Free-Response Tasks in Primary Mathematics: A Window on Students' Thinking

Jodie Hunter
Massey University
<J.Hunter1@massey.ac.nz>

Ian Jones
Loughborough University
<I.Jones@lboro.ac.uk>

We administered specially-designed, free-response mathematics tasks to primary students (*N* = 583, ages five to 12 years old). Our focus was on whether (i) the children's responses could be reliably assessed, and (ii) the responses could provide insights into children's mathematical thinking. We used a contemporary comparative judgement technique, interviews with four teachers, and analysed a sample of six responses to make inferences about the students' mathematical thinking. We found that the sampled responses' scores, interviewees' comments and qualitative features of the sampled responses led to consistent insights on the children's mathematical thinking. We argue that free-response tasks should supplement traditional assessments in primary mathematics.

Assessment tasks are a ubiquitous method for understanding students' thinking in primary mathematics. In New Zealand common assessment tools include Progressive Assessment Tests (PATs) which are standardised multiple-choice item tests. However, a common criticism of such tests is the predominance of short, closed items that emphasise the recall and application of isolated facts and procedures (Berube, 2004). Such tests risk promoting a narrow and arguably distorted view of students' mathematical thinking (NCETM, 2009). In comparison, consider tasks designed to assess writing ability. These might include short items such as spelling or grammar, but also free-response items that ask students to produce a piece of descriptive writing. This latter is invaluable to primary teachers as a window onto students' thinking in the context of writing (Brindle, Graham, Harris & Hebert, 2016), yet free-response tasks in mathematics assessment are rare.

We investigated the use of specially-designed, free-response tasks in primary mathematics classrooms. The tasks comprised a short prompt (see Table 1) followed by a blank page for the student's response, analogous to a writing task in language lessons. A traditional barrier to the use of such assessments in mathematics is that they are difficult to score in a meaningful and reliable manner (Laming, 1994). We applied an emerging assessment technique, based on comparative judgement (Pollitt, 2012), that has been reported to overcome this barrier in secondary school and university mathematics (Bisson, Gilmore, Inglis, & Jones, 2016; Jones & Alcock, 2014; Jones & Inglis, 2015). The comparative judgement technique was used to address our first research question: *Can primary children's free-response mathematics work be reliably assessed?*

A reported advantage of free-response mathematics tasks, at the secondary level, is that they can provide a window onto children's mathematical thinking (Jones & Karadeniz, 2016). This hypothesis formed our second research question: *What insights into primary children's mathematical thinking do the assessed responses provide?* We explored this research question using two methods. First, primary teachers involved in assessing the children's responses were interviewed to investigate what features they focussed on when making assessment decisions. Second, we sampled six responses, two low, two medium and two high-scoring, and analysed their features. Our analysis investigated the consistency

2018. In Hunter, J., Perger, P., & Darragh, L. (Eds.). Making waves, opening spaces *(Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia)* pp. 400-407. Auckland: MERGA.

between the scores of the six sampled responses, the teachers' comments during the interview, and qualitative analysis of the responses themselves.

Methodology

Context and Participants

The assessment tasks were administered to students (N = 583, ages five to 12 years old) in a low socio-economic primary school. Following this, twenty teachers participated in a staff meeting where they undertook the process of comparative judgement (described later in the paper). Four teachers agreed to be interviewed after the staff meeting to investigate their perceptions of the assessment process. They included a Year 2–3 teacher, a Year 6 teacher, and two Year 7–8 teachers.

The Tasks and Administration

Fourteen tasks (see Table 1 for examples) were designed by the researchers and selected by members of the school's senior leadership team as relevant to the areas of focus in mathematics lessons during the previous term. Students in Years 0–3 completed three tasks each and students in Years 4–8 completed four tasks each. Over a week, students were provided with 10–15 minutes to complete one task each day individually. Teachers wrote the responses for the children in Year 0–2 when required.

Table 1
Example assessment task prompts and year levels

ID	Task prompt	Years
1	Write and draw everything you know about addition and subtraction.	0 – 4
2	Write one or more tricky word problems for a friend involving multiplication or division. Show how you would solve them.	3 – 8
3	Write and draw everything you know about the operations (addition, subtraction, multiplication, division).	4 – 8
4	This is a graph of favourite fruit of one class of children. [pictogram of fruit shown here] What statements can you make about the children's favourite fruit?	0 – 2
5	These graphs show boys and girls favourite winter sports. What statements can we make about the girls and boys favourite winter sports? [two bar graphs shown]	3 – 4
6	This graph shows how many hours people in two different classes watched television and did their homework over the week. Think about things such as the mean, mode, median and range. [two dot plots shown here] What statements can we make about the two different classes of children and how much time they spend watching television and doing homework?	5 – 8

Assessing the Tasks

Free-response mathematics tasks do not lend themselves to scoring. This is partly because rubrics assume a set of pre-defined response types (Jones & Inglis, 2015), whereas the tasks used here were designed to generate a wide range of responses without anticipating

the responses in advance. Moreover, even if a rubric could be designed, the marking is likely to be unreliable; that is assessors would judge the extent to which a given response matches the rubric inconsistently (Murphy, 1992). Therefore an alternative assessment method was required and we used comparative judgement (Pollitt, 2012).

Comparative judgement requires no rubric or scoring and instead assessors are presented with two responses and are simply asked which student has demonstrated the 'better' mathematical thinking. Many such binary decisions are collated from a pool of assessors, and are used to estimate a parameter for each script using the Bradley-Terry model (Bradley & Terry, 1952). The parameters are then scaled and can be treated as scores for the test responses (Jones & Alcock, 2014). Recent research has demonstrated that using comparative judgement to assess free-response mathematics tasks produces reliable and valid outcomes that are robust across a diversity of learners from school students to undergraduates (e.g. Bisson et al., 2016). A feature of comparative judgement is that assessors can compare responses to the different tasks and still make consistent and valid judgements about which student is 'better'. This feature is important here due to the different tasks administered to students, as exemplified in Table 1.

Due to space constraints we do not detail or make the case for using comparative judgement here. Readers are referred to technical explanations on the use of comparative judgement for assessment (e.g. Bramley, 2007; Pollitt, 2012).

Assessment Outcomes

The 1912 test responses were anonymised, scanned and uploaded to a comparative judgement website (nomoremarking.com) for assessment. Twenty teachers from the school and 10 other individuals (with a teaching background who now work within professional learning and development) were recruited to comparatively judge the responses, and they completed between 72 and 1400 judgements each, resulting in a total of 12888 judgements. For each pairwise judgement the teachers were instructed to select the 'better response'. The binary decision data was statistically modelled (Pollitt, 2012) to produce a parameter estimate of the 'quality' of each response. The parameter estimates were then scaled to produce a set of scores ($\mu \approx 50$, $\sigma \approx 15$).

The reliability of the assessment outcomes were investigated using standard techniques (Bramley, 2007; Pollitt, 2012). First we calculated the Scale Separation Reliability (Bramley, 2007), a measure of the consistency of teachers' judging considered analogous to Cronbach's alpha for traditional scoring procedures, and found that this was satisfactory, $SSR = 0.83$. We then calculated a misfit statistic for each judge and each test response to investigate whether any individual teachers had judged anomalously or any responses been judged inconsistently by different teachers. Following the standard practice of considering any misfit statistic greater than two standard deviations above the mean misfit statistic to be misfitting (Pollitt, 2012), we found that none of teachers, and just 59 test responses (0.3%) were misfitting. Taken together these measures suggest the assessment outcomes were reliable.

Data Collection and Analysis: Interviews

Four self-selected teachers were interviewed individually for between 20–25 minutes. The interviews were audio-recorded, wholly transcribed and analysed using grounded theory (Corbin & Strauss, 2007) to identify themes. We present our findings with respect to how the teachers made their judgements, and how they interpreted student understanding of two areas prominent in the data: numeracy and statistical literacy.

Results

Students' Thinking: Teachers' Observations

Three of the teachers began their interviews by reflecting on the open-endedness of the tasks. They described how this resulted in a wide variety of responses while also allowing different entry points and levels on which the tasks were answered. One teacher noted that the format of the task meant students "can show a lot more of what they know [compared to traditional tasks]".

Analysis suggested there were three key themes in relation to the criteria that teachers used to make pairwise judgement decisions. First, all four interviewed teachers stated that they preferred student responses that were mathematically sound rather than a social response; for example one teacher said "some of the responses were very personal, and so you knew that, that was not a mathematical answer". Second, three of the teachers' expressed a preference for responses that used examples, illustrations, and explanations to provide evidence of student thinking. For example, one teacher said "looking at the evidence, how they prove it". Third, three of the teachers said they preferred student responses that were meaningful in terms of the task prompt. For example, one said "some type of link that linked to the question". Another teacher said that when judging the statistical literacy tasks she was "looking for a statement that did actually relate to the data … rather than just writing out a whole bunch of numbers".

Teacher Judgements of Responses to Number Questions

Our analysis of the interviews revealed that the teachers commented frequently on the arithmetical component of student responses. Three teachers said that responses to the numeracy tasks often demonstrated an understanding of the properties of operations. In particular, three of the teachers noted that students commonly identified the relationship between repeated addition and multiplication and two described how students identified the inverse relationship of addition/subtraction and multiplication/division.

These teachers also observed that students had good recall of basic facts and were confident with addition. However, some students appeared to misinterpret the prompts and *"*just wrote down hundreds of plus problems, but they couldn't actually say, you know, which words show that it's addition". Similarly, two teachers noted that in many responses the students "don't really know how to write a division or a subtraction word problem and solve it that well". The teachers also commented that most students appeared to equate "tricky" with large numbers and then wrote problems that they could not solve.

Teacher Judgements of Responses to Statistical Literacy Questions

Our analysis also revealed that the interviewed teachers commonly referred to responses to statistical literacy tasks. Three of the teachers noted that many students were able to interpret simple graphs such as pictograms and bar graphs and make simple statements from these. They also noted that many students struggled with the more complex graphs and having to reconcile two graphs. For example, one teacher said that "certain types of graphs and data displays were consistently bad". All four teachers said that the students tended to make general observations about the data but in many cases had difficulty constructing specific statements. For example, one teacher observed there was "[only] a handful that actually were able to make a good statement saying that this graph is showing". Three

teachers also noted common written student phrases of the type “oh that one’s the most, that one’s the least on a graph” when attempting to make a statement from the graphs.

Two teachers (both of Year 7–8) noted that some of the students appeared to know what the mean, median, and mode were but did not use these in a meaningful way to convey information about the set of data. For example, one teacher said “it was irrelevant to actually what the question was asking them”. One of these teachers reflected that “they [students] don’t often think that deeply about it or understand what statistics is for”.

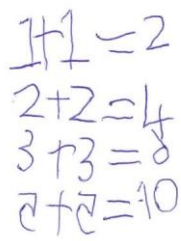
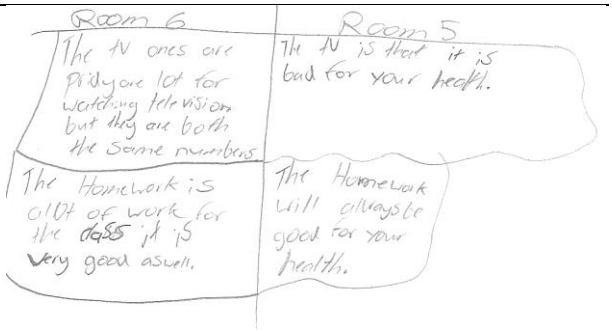
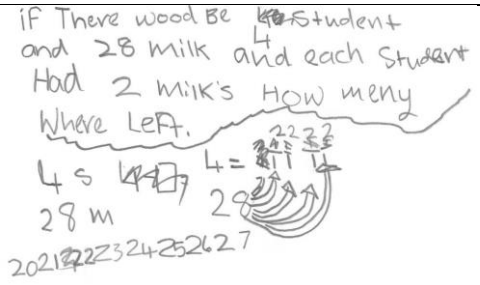
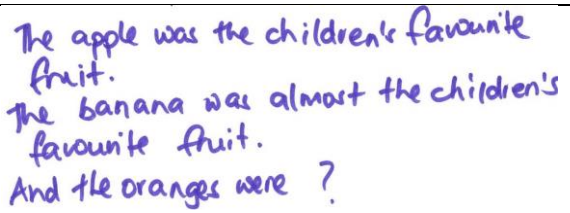
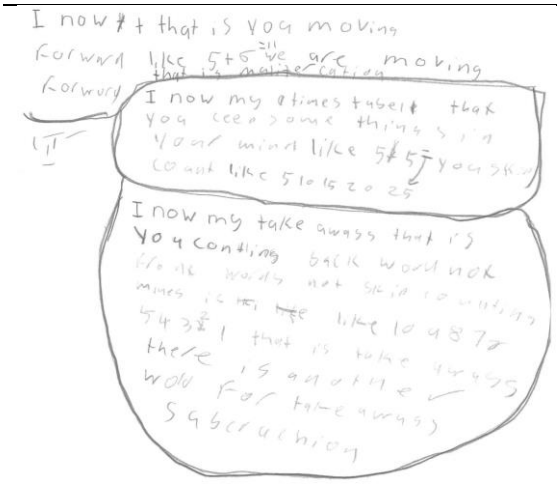
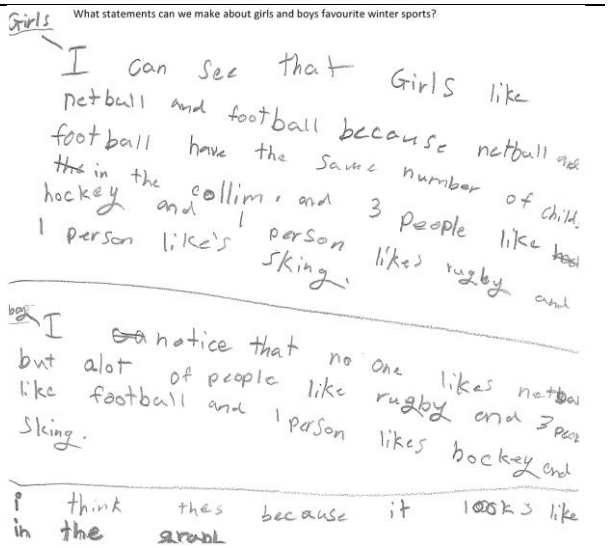
Numeracy tasks	*Statistical literacy tasks*
	
Task 1, 25th percentile, Year 2	Task 6, 25th percentile, Year 6
	
Task 2, 50th percentile, Year 5	Task 4, 50th percentile, Year 2
	
Task 3, 75th percentile, Year 5	Task 5, 75th percentile, Year 4

Figure 1: Sampled responses to the tasks in Table 1.

Students' Thinking: The Tasks

To gain further insights into students' mathematical thinking we sampled six responses. In light of the interview analysis just reported we sampled responses to five tasks from across the age ranges that focussed on numeracy (Tasks 1 to 3 in Table 1) and statistical literacy (Tasks 4 to 6). For each task type (numeracy/statistical literacy) we sampled the three responses closest to the 25th, 50th and 75th percentiles of the assessment scores. The sampled responses are shown in Figure 1.

Numeracy Tasks

Figure 1 displays increasing sophistication from the 25th to the 75th percentile in the responses to the numeracy questions. The response at the 25th percentile shows four addition equations, with no written explanations, drawings or any focus on subtraction as requested in the task prompt (Table 1). In terms of the general themes of the interviewed teachers preferring the use of illustrations and links to the task this response was lacking. The additions are all of the form $x + x = 2x$, consistent with young children's informal reported strategy of doubling (Ter Heege, 1985). The responses at the 50th and 75th percentiles are also partial responses to the task prompt: only one example is given in the response to Task 2, and multiplication and division is omitted from the response to Task 3. However they show increasing sophistication: the Task 2 response implicitly involves multiplication or division, makes use of context (although the use of context was not mentioned by the interviewed teachers), and diagrams; the Task 3 response contains a written metaphor of movement for addition and subtraction, and this metaphor implies the inverse nature of operations which the teachers cited as influencing their judgements. Although the response does not overtly describe multiplication or division, repeated addition and subtraction are in evidence. Therefore the increasing sophistication of the numeracy tasks sampled show good consistency with our analysis of the teacher interviews. However, none of the sampled tasks showed evidence of students using arithmetical examples that were too complicated for them to calculate.

Statistical Literacy Tasks

Similar to the numeracy tasks, Figure 1 suggests increasing sophistication in the student responses to the statistical literacy prompts from the 25^{th} percentile to the 75^{th} percentile. The response at the 25^{th} percentile shows a social response to the prompt, comparable to what Watson and Callingham (2005) classified as an idiosyncratic response in their statistical literacy construct. In this type of response, personal beliefs and experiences dominate, as illustrated by the student's statements "The TV is that it is bad for your health" and "The homework will always be good for your health". The student response aligns with the teachers' observations of student difficulties in interpreting and understanding the more complex graphs. The Task 4 and Task 5 responses both provide statements that are related to the prompt and task. This aligns with the teachers' preference for responses linked to the question. They both show similarities to the category of informal responses on Watson and Callingham's construct with interpretations of basic one-step graphs provided but little justification for these "The apple was the children's favourite fruit". However, while the Task 4 response provided two simple statements, the Task 5 response made a range of statements about the data and included some basic data reading from the graph. The responses also reflected the teachers' preference for evidence that students could interpret simple graphs and make statements about these. However, none of the sampled responses

showed evidence for the teachers' observations that some students misuse and appear to misunderstand averages.

Discussion

Mathematics assessments are often criticised for privileging a narrow, fragmented view of children's mathematical thinking. We addressed this criticism by designing free-response tasks and administering them to primary students of various ages.

To inform the first research question we applied a comparative judgement technique to assess the responses. The outcomes were found to be reliable, and this finding is consistent with the use of free-response tasks with older groups of students (Bisson et al., 2016; Jones & Alcock, 2014; Jones & Inglis, 2015; Jones & Karadeniz, 2016). Moreover, a comparison of the scores and qualitative analysis of six sampled responses, triangulated with the comments of four interviewed teachers, supported the validity of the assessment method. Specifically, the highest scoring responses showed more mathematical sophistication and better reflected the teachers' comments on what they valued when assessing than the medium scoring responses; and, likewise, the medium scoring responses compared to the low scoring responses.

To inform the second research question, we interviewed a sample of four teachers after they had completed their comparative judgements, and undertook a qualitative analysis of six sampled responses. For both types of task – numeracy and statistical literacy – there was consistency between the scores, the qualitative features of the sampled responses, and the interviewed teachers' comments: higher scoring tasks better addressed the task question, were mathematical rather than social, and made use of examples and illustrations. For the numeracy tasks the scores and qualitative features were consistent with teachers' preference for sophisticated use of arithmetic operations, including evidence of knowledge of the reversibility of operations. For the statistical literacy tasks the scores and features were consistent with teachers' preference for meaningful written interpretations of graphical representations that were accurate and mathematical (rather than social).

Taken together, our findings provide support for the use of free-response mathematics tasks with primary students both for summative assessment, if scored using a reliable holistic assessment technique such as comparative judgement, and for providing teachers with a window onto children's mathematical thinking.

Limitations

We are aware of three limitations with the present study. First, no prior mathematics achievement data were available due to ethical constraints on data collection, and so we could not further validate the assessment outcomes. Second, only four teachers were interviewed and six responses sampled, threatening generalisation. Third, while free-response tasks have been reported to be well-suited to assessing conceptual knowledge (Bisson et al., 2016), they are less appropriate to providing insights about procedural knowledge. Therefore free-responses methods should be combined with traditional assessment approaches to provide a fuller picture of children's mathematical thinking (Jones & Inglis, 2015).

Conclusion

Traditional assessments have the advantage that they can be reliably scored, but provide only a narrow window onto children's mathematical thinking. Free-response tasks have the

potential to provide richer insights, but historically have been extremely difficult to score reliably. Using comparative judgement we have provided evidence that this barrier can now be overcome. We recommend teachers and researchers consider using free-response tasks to supplement traditional assessments, thereby providing reliable and valid insights onto the mathematical knowledge of primary students.

References

Berube, C.T. (2004). Are standards preventing good teaching? *Clearing House, 77*, 264–267.

Bisson, M. J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education, 2*, 141–164.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*, 324–345.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 264–294). London: Qualifications and Curriculum Authority.

Brindle, M., Graham, S., Harris, K. R., & Hebert, M. (2016). Third and fourth grade teacher's classroom practices in writing: a national survey. *Reading and Writing, 29*, 929-954.

Corbin, J., & Strauss, A. (2007). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory (3rd ed.). Thousand Oaks, CA: Sage Publications.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*, 1774–1787.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*, 337–355.

Jones, I. & Karadeniz, I. (2016). An alternative approach to assessing achievement. In C. Csíkos, A. Rausch & J. Szitányi (Eds.), *The 40th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 51-58). Szeged, Hungary.

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology, 37*, 152–183.

Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52*, 58–63.

NCETM. (2009). *Mathematics Matters: Final report*. London: National Centre for Excellence in the Teaching of Mathematics.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles Policy and Practice, 19*, 281–300.

Ter Heege, H. (1985). The acquisition of basic multiplication skills. *Educational Studies in Mathematics, 16*, 375-388.

Watson, J. M. & Callingham, R. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education. International Association for Statistical Education (IASE) Roundtable*, Lund, Sweden, 2004 (pp. 116–162). Voorburg, The Netherlands: International Statistical Institute.