Towards a Framework for Numeracy Assessment

Rosemary Callingham
University of Tasmania
<Rosemary.Callingham@Central.tased.edu.au>

Patrick Griffin
University of Melbourne
<p.griffin@edfac.unimelb.edu.au>

One approach to assessing numeracy is through performance assessment that requires students to create a response. Four specially designed performance assessment tasks, addressing different aspects of mathematics, were used to assess students' numeracy achievement. These tasks were considered for their educational measurement characteristics and the underlying variable interpreted from a numeracy perspective. An overall developmental sequence of numeracy was found, with associated 'productive elements'. From these, a framework for numeracy assessment is proposed and implications for assessment and classrooms are examined.

Current Australian definitions of numeracy all stress the practical and contextual nature of the concept (e.g., AAMT, 1997a; AAMT, 1997b). As well as the application of mathematical knowledge, numeracy often requires the communication of the mathematical ideas to other people (DEA, 1995). Assessing numeracy, therefore, requires demonstration of not just the mathematical skills, but also the communication of ideas and thinking within a grounded context. Performance assessment, defined as 'assessments in which pupils create an answer or product that demonstrates their knowledge and skill...' (Airasian, 1994 p. 228) is one way of achieving this.

Judging performance assessment is generally done by comparison against some pre-defined over-arching criteria (e.g., Beesey et al., 1998) or by using a scoring rubric developed by considering students' responses to the task during trialing or administration (e.g., Harmon et al., 1997). These scales may be termed holistic or analytical respectively. Rarely does performance assessment provide an analytical scale within an holistic sequence of development of thinking or skill (Moskal, 2000).

This paper reports the outcomes from a rigorous assessment program using four different performance assessment tasks. The assessment was intended to evaluate the effectiveness of a teacher professional development program aimed at enhancing numeracy outcomes of Indigenous students in lower secondary years—the INISSS program (Improving Numeracy for Indigenous Students in Secondary Schools) (Nicholson, 1999). The tasks were developed specifically for the program to match the teaching approaches being encouraged in the professional development sessions (Callingham, Griffin & Corneille, 1999). Each task involved students in a short, structured investigation around different mathematical concepts. The targeted ideas were: identifying relationships (*Street Party*); exploring number patterns (*Bean Counters*); variation and bias (*Come in Spinner*); and area (*Newspaper by the Metre*). Each task provided both holistic and analytical scales, the holistic scale being intended to define overall concept development for teaching purposes and the analytical intended to provide an easily used scoring system for consistent rating of students by teachers.

Research Questions

The four tasks used each addressed a different area of the mathematics curriculum. With this in mind, the first research question addressed was:

- Do the tasks measure the same variable in a consistent and reliable manner?

If this were established, then the second question became:

- How could the variable be interpreted in the context of students' demonstrated responses and classroom practice?

Methodology

Nearly 2000 year 8 students in the nineteen (19) INISSS project schools attempted the tasks. A pre-test/post-test model was used, with each student completing two tasks early in the 1999 school year and the other two in October 1999. Classroom teachers administered the tasks to their classes in mathematics lessons. Teachers presented each task as an assessment task, but using their normal teaching and learning activity such as scaffolding through open-ended questioning and reading the questions to poor readers. Any tools normally available in the classroom, including calculators, could be used except where restrictions were specified.

Each task was based around a story shell and provided a graded set of items that targeted a developmental sequence within a particular area of the mathematics curriculum. The text of *Street Party* is provided in the appendix, other tasks were similar in structure. Each question within each task was also provided with an analytical scoring scale that was used for marking student responses (Callingham, 1999). Thus each task provided teachers with an holistic scale of development that covered the whole task, and a detailed analytical scale for each item. An example of these scales is provided in Table 1 for question 6 from *Street Party*, a task addressing early algebra concepts. Students were asked to make a long table by placing small tables end-to-end, and to find out how many people they could seat for given numbers of tables. The analytical scale is placed approximately where the different stages would fall on the holistic scale that applied to the whole task.

Table 1
Holistic and Analytical Scales for Street Party Question 6:

How could you work out how many people they could seat for any number of tables? Write to Dean explaining your method in the space on your record card.

Holistic scale	Analytical scale	Score
	No explanation or irrelevant answer.	**0**
Students recognise patterns in pictures and diagrams, and continue these accurately. They use this ability to solve simple problems involving direct relationships only.	Explanation **based on patterns only** e.g., it always goes up by 2.	**1**
Students can describe straightforward relationships in words, and relate these to a real situation. They use this ability to solve problems involving direct relationships where the numbers involved are too large to allow just a continuation of the pattern. Students can describe straightforward concrete relationships in words or symbols and manipulate these effectively. They use these skills to solve problems involving direct and inverse relationships.	Explanation based on a **generalised relationship** expressed in words or symbols.	**2**

By providing both holistic and analytical scales it was hoped that marking the tasks would be clear and unambiguous and interpretation of the students' responses would be enhanced. The 41 teachers who participated in the professional development sessions received training in the use of the scoring rubrics; other teachers relied on the advice of these colleagues.

Following a feedback session with teachers after the first administration of the tasks, and initial data analysis, some modifications were made to some scoring rubrics, to make them easier for teachers to use and more explicit about the nature of the desirable response. For each task, at least four rubrics were left unchanged to provide an anchor for the second administration analysis. Some minor modifications were also made to the student answer sheets to allow students more room for their responses, by presenting these as folded A3 rather than A4 size. These changes were not expected to influence the substance of students' answers (Callingham, Griffin & Corneille, 1999).

Findings

Measurement Characteristics of the Tasks

Person ability and item difficulty. Rasch analyses using the QUEST computer program (Adams & Khoo, 1995) allowed the relative positions of person ability and item difficulty to be plotted on a single scale. This is shown in Figure 1. The variable map indicates that each task has items ranging from very low difficulty to very high difficulty, as shown by the positions on the scale. Further, it matched the range of ability in the year 8 students who completed the tasks, shown on the left-hand side of the figure. Thus the tasks allowed all students, from the weakest to the most able, to demonstrate achievement.

Fit to the model. To establish the uni-dimensionality of the variable, as required by the Rasch model, the fit of the data to the model was determined. After the first administration, one group of items in the *Newspaper by the Metre* task did not appear to fit the model. The scoring rubrics for these were changed to make them more explicit and understandable to the teachers who had to mark them. The fit map for the second administration is shown in Figure 2.

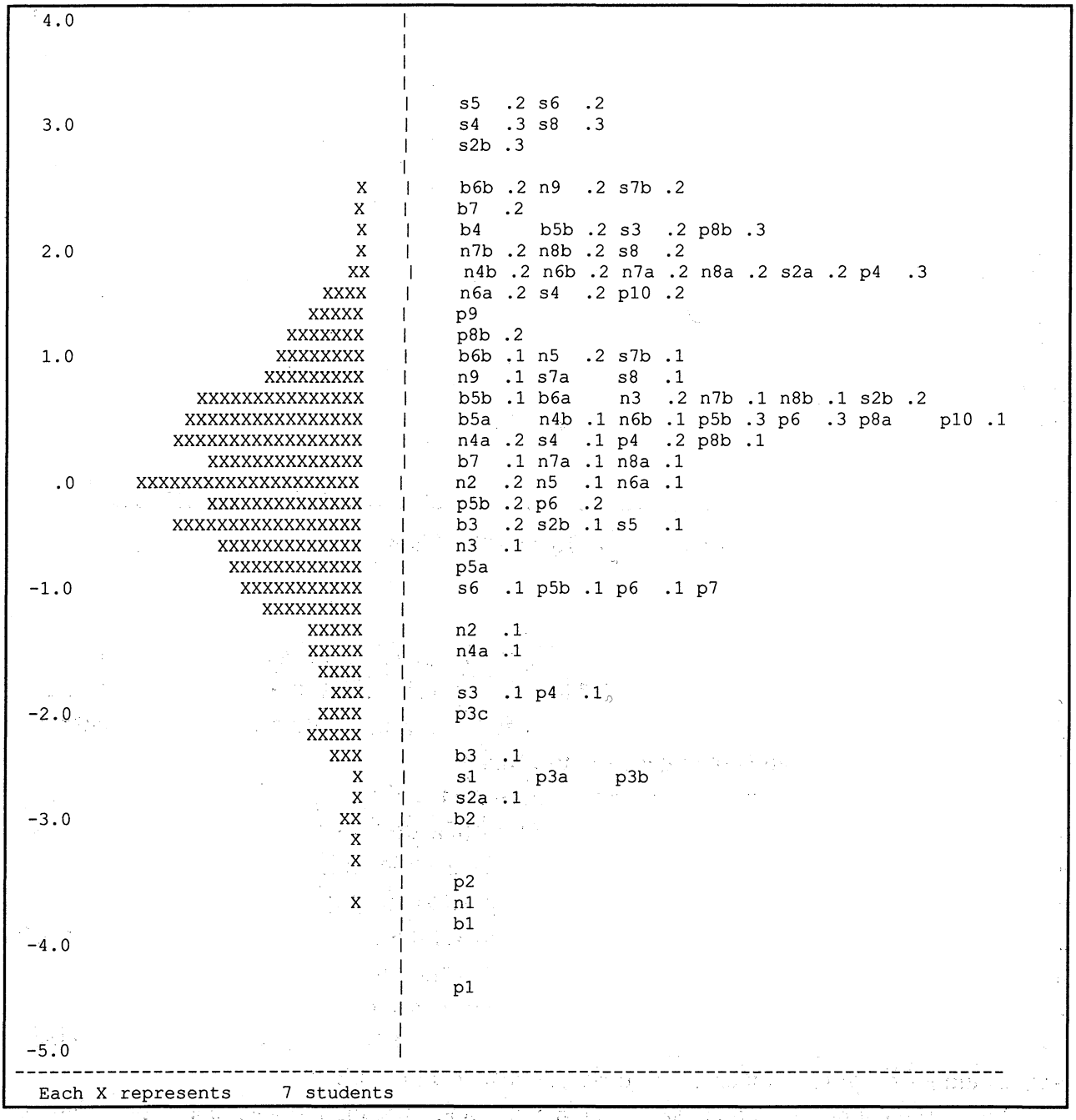
Although there is some misfit of data to the model, shown by the markers that fall outside the accepted limits, this can mostly be explained by students finding the items easier than expected at the second administration. The earlier misfit in the newspaper task items had disappeared, suggesting that the application of the revised scoring rubrics had improved the fit of these items to the model. Overall, the fit to the model is acceptable, and thus it can be concluded that the tasks were measuring the same variable, which was defined as numeracy.

Interpretation of the Variable

Since it was evident that the tasks were all measuring the same variable, it was important to identify common skills and understandings across all tasks. This was done by a qualitative content analysis of clusters of items having similar difficulty levels. For example, the easiest items required students to solve simple, one-step problems with the concrete materials provided if needed. *Bean Counter* 1 asked students to solve an addition square involving single-digit numbers, using beans to model this if they wished. *Street Party* 1 required students to make a long table from two small tables and work out how many people could sit at it. *Newspaper by the Metre* 1 asked students to construct a square metre out of newspaper. The

cognitive processes involved appeared to require pattern recognition in straightforward contexts—the pattern in the beans in the squares, in the number of people and in a square shape. A similar procedure was undertaken for all items.

At the lowest level, the items required pattern recognition and use, further up they demanded identification and application of rules and, at the top level, generalisation to a range of situations. Thus a sequence of pattern → rule → generalisation was identified. This provided a hypothesised overall holistic scale that was derived from the clusters of students' responses, determined by application of the analytical item scales.



Key to Items: **p** Street Party **n** Newspaper by the metre **s** Come in spinner **b** Bean counter

Figure 1. Variable map for INISSS numeracy tasks

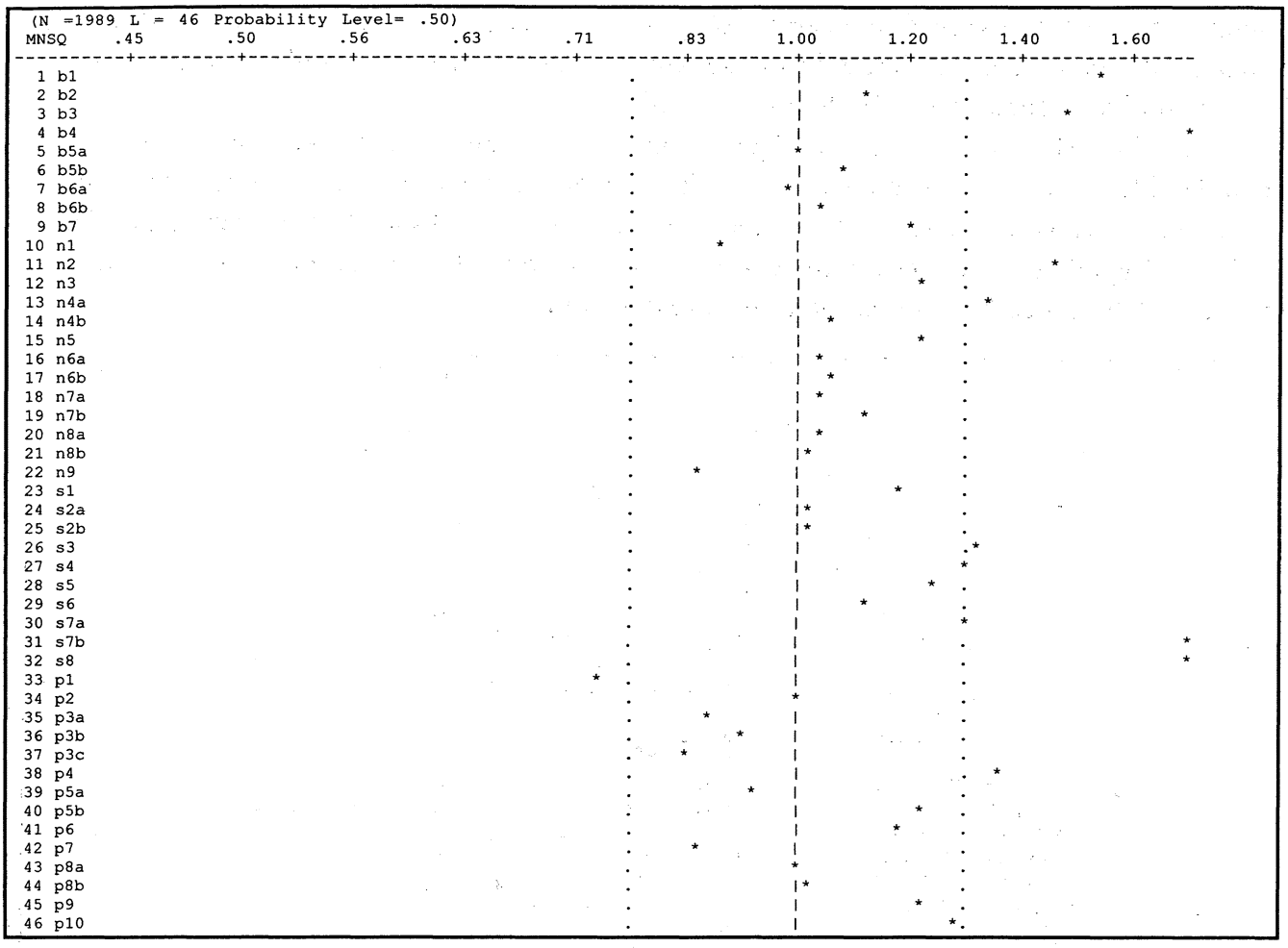


Figure 2. Fit map for INISSS items

Students' Responses

In order to verify this sequence, consideration was given to actual students' responses produced during the assessment process. While the responses fitted the overall hypothesised cognitive developmental sequence, it appeared that there were differences in the manner of responding to the tasks that were aligned to this sequence.

Students who were at the lower end of measured ability typically provided minimal written responses. Their written explanations were often irrelevant such as '...youseing my brans (*sic*)'. Some of these students had been videotaped while undertaking the tasks and they were revealed using strategies such as counting by twos or physical manipulation of the concrete materials supplied. Students operating at around the middle level of the ability range tended to write their explanations in natural language while those at the upper end could use symbolic notation and mathematical language. These response modes were termed 'productive elements'. A summary of the 'cognitive development' and 'productive elements', together with examples of typical students' responses, is provided in Table 2.

This combination of 'cognitive development' and 'productive elements' provided a framework for the interpretation of students' demonstrated outcomes. It described increasingly sophisticated mathematical concepts that were matched by a sequence of response modes and this formed the basis for reporting students' progress to teachers.

Table 2.
Hypothesised Developmental Framework for Numeracy Assessment

Cognitive	Productive	Party	Newspaper	Bean	Spinner
Relationship/ generalisation use	Summary such as graphs or symbolic language.	Expresses rule in symbolic language e.g., 2n + 2 or similar and uses this to solve extended and inverse problems.	Estimation of area of larger and smaller objects in square metres based on understanding of relationship between units and personal benchmarks.	Maximum number of beans is 35 (with associated reasoning).	Two factors affecting spinner - physical such as pencil not being centred, introduces bias, and need for large sample - as gets bigger more likely to centre around expected value.
Rule extension			Estimation using personal benchmarks based on familiar objects e.g., exercise book.	"Because you must have at least one bean in each square the number of beans in other squares is limited."	
Rule application	Written explanation - natural language	"For any number of tables you double and add two for the ends."	Recognition of size by direct comparison with square metre.	Completes all possible solutions to a repetitive problem.	"It's a 50:50 spinner therefore for 1000 spins you'd expect 500 darks."
Rule recognition		"It goes up by 2's."	"You can cut up the square and stick it together again and it will still be a square metre."	"As one goes up the other goes down."	"It's half and half."
Pattern use	Verbal explanation in natural language.	Counts by 2 orally.	Measuring to determine size.		
Pattern recognition			Use of square.	"We juggled the beans, the numbers add up."	

Classroom practice. Considering the measured variable in the context of classroom practice provided further verification of the hypothesised framework. Since most of the professional development sessions had been videotaped, these records were examined to relate the teaching approaches that were being advocated to the numeracy assessment framework.

The teaching approaches involved open-ended investigations with concrete materials. Following the initial assessment, teachers reported being disappointed with the ways in which their students completed the tasks. In particular, they commented on the poor explanations provided by students in their responses, and undertook to develop these skills explicitly. Student talk and discussion was encouraged, and many teachers reported increasing the emphasis on literacy skills. Videotaped classroom sessions indicated that teachers were doing this through modelling and reinforcing the importance of providing explanations, especially in written form.

Feedback from teachers after the second assessment session showed how much classrooms had changed. Students not only undertook the assessment tasks with greater reported levels of skill and confidence, they also persisted for longer and wanted to spend more time on the tasks. Some teachers reported spending up to six lessons developing the ideas behind the tasks. This may well be a case of assessment driving practice. However, in this instance the assessment process was specially designed to match the teaching and learning strategies being advocated.

Discussion

The performance assessment tasks discussed here provided a sound measurement approach to the assessment of a single construct defined as numeracy. These tasks supplied a means of assessing underlying mathematical ideas and skills through an analytical scoring rubric for each item within an holistic cognitive developmental sequence. In addition, the manner in which students responded to the tasks, suggested that different 'productive elements' were involved—talk, writing in natural language and the use of symbols and mathematical language. Combination of the cognitive sequence and 'productive elements' provided a framework for assessing numeracy, based on quantitative data from Rasch modelling and a qualitative analysis of the item content and students' responses. This framework has the following potential for teachers and administrators.

Firstly, reporting against this framework indicates at what level students are working, measured using assessment tasks that match teaching practice. The close link between teaching and assessment provides information that can be used to plan appropriate teaching programs, for intervention or extension. Further, the 'productive elements' aspect suggests that effective teaching should both target the cognitive level and be encouraging work production through a cycle of talk, record, symbolise. The framework explicitly targets a level and describes how teachers might intervene.

Secondly, this framework could be used to develop assessment tasks. The overall sequence provides the scope of the task and the analytical scales for each item can be designed to address different levels within the holistic sequence.

Further work is needed to confirm the components of the proposed framework, both cognitive and 'productive elements'. In addition, the feasibility of using the framework in these ways needs to be established by additional work with teachers. However, these early findings are encouraging, and this appears to have the potential to be a useful, rigorous assessment tool for teachers.

References

AAMT (Australian Association of Mathematics Teachers). (1997a). *Policy Statement on Numeracy in Australian Schools.* Adelaide: Author

AAMT (Australian Association of Mathematics Teachers). (1997b). *Numeracy = everyone's business. Report of the Numeracy Education Strategy Development Conference, May 1997.* Adelaide: Author.

Adams, R.J. & Khoo, S.T. (1995). *Quest: Interactive item analysis.* Melbourne: ACER.

Airasian, P. (1994). *Classroom assessment.* New York: McGraw-Hill.

Australian Association of Mathematics Teachers (AAMT). (n.d.). *AAMT discussion paper on assessment and reporting in school mathematics.* Adelaide: Author.

Beesey, C., Clarke, B., Clarke, D., Gronn, D., Stephens, M., & Sullivan, P. (1998). *Effective assessment for mathematics.* Melbourne: Victorian Board of Studies. Available: http://www.bos.vic.edu.au/.

Callingham, R., Griffin, P., & Corneille, K. (1999, November). *Using performance assessment tasks to assess numeracy outcomes: The INISSS project assessment process.* Paper presented at the Australian Association for Research in Education Conference, Melbourne.

Callingham, R.A. (1999). Developing performance assessment tasks in mathematics: a case study. In J.M. Truran & K.M. Truran (Eds.) *Making the Difference. Proceedings of the 22nd Annual Conference of the Mathematics Education Research Group of Australasia.* MERGA: Adelaide, SA. 4-7 July. pp. 135-142.

Harmon, M., Smith, T., Martin, M., Kelly, D., Beaton, A., Mullis, I., Gonzalez, E., & Orpwood, G. (1997). *Performance assessment in IEA's third international mathematics and science study (TIMSS).* Boston: TIMSS International Study Center.

Moskal, B.M. (2000). Scoring rubrics: What, when and how? *Practical assessment, Research and Evaluation,* 7 (3). Available online: http://ericae.net/pare/getvn.asp?r=7&n=3.

Nicholson, V. (1999, November). *The INISSS program.* Paper presented at the Australian Association for Research in Education Conference, Melbourne.

Appendix: Text of Street Party Task

Dean's community is planning a street party to celebrate the year 2000. They have lots of small square tables. Each table seats 4 like this:

The community decides to put the tables in an end-to-end line along the street to make one big table.

1. Make a line with 2 tables. How many people will be able to sit at it?
2. Make a line of 4 tables. How many people will be able to sit at it?
3. Make a line of tables that would seat
   (a.) 8 people. (b.) 12 people. (c.) 20 people.
4. Find two ways of showing **all** your results so far.
5. Dean says they can borrow 99 tables. How many people could they seat? Write a note to Dean explaining how you got your answer.
6. How could you work out how many people they could seat for any number of tables? Write to Dean explaining your method in the space on your record card.
7. Jen wants to use a small table that is not square. Make a different shape for a small table. Draw your small table for Jen showing the people sitting round it in the space on your record sheet.
8. Use your small table to make some big tables by putting small tables together. Draw three diagrams in the space on your record sheet showing how the big table grows, and the number of people that can sit around it as it grows. Explain to Jen what happens as your table grows by showing your findings in another way.
9. How many of your small tables would you need for 200 people?
10. Find a rule to work out how many of your small tables you would need for any number of people at the party. Write a note to Jen explaining your rule.