Levels of Attainment in Number for Pupil Performance Profile in the Primary School

Foong Pui Yee
National Institute Of Education
Nanyang Technological University
Singapore

Abstract

This paper reports a part of an ongoing Pupil Profiling Project in mathematics, which aims to produce a set of six test instruments on number, measurement and geometry for Primary 3 and 5. Item response theory was used for analysis of the test results to provide profiles of pupil performance on their levels of attainment of basic skills and processes across the mathematics curriculum. Such profiles served as semi-diagnostic information for teachers to assess in greater details the weaknesses and strength of individual pupils. The development of one of the tests on Number for Primary 3 will be presented in this paper. For each test a sequence of attainment levels was identified. The lowest levels consisted of elementary knowledge and skills while the higher levels reflect increasing complexity of understanding and the use of more advanced problem solving processes. The two objectives of this study are:

i. to develop and validate tests on Number for use in Primary 3 school-based assessment,

ii. apply item response theory to analyse data and to present results in a form of "Kidmaps" that are meaningful to teachers and pupils.

Introduction

Information on test item responses is necessary if a test researcher wants to design a test of which the item difficulties match the individual's ability. Conventional educational testing depends on the sample of persons from which the item statistics were obtained. The basic unit of measurement is the test as a whole and not the test item as such. For example when a pupil takes a test, we realise that the raw score obtained by the pupil depends both on a the ability of the pupil as well as the difficulty level of the test. Thus, a pupil with a test score of 58 in one test may have the same ability level as another pupil who has a test score of 72 in another test of the same kind, if the first test is more difficult than the second test. This is to say that the meaning of the total test score is uncertain unless more information is known about the items in the test.

Seen in another way, the essential feature that is missing in most educational testing is constancy of scale. Each test that we create defines a new scale that is unique to itself, and the scores that we obtain are bound to that scale and that test only. To overcome this problem, the purpose of this research is to apply modern measurement theory, particularly the Rasch Model formulation of Item Response Theory (Kwan and Shannon, 1989; Pollit, 1990, Willmott and Fowles, 1974). The central assumption in the Rasch Model is that the set of people to be measured, and the set of items used to measure them, can each be uniquely ordered in terms respectively of their ability and difficulty.

The development of the Number tests in this research have been strongly influenced by the success of the Basic Skills Testing Program (BSTP) in New South Wales, Australia (Masters et al, 1990). The BSTP program used the item response theory techniques to map student achievement with respect to a set of defined skill levels and to provide feedback to parents and teachers in a form that will bd useful in helping students to build upon their current achievement.

In conventional educational testing, the measuring instrument is a test which comprises a set of questions or test items. This test or measuring instrument will be different if some items are added, omitted or replaced. This is to say the most educational measurement is lacking in "specific objectivity". To achieve objectivity similar to that of physical measurement, at least two conditions must be met, namely, (a) estimates of attaintment of a pupil are independent of the particular set of items which comprise the test, and (b) the characteristic of the test items such as ease or difficulty is also independent of the distribution of attainment of the pupils who are given the test.

In physical measurement these conditions are met because a common standard for scaling can be first fixed before a measuring instrument is constructed. For example, the Celsius temperature scale is defined arbitrarily first and then all measurements are expressed on that scale no matter what kind of thermometer is used. In educational testing however, the measuring scale depends on the type of test (measuring instrument) used simply because frequently we first construct a test and then derive from it a measuring scale. This process is exactly the reverse of that used in physical measurement. Hence, if one desires to achieve objectivity in educational measurement, it will be necessary to solve two problems: (i) find a common standard for scaling, and (ii) find a way to calibrate the measuring instrument

The second problem can be solved through the use of the Rasch Model. The first problem is much more intractable since it depends on a clear operational definition of what the test is supposed to measure whether it be basic skills or more complex problem solving processes. The test items if they have been properly constructed can be ranked from easy to difficult and so they define a scale of attainment corresponding to their levels of difficulty. To measure a pupil's level of attainment is equivalent to estimating an appropriate location on this scale, and this can be done by making use of his or her responses to the test items.

Since when a pupil attempts to answer a test item, a correct response depends as much on his/her ability as on the item difficulty, it follows that we can expect a pupil's chance for success to increase with ability but to decrease with item difficulty. This is central to the item response theory which allows for the estimation of person abilities on the same latent continuum scale, independent from the subset of items that have been designed to fit the model. The basic unit of measurement of the item response models is the item. This is to say that, if a pupil's ability is higher than the level of difficulty of a item, we can expect a correct response; on the other hand, if the pupil's ability is lower than the level of difficulty of the item, then we can expect an incorrect response.

Stages in the Development of the Tests

Two stages were involved in the construction of the test items and their levels: a pre-pilot and pilot trials where the instruments developed were administered to pupils to check for content and construct validity of the test items. The scope and objectives of the relevant sections of the Singapore primary mathematics determined the content validity of the tests. School teachers were consulted in selecting and constructing items which would be relevant to the contents and processes that pupils were being exposed to. This was also to establish various aspects of the curriculum to be represented; and to identify topics and its relative emphasis.

The delineation of the attainment levels was more problematic. Since the intention was to construct a cognitive hierarchy of attaintment levels for each of the test, there was a need to establish and match levels of difficulty for the items. This implied that the basis of the assignment of items to level must be in terms of success rate - a norm-referenced notion - at least in the initial stages of establishing a system of levels.

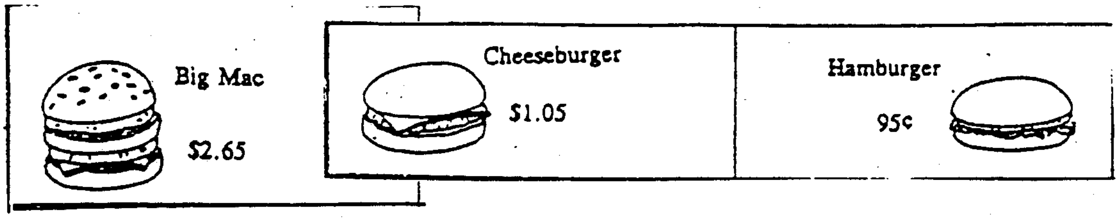
Bloom's taxonomy of levels of cognitive domain for number was used initially as guidelines in determining a prior basis for assigning tasks to the various attainment levels. However from the pre-pilot and pilot trials with the pupils, it was found that different emphases of teaching and curriculum could alter the difficulties of some of the test items selected in the beginning. For example, some items which test knowledge and conceptual understanding designed for the lowest level, were found to be more difficult for pupils than those items on routine skills like one-step or two-step word problems. As the emphasis in the classroom was on application of routine skills, the pupils' performance of routine skills was found to be more advance than their understanding of underlying concepts. Nevertheless, with curricular change in the new syllabus to teaching for meaning on the development of concepts, this would eventually change the balance. Hence the final attaintment levels for the tests items were established through empirical findings in the pilot trials. At the start, an item bank of 30 multiple-choice items assigned to their respective prior levels, were collected. Before the instruments were put together for the piloting, a pre-trial run was conducted on one class of about 40 pupils for each of the instruments in another school. A classical item analysis was done to improve or eliminate items in which suitability of language, structure and distractors in the multiple-choice format were problematic to responders. Items which were poor discriminators or unexpected in level of difficulties were reviewed and rewritten.

A workshop was then arranged for teachers who were teaching primary three in the two pilot schools to evaluate and improve on the items. The teachers had to rate the items on whether they were appropriate for P3; and also the degree of difficulty of the questions. At an assigned date, the tests were given to 151 P3 pupils of the two pilot schools. Pupils were expected to complete the test within 45 minutes. Generally majority of the pupils were able to complete a test within 30 minutes.

Item Calibration and Attainment Levels

The topic on Number at P3 deals mainly with concepts of number and the ways we write them; place value and number computation with operations of addition, substraction, multiplication and division. Pupils are expected to deal comfortably with number situations they are likely to meet in daily lives. For example, some problems were set within contexts of reading a menu to test computational skills involving making purchases which are real life situations. Pupils should have a feeling for correctness or reasonableness of an answer based on an understanding of when a particular arithmetic operations should be used. At a higher level they should have developed estimation skills and use strategies to solve non-routine and unfamiliar problems.

The test data collected from the pilot schools were analysed using a Rasch analysis program called Titan (Adams & Khoo, 1991). Figure 1 shows the calibration of attainment levels after Titan analysis, and the properties of the items in each level. The attainment levels are calibrated along a *logit* scale of -1.0 to 3.0. For instance, up to scale of -1.0, five items were calibrated at the lowest attainment level 1. These were items that 95.4% of the pupils were likely to complete all correctly. They were representative of questions that test basic number sense and use of the four operations in simple one-step problems, in this case in context of a situation which is real to children at this age in P3, i.e. "paying for a meal at 'McWonut'". The understanding of the questions is well facilitated in a format complete with pictures, where the pupils can easily related to without the difficulty of language structure or semantics. What is required are computation skills in numbers involving money using word clues such as 'difference" and "how much less". These fundamental skills were emphasised at the lower primary levels. Examples of the items:



*. Farah paid for a Big Mac with a $5.00 note. How much change should she receive?

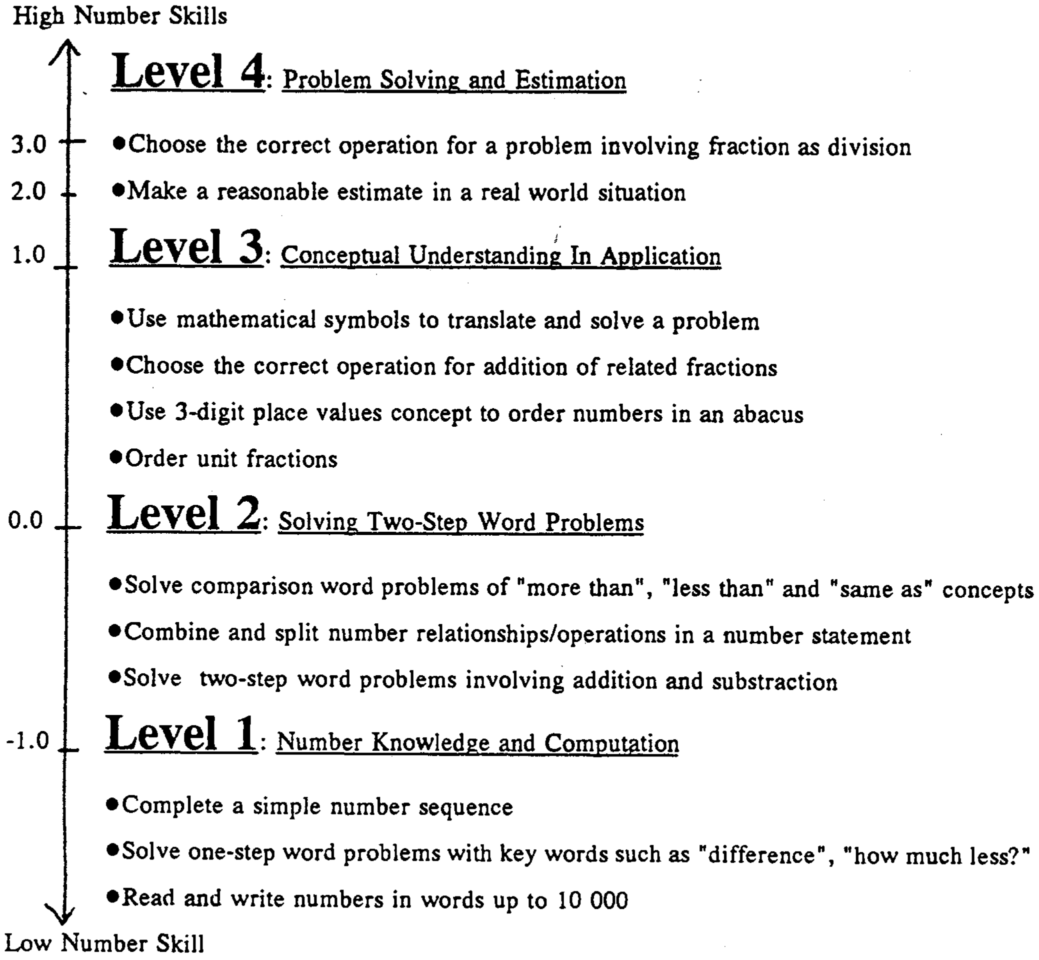
A. $3.65 B. $2.35 C. $7.65 D. $1.35

*. What is the difference in price between a hamburger and a cheeseburger?

A. $2.00 B. 90¢ C. 10¢ D. $1.90

Figure 1

Skill Levels in Number at Primary Three



Level 4: Problem Solving and Estimation

- Choose the correct operation for a problem involving fraction as division
- Make a reasonable estimate in a real world situation

Level 3: Conceptual Understanding In Application

- Use mathematical symbols to translate and solve a problem
- Choose the correct operation for addition of related fractions
- Use 3-digit place values concept to order numbers in an abacus
- Order unit fractions

Level 2: Solving Two-Step Word Problems

- Solve comparison word problems of "more than", "less than" and "same as" concepts
- Combine and split number relationships/operations in a number statement
- Solve two-step word problems involving addition and substraction

Level 1: Number Knowledge and Computation

- Complete a simple number sequence
- Solve one-step word problems with key words such as "difference", "how much less?"
- Read and write numbers in words up to 10 000

At the highest Level 4, P3 pupils generally found questions that require them to estimate an answer difficult. About 70% of the pupils had a 50% chance of not getting the correct answer. Example of such items in Level 4 are as follows:

*. How many Big Mac can you buy with $10?

A. 3 B. 4 C. 5 D. 6

*. The farmer collected 43 eggs and wanted to pack them into the egg trays. One egg tray can contain only 10 eggs. How many egg trays are needed for packing all the eggs ?

A. 3 B. 4 C. 5 D. cannot be done

As the pupils were so used to computing numbers and getting an exact answer, it is no wonder that 48% of them chose the last option "cannot be done" for the egg problem. Whereas for the Big Mac problem, they tried to compute the division but incorrectly. Only one-third of the pupils were correct and an equal number choose '4' as the answer.

As all 18 items were calibrated and all 151 pupils are measured on the same scale, it was possible to interpret pupil's scores in terms of the kind of skills that typified those achievement levels. The Rasch model techniques was able to provide individual pupil profiles known as 'kidmaps' which are graphical representations of individuals' ability estimates and their response patterns. Figure 2 shows a kidmap of a pupil whose ability was estimated on the *logit* scale as -0.29 with a percentage score of 44.44%. The kidmap was constructed such that items were plotted in order of difficulty on the left hand side of the profile if the pupil has answered them correctly and on the right hand side for the incorrect items. To interpret these meaningfully, the pupil's achievement level is marked 'XXX' (in Figure 2), as a skill band on level 2 which gives a "best" estimate of his or her level of achievement. At a glance it show that he/she has high probability of getting items on number knowledge, computation and solving simple word problems correct, but he would have difficulty in understanding underlying mathematics concepts and problem solving which are at higher levels of 3 and 4.

Conclusion

This study is in the preliminary process of exploring a logistic test model such as the Rasch Model for a basis of criterion-referenced testing based on item banking. There are lots of room for increasing, refining and validating the pool of test items that will fit the theoretical model for this study over a much larger sample. In using items response theory for analysis, the challenge is to develop better test instruments for more meaningful profiling of pupils learning of mathematics which covers a diverse range of concepts, skills and processes. The product of the Pupil Profiling Project at NIE, of which this study is a part, is the development of a scale which the items of the tests in Number, Measurement and Geometry can be represented in terms of pupils' ability or achievement at the various stages of their primary education. The *kidmaps* that are produced will also be helpful for teachers prior to instruction, in order to enhance their understanding of where their pupils' strength and weaknesses lie in the various aspects of the mathematics curriculum.

Figure 2

```
Titan Analysis for NUMBER P3 Run1
------------------------------ K I D M A P------------------------------
Candidate: 13105                                       ability:  -.29
group:    All                                        fit:       .80
scale:    All                                        % score:  44.44

-----------Harder Correct ----------------------Harder Incorrect -----------
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |   18(2)
                               |   |
                               |   |   17(2)
                               |   |
                               |   |
                               |   |   16(3)
                               |   |   15(1)
                               |   |
                               |   |   14(1)
                               |   |   13(1)
                               |   |   11(3)  12(3)
.........................................
                       10      |   |
                               |   |    9(1)
                 6     8       |   |
                               |XXX|
                       5       |   |
                               |   |
                               .........................................
                               |   |
                               |   |
                       4       |   |
                               |   |
                 3     2       |   |    7(1)
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                       1       |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
                               |   |
-----------Easier Correct ----------------------Easier Incorrect -----------
```

Reference

Adams, R.J. & Khoo, S.T. (1991) Quest: the interactive test analysis system, Hawthorn, Victoria: ACER.

Kwan,P.Y.K., & Shannon A.G. Objective tests and latent trait theories. International Journal of Mathematical Education in Science and Technology Vol.20, 3 , 457-467, 1989.

Masters,G., Lokan,J., Doig, B., Khoo, S.T., Lindsey,J., Robinson,L.,& Zammit,S. (1990) Profiles of Learning: The basic skills testing program in New South Wales,1989, Hawthorn, Victoria:ACER

Pollit, A. (1990) Diagnostic Assessment Through Item Banking. In Entwistle, N. (Ed) Handbook of Educational Ideas and Practices. Routledge.

Willmott, A.S. & Fowles, D.E. (1974) The Objective Interpretation of Test Performance - the Rasch Model Applied. NFER.