An alternative approach to assessing achievement

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Metadata Record: [https://dspace.lboro.ac.uk/2134/21090](https://dspace.lboro.ac.uk/2134/21090)

Version: Accepted for publication

Publisher: IGPME (© the authors)

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: [https://creativecommons.org/licenses/by-nc-nd/4.0/](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Please cite the published version.
AN ALTERNATIVE APPROACH TO ASSESSING ACHIEVEMENT
Ian Jones and Ilyas Karadeniz
Mathematics Education Centre, Loughborough University

Traditional exams typically assess general achievement by testing procedural knowledge across a sample of mathematical domains. Here we explore whether achievement can be assessed by testing conceptual understanding across domains. This follows previous work in which we showed that comparative judgement, based on pairwise expert judgements of students’ work rather than rubrics and scoring, can be used to measure understanding of a specific concept (e.g. fractions). In the present study, school students (N = 197) sat open-ended tests sampling a range of concepts, and their responses were comparatively judged. Analysis supported the validity of the approach for assessing general achievement. We conclude that comparative judgement could help improve the assessment of mathematics.

INTRODUCTION

Summative mathematics tests, such as examinations at the end of secondary schooling, typically comprise short questions sampling across mathematical domains. The question scores are then summed to produce a measure of overall mathematical achievement. However, questions tend to test the recall and application of facts and algorithms, thereby privileging procedural knowledge (e.g. Noyes, Wake, Drake & Murphy, 2011). Few questions test understanding of concepts and their interconnections, and as such conceptual understanding is underrepresented in assessments of overall mathematical achievement (Burkhardt, 2009).

One reason for this is that conceptual understanding is difficult to assess reliably. Scientific instruments, such as the Mathematical Equivalence Assessment (Matthews, Rittle-Johnson, McEldoon & Taylor, 2012), require painstaking work to develop and validate, and such a process is not practical for routine test production. In previous work we proposed a novel and efficient approach to measuring understanding of a given concept (Jones, Inglis, Gilmore & Hodgen, 2013). Students were administered tests that contained a short prompt followed by a blank page for a response (an example prompt and student response from the present study is shown in Figure 1). We observed that students produced a wide variety of response types, making use of symbols, diagrams and natural language to express their understanding.

Such open-ended tests can be assessed validly and reliably using a novel comparative
judgement technique (Jones & Alcock, 2014; Jones et al., 2013). Mathematicians are presented with pairs of student responses and asked to decide which is “better” in terms of a global construct such as “conceptual understanding”. The decision data is fitted to a logistic model (Pollitt, 2012) to produce a parameter estimate of the “quality” of each response as collectively perceived by the judges. The parameter estimates can then be used for routine assessment procedures such as evaluating validity and reliability (Jones et al., 2013), and assigning grades (Jones & Alcock, 2014). Comparative judgement requires no scoring rubrics, and instead validity is grounded in the collective expertise of the judges making direct comparisons of students’ work. The approach is supported by a corpus of psychophysical research demonstrating that human beings are more consistent when judging one object against another than when judging a single object in isolation (see Laming, 1984).

Whereas in previous work we focussed on measuring understanding of single concepts (Jones & Alcock, 2014; Jones et al., 2013), here we explored aggregating parameter estimates across several tests to produce an assessment of overall mathematical achievement. The method described below was designed for research purposes and is not proposed as directly usable for routine summative assessment, nor were the open-ended tests designed as wholesale replacements for procedural questions. Rather, we evaluated whether a comparative judgement approach can yield a meaningful assessment of overall mathematical achievement, and so offer a complement to traditional testing procedures.
METHOD

Materials. Twenty-five open-ended test questions were written by a researcher, covering the topics listed in Figure 2. An example is shown in Figure 1.

Participants and test administration. The tests were administered to students \((N = 197)\) aged 12 to 14 years in a large secondary school with a culturally and socioeconomically diverse intake. Students sat the tests in a single lesson under examination conditions and the supervision of their mathematics teachers. Teachers selected which tests to administer to their classes, and were advised to allow at least ten minutes per test. The outcome was 686 completed tests, and each student completed between 1 and 5 tests \((\text{mode} = 4)\) over the course of 50 minutes. In addition, national test scores for mathematics and reading were obtained for 163 of the participants.

Comparative judgement. The 686 student responses were anonymised, scanned and uploaded to an online comparative judgement engine (www.nomoremarking.com). Eleven mathematics PhD students were recruited, all of whom had undertaken judging work for previous studies, and allocated 600 pairwise judgements each. For each pair they were asked to decide, based on the evidence in front of them, which student was “the better mathematician”. Most pairings presented to the judges were student responses to two different test questions.

The judgement decision data were fitted to the Bradley-Terry model (Pollitt, 2012) to produce a final parameter estimate \((\text{mean} = 0, \text{SD} = 2.1)\) for each test response. The internal consistency of the parameter estimates was calculated using the Scale Separation Reliability (analogous to Crobach’s \(\alpha\)) and found to be acceptably high, \(\text{SSR} = .87\). Inter-rater reliability was estimated using a split-halves technique \((\text{iterations} = 100)\) whereby the judges were randomly split into two groups, the decision data refitted to the Bradley-Terry model, and the resulting two groups of parameter estimates correlated. Inter-rater reliability was found to be acceptably high, \(r = .75\). Note that this is an underestimate because the split-halves technique involves effectively discarding half the data in order to calculate a correlation coefficient.

Response analysis. To help evaluate the validity of the outcomes we investigated how judges made their decisions. In previous work we have done this through post-judging surveys and interviews (Jones & Alcock, 2014; Jones & Inglis, 2015). However, these methods were unable to access possible subconscious processes, and findings could not be linked directly to the actual judgement decisions used to produce student scores. In the present study we sought to overcome these limitations by classifying the student responses using an adapted coding scheme (Hunsader et al., 2014). For brevity the coding scheme is only summarised here and readers are directed to Hunsader et al. (2014) for further details.
Coding scheme.

Real world. Response uses a context outside of mathematics (no coded 0, yes coded 1). The mean score for real world across all responses was 0.30.

Connections. Response introduces a relevant concept not explicitly prompted in the test question (no coded 0, yes coded 1). For example, if a student connected percentages and decimals in the question about percentages, this would be coded 1. The mean score for connections was 0.07.

Graphics (graphs, pictures, tables). No use of graphics (coded 0). Use of graphics that are superfluous to the mathematics (coded 1), explicitly illustrate the mathematics (coded 2), are required to interpret the mathematics (coded 3). The mean score for graphics was 0.71.

Numbers. Numbers may be present but not as part of an expression of equation (coded 0). Numbers used in an expression (e.g. $3 + 2$, coded 1), or an equation (e.g. $3 + 2 = 5$, coded 2). The mean score for numbers was 0.69.

Letters. Letters may be present but not as part of an expression or equation (coded 0). Letters used in an expression (e.g. $x + 1$, coded 1), or an equation (e.g. $y = x + 1$, coded 2). The mean score for letters was 0.20.

For illustration, the response shown in Figure 1 was coded “1” for the categories real world and connections because there is a non-mathematical context and interconnections between the concepts equation and ratio respectively. It was coded “0” for the category graphics because there are no graphs, tables or pictures. It was coded “2” for the categories numbers and letters because the response contains equations such as $5q = 10$. The codes were intended to be hierarchical, such that a score of (say) 2 for letters reflects a “more mathematically sophisticated” response than a score of 0 or 1. This enabled a total score (min. = 0, max. = 9) to be calculated for each response by summing across the codes (for example, the response in Figure 1 scored 5). The Spearman rank-order correlation between comparative judgement parameter estimates and coding scheme total scores was moderate, $\rho = .34$, $p < .001$. This provides support for a meaningful relationship between the two measures.

ANALYSIS AND DISCUSSION

The outcomes were evaluated in terms of criterion validity, the basis of judges’ decisions, and the performance of individual test questions.

Criterion validity. We explored the extent to which the parameter estimates reflected mathematical achievement rather than a mathematics-irrelevant construct such as written communication skills. To do this an achievement score was assigned to each student by calculating the mean parameter estimate across tests. Students who
completed fewer than three tests ($N = 23$) were removed from this analysis because our interest was in whether sampling across topics can be used to assess general achievement. Mathematics and reading scores based on national tests at the end of primary schooling were available for 148 of the students who sat three or more tests. We hypothesised that mathematics scores, but not reading scores, would be a significant predictor of mean parameter estimates. Multiple linear regression explained 22% of the variance, $R^2 = .22$, $F(2, 145) = 20.04$, $p < .001$. Mathematics scores significantly predicted parameter estimates, $\beta = .40$, $t(145) = 4.94$, $p < .001$, but reading scores were not a significant predictor, $\beta = .10$, $t(145) = 1.02$, $p = .310$. This result lends support to the criterion validity of the assessment.

However, most of the variance in the data (78%) was not explained by this analysis. Three limitations of the available achievement data may have contributed to this. First, the national tests were taken two or three years prior to the open-ended tests. We repeated the multiple linear regression for the younger ($N = 61$) and older children separately, and found these analyses explained 34% and 13% of the variance respectively. This suggests that, unsurprisingly, the national test data becomes less informative as children progress through secondary school. Second, the reading scores acted as a proxy for written communication skills and their suitability for this purpose could not be verified. Third, the mathematics scores were based on largely procedural tests whereas the parameter estimates were based on conceptual tests, and so the two sets of scores were intended to measure different, albeit related, constructs.

**Response analysis.** A multiple linear regression analysis was conducted to investigate the extent to which the classification of responses (coding scheme) predicted parameter estimates (comparative judgement). We included two further predictors that were evident to the judges when making their decisions. The first was test question, which can be expected to impact on assessors’ perceptions of the mathematical quality of a response (Good & Cresswell, 1988). The second was file size (of the scanned responses), which acted as a rough proxy for the quantity written for each test response. The analysis explained 35% of the variance, $R^2 = .35$, $F(7, 678) = 52.39$, $p < .001$, and the results are summarised in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>File size</td>
<td>0.50***</td>
<td>[0.01, 0.02]</td>
</tr>
<tr>
<td>Numbers</td>
<td>0.15***</td>
<td>[0.18, 0.50]</td>
</tr>
<tr>
<td>Graphics</td>
<td>0.12***</td>
<td>[0.12, 0.41]</td>
</tr>
<tr>
<td>Connections</td>
<td>0.11</td>
<td>[-0.04, 0.99]</td>
</tr>
<tr>
<td>Letters</td>
<td>0.05</td>
<td>[-0.04, 0.43]</td>
</tr>
<tr>
<td>Test question</td>
<td>-0.04</td>
<td>[-0.02, 0.00]</td>
</tr>
<tr>
<td>Real world</td>
<td>0.03</td>
<td>[-0.17, 0.45]</td>
</tr>
</tbody>
</table>

Table 1: Multiple linear regression results for features of student responses as predictors of parameter estimates. ***$p < .001$. 
Jones, Karadeniz

File size was the strongest predictor. This is unsurprising given that the more that is written the more mathematics is visible to judges to influence their decisions. Student responses with the smallest file sizes typically contained little or no mathematics, such as “I don’t know”. However caution must be exercised. File size is only a proxy for “quantity communicated”. Moreover, variance in the length and style of question prompts will have contributed to variance in file size.

Connections was not a significant predictor, (although it was marginal at $p = .70$). This has relevance for the type of conceptual understanding being assessed using these tests, which mainly focus on individual concepts rather than interconnections between concepts. The test questions used here perhaps should be enhanced by the use of unstructured or semi-structured problem-solving tasks that require students to draw on multiple mathematical domains (see Jones & Inglis, 2015). Another option is to develop more open-ended test questions similar to those in Figure 1, but which require the explicit connection of two or more concepts. In the present study, seven of the 25 questions were of the type “What are the differences between multiplication and division? Give as many examples as you can as to how and why they are different.” (Note that responses to these tests were only coded “1” for connections if students introduced a concept not prompted by the question).

As expected, the use of numbers was a significant predictor, as was the use of graphics. However, the use of letters was not a significant predictor, which is perhaps surprising given the high status of symbolic algebra in school mathematics. This may have been due to the relatively rare use of letters in the responses. Two test questions focussed on equations (as in Figure 1), and we might expect letters to be more prominent in the corresponding responses than for other questions. Indeed, the mean score for letters for the two equations tests ($N = 73$) was 1.00, and the mean for the non-equations tests ($N = 613$) was 0.11. The Spearman rank-order correlations between parameter estimates and letter scores for the responses to equations and non-equations tests were significantly different, $\rho = .34$ and $\rho = .08$ respectively, $z = 2.17$, $p = .03$. Therefore we expect the use of letters to be more influential on judges’ decisions for test questions that explicitly focus on symbolic algebra.

Test question was not a significant predictor, suggesting a given student’s parameter estimate was not strongly dependent on which questions he or she sat. Figure 2 shows parameter estimates by question and reveals some variation. A one-way ANOVA revealed a significant difference across mean question parameter estimates, $F(24, 661) = 5.07, p < .001$, $\eta^2 = 0.155$, and post-hoc tests (Bonferroni corrected) suggested this difference was mainly driven by four questions close to the extremes of Figure 2 (namely “translate/rotate”, “equation 1”, “Pythagoras” and “enlarge”). This reinforces the importance of aggregating across several questions when estimating overall achievement. While “extreme” questions should be avoided, some variance in question performance is to be expected, and can be monitored.
Real-world applications have an important role in many mathematics curricula around the world, but we found that student use of real-world context was not a significant predictor of parameter estimates. This might be due to only a few of the test questions explicitly requesting real-world examples (three out of 25), or the research mathematicians who undertook the judging valuing pure over applied mathematics. These two hypotheses will be investigated in further work.

We were unable to evaluate the possible influence of other features, including mathematically-irrelevant constructs such as neatness, or sought-after learning outcomes such as mathematical creativity. Such features will be investigated in future studies using adaptations of the methods reported here, along with standardised instruments from the literature (for example the Creative Behaviours in Mathematics Questionnaire, see Leu & Chiu, 2015).

**CONCLUSION**

We investigated whether students’ general mathematical achievement can be assessed using open-ended conceptual test questions and a comparative judgement technique. Analysis supported the validity of the approach in terms of students’ national test scores and researcher classification of the test responses. Further work is required to better understand the features of responses that are most valued by expert judges. This will enable the design of test questions and judging procedures that maximise the validity of the approach, and therefore the confidence of stake-holders that outcomes are legitimate measures of overall mathematical achievement.

Assessment processes should match the objectives of curricula. There is a current focus in many countries on improving students’ conceptual understanding of
Jones, Karadeniz

mathematics, and as such assessments should capture conceptual understanding. An approach based on open-ended test questions evaluated using comparative judgment has been described here, and our findings offer promise for complementing and enhancing the common practice of aggregating students’ scores on procedural questions. This could lead to richer, more valid examination practices for which outcomes are based on both procedural knowledge and conceptual understanding. This in turn might lead to mathematics teaching and learning that better reflects what is valued by teachers, policy-makers and other stake-holders of schooling systems.

References


