Using quantitative and qualitative data to construct a developmental learning framework in the context of elementary high school algebra.

Judith Falle

Abstract

The study for a PhD thesis aims to identify linguistic characteristics of students' utterances that indicate their levels of understanding of elementary algebra. To do so, both quantitative and qualitative data was collected. Secondary school students were asked to complete a forty-item algebra test, and then a selection of students was interviewed.

The responses of the students (n = 222) on the algebra test were analysed using the Rasch model. The results display groupings of types of algebra items that indicate developmental levels. These levels may be described in terms of a qualitative framework such as the SOLO taxonomy, as well as in terms of the mathematical characteristics of the items.

Interview data obtained from students (n = 32) is also to be used to explore their levels of understanding, and to support findings from the analysis of the test items.

In this paper, issues arising from the use of the Rasch models, and in devising and obtaining valid and reliable interview data are posed and discussed. Some issues are the strengths and limitations of the Rasch model, the structure and conduct of interviews and ways in which interview data may be analysed.

Research into human behaviour faces a perennial problem of providing convincingly objective, valid and reliable data. Unlike the physical sciences there is not usually recourse to repeated, and hence replicable experiments. Quantitative data, although objective, and able to be validated and shown to be reliable through repeated applications of the data collection instrument, may provide insight into a limited set of behaviours – those which can be enumerated. Reliance on qualitative data alone leads to the problems of establishing reliability and validity. Consequently both quantitative and qualitative data-collection instruments need to be used in such research in such a way that one complements the other. Quantitative data informs the researcher about 'how much' mainly through counting techniques that are assumed to be free of the influence of the researcher. Qualitative data provides information about 'how well', but is often suspected of being unduly influenced by the will of the researcher. Leaving aside the philosophical and psychological questions

these last sentences raise, the use of an analytical tool that takes account of 'how much' and 'how well' provides researchers with a framework of particular value in the field of education.

One such analysis tool is the Rasch Model, developed by Georg Rasch and first published in 1980. The model provides a framework of developmental pathways based on objective measures. Like all models, it makes certain assumptions and simplifications that limit the extent of inferences that might be drawn from it. The use of the Rasch model and some of the limitations encountered will be discussed in this paper in the context of research into connections between language use and cognitive development in mathematics. Firstly a brief outline of the research is provided, then a general overview of the Rasch model, as used in the study. Data from the study will then be used to illustrate the power of the model to use quantitative data to provide insight into 'how well' students understand the mathematics and to illustrate some of the consequent questions.

Outline of the research: Context for application of the Rasch model.

When listening to students explain their ideas, teachers often make qualitative, but intuitive, judgements about the understanding of those students of the subject under discussion. What are the bases for these judgements? It is often not strictly the content of the statements, but more subtle characteristics that serve to inform such judgements. It is these characteristics that my study sets out to identify.

The study was based on research by Bills and Gray (2001), and Bills (2002) in Britain. They analysed the verbal responses of children from age six years to ten years old as they described their thinking as they carried out mental computations. From this analysis Bills and Gray identified several features of the children's utterances that marked their success, or lack of success. Another study by Boero, Douek and Ferrari (2002), involving analysis of written explanations of elementary algebraic understanding by university students, found similar characteristics that were indicative of successful (and unsuccessful) completion of first year mathematics.

The research focus was on constructing a model of the cognitive development of students' understanding of elementary algebra (that which is taught in the junior years of secondary school – Grades 7 to 10) that linked linguistic features of students' explanations of their thinking as they carried out algebraic processes to their success. Hence, some objective basis for the intuitive judgements made by teachers might be established. Further research might also find that this model could be more widely applicable. Students from Years 7 to 10 from three secondary schools – two independent, single-sex schools and one Catholic, systemic, co-educational school – participated in the study.

The first task was to establish the mathematical success of the participants. The 'objective' measures of school-based tests or examinations were not comparable across the participating schools, and the SNAP

tests (state wide tests of basic skills conducted in Years 7 and 8) had too broad a focus, and did not treat the formal algebra that my study was to use. The participants, (all students from Years 8 and 9 in the participating schools, n = 222) were asked to complete a 'survey' (test) consisting of forty items that required them to use elementary algebraic processes to manipulate expressions and solve linear equations. The items were based on examples given in the syllabus documents (Board of Secondary Education 1988, Board of Studies 2002, Board of Studies NSW 1996a, 1996b, 1996c, 1999, 2002), and text books used by the participating schools. This survey was used to assess the 'ability' of the students.

Traditionally, assessment of mathematics is seen in strictly quantitative terms. Students are marked as 'correct' or 'incorrect' on test items. The sum of the number of correct items is used as the indicator of their mathematical understanding. Regardless of the test format, or the nature of the test items, this simple aggregation model predominates. However, not all items in a test are of equal difficulty, and hence not all can be answered correctly by all pupils – regardless of the skill with which the test might be constructed. The problem, then, was to find a way of evaluating the items in the survey that reflected the level of difficulty and consequently the ability of the participating students.

Hence, the data collected for the study needed to provide information on the status of the algebraic knowledge of the students involved in the study as well as insight into their thinking and their use of language. Assessment instruments such as the survey reveal little of students' thinking about the mathematics with which they are to engage, so, from the test results, students were selected for interview (n = 32) so that there was a representative range of abilities. The Rasch model was chosen to analyse the data from the survey because the resultant order of item difficulty and student ability estimates are assumed to reflect a developmental pathway. That is, the more difficult items a student is able to respond to correctly, the better that student's understanding of the [mathematical] concepts under examination.

Overview of the Rasch Model.

The probability of an individual's success on an item can be articulated as a function of the difference in ability of an individual and the degree of difficulty of an item. The Rasch model enables relationships between independent variables on an equal interval scale to be illustrated in such a way that the distances between what are initially ordinal values to be made equal and meaningful (Bond & Fox 2001).

The model is: sensitive to the developmental order of the skills or abilities under investigation; estimates the developmental differences between the ordered skills or persons in the population being studied; and allows the verification of general developmental patterns. The Rasch model is designed to analyse the one construct (unidimensional) using dichotomous data. In this study, the construct was that of elementary algebra, and the data was the 'correct' or 'incorrect' responses to the questions.

The model aligns participant ability with item difficulty on the same linear scale. This means that a participant with a particular ability rating has a 50% chance of correctly responding to items that have a difficulty estimate at the same numerical value. Such a participant has less chance of responding correctly to items with a difficulty estimate greater than the ability rating and a greater chance of answering correctly those items with a difficulty estimate below that of the ability rating.

The participant ability rating and the item difficulty estimates are calculated using the natural logarithm of the odds of a successful response to the item. This number is termed a *logit*. The calculation takes the ordinal value of the percentage number of correct responses and maps this onto a log-linear scale where equal intervals have equal values. In this way differences in abilities of participants and differences in item difficulties are represented by interval differences.

As with all statistical inferences, the more extensive the available data the less the possible error. Thus, the precision of the estimates of individual ability and item difficulty depends on the data available at a particular point. If insufficient items are found to be at the ability level of a participant, then that ability estimate is subject to greater error than one where there are a number of items of a difficulty equal to that ability level. This means that the extremes of the ability levels or item difficulty ranges may be subject to a greater error if the test does not include items that will provide the information. In part, this means that any test should have no items that allow all participants to give a correct response, and all items should allow at least one correct response.

The Rasch model is based on the premise that ability levels and item difficulty can be represented as a linear progression of development of a single construct. Factors other than that of the construct being modelled will result in *misfits* of items or ability. Items fit the model if they lie within the acceptable ±2 standard deviations, at the 95% confidence level. There are two types of 'fit' described in terms of chi-square ratios. Infit statistics are calculated as weighted mean square values of the residuals (differences between the actual result and the result expected), and standardised to *t* values. There is an expected value of +1. 'Infit mean square' values greater than +1 indicate that there is more variation than expected between the model and the data. 'Outfit mean square' values less than one, and greater than zero, indicate that there is less variation than predicted by the model. The one of most concern in this case is that of the 'infit', where items with a high difficulty estimate are answered by participants with a lower ability estimate. In these cases, factors other than those involved in the construct being modelled may be operating, and these items do not therefore reliably test the construct.

Construct reliability is dependent on the reliability of the estimates of item difficulty and the ability levels. The mean for these estimates is set, at a default of 0.00. Reliability is measured on a 0 to 1 scale. Thus the

nearer the reliability is to 1 the more confidence that can be placed in the replicability of the item placement or ability estimates if the test were given to other suitable samples.

Figures 1 and 2 provide an illustration of how the Rasch model can be related to more conventional ways of apportioning ability estimates.

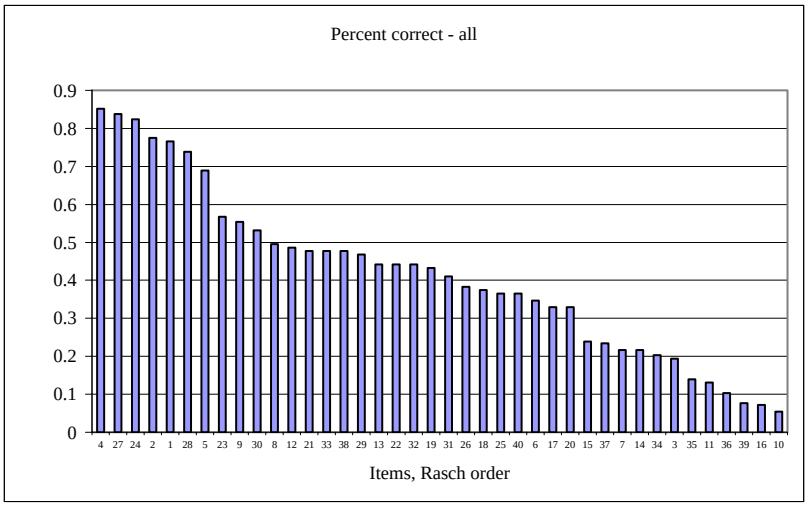


Figure 1: *Percentage of participants correctly answering each item.*

Figure 1 illustrates how the number of students correctly responding to each item, correlates to the level of difficulty, calculated as the odds of the item being correctly answered i.e. the more difficult the item, the fewer participants who will correctly respond. The converse, of course, is also true.

If the items are arranged in order of difficulty and the students in order of ability estimates, the model can be illustrated by the matrix in Figure 2. The dark patches represent correct responses to each item by individuals. Ideally a diagonal line from the bottom left to the top right of the matrix would be the boundary of correct responses. All correct responses would lie to the left of that line. But we are dealing with human behaviour. These variations are discussed in a later section.

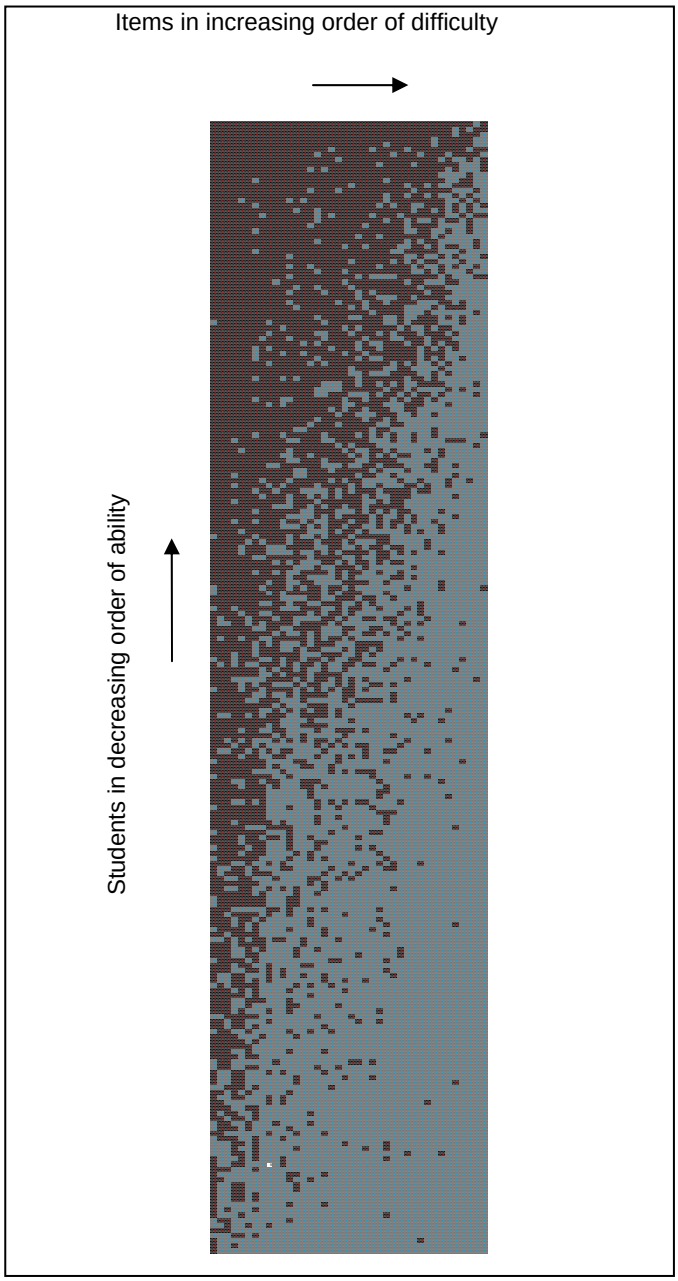


Figure 2: *Matrix of student ability against item difficulty*

That is the basis, in general terms, of the model. It is widely used to establish the reliability of state-wide assessment instruments, and international assessments such as PISA, and can be powerful in situations such as my research. However, some questions arose as the model was used to analyse the data, in particular the need to aggregate the 'non-attempts' (blanks on the survey) with the 'incorrect' attempts, and the need for criteria by which partial credit could be attributed for some responses.

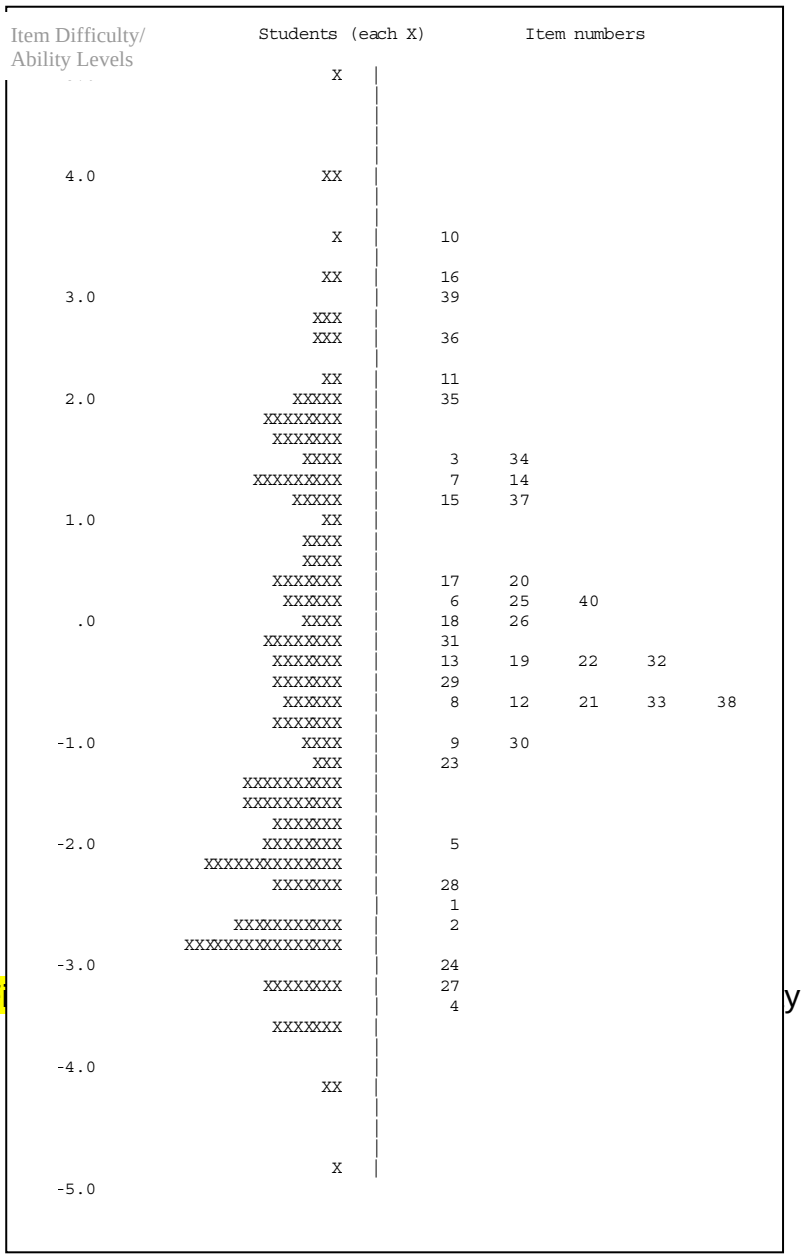
To illustrate the model as well as these 'problems', a discussion of aspects of the analysis of the algebra survey conducted by the author follows.

The Rasch Model in Practice.

The researcher collected and marked the survey items as ‘correct’, ‘incorrect’ or ‘no attempt’, in order to recognise an attempt at a question, even though the response was, in the end, incorrect. This preliminary marking allowed the identification of students who were confident to make an attempt at a question, even if incorrect. From a pilot study, it was found that students who presented a paper with few items attempted were unlikely to make informative interview subjects, and so, identification of these students was important for the second phase of data collection – the interviews.

The data coded in this way [‘correct’ (2), ‘incorrect’ (1) and ‘no attempt’ (0)] and analysed using Quest software, resulted in the map of item difficulty and ability estimates similar to that in Figure 3, but the clusters of items were not so focussed. Although the numerical values attributed to item difficulty estimate changed, the order did not, nor did the interval differences between the correctly answered items. The distribution of student ability estimates did change, because credit was being given to incorrect responses. In order to clarify the model, item responses were then coded only as ‘correct’ (1) or ‘incorrect’ (0) for the purposes of Rasch modelling. Herein lay the first ‘problem’.

Correct responses were only those which were considered to be mathematically complete, (e.g. fractions expressed in lowest terms, all like terms collected, equations completely solved). Some items in particular attracted a variety of partially correct responses, but as students were not required to show working, and many did not, coding these on a spectrum of ‘correctness’ required some ‘second guessing’ about the intention of the students. As this could not be validated, because of the anonymity of the scripts, this coding strategy was discarded, but this has lead to another, important ‘problem’. (These ‘problems’ will be discussed later in this paper.) The resultant map that detailed item difficulty and student ability is illustrated in Figure 3. The item numbers appear to the right of the vertical line, and each ‘X’ represents one student.



Fi y levels.

Figure 3: *Rasch model of item difficulty and student ability levels*

The items appear in clusters, indicating that they are of similar difficulty. The extent of the distance between items, and between clusters is indicative of the extent to which the difficulty changes. The task for the researcher (or teacher, using this model) is to identify the possible reasons for these differences. The information presented by the model is, in this particular research, supplemented by an analysis of the errors made in the responses to the items, and data from interviews.

The use of a particular model depends on how well it reflects the actual situation being investigated. For the Rasch Model to be reliable, there needs to be sufficient items that take account of the total range of abilities

(Bond & Fox 2001). Thus the ability estimates for three students at the top of the scale (4.0 to 5.0 logits) and at the bottom of the scale
(-3.7 to -4.8 logits) are subject to considerable error, because there is too little data. Deviations from the expected need to be investigated as to whether they are idiosyncratic, individual responses to a particular item, or whether they represent more common misconceptions. The data summarised in Figure 3 may also be presented in a similar way for each participant, so one may examine whether a particular student's responses fit the overall pattern, as well as ways in which they do not.

To return to Figure 2, theoretically, a diagonal line from the bottom left (student of least ability) to top right (item of greatest difficulty) could be drawn. On the right of the line would be no responses. In other words, the most able students would correctly respond to most items, only failing to do progressively so as the items increased in difficulty. The least able students would correctly answer only the few items of least difficulty. This is the overall pattern. The complexities and idiosyncrasies of human thinking are illustrated by the scattering of correct responses and incorrect responses where theory does not expect them to be. The statistical measures of significance however indicate that this scatter is within reasonable limits of variation, with one or two exceptions.
When items are mapped to illustrate where they fit within the statistically significant ± 1 Standard Deviation from the mean, the items that do not 'fit' become obvious. In Figure 4, the two items of concern are Items 5 [Simplify 2ab + 3b + ab] and 12 [Simplify 2/a x 3/b] which lie outside the +1SD(=1.3) boundary. Both these items tended to be answered successfully by students whose ability scores were below the difficulty estimates for the items. Interviews and error analyses demonstrated that these students were using 'rules' that were incorrect, but which, felicitously for them, worked in these cases. The more able students either answered correctly, or became confused as to what to do and left the item blank, or offered a variety of idiosyncratic responses.

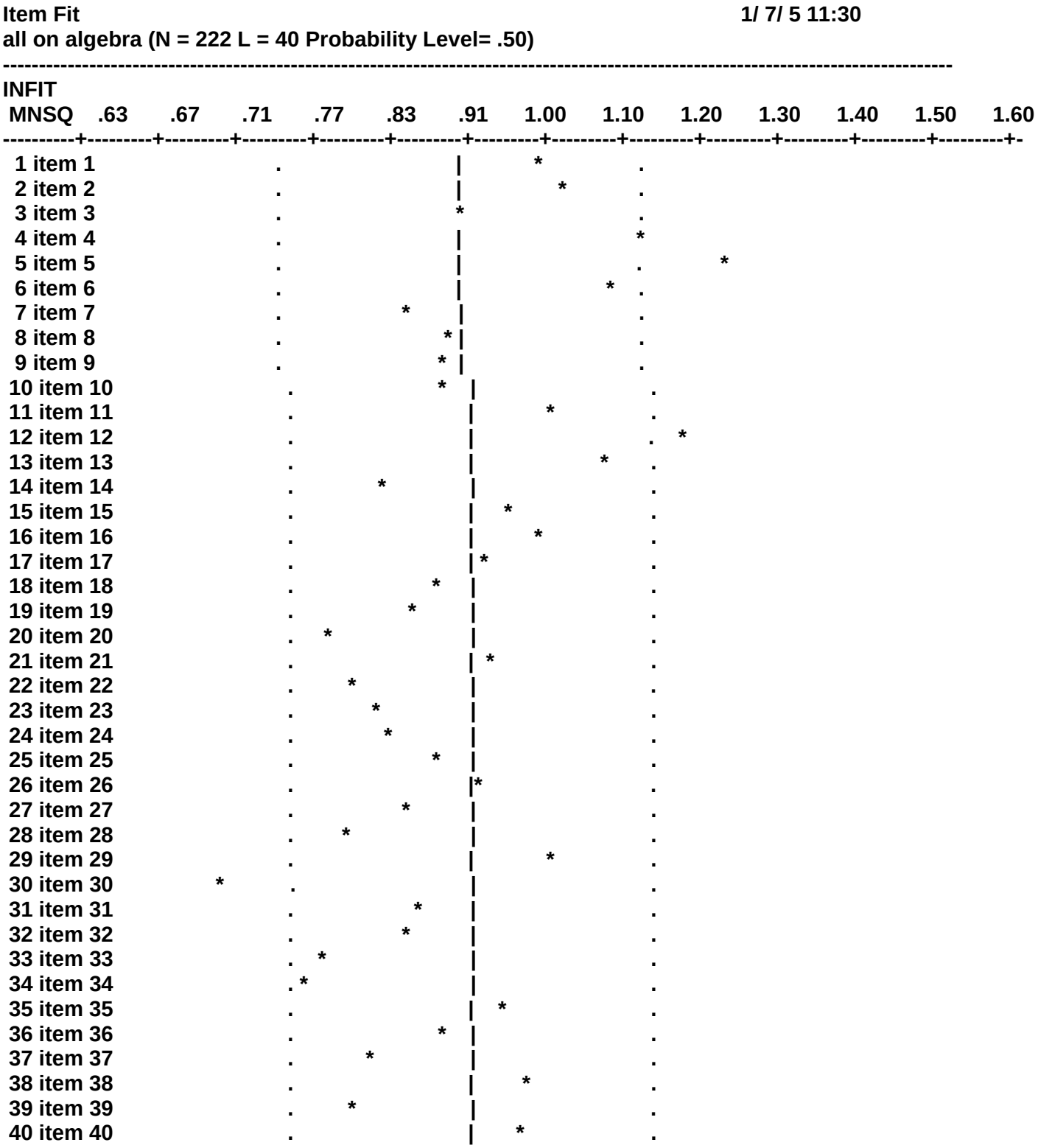


Figure 4: *Item fit map*

Items such as these do not 'fit' the model – other aspects other than algebraic thinking may be operating, and so these items may be excluded from further tests of this nature, or investigated in order to discover the factors operating. This is one of the assumptions of the model – that incorrect thinking, or thinking that relies on knowledge and understandings outside those tested will result in such misfits.

However, Stacey and Steinle (2006) found that correct answers do not necessarily imply correct thinking, and that this is not always reflected in the Rasch model. In a study investigating students' understandings of decimals, they found that even within the expected range of variation, students were responding correctly to items, but for the wrong reasons.

Does this matter? Correct thinking about mathematical procedures enables one to generalise to other situations. Thinking that uses incorrect understandings such as 'large decimal has more numbers', or 'you operate arithmetically on the top then the bottom of fractions' leads one to obtain correct answers sometimes, and at others, not. Like Pavlov's dog, intermittent reinforcement leads to long-term memory retention – of procedures that are mathematically useless and which contribute to the perception that mathematics is a mysterious and serendipitous affair.

Two other aspects of the Rasch model that provided food for thought were the influence of non-attempts on the estimate of the degree of difficulty of items and ways in which partially correct attempts could be coded and the influence of this. The latter problem is discussed first. The Rasch model allows for 'partial credit' to be given in clearly defined situations. An example provided by Bond and Fox (2001) is that of coding Likert Scale responses [e.g. Agree Strongly/Agree/ Disagree/ Disagree Strongly]. However, instances such as the response to an item requiring solution of an equation, or simplification of an expression are more difficult to compartmentalise in such a way. There is usually no unique solution strategy or pathway, nor general criteria for establishing degrees of success along that pathway. Such criteria might be established through error analyses, examination of student work or interviews and then used to construct test situations that elicit the data. There is much work to be done here, and is beyond this present study. It is mentioned here because two examples of the first problem of how to deal with non-attempts serve to illustrate this point also. They will be discussed below.

The question of the influence of non-attempts on an item is also worth considering. One thought is that the greater the number of non-attempts, the more difficult the item is likely to be. If this is so then perhaps the non-attempts in the instance of the survey, are the result of students not having been taught about equations. Alternatively, given that some of these items appear late in the survey, some students may have simply given up, or run out of time.

The number of non-attempts at various items increases as the item difficulty increases, but not in all cases (see Figure 5). The two most difficult items (Q10 [$(x+y)^2$], and Q16 [*Simplify* $2a^2 \quad 5a/4$]) were attempted by the majority of participants. Only 24 students failed to provide a response to Item 10. Of the remaining 198 responses, 186 were incorrect, leaving 12 correct responses. Of the 186 incorrect responses, 86 offered the same answer [$x^2 + y^2$]. Item 16, was not answered by 77 students, and only 16 students answered correctly. The incorrect responses showed a range of understandings and partly correct manipulations. It was this variety of partly correct responses that suggests the mapping of solution strategies might reveal stages of cognitive growth.

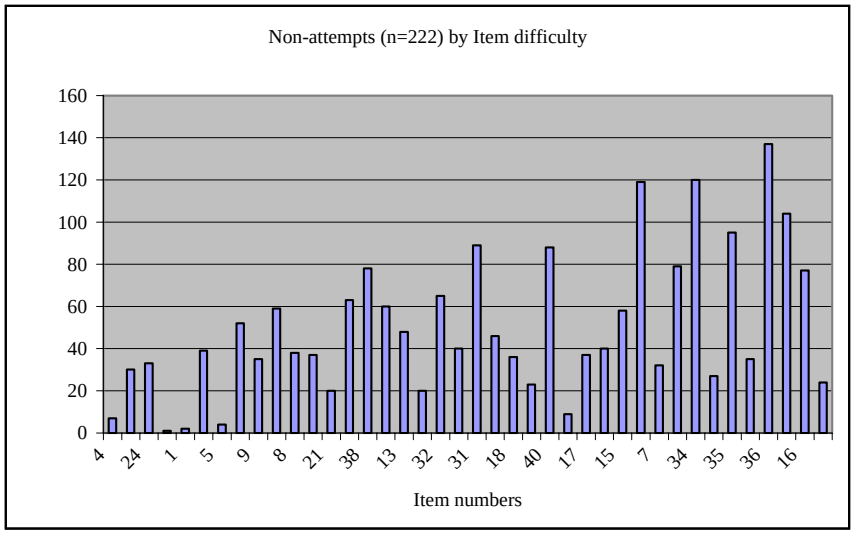


Figure 5: *Number of students not attempting survey items*

The examples of these two items illustrate some of the questions that can be raised if the data provided is interrogated closely. Although a majority of students responded to these items, albeit incorrectly, error analysis revealed two different stories. Responses to Item 10 reflected an extensively studied and well-documented misconception (e.g. Matz 1982), where students 'distribute' the square sign across the two variables in the brackets, without attending to the meaning implied by the symbols. Responses to Item 16 provided the author with the problem of how to attribute credit for partly correct responses, or even whether this should be done. In the end, as mentioned earlier, this strategy was discarded. Understanding of the reasons for the responses to Item 16 was gained through interview data.

Conclusion

The Rasch Model is a sophisticated statistical model that uses scores on test items to develop measures of item difficulty and participant ability (probability of success on a particular item) on the same scale. Like all mathematical models used to represent human behaviour, certain assumptions and simplifications are necessary. In evaluating the use of such models, these have to be accounted for, and such aspects investigated by other means. In other words, powerful as this model is, the clarity of its representation of performance of both items and respondents, often leads to questions that cannot be answered by the model alone. Although the model does presume to represent a progression of cognitive development, some patterns of learning may remain hidden. Stacey and Steinle (2006) caution that the model may measure development of mastery of skills but not provide a picture of conceptual learning, particularly if items, or groups of items, are regarded in isolation from the overall pattern of a student's performance. This points to the necessity of using other sources of related data, such as interviews, or the close analysis of individual performances also provided by the Rasch Model in order to construct a framework of cognitive development – or any other model devised to explain human behaviour.

REFERENCES

Bills, C. 2002, 'Linguistic pointers in young children's descriptions of mental calculation', Paper presented at the *26th Annual Conference of the International Group for the Psychology of Mathematics Education*, Norwich.

Bills, C., & Gray, E. 2001, 'The 'particular', 'generic' and 'general' in young children's mental calulations', Paper presented at the *25th Annual Conference of the International Group for the Psychology of Mathematics Education*, Utrecht.

Board of Secondary Education. 1988, *Mathematics Syllabus: Years 7 and 8*, 2nd edn, Board of Secondary Education, Sydney.

Board of Studies. 2002, *Mathematics Years 7 - 10 Draft Syllabus*, Board of Studies NSW, Sydney.

Board of Studies NSW. 1996a, *Mathematics Years 9 - 10 Syllabus: Advanced Course*; Stage 5, Board of Studies NSW, Sydney.

Board of Studies NSW. 1996b, *Mathematics Years 9 - 10 Syllabus: General Course*; Stage 5, Board of Studies NSW, Sydney.

Board of Studies NSW. 1996c, *Mathematics Years 9 - 10: Intermediate Course; Stage 5*, Board of Studies NSW, Sydney.

Board of Studies NSW. 1999, Mathematics Years 7 - 8 Syllabus Outcomes (incorporating linked outcomes, Stages 3 to 5), Board of Studies NSW, Sydney.

Board of Studies NSW. 2002, *Mathematics 7 - 10 Syllabus*, Board of Studies NSW, Sydney.

Boero, P., Douek, N., & Ferrari, P. 2002, Developing mastery of natural language: Approaches to theoretical aspects of mathematics, in *Handbook of International Research in Mathematics Education,* ed L. English, Lawrence Erlbaum, London, pp.241-267.

Bond, T., & Fox, C. 2001, Applying the Rasch Model: Fundamental measurement in the human sciences, Lawrence Erlbaum Associates, Mawah.

Matz, M. 1982, Towards a process model for high school algebra errors, in *Intelligent Tutoring Systems*, eds D. Sleeman & J. Brown, Academic Press, London, pp.25-50.

Stacey, K. & Steinle, V. 2006, 'A case of the inappl;icability of the Rasch Model to mapping conceptual learning', *Mathematics Education Research Journal*.