Australian Year 7 Benchmarks in Numeracy: Student Performance on Matching TIMSS Items

John Lindsey
Australian Council for Educational Research
<lindsey@acer.edu.au>

Barry McCrae
Australian Council for Educational Research
<mccrae@acer.edu.au>

TIMSS items that matched draft Australian Year 7 numeracy benchmarks were identified and the performance of students from Australia and other countries on these items was examined. It was concluded that at least two-thirds of Australian Year 7 students would achieve them in general, and that many would be achieved by at least four-fifths of Year 7 students. Items apparently assessing the same benchmark were spread along the TIMSS item difficulty continuum, indicating a likely difficulty in constructing tests to assess achievement of the benchmarks.

Benchmarks in numeracy at years 3, 5 and 7 were approved by Ministers for Education in States, Territories and the Commonwealth in April 2000 (Curriculum Corporation, 2000). The first draft of the Year 7 numeracy benchmarks was released during October 1998 for consultation, beginning a process of review and revision. A revised draft of these benchmarks was released in February 1999, and was followed by still further review and revision. The work reported in this paper was part of a project (Lindsey, Pearn, Lokan, Doig & O'Connor, 1999) undertaken by the Australian Council for Education Research, on behalf of the Commonwealth Department of Education, Training and Youth Affairs, to contribute to formative reviews during the benchmarks' development.

The numeracy benchmarks cover three aspects of numeracy: *Number Sense*, *Measurement and Data Sense*, and *Spatial Sense*. They are performance indicators that articulate nationally agreed *minimum* acceptable standards for Years 3, 5 and 7. That they describe *minimum* standards is emphasised, for example, by their incorporation as learning outcome indicators at levels 2, 3 and 4 in Victoria (Board of Studies, 2000) — thus they are expected to be achieved by the majority of Victorian students at the end of Years 2, 4 and 6 respectively.

The ACER study focussed on the February 1999 version of the draft Year 7 numeracy benchmarks. It had two main objectives:

- to compare the benchmarks with expectations in other countries; and
- to examine the performance of students from Australia and other countries on each of the Third International Mathematics and Science Study (TIMSS) (Lokan, 1999) items that could be matched to the benchmarks.

In the first part of the ACER study, the February 1999 benchmarks were compared with statements of expectations from twelve other education systems from eight countries. For each of the twelve performance documents, matches (i.e. closely corresponding statements) were found for than two-thirds of the benchmarks in each of the three strands (Number Sense, Measurement and Data Sense, and Spatial Sense). In general, where matches were found, almost all were where the other country, state or province had a higher expectation for the achievement of that standard. However, only two of the other documents contained expressly minimum standards and nor was it possible to obtain information on the actual performance of students against their countries' expectations.

Given the uncertainty about the basis for comparison of the Australian benchmarks with performance descriptions that were not developed as statements of minimum

acceptable achievement, it was concluded (Lindsey et al., 1999) that the TIMSS data could be considered to provide more useful reference points. It is this aspect of the ACER study that is reported in detail here, with particular reference to the Number Sense benchmarks.

All of the final Year 7 numeracy benchmarks have exact or very close equivalents in the February 1999 draft version. Of the thirteen final Number Sense benchmarks, twelve have direct equivalents in the February 1999 version and the other one (N7) is a combination of two February 1999 version benchmarks. One of the fifteen February 1999 Number Sense draft benchmarks (Interpret and use negative whole numbers in contexts relevant to students' everyday lives) has no corresponding final benchmark.

Methodology

Each TIMSS Population 2 (lower secondary level) item was carefully analysed by mathematics test developers to find the knowledge and skills required to give a correct answer. Depending on how closely each item appeared to be assessing the same content and skill as a draft benchmark, the item was classified as being a 'close match', a 'good match', a 'reasonable match', or 'no match' to the benchmark. About a third of the TIMSS items are secure: however, all items were examined in the matching exercise and results on all that were identified as being close, good or reasonable matches were included in the analysis.

Where one or more TIMSS items matched a benchmark, the TIMSS performance data on those items by Australian Year 7 students provided an estimate of the difficulty of the benchmark. A comparison of the difficulty estimates for the various benchmarks was then made.

The local TIMSS data cannot give a full Australia-wide picture of achievement for Year 7 because of the way the sampling had to be done to satisfy the TIMSS international criteria of using the two adjacent grades containing the majority of 13-year-olds. While in almost every other country the two lower secondary grades sampled in Population 2 were seventh and eighth grades, in Australia they were Year 7 and Year 8 in some states and Year 8 and Year 9 in others, depending on school starting age policies. Thus the youngest Year 7 students (and the oldest Year 9 students) were not included in the Australian TIMSS Population 2 sample. The Year 7 students for whom results were used came from New South Wales, Victoria, Tasmania and the Australian Capital Territory.

In the course of matching the draft Year 7 benchmarks to the TIMSS items, a few correspondences between the benchmarks and items in the Population 1 (10 year olds) test were located. Matches to these items were also used in the analysis on the rationale that numeracy knowledge and skills are gained cumulatively, and a standard expected at a lower grade level would be expected to be achieved also by students at a higher grade level. Details can be found in Lindsey et al. (1999).

Results

Twenty-six of the 157 Population 2 (lower secondary) TIMSS items were considered to be a close match to the draft benchmark statements, fifteen to be a good match and a further eleven to be a reasonable match. There were sufficient matches found to consider each benchmark strand separately.

Number Sense

In the Number Sense strand there were 22 Population 2 items (13 of which are released) matching a Year 7 draft numeracy benchmark. A further eight items from the Population 1 test were found to be good or close matches also. Nine of the 15 Number Sense benchmarks had at least one corresponding TIMSS item, with seven benchmarks having two or more items matched to them. Draft benchmarks N01, N05, N06, N07, N08 and N09 had no items matched to them, and so this analysis could give no information about their level of difficulty.

Measurement and Data Sense

In the Measurement and Data Sense strand there were 20 items (13 released) at Population 2 matching a Year 7 benchmark and a further six items at Population 1. Ten of the 12 benchmarks had at least one corresponding TIMSS item and eight benchmarks had two or more items matched to them. Draft benchmarks M06 and M08 had no matching items.

Spatial Sense

In the Spatial Sense strand there were 10 items (seven released) at Population 2 and five items (four released) at Population 1 which matched a Year 7 draft benchmark. Five of the seven benchmarks had at least one corresponding TIMSS item, with four benchmarks having two or more items matched to them. Draft benchmarks S03 and S07 had no items matched to them.

Estimating Benchmark Difficulty

The performance of Year 7 students on each of the matching released items for the Number Sense strand is shown in Table 1 below. Using these matches and the percentage correct data for Australian Year 7 students, an estimate of the difficulty for each benchmark was made.

The process used is illustrated in the following three examples:

Draft Number Sense Benchmark 3 (N03): "*Read, record and compare simple common fractions in contexts relevant to students' everyday lives*", had a total of 5 matches, one of them at Population 1, four of the five being good matches. Item 13 in Table 1 was the most difficult of the set with the other items clustered just below it. There were no close matches in this set because none of the items is set "in contexts relevant to students' lives". If the proposition is accepted that setting these items in a relevant context would not have increased their difficulty, then this evidence gives a facility estimate for Australian Year 7 students at 65% or easier for this benchmark.

Draft Number Sense Benchmark 11 (N11): "*Estimate answers to computations in contexts relevant to students 'everyday lives*", had a total of seven matches, all of them close, and all but one of them at Population 2. Items 19 and 1 in Table 1 set the lower and upper boundaries, respectively, for the facility range, with most items clustered around item 6 at 73% facility.

Draft Number Sense Benchmark 14 (N14): "*Use number sense, appropriate strategies, computational skills and key information to solve problems in contexts relevant to students' everyday lives*", had a total of four matching items, all at Population 2 and one a close match (item 9 in Table 1) which was also a good representative of the set. It had a facility at Year 7 of 70%.

Table 1

Year 7 Numeracy Benchmarks and Matching TIMSS Items

Number Sense benchmark		Feb 1999 b'mark	TIMSS item no[a]	TIMSS item code[b]	Int'l Yr 7 % correct	Aust Yr 7 % correct
N1	Read, record, compare & order whole numbers	N01				
N2	Use place value knowledge to interpret & model whole numbers & decimal to 2 places	N02	3	L9	82	82
			17*		53	57
N3	Read, name, record & compare simple common fractions	N03	7*		63	72
			11	K1	65	66
			12*		51	66
			13	N14	62	64
N4	Recognise & use equivalences between percentages & fractions	N04	14*		55	63
N5	Continue number patterns involving whole numbers, fractions & decimals to 2 places	N06 N07				
N6	Create & describe number patterns	N06 N07				
N7	Calculate accurately with whole numbers, decimals & unit fractions in mental, written or calculator method	N10	16	P13	61	61
		N12	20	K6	35	46
N8	Use the inverse relationships between addition & subtraction, & multiplication & division	N08				
N9	Represent, interpret & solve problems involving division	N09				
N10	Estimate answers to computations in contexts familiar to students	N11	1	N11	79	88
			4*		67	81
			5*		72	73
			6	V1a	48	73
			8	I7	59	72
			19	U1a	31	48
N11	Solve problems involving simple ratios with whole numbers & money	N13	15*		66	63
			18*		51	51
			22	R14	32	24
N12	Solve one- and two-step problems in contexts familiar to students	N14	2	R12	86	83
			9	I5	68	70
			10*		55	67
			21	R7	43	34
N13	Apply knowledge of numbers & their properties when calculating	N15				

[a]Items in Lindsey et al. (1999) were numbered in order of increasing difficulty for Australian Year 7 students

[b]Items can be found using these codes at <http://nces.ed.gov/timss/timss95/resources.asp>. Asterisked items have not been publicly released.

This process was applied to each benchmark in each strand. For the Number Sense benchmarks the weight of evidence, especially from benchmarks that appear to have some closely matching TIMSS items, was that the draft benchmarks were pitched at a facility level of around 70% or easier for Australian Year 7 students.

For the Measurement and Data Sense benchmarks, the weight of evidence, especially from the items that are close matches, was that the benchmarks were pitched at a facility level of 60% or easier for Australian Year 7 students (and, for some of the benchmarks, much easier).

For the Spatial Sense benchmarks, the weight of evidence, especially from the items that were close matches, was that the benchmarks are pitched at a facility level of 65% or easier for Australian Year 7 students.

Taking all three strands together, the overall picture both across and within the strands is reasonably uniform. The information about matching items in all strands suggests that the benchmarks are pitched at a facility level where two-thirds of Australian Year 7 students would achieve the benchmarks, and in many cases a much higher proportion would do so.

The thirteen released TIMSS items that were matched to the February 1999 draft Year 7 numeracy benchmarks Number Sense strand are shown in Figure 1 in order of difficulty. Similar figures for Spatial Sense and for Measurement and Data Sense are in Lindsey et al. (1999).

A feature to note from Figure 1 is the spread of difficulties for items that have each been matched to the Number Sense benchmarks. It is also interesting to look at the range of difficulties in the cases where more than one item was identified as assessing the same benchmark. In some cases, for example, items 11 and 13 (final benchmark N3), the difficulties are closely comparable, while in others, for example, items 2, 9 and 21 (final benchmark N12) they are widely spread. The four released items that were found to be a close match to final benchmark N10 (items 1, 6, 8 and 19 in Table 1 and Figure 1) had the largest spread of difficulty of 40%; they are shown in Figure 2.

This underlines a problem that will be encountered by those constructing test items for inclusion in statewide assessment programs to assess benchmark level skills: There are usually many test items that can be developed to assess a given skill, especially when there is only a verbal description to rely on, and it is hard to write items to assess skills at a given difficulty level.

Conclusions

TIMSS items that appeared to correspond to (‘match’) the draft benchmarks were identified, performance data on these items were examined, and the appropriateness of the placement of the draft benchmarks at Year 7 was considered in relation to the data. In summary the findings were:

- It was possible to match TIMSS items to the draft benchmarks, and to determine how well they matched.
- A moderately large number of TIMSS items in each strand was found with some degree of match to the benchmarks. Although not all benchmarks were matched, several in each strand had more than one matching item, enabling an estimate to be made of their difficulty level. These difficulty level estimates were reasonably uniform across all three strands and suggest that the benchmarks are pitched at a level where at least two-thirds of Year 7 students would achieve them, and in many cases, four-fifths or more would do so.

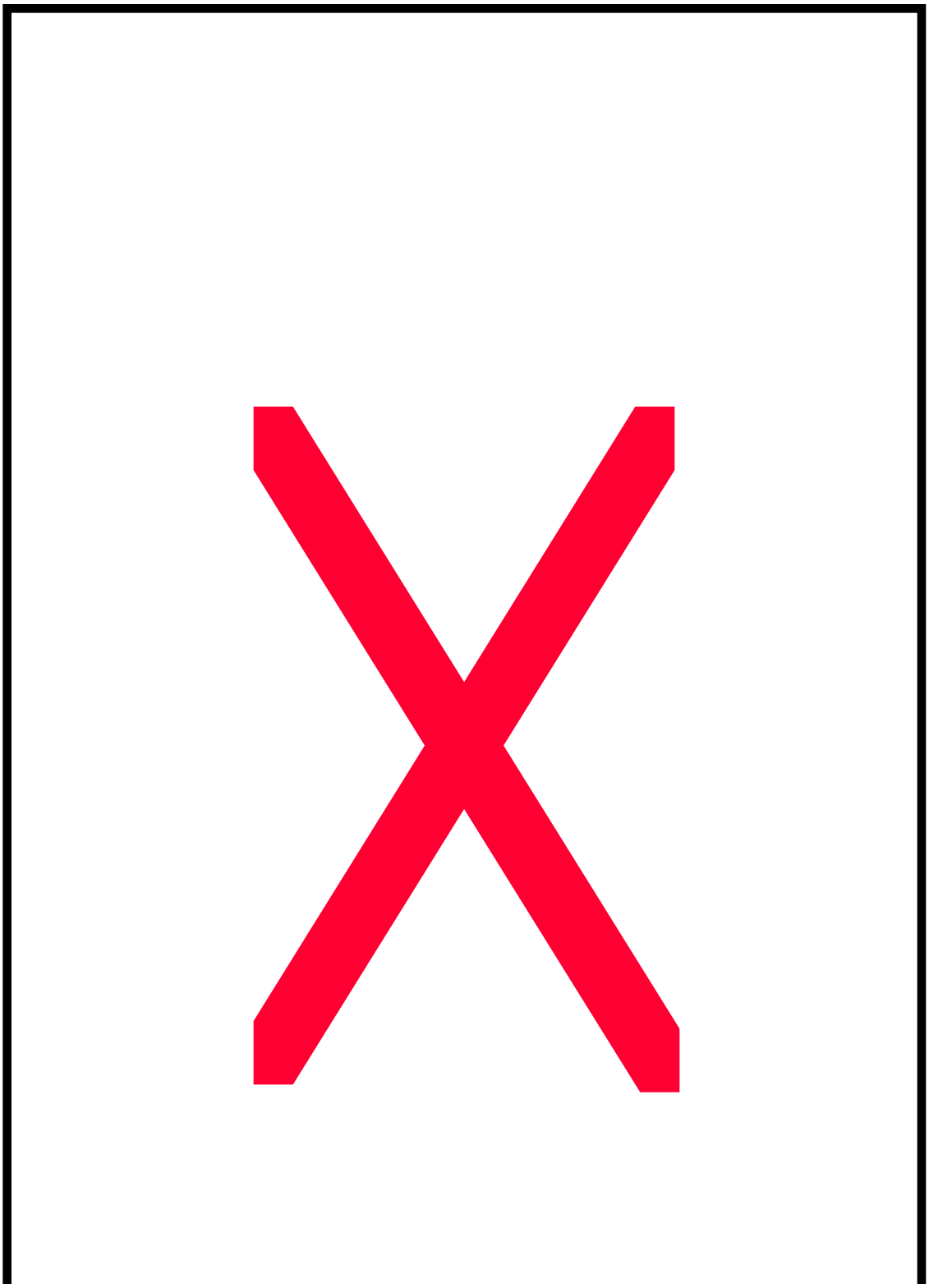


Figure 1. Difficulties of released TIMSS items that match February 1999 draft Number Sense benchmarks.

1. A newspaper reported that about 18 200 trees had been planted in the park. The number was rounded to the nearest hundred.

   Which of these could have been the actual number of trees planted?

   A. 18 043
   B. 18 189
   C. 18 289
   D. 18 328

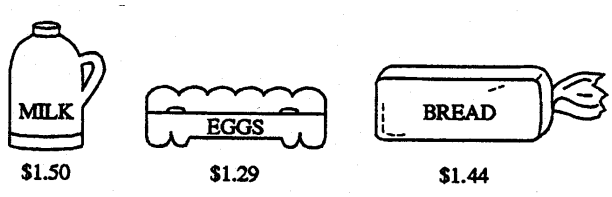
6. Rounded to the nearest 10 kg the weight of a dolphin was reported as 170 kg.

   Write down a weight that might have been the actual weight of the dolphin.

   Answer: _____

8. Prabhu had $5 to buy milk, bread, and eggs.

   When he got to the shop he found that the prices were those shown below:



   At which of these times would it make sense to use estimates rather than exact numbers?

   A. when Prabhu tried to decide whether $5 was enough money
   B. when the shopkeeper entered each amount into the cash register
   C. when Prabhu was told how much he owed
   D. when the shopkeeper counted Prabhu's change

19. Teresa wants to record 5 songs on tape. The length of time each song plays for is shown in the table.

Song	Amount of Time
1	2 minutes 41 seconds
2	3 minutes 10 seconds
3	2 minutes 51 seconds
4	3 minutes
5	3 minutes 32 seconds

   ESTIMATE to the nearest minute the total time taken for all five songs to play and explain how this estimate was made.

   Estimate: _____

Figure 2. Released TIMSS items that match Number Sense final benchmark N10.

The likely difficulty for test developers of constructing assessments for pre-defined standards is highlighted by some of the results obtained in this project, where items apparently assessing the same benchmark were spread along the TIMSS item difficulty

continuum. The same phenomenon was reported by O'Connor, G., Doig, B., Lindsey, J., Pearn, C., & Lokan, J. (1999) for the Year 3 and Year 5 benchmarks.

Acknowledgements

The research reported in this paper was funded by the Commonwealth Department of Education, Training and Youth Affairs.

The contributions of our colleagues Cath Pearn, Jan Lokan, Brian Doig and Gayl O'Connor to the conduct of this research is gratefully acknowledged.

Barry McCrae is a Principal Fellow of the Department of Science and Mathematics Education at The University of Melbourne.

References

Board of Studies. (2000). *Mathematics Curriculum and Standards Framework II*. Melbourne: Author.

Curriculum Corporation. (2000). *Numeracy benchmarks:Years 3, 5 & 7*. Melbourne: Author.

Lindsey, J., Pearn, C., Lokan, J., Doig, B., & O'Connor, G. (1999). *Comparison of Australia's revised draft year 7 numeracy benchmarks and international standards*. Melbourne: Australian Council for Educational Research.

Lokan, J. (1999). Overview of the Third International Mathematics and Science Study (TIMSS) in Australia. In *Raising Australian standards in mathematics and science: Insights from TIMSS* (Proceedings of the ACER National Conference 1997, Melbourne, pp. 7-29). Melbourne: Australian Council for Educational Research.

O'Connor, G., Doig, B., Lindsey, J., Pearn, C., & Lokan, J. (1999). *Comparisons: Australia's revised draft Year 3 and Year 5 numeracy benchmarks and international standards. Final report*. Melbourne: Australian Council for Educational Research.