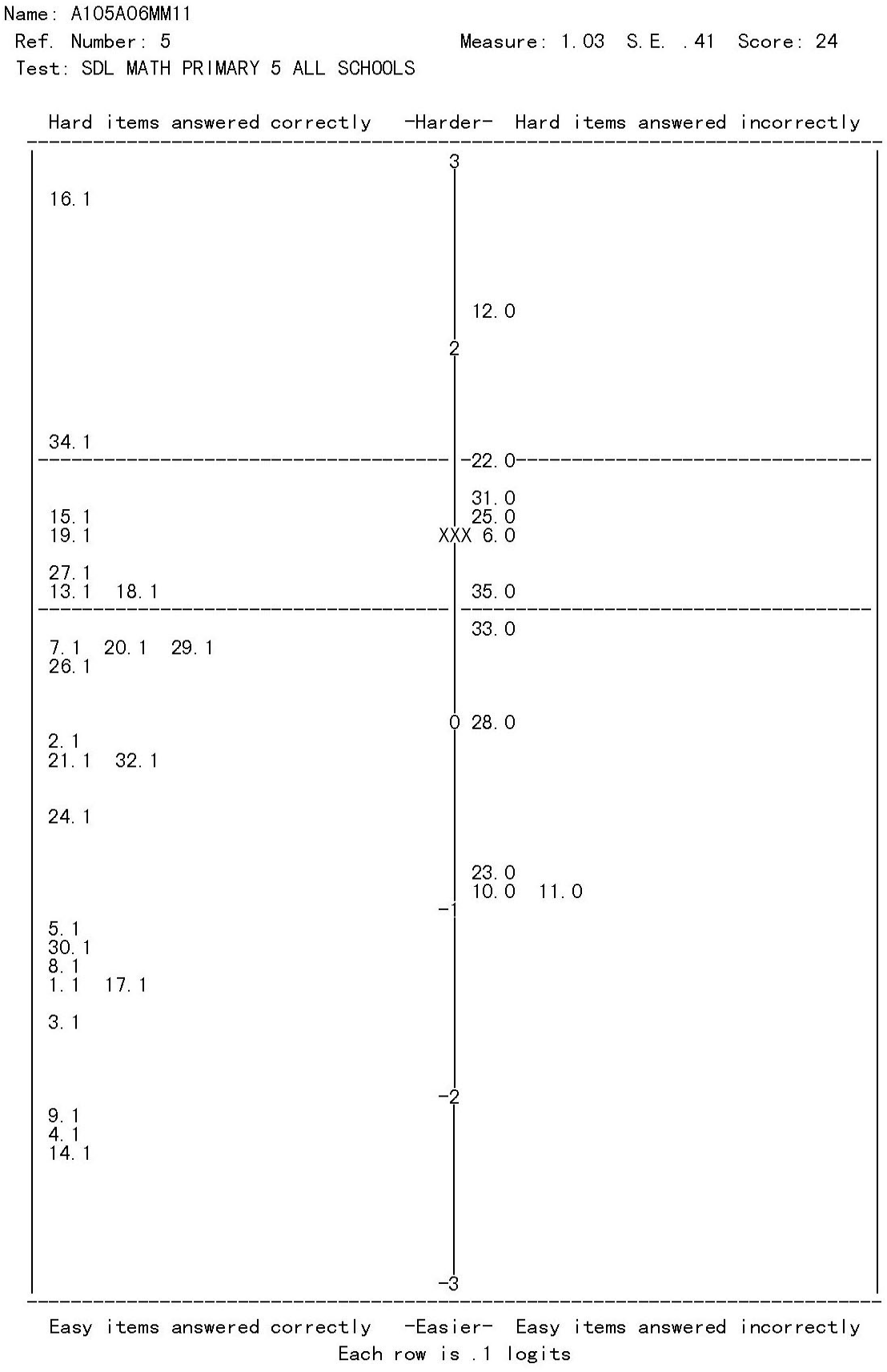


Figure 2. Person Diagnostic PKMAP

3.5.2 Person Keyforms

Person Keyforms is another graphical device produced by the Winsteps software (Linacre, 2011) that provides diagnostic feedback information for individual students. An example from this study is presented in Figure 3. In the Person Keyforms, the Rasch measurement scale is presented horizontally and extends from left (less able students) to right (more able students). Actual responses of the student to items in the assessment are printed vertically at the student's ability location.

In the example from this study, student number five has a person measure of 1.03 logits. Based on the student's person measure, the expected responses to Q16, Q12, Q34 . . . Q14 are, A, B, A . . . A, respectively and these predicted responses are printed in a column in the Person Keyform of this student at the location of their ability estimate. Observed responses to the right of the estimated ability column are those items that are more difficult than the ability of the student. The further away an observed response from the estimated ability column, the more of a discrepancy there is between the predicted and actual response.

The actual responses made by the student are printed in two forms; namely, either (a) with a period before and after the character of the response (for example, .B.) in situations where the actual response is not equal to, but not too highly unexpected, from the predicted response, or (b) with round brackets before and after the character of the response (for example, (B)) in situations where the actual response is highly unexpected compared to the predicted response. For instance, in the example presented in Figure 3, the student is predicted to choose option A for Q16, but the student has chosen option B instead, and so the symbol (B) is used in the figure. By referring to the horizontal ability scale, it can be seen that only students with an ability estimated at about 3.8 logits are predicted to choose option B in Q16, so it is highly beyond expectation that student number five, whose ability is only 1.03 logits, would choose this option. The teacher might want to investigate the reasons behind such a large discrepancy. Would cheating, luck, or other reasons be the answer?

Similarly, according to the Rasch model, student number five is estimated to choose option A for Q10, but instead the student has chosen option D. Only

students with ability estimated at around -2 logits, which is much lower than the ability estimate of student number five, would make such a choice for Q10. This choice of student number five is therefore highly unexpected and the choice is represented by (D) in Figure 3. Again, the teacher might want to find out more about such a large discrepancy. Would the reason be carelessness, under-preparation, lack of test-taking skills, previously unidentified misconceptions, or special learning difficulties?

By inspecting the actual and predicted options of each student, the teacher would get very specific information, based on which the teacher is then able to strategise the next course of action in support of the student's learning.

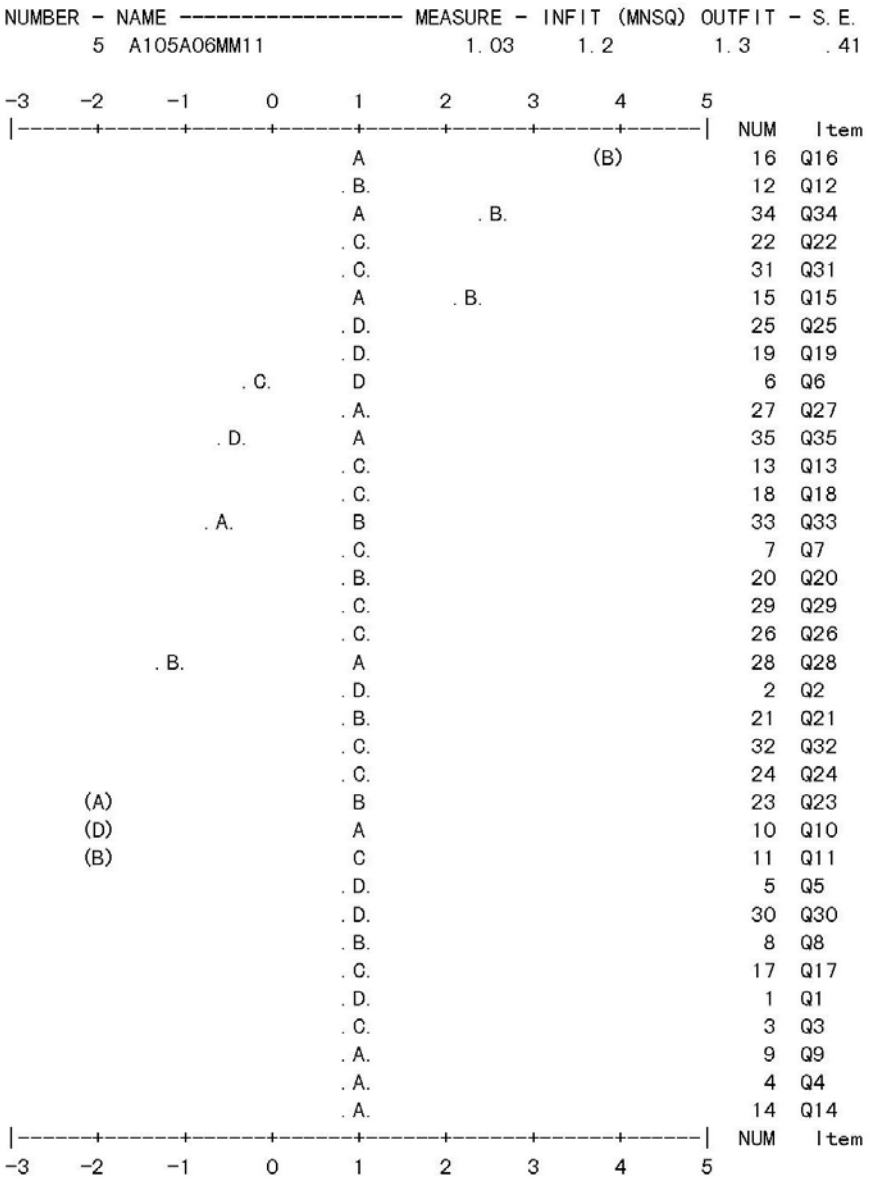


Figure 3. Person Keyform of Student Number Five

4. Conclusion

This study was part of a larger longitudinal study on the effect of feedback and self-regulated learning on mathematics achievement of primary students. The current study focused attention on the development of mathematics assessment for Primary 5 students in Hong Kong. It was undertaken to optimally use diagnostic information generated from Rasch analysis of assessment data in order to inform teaching and learning. The Rasch analysis conducted in this study made use of the Winsteps software (version 3.72.3) (Linacre, 2011) and some of the outputs were unique features of the software.

The Rasch measurement approach was used to develop and validate a 25-item mathematics assessment for Primary 5 students in Hong Kong in this study. Data analysis with Winsteps (version 3.72.3) (Linacre, 2011) showed the mathematics assessment is underpinned by a unidimensional construct, has acceptable item and person reliabilities, satisfactory item fit indices and item difficulties, good alignment between item difficulty and student ability, and has no gender DIF. The analysis undertaken demonstrates the procedures necessary for scientific inquiry into the validity of test scores, which are of key importance in all forms of testing (Messick, 1989; Ariffin, Omar, Isa, & Sharida, 2010). Establishing validity in test scores is particularly important to teachers in their implementation of assessment for learning because test scores form the basis of subsequent instruction.

As illustrated in this study, Rasch analysis can generate rich and imperative information for teachers about the assessment items and the students taking the assessment. Multiple frames of reference are available to the teachers to get both specific and holistic understanding of each student’s performance profile. The frames of reference include each individual student, performance of the entire group being assessed, difficulty of individual items, as well as all the items that constitute the test.

As an illustration, the study discussed selected items at the two extreme ends (most difficult and easiest items), as well as selected student responses in

order to show teachers how to detect issues arising from the item- or student-levels, and how to collect information for further teaching and remedial instruction. Amongst the information generated from the Rasch analysis, the diagnostic information provided by the Person-Kid-Map (PKMAP) and person Keyforms are most crucial for the identification of evidence regarding individual students' achievement. Through the PKMAP and the person Keyforms, teachers could get to know the Zone of Proximal Development of each student, areas of mastery and areas needing remediation.

5. References

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability multiple intelligent item using Rasch measurement model, *Procedia Social and Behavioral Sciences, 9*, 729-733.

Arslan, S., & Ozpinar, I. (2010). Assessment in mathematics course and evaluating primary school coursebooks in terms of assessment, *Procedia Social and Behavioral Sciences, 2*, 4157-4163.

Babakhani, N. (2011). The effect of teaching the cognitive and meta-cognitive strategies (self-instruction procedure) on verbal math problem-solving performance of primary school students with verbal problem-solving difficulties. *Procedia Social and Behavioral Sciences, 15*, 563-570.

Berry, R. (2008). *Assessment for learning*. Hong Kong: Hong Kong University Press.

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

Bulut, M. (2007). Curriculum reform in Turkey: a case of primary school mathematics curriculum. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(3), 203-212.

Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66.

Earl, L., & S. Katz. (2013). Getting to the core of learning: Using assessment for self-monitoring and self-regulation. In M. M. C. Mok, (Ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific,* pp. 123-137. Dordrecht Heidelberg, New York, London: Springer.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA design. *Applied Psychological Measurement, 20*, 201-212.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112.

Kluger, A., & DeNisi, A. (1996). The effect of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Lee, K. O. (2012). Physical Education in higher education in Hong Kong: The effects of an intervention on pre-service sports coaches' attitudes towards assessment for learning used in sports. In M. M. C. Mok (Ed.), *Self-directed Learning Oriented Assessment in the Asia-Pacific*. New York: Springer.

Linacre J. M., & Tennant A. (2009). More about Critical Eigenvalue Sizes (Variances) in Standardized-Residual Principal Components Analysis (PCA). *Rasch Measurement Transactions, 23*(3), 1228.

Linacre, J. M. (2011). *A user's guide to Winsteps/Ministep Rasch-model computer program*. Chicago, IL: Winsteps.com.

Mok, M. M. C. (2010). *Self-directed Learning Oriented Assessment: Assessment that Informs Learning & Empowers the Learner*. Hong Kong: Pace Publications Ltd.

Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745-783). Mahwah, NJ: Lawrence Erlbaum.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegermann, D. Leutner, & R. Brunken (Eds.), *Instructional design for multimedia learning* (pp. 181-195). Munster, NY: Waxmann.

National Mathematics Advisory Panel (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. U.S. Department of Education: Washington, DC.

OECD (2010). PISA 2012 Mathematics Framework. Paris: OECD Publications. Retrieved 23 May 2013 from http://www.oecd.org/dataoecd/8/38/46961598.pdf

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046.

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*(1),1012.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Savage, J. (2011). *Cross-curricular teaching and learning in the secondary school*. New York: Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Wang, W. C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9*(4), 387-408.

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: validation of the short form of a prekindergarten and kindergarten mathematics measure, *Educational Psychology: An International Journal of Experimental Educational Psychology, 32*, 311-333.

Authors' email: Jingjing Yao jingjing@ied.edu.hk

Magdalena Mo Ching Mok (Corresponding Author) mmcmok@ied.edu.hk