

Modelling Mathematics Problem Solving Item Responses Using a Multidimensional IRT Model

Margaret Wu and Raymond Adams

University of Melbourne

This research examined students' responses to mathematics problem-solving tasks and applied a general multidimensional IRT model at the response category level. In doing so, cognitive processes were identified and modelled through item response modelling to extract more information than would be provided using conventional practices in scoring items. More specifically, the study consisted of two parts. The first part involved the development of a mathematics problem-solving framework that was theoretically grounded, drawing upon research in mathematics education and cognitive psychology. The framework was then used as the basis for item development. The second part of the research involved the analysis of the item response data. It was demonstrated that multidimensional IRT models were powerful tools for extracting information from a limited number of item responses. A problem-solving profile for each student could be constructed from the results of IRT scaling.

Researchers in the field of educational assessment are continually developing new approaches to improve the efficiency of assessments. They are often concerned with methodologies that can extract the most useful and accurate information from students' responses to test items. The advances in this area can be identified in two directions at least. On the one hand, in the past decade, psychometricians have called for closer links between psychometric models and cognitive processes (e.g., Embretson, 1997; Frederiksen, Mislevy, & Bejar, 1993; Masters & Doig, 1992). On the other hand, improved mathematical modelling and estimation methods in Item Response Theory (IRT) aim at extracting more information from existing data, particularly with multidimensional modelling and distractor analyses (e.g., Adams, Wilson, & Wang, 1997; Embretson, 1991; Wang, 1998). Many of these methods have been made possible through the increased power of personal computers.

This paper examines students' responses to mathematics problem-solving tasks and applies a general multidimensional IRT model at the response category level. In doing so, cognitive processes can be linked with an IRT model to extract more information than would be possible using conventional practices in scoring items. With a limited number of test items, student problem-solving profiles can be constructed that are informative for both students and teachers. This approach has not been undertaken previously in the area of problem solving, although unidimensional IRT models have been used to build profiles of problem solving dating back to the 1970s (Collis & Romberg, 1992; Cornish & Wines, 1977; Malone, Douglas, Kissane, & Mortlock, 1980; Stacey, Grove, Bourke, & Doig, 1993; Willmott & Fowles, 1974).

Mathematical Problem-solving Tasks and Cognitive Processes

It can be argued that, in order to teach mathematical problem-solving skills effectively, one needs to link the demands of problem-solving tasks to the cognitive processes involved. From within the field of teaching and learning, Wu (2004) divided the research into two approaches to identifying problem-solving cognitive processes: the factor-analytic approach and the information processing approach.

Factor-Analytic Approach

This approach is generally empirical in that different factors that distinguish and characterise different kinds of abilities are identified through the use of exploratory factor analysis. Based on the factor loadings for various mathematical tasks, one can draw possible conclusions about the nature of mathematical thinking associated with the factors. Carroll (1993, 1996) analysed numerous published datasets and identified a three-strata theory whereby cognitive abilities can be hierarchically classified in terms of general, broad and narrow factors. At the top level, the general factor is General Intelligence. At the second level there are broad factors classified as Fluid Intelligence, Crystallised Intelligence, General Memory and other factors such as Visual and Auditory Perception. Mathematical abilities are mostly associated with Fluid and Crystallised Intelligences, as Fluid Intelligence covers third level factors such as various reasoning abilities, while Crystallised Intelligence covers third level factors such as Language, Reading Comprehension, and Communication Abilities. General Memory, a broad second level factor, also plays an important role in mathematics, as the capacity of one's working memory has an impact on the ability to solve complex, multi-staged tasks.

Exploratory factor analysis has been criticised by some researchers in that the interpretation of factors can be difficult and somewhat open to debate. For example, exploratory factor analysis may identify factors that are irrelevant to cognitive processes (e.g., Carroll, 1945; Hambleton & Rovinelli, 1986; Heim, 1975; McDonald & Ahlwat, 1974; Nandakumar, 1994). Some researchers prefer to use confirmatory factor analysis (Jöreskog & Sörbom, 1979) to test hypotheses about factors associated with cognitive processes, as they believe that confirmatory factor analysis has a more rigorous basis than exploratory factor analysis, in that the evaluation of statistical significance can be carried out in conjunction with theoretical underpinnings in the relevant field.

Information Processing Approach

The information processing approach to identifying problem-solving processes focuses on the sequential steps of cognitive demands required in solving a mathematical problem. Many classification schemes for problem-solving processes are derived from Polya's conception of mathematics

problem solving as a four-phase heuristic process:¹ understand the problem, devise a plan, carry out the plan, look back and check (Polya, 1973). While these broad processes cover most aspects of problem solving, they do not provide sufficient detail and guidance for teachers to give specific instructions to improve students' problem-solving skills.

Schoenfeld (1983) developed a model based on findings from research by information-processing theorists. His model incorporated Polya's structure and described mathematical problem solving in five episodes: reading, analysis, exploration, planning/implementation and verification. Mayer and Hegarty (1996) examined mathematics problem solving in terms of four components: translating, integrating, planning and executing. They hypothesised how expert problem solvers use different strategies from novice problem solvers in these four components.

In general, information processing theorists are not so concerned about the existence of separate ability factors corresponding to the stages of problem-solving processes. Rather, they identify the stages of problem solving so that these can be used to teach students how to approach a problem and what to do when they encounter difficulties. In particular, these problem-solving stages can serve as useful prompts for students to monitor and evaluate their own thought processes (Silver, 1982). Without clearly identified problem-solving stages, the problem-solving activities carried out in the classrooms can be somewhat ad hoc and disorganised. Approaching problem solving in a systematic way using steps defined through the information processing approach can help students acquire skills that are transferable to a wider range of problems.

In summary, the factor-analytic approach attempts to identify distinct abilities as required in problem-solving tasks, but these abilities are not necessarily required in sequential order in the steps of problem solving. In contrast, an information processing approach identifies the cognitive processes required in sequential steps in the problem-solving process, and these cognitive processes may not be distinct in the different steps of the problem-solving process. Both approaches attempt to isolate various components of cognitive processes involved in problem solving in systematic ways so that the components can be examined for each person individually to address strengths and weaknesses.

Problem-solving Framework for this Research

In developing a problem-solving framework for this research, the principles that underlie both the factor-analytic and information processing approaches were combined. That is, on the one hand, the study aimed to

¹ In the Encarta dictionary, Heuristics is defined as "problem solving by trial and error", or, more explicitly, "a method of solving a problem for which no formula exists, based on informal methods or experience, and employing a form of trial and error".

identify separate ability factors important for solving mathematical problems, but at the same time, these factors ought to correspond to tangible cognitive processes useful for teaching and learning, and for identifying specific weaknesses students have in solving mathematics problems.

This starting point was rather ambitious, as it involved a long and iterative process of developing a theoretical framework and validating it. Initial attempts to analyse a number of existing problem-solving data sets using factor analysis produced dismal results. The factor analysis results were unstable, with the factors and factor loadings varying considerably depending on the number of items selected, the number of factors extracted, and the test forms used. Three possible reasons were identified for this: (a) The factor analyses identified item difficulty as one factor, even when tetrachoric correlations were computed as the basis for the factor analyses; (b) the tests had a speededness effect, so that many missing responses at the end of the tests were not-reached items, and they did not reflect students' inability to answer the questions; and (c) many items involved multiple cognitive processes which were not captured with a correct/incorrect scoring (Wu, 2004).

These findings highlighted the importance of developing assessment instruments that clearly reflected the theoretical aspects of problem-solving processes, and that had scoring guides that captured the different processes. Consequently, new test instruments were developed for this study, specifically designed to capture the information required that matched the theoretical basis set out in the framework of this study. The difficulty, however, was 'how does one design a test instrument based on a theory that is yet to be developed through the empirical validation of the instruments'? The starting point was to gather information from published research studies as well as studying student responses from existing problem-solving tests developed in the past. An iterative process followed with the development and validation of the theory.

As the goals were to improve students' problem-solving proficiencies, it would be helpful to identify factors that could be translated directly into instructional practices. The factors identified through the factor analytic approaches such as Fluid and Crystallised Intelligences are somewhat removed from direct classroom applications. How does one improve one's Fluid and Crystallised Intelligences even if the levels of these are measurable? The problem-solving stages identified using the information processing approach are more likely to be of practical use in the classrooms. However, these stages are often formal concepts that are not always easily translated into instructions. For example, how does one devise a plan, and how does one carry out a plan? Is it always necessary to devise a plan? Do different problems require different ways to devise a plan? In addition, problem-solving processes such as planning, analysing, exploration are concepts that are difficult to observe, and hence difficult to measure through an assessment instrument. It is also doubtful that planning, analysing, exploration, and implementation would each be associated with different ability factors. These are labels for problem-solving heuristics, not necessarily distinct factors of

cognitive abilities. For example, in each step of the problem-solving process, reasoning skills are often required. Consequently, stages of problem solving as identified by information processing theorists involve an overlapping set of cognitive abilities.

A study of students' problem-solving item responses showed that students' failure to find the correct answers was not always due to their inability to follow a formal problem-solving process. Most of the time, the failure was caused by some errors made by the students, rather than a complete failure to approach a problem. By studying common errors students made, we identified the cognitive processes that were important in solving mathematical problems, with the belief that, if students were taught how to avoid common errors, they would be better problem solvers. For example, we found that many incorrect answers were the result of misunderstanding of the problem texts or carelessness of computation. There were, of course, also incorrect applications of mathematical concepts and formulation. Identifying the sources of error can be helpful in providing instructional materials that teach students how to avoid the errors. Furthermore, if these sources of error are found to form different ability *dimensions*, then we have reasons to believe that these sources of error relate to different cognitive processes. A report of students' *profiles* based on these dimensions could be useful to identify individual differences in their strengths and weaknesses, with possible remedial plans.

Four dimensions of problem solving eventually were identified as the basis for the problem-solving framework of this study, and these were tested through a number of trials in the schools. The identification of the dimensions was based on three principles: (a) that the dimensions provide useful information for teachers and students, (b) that a student's behaviour associated with the dimensions is observable through a test instrument, and (c) that the response data from the test instruments can be modelled and analysed using available software. The dimensions are described below with supporting theoretical and empirical rationale.

Dimension 1: Reading/Extracting all information from the question

Both the factor-analytic and information processing approaches identified reading as an important component in the problem-solving process. Clearly, one cannot proceed with solving a problem until one understands what the task is. In our current setting of paper-and-pencil tasks, the first step is to read and understand the given task. The definition of this dimension goes beyond reading comprehension of the words. It includes understanding the problem situation. Consider the following item.

Michael drives to work. The distance between his home and work is 25 km.
If Michael works five days a week, what is the total distance Michael travels between his home and work in a week?

To understand the problem situation fully, one needs to know not only about the distance between Michael's home and work, and the number of working

days in a week, but also the fact that Michael needs to make two trips per day—drive to work and back from work. This latter information is not explicitly stated in the question, but nevertheless it forms part of the information required for solving the problem. A full understanding of the problem situation will include the number of trips required per day. Without a clear understanding of all the parameters related to the problem situation, one cannot solve the problem correctly.

There are many examples that demonstrate that the failure in solving a problem correctly is often due to an incomplete, or incorrect, understanding of the problem situation. Whimbey and Lochhead (1991, p. 26) place a strong emphasis on the importance of understanding a problem:

Good problem solvers take great care to understand the facts and relationships in a problem fully and accurately Quite often (poor problem solvers) could have found out (the solution) if they had been more careful. But poor problem solvers have not learned how important it is to try to be completely accurate in understanding all the ideas of a problem.

The following is an example where about one quarter of the students failed to obtain the correct answer due to incorrect reading:

Here is a sequence of numbers:

2, 5, 8, 11, 14, 17,

What is the 10th number?

- A 20
- B 21
- C 28
- D 29
- E 47

In a trial, while 67% selected the correct answer, 23.7% of the students chose the incorrect distractor A. It is conjectured that these students have not read the sentence “What is the 10th number?” They proceeded to give the answer for the next number in the sequence. These students have made a reading error rather than an error of mathematics. In contrast, the following item had a much higher percentage correct (85%) when the next number in the sequence was not one of the distractors.

Starting from 2 and counting by 7, Tim gets the following numbers:

2, 9, 16, 23, 30, 37, ...

What is the 10th number?

- A 20
- B 21
- C 28
- D 29
- E 47

The second example is expected to be more difficult than the first example, as it involves counting by 7 instead of counting by 3. The percentage correct, however, is higher than that for the first example. One possible explanation for this is that in the second example students who thought the answer was the next number in the sequence had to re-read the question, since the next number in the sequence was not one of the answer options. Clearly, understanding the problem situation is one important factor in problem solving.

Dimension 2: Real-life and Common Sense Approach to Solving Problems

One aspect of mathematical problem solving that is most disturbing in recent years is that researchers have found that, in some instances, school mathematics was taught in such a way that school children, after receiving instruction in mathematics, regarded mathematics purely as an academic discipline divorced from real-world problem solving (Verschaffel & de Corte, 2000). Neshet (1980) gave the following problem to grade five students. "What will be the temperature of water in a container if you pour 1 jug of water at 80°F and 1 jug of water at 40°F into it?" Many children answered "120°F"! Other studies also found that many children are quite happy to give answers such as 5.2 buses or 8.4 balloons as they accurately carry out long divisions or multiplications.

Termed "suspension of sense-making" Schoenfeld (1991, p. 316), this phenomenon is widespread around the world. Cai and Silver (1995) found that while Chinese students are better at the execution of the division computation procedure, US students outperform Chinese students in giving "sense-making" answers. (Schoenfeld p. 316) commented that: "There is reason to believe that such suspension of sense-making develops in school, as a result of schooling."

In a trial test, we gave the following item to grade 5 and 6 students.

A small hose can fill a swimming pool in 12 hours, and a large hose can fill it in 3 hours. How long will it take to fill the pool if both hoses are used at the same time?

- A 2.4 hours
- B 4.0 hours
- C 7.5 hours
- D 9.0 hours
- E 15.0 hours

The only sense-making answer is A, where the time is shorter than the time for each hose alone. However, more than half of grades 5 and 6 students chose an incorrect answer. In particular, a number of high ability students chose distractor C, giving the average time for the two hoses.

These observations suggest that students' proficiency on traditional school mathematics topics like computational procedures do not necessarily

reflect their ability to solve real-world problems. In designing a test instrument, we included a number of items specially designed to tap into making sense of the problem situation and evaluating answers.

Dimension 3: Mathematics concepts, mathematisation and reasoning

In a test of mathematics problems solving, there obviously needs to be a dimension that taps into the heart of mathematics itself. It is intended to measure students' knowledge and skills in mathematics concepts, their ability to turn a problem situation into a mathematical model, and their ability to reason. To this end, items were designed covering arithmetic and measurement, as well as logical reasoning. In most cases, the items require higher order thinking skills than merely recall. This may be one distinction that was made in focusing on mathematics problem solving than just mathematics. In many cases, some degree of *mathematisation* is required. "Mathematisation" (Treffers, 1986) is the process of turning a problem situation into an appropriate mathematical formulation; a skill that is often lacking in primary school students, particularly when mathematics is taught in a rote manner. In this study, items were designed using real-world contexts as much as possible, as research studies showed that problem contexts could make a difference to the way students solve a problem. The following is a sample item.

At Nick's school, a round-robin tennis tournament is organised where every player plays every other player once. There are 6 players in the competition. The school has 3 tennis courts. The matches are scheduled every evening, with one match played on each tennis court. How many evenings are needed to schedule all the matches for the tournament? Show how you found your answer.

Dimension 4: Standard computational skills and carefulness in carrying out computations

In analysing some problem-solving trial data, it was found that a number of able students who understood complex mathematical concepts failed to carry out the computations correctly. From time to time, as teachers marked test papers, there was often a sigh: "This is a good student, but so careless!"

The following item is an example.

Megan obtained an average mark of 81 for her four science tests. The following shows her scores for Tests 1, 3 and 4. What was her test score for Test 2? Show how you found your answer.

Test 1	Test 2	Test 3	Test 4	Average Mark of 4 tests
84	?	89	93	81

An item analysis showed that 41% of students obtained the correct

answer and 6% of students used the correct method but made a computation error. The interesting observation was that these two groups of students had the same estimated average ability on this test. We also made the same observation with a number of other items where the item required a conceptual understanding as well as carrying out computation.

These observations suggest that there perhaps is a ceiling to the ability to carry out computation, and that computational skills do not continue to grow in parallel with the growth of concepts and reasoning skills. There is also an element of carelessness that plays out more in computation than in conceptual understanding. While carelessness may be a personality trait rather than learned mathematical proficiency, carelessness is one reason for students' failure to obtain the correct answer. From this point of view, it is important to track and distinguish between the nature of the error, whether it is in conceptual understanding or carelessness in computation.

Discussion of the Dimensions

The four dimensions as defined above place an emphasis on problem-solving processes rather than on traditional mathematics content strands. This approach is consistent with four recent directions for mathematics education.

First, the document, *Curriculum and evaluation standards for school mathematics* (National Council of Teachers of Mathematics, 1989), lists a set of goals for mathematics education that moves the curriculum away from the traditional emphasis on decontextualised mathematical knowledge towards the processes of problem solving and doing mathematics. Under this "holistic" approach, the traditional mathematics content strands are subsumed under the broader approach of problem solving.

Second, the theory of Realistic Mathematics Education (RME) developed in the Netherlands (de Lange, 1996; Gravemeijer, 1999) over the past 30 years is gathering support from around the world (de Lange, 1996; OECD, 2003; Romberg & de Lange, 1998). Two principles underlie RME: (a) Mathematics must be connected to the real-world; and (b) mathematics should be seen as a human activity.

Third, the concern with the observed phenomenon "suspension of sense-making" has prompted mathematics educators to advocate an approach that takes mathematics out of the classrooms and into the real-world (e.g., Bonotto, 2003; Verschaffel, Greer, & de Corte, 2000). This approach takes mathematics to the real-world, in contrast to RME that takes the real-world to mathematics. The emphasis is to solve real-world problems using mathematics, rather than to teach mathematical concepts using real-world problems. Real-world problems come in all different forms, and they are certainly not confined to the boundaries of traditional mathematics topics such as arithmetic, algebra, geometry and statistics. The PISA (Programme for International Student Assessment) mathematics framework (OECD, 2003) also takes this approach.

Fourth, Ellerton and Clarkson (1996) described a widely-used research methodology known as "Newman research":

According to Newman (1977, 1983), any person confronted with a written mathematics task needs to go through a fixed sequence: Reading (or Decoding), Comprehension, Transformation (or Mathematising), Process Skills, and Encoding. Errors can also be the result of unknown factors, and Newman (1983) assigned these to a composite category, termed “careless”. (p. 1000)

The Newman processes described above closely resemble the four dimensions derived for this research.

While the four dimensions in this study were chosen with some theoretical and empirical underpinning, it remained to be seen whether an assessment instrument could be designed to operationalise these dimensions as separable latent constructs. If successful, the reporting of students’ proficiencies along these separate dimensions would be helpful to teachers and students. Instead of reporting a single proficiency score of problem solving, one could identify specific areas of weaknesses and strengths. For instance, one might find that a student had good conceptual understanding of mathematics, but lacked meticulousness in reading and carrying out computation. Such profile building for each student would provide useful information. On the other hand, if the four dimensions were all highly correlated, there was little gain in reporting scores on separate dimensions. This validation of the dimensions of the framework formed the key tasks in this research, as described below.

Designing the Test Instrument

One problem mentioned before in analysing existing tests for dimensionality is that the items were not designed to provide information on the separate dimensions, particularly when items required multiple latent traits within one single task. For this reason, a clear articulation of the underlying dimensions described above helps with the construction of the items. The following is an example to demonstrate how an item may be constructed to provide information on the cognitive processes. Consider the following item.

Mrs Lang had a box of 12 icy poles. She gave 3 of them to the children. How many icy poles did Mrs Lang have left?

- A 3
- B 4
- C 8
- D 9
- E 11

In a trial test, 68% of students obtained the correct answer D, 24% chose A, and a small percentage chose C and E. The average estimated ability of students choosing D is the highest (where ability is estimated via a unidimensional Rasch model), followed by the group choosing A, and the lowest ability group of students chose E. It is hypothesised that the students

who chose A have most probably missed the last word in the question, *left*, and proceeded to answer the question: “How many icy poles did Mrs Lang give to the children?” This mistake is related to reading the question rather than computational skills. On the other hand, the students who chose E did not understand the question nor the mathematical concept involved in this question. It is hypothesised that these students applied a “keyword approach” to learning mathematics rather than the understanding of mathematical concepts. The keyword, *left*, prompted these students to use subtraction, as subtraction is commonly associated with word problems involving the word, *left*.

Not every item can use multiple-choice format to capture all the information required. Overall, about one quarter of the items were in multiple-choice format, the remaining items were open-ended, and scoring was needed to capture the different kinds of errors students made, in particular, whether the error was in computation or method. For example, for the item on the average score of four science tests (see Dimension 4 descriptions above), two variables were used to capture student responses. The first variable recorded the method used including correct conceptual approach, trial-and-error approach, or an incorrect approach. The second variable recorded whether computation was carried out correctly, irrespective of the method used.

Owing to practical constraints in the schools for trialing the test, four linked trial forms were prepared, where each form was designed for 30 minutes of administration time (approximately 20 items). A pilot test was first administered to a small sample of students. Data were analysed and items were revised to circumvent problems associated with the clarity of the wording of questions. A larger trial took place with 951 students in the sample. The students were from grades 5 and 6, with about an equal number of boys and girls, from a number of suburbs in both Sydney and Melbourne. In all, 48 items were trialed, but each student answered around 20 items only.

Unidimensional Rasch Analysis

A unidimensional Rasch analysis was carried out using the Random Coefficient Multinomial Logit Model (Adams, Wilson, & Wang, 1997). This model is implemented in the software ConQuest (Wu, Adams & Wilson, 1998). In its simplest form, this model is the Rasch model (Rasch, 1960). For this study, an extension of the Rasch model, the partial credit model (Masters, 1982), was used. Note that, in this paper, unidimensionality refers to a single latent variable being tested through all the items.

The unidimensional analysis showed that the tests had a reliability of 0.8 (test length approximately 20 items). Eight items had poor unweighted fit t values and two items had poor weighted fit t values (> 3.0 or < -3.0)²,

² For definitions of unweighted and weighted fit statistics, see Wright and Masters (1982).

indicating that the items did not all fit a unidimensional scale. However, in general, the residual based fit statistics are not very sensitive to departures from unidimensionality, particularly when there is an equal mix of items from different dimensions (Smith & Miao, 1994).

Confirmatory Factor Analysis and Multidimensional IRT Model

To find evidence of dimensionality in the data, a number of multidimensional Item Response Theory analyses were carried out and the goodness-of-fit of different models were examined to identify the best model. These multidimensional IRT analyses were essentially confirmatory in nature, as items were pre-assigned to dimensions, based on some theoretically grounded hypotheses.

A general form of the Multidimensional Random Coefficient Multinomial Logit Model (Adams, Wilson & Wang, 1997) was fitted, with *between-item dimensionality*. This means each item was loaded on a single latent dimension only so that different dimensions contained different items.

A three-dimensional model, a two-dimensional model and a one-dimensional model were fitted in sequence. The three-dimensional model assigned items into three groups. Group 1 consisted of items that had a heavy reading and extracting information component. Group 2 consisted of items that were essentially common-sense mathematics, or non-school mathematics. Group 3 consisted of the rest of the item pool, consisting of mostly items that were typically school mathematics, as well as logical reasoning items. In this IRT model, Dimensions 3 and 4 of the framework, mathematics concepts and computation skills, had been combined to form one IRT dimension. This was because there were no items that tested computation skills alone. So, for a between-item multidimensional IRT model, it was not possible to separate Dimension 4 of the framework, *computation skills*, as a separate dimension to be modelled in the IRT analysis.

The two-dimensional model was formed with Groups 2 and 3 combined and the one-dimensional model placed all items in a single dimension. To assess the relative model fit of the three models, the *deviances* from fitting the three models were compared. The deviance provides a degree of goodness-of-fit, and the smaller the deviance, the better the fit of the model. The difference between the deviances from fitting two models can be used to carry out a significance test to determine if the model fit has become significantly worse when the model was simplified with fewer parameters. Table 1 shows that the three-dimensional model fitted the best with the smallest deviance (and so it should, because there were more parameters fitted), and the model fit was worse when the model was reduced to two- and one-dimensions; that is, the three-dimensional model represented the structure of the item response data more appropriately than the two-dimensional and one-dimensional models.

The estimated correlations between the three dimensions are given in Table 2. These correlations are direct estimates of the population correlations;

that is, they are not attenuated by measurement error. To put the magnitudes of the correlations into perspective, in the OECD PISA³ project (Adams & Wu, 2002), the correlation between reading and science was 0.89, the correlation between science and mathematics was 0.85, and the correlation between reading and mathematics was 0.82. Using these correlations as a crude reference, one would not expect the correlations between the subscales of problem solving to be much lower than 0.82, the correlation between mathematics and reading in the PISA project (See Chapter 12, Bond & Fox, 2007).

Table 1
Comparison of Fit between Unidimensional and Multidimensional Models

Model	Deviance	Number of parameters	Change in Deviance	Change in Degrees of Freedom	p (significance)
3-Dimension	28164.9	70			
2-Dimension	28182.5	66	17.6	4	$p < 0.005$
1-Dimension	28196.7	63	14.2	3	$p < 0.005$

Table 2
Estimated Correlations among Dimensions

Dimension	Dimension		
	1(Reading)	2(Sense Making)	3(Others)
Dimension 1 (Reading)			
Dimension 2 (Sense Making)	0.91		
Dimension 3 (Others)	0.94	0.83	

To check whether the results obtained in the three-dimensional analysis could simply be due to chance, a three-dimensional model was run with items allocated to the dimensions in an arbitrary way. That is, items 1, 4, 7, 10, ... were loaded on dimension 1, items 2, 5, 8, 11, ... were loaded on dimension 2, and items 3, 6, 9, 12, ... were loaded on dimension 3. The deviance from this run was 28187, a figure of 22 more than the deviance from the first three-dimensional analysis where items were allocated according to hypothesised cognitive demands. A comparison of deviances between the three-dimensional run (with random allocation of items) and the one-dimensional run showed a reduction of 9 in deviance with 7 degrees of freedom, indicating that there was no significant decline in terms of fit when a unidimensional model was fitted. This indicates that the unidimensional model fitted just as well compared to this

³ PISA stands for Programme for International Student Assessment. It is an OECD project. In 2000, around 35 countries participated in this project with assessments in Reading, Science and Mathematics.

three-dimensional model where the items were assigned to dimensions at random. The estimated correlations among the three arbitrary dimensions are given in Table 3. These estimates of correlations clearly showed that the three dimensions were not distinguishable.

Table 3
Correlations among three Arbitrary Dimensions

Dimension	Dimension		
	1 (Random assignment)	2 (Random assignment)	3 (Random assignment)
Dimension 1 (Random assignment)			
Dimension 2 (Random assignment)	0.996		
Dimension 3 (Random assignment)	0.993	0.993	

These results give us some assurance that the presence of multidimensionality in the first three-dimensional model (where items were allocated to dimensions according to some theoretical basis) was not likely to be the result of chance alone. In fact, with the high correlations between all main subject domains as shown by PISA, it was extremely difficult to find groups of items within a subject domain to demonstrate multidimensionality. For example, the correlations between Reading subscales in PISA were 0.97, 0.89, and 0.93.

Modelling Within-item Dimensionality

To model all the cognitive demands of an item, one needs to examine *within-item dimensionality* (Adams, Wilson, & Wang, 1997) at the level of the response categories of the items. Whereas, between-item dimensionality models can only assign items to one 'dominant' dimension, within-item dimensionality allows for an item to be assigned to more than one dimension. Given that we have recorded more information than just right/wrong responses for a number of items, the data could be fitted to a more 'accurate' Item Response Theory model, according to the four hypothesised dimensions when items were designed.

Consider the following example:

After spending \$14.70 on meat and \$7.30 on vegetables in a supermarket, Anne had \$39.20 left in her purse. How much money did Anne have before going to the supermarket?

Coding the responses 1 for correct method but incorrect computation, and 2 for correct method and correct computation, one can set up the following within-item dimensionality IRT model (Multidimensional Random Coefficient Multinomial Logit Model), where $P(x)$ is the probability of the response category x :

$$P(0) = \frac{1}{D}, P(1) = \frac{\exp(\theta_1 - \delta_1)}{D}, P(2) = \frac{\exp((\theta_1 + \theta_2) - (\delta_1 + \delta_2))}{D}$$

where θ_1 is a student's ability of mathematics concept, θ_2 is computation ability. δ_1 is item difficulty parameter for mathematical concepts for this item, δ_2 is the item difficulty parameter for computation for this item⁴, and D is the normalising divisor, being the sum of the three numerators.

The decision as to which dimensions on which a response category should be loaded was largely determined by two considerations. The first consideration was the hypothesised cognitive demands when an item was designed. The second was a review of the fit of the item and the change in deviance for the overall model. From this point of view, a confirmatory factor analysis was essentially carried out, or to some extent, exploratory factor analysis, in the sense that there was an iterative process where a hypothesis was formed and tested, and then adjusted and further tested. It should be noted that in cases where the category scores had changed from one model to another, the model could no longer be considered as a sub-model of the earlier model. In such cases, one could not compare the deviances of the two models.

The four-dimensional model produced item parameters with reasonable fit values, with no unweighted fit t or weighted fit t values greater than 3. That is, all items seemed to fit the dimensions they were assigned. The estimated correlations between the four dimensions are shown in Table 4.

⁴ While $P(2)$ appears to be in the form of a compensatory multi-dimensional IRT model, the 'local step' between 0 and 1 involves only θ_1 and δ_1 , and the local step between 1 and 2 involves only θ_2 and δ_2 . From this respect it seems the model is not compensatory. As the local steps have a simple Rasch model, the equations have the form of a multi-dimensional partial credit model. This model differs from the multi-component model (Embretson, 1997) or the steps model (Verhelst, Glas, & de Vries, 1997) in that the probabilities of the two 'steps' are not multiplicative. With the multi-component or the steps model, the step probabilities are multiplied to form the likelihood of a response pattern, thus making an assumption of independence between the two steps. In contrast, the model applied here does not assume independence. While it maybe likely that internal steps within a task are not independent, the implicit dependency can make the interpretation of item parameters difficult (Verhelst, Glas, & de Vries, 1997). In this project, the main focus, however, is to give feedback to students rather than to analyse item features. So the interpretation of item parameters is not a main concern.

Table 4
Correlations among Four-dimensional Within-item Model

Dimension	Dimension		
	1 (Reading /extract information)	2 (Sense -making)	3 (Maths concepts)
Dimension 1 (Reading /extract information)			
Dimension 2 (Sense-making)	0.95		
Dimension 3 (Maths concepts)	0.92	0.79	
Dimension 4 (Computation)	0.82	0.80	0.85

These estimates suggest that the correlation between *reading/extracting information* and *sense-making* is the highest. This conclusion seems plausible as the ability to fully comprehend and extract information is closely related to making sense of the problem situation. However, the correlation between *sense-making* and *mathematics concepts* or *computational skills* is relatively lower. This result confirms the studies of researchers such as Verschaffel and de Corte (2000) who found that sometimes the more school mathematics is taught, the more students divorce mathematics from the real-world. *Computation*, on the other hand, has relatively lower correlations with *reading* and *sense-making*, but slightly higher correlation with *mathematical concepts*. This result is not surprising as computation involves basic mathematical concepts; but the fact that *computation* and *mathematical concepts* do not have a correlation close to 1 suggests that the development of mathematical concepts does not always go hand-in-hand with that of computation.

Reporting Students' Problem-solving Proficiency

This model of within-item dimensionality enables us to build a profile for each student, identifying areas of strength and weakness. For example, some students have a good grasp of mathematical concepts but fall down in extracting all information from the question, while others need to improve their mathematical knowledge and skills. IRT modelling enables us to provide such profiles for students from a relatively short test, as we extract as much information as possible from all item responses, and not just correct or incorrect answers. Figure 1 shows examples of student profiles. The horizontal axis shows the four dimensions. Each student's proficiency level on each dimension is shown on the vertical axis. The mean and standard deviation of abilities on each dimension have been standardised to 0 and 1 respectively, to make the scales comparable across the four dimensions.

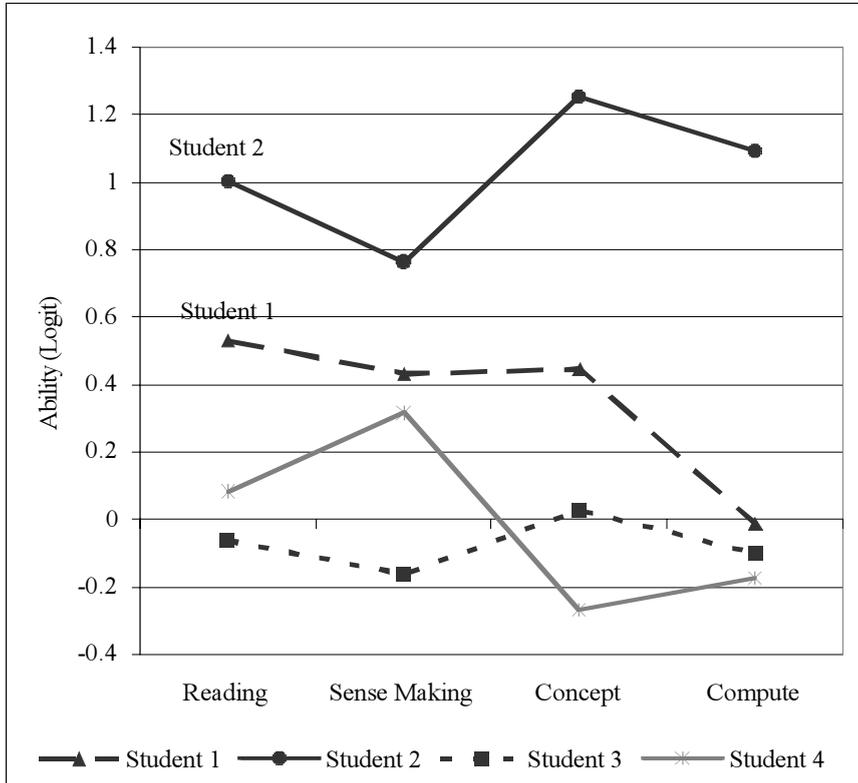


Figure 1. Students' Ability Profiles on the four dimensions in standard deviation units.

It can be seen that the computation skills of Student 1 could be improved in relation to the other three skills, while Student 2 has strengths in computation and mathematical concepts but should take a closer look at a common sense approach to solving mathematics problems.

A formal validation of the students' reported profiles is not easy. However, checking the reported profile patterns with the students' test papers and item responses generally showed agreements. For example, a marker has placed a note on the test paper of Student 1 indicating that the student has made some careless computational errors.

More generally, a profile of problem-solving proficiency can provide teachers and students with information about appropriate remedial measures. For example, a student with a relatively lower score on the reading/extracting information dimension can focus on improving skills such as reading comprehension, visual imaging, organising information and so on. A student with a relatively lower score on the common sense approach dimension can focus more on checking and evaluating answers. The fact that the four dimensions are reflecting somewhat different ability factors suggests

that specific lessons can be devised to focus on each problem-solving dimension separately, and targeted intervention should more effectively improve students' problem-solving skills. In many cases, problem-solving skills can be improved just through an awareness of one's own areas of weaknesses when these weaknesses are clearly identified in the process of problem solving. Thus, the usefulness of the problem-solving framework as defined in this study lies in the fact that distinct skills are identified, and, at the same time, students can clearly relate to these skills in the sequential process of solving problems.

Conclusions

This research showed that multidimensional Item Response Theory models are powerful tools for extracting information from a limited number of item responses. A within-item multidimensional IRT model allowed for response categories to be loaded on different dimensions, while factor analysis modelled item responses at the item level only. In addition, factor analysis was prone to idiosyncratic disturbances in the item features. It failed to extract relevant factors particularly when the items came from linked test forms. In contrast, IRT models were able to deal with linked test forms, at the same time allowing for confirmatory factor analysis to be carried out.

Although the results obtained in this study clearly indicated the presence of multidimensionality in the data, as seen from estimated correlation coefficients between the dimensions and the fit indices, there was, however, no evidence that the model fitted was the best model one could find. In fact, there would certainly be models with different loadings on the dimensions that would improve the fit further. What we have demonstrated is a methodology that can help disentangle complex cognitive processes embedded within a task. There is likely to be no 'true' and unique identification of cognitive factors. Any method of modelling the dimensions is worthwhile provided that the results are useful. The strongest message, however, is that the test items must be designed with a sound theoretical underpinning in cognitive processes, without which there is little chance of uncovering useful ability profiles for reporting back to teachers and students.

References

- Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1—23.
- Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.
- Bond, T.G., & Fox, C. M. (2007) *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed). Mahwah, NJ: Lawrence Erlbaum.
- Bonotto, C. (2003). Suspension of sense-making in mathematical word problem solving: A possible remedy. Retrieved August 16, 2003, from <http://math.unipa.it/~grim/Jbonotto>
- Cai, J., & Silver, E. A. (1995). Solution processes and interpretations of solutions in solving division-with-remainder story problems: Do Chinese and U.S. students have similar difficulties? *Journal for Research in Mathematics Education, 26*, 491—497.

- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 10, 1—19.
- Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.
- Carroll, J. B. (1996). Mathematical abilities: Some results from factor analysis. In R. J. Sternberg & B.-Z. Talia (Eds.), *The nature of mathematical thinking*. Mahwah, NJ: Lawrence Erlbaum.
- Collis, K. F., & Romberg, T. A. (1992). *Collis-Romberg mathematical problem solving profiles*. Melbourne: Australian Council for Educational Research.
- Cornish, G., & Wines, R. (1977). *Mathematics profiles series: Operations test teachers handbook*. Melbourne: Australian Council for Educational Research.
- De Lange, J. (1996). Using and applying mathematics in education. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 49—98). Dordrecht, The Netherlands: Kluwer.
- Ellerton, N. F., & Clarkson, P. C. (1996). Language factors in mathematics teaching and learning. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 987—1033). Dordrecht, The Netherlands: Kluwer.
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56, 495—515.
- Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag
- Fredriksen, J., Mislevy, R. J., & Bejar, I. (Eds.) (1991). *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum.
- Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning: An International Journal*, 1(2), 155—177.
- Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, 10, 287-302.
- Heim, A. W. (1975). *Psychological testing*. London: Oxford University Press.
- Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt.
- Lincare, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2(3), 266—283.
- Malone, J. A., Douglas, G. A., Kissane, B. V., & Mortlock, R. S. (1980). Measuring problem-solving ability. In S. Krulik & R. E. Reys (Eds.), *Problem solving in school mathematics* (pp. 204—215). Reston, VA: NCTM.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149—174.
- Masters, G. N. & Doig, B. A. (1992). Understanding children's mathematics: Some assessment tools. In G. Leder (Ed.), *Assessment and learning of mathematics*. Melbourne: Australian Council of Educational Research.
- Mayer, R. E., & Hegarty, M. (1996). In R. J. Sternberg & B.-Z. Talia (Eds.), *The nature of mathematical thinking*. Mahwah, NJ: Lawrence Erlbaum.
- McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82—99.
- Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items — Comparison of different approaches. *Journal of Educational Measurements*, 31, 1—18.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

- Nesher, P. (1980). The stereotyped nature of school word problems. *For the Learning of Mathematics*, 1(1), 41—48.
- Newman, M. A. (1977). An analysis of sixth-grade pupils' errors on written mathematical tasks. *Victorian Institute for Educational Research Bulletin*, 39, 31—43.
- Newman, M. A. (1983). *Strategies for diagnosis and remediation*. Sydney: Harcourt, Brace Jovanovich.
- OECD (2003). *The PISA 2003 assessment framework*. Paris: OECD.
- Polya, G. (1973). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: Princeton University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Romberg, T., & de Lange, J. (1998). *Mathematics in context*. Chicago: Britannica Mathematics System.
- Schoenfeld, A. H. (1983). Episodes and executive decisions in mathematical problem solving. In R. Lesh & M. Landau, M. (Eds.), *Acquisition of mathematics concepts and processes* (pp. 345—395). New York: Academic.
- Schoenfeld, A. H. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. E. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 311—343). Hillsdale, NJ: Lawrence Erlbaum.
- Silver, E. A. (1982). Knowledge organisation and mathematical problem solving. In F. K. Lester & J. & Garafalo (Eds.), *Mathematical problem solving: Issues in research*. Philadelphia, PA: Franklin Institute Press.
- Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 316—327) Norwood, NJ: Ablex.
- Stacey, K., Groves, S., Bourke, S., & Doig, B. (1993). *Profiles of problem solving*. Melbourne: Australian Council for Educational Research.
- Treffers, A. (1986). *Three dimensions*. Dordrecht, The Netherlands: Reidel.
- Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag.
- Verhelst, N. D. (2001). *Some thoughts on reliability*. Unpublished manuscript.
- Verschaffel, L., Greer, B. & de Corte E. (2000). Making sense of word problems. Lisse, Switzerland: Swets & Zeitlinger.
- Wang, W. (1998). Rasch analysis of distractors in multiple-choice items. *Journal of Outcome Measurement*, 2(1), 43—65.
- Whimbey, A., & Lochhead, J. (1991). *Problem solving and comprehension* (5th ed.) Hillsdale, NJ: Lawrence Erlbaum.
- Willmott, A. S., & Fowles, D. E. (1974). *The objective interpretation of test performance*. Windsor, UK: National Foundation for Educational Research Publishing.
- Linacre, M. J., & Wright, B. D. (2000). WINSTEPS Rasch measurement computer program [Computer software]. , Chicago: MESA Press.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wu, M. L. (2004). *The application of item response theory to measure problem-solving proficiencies*. Unpublished doctoral dissertation, The University of Melbourne.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Multi-aspect test software [Computer software]. Melbourne: Australian Council for Educational Research.

Authors

Margaret Wu, Assessment Research Centre, University of Melbourne, Parkville, VIC Australia 3010. E-mail: <m.wu@unimelb.edu.au>

Raymond J. Adams, Assessment Research Centre, University of Melbourne, Parkville, VIC Australia 3010. E-mail: r.adams@unimelb.edu.au