Exploring diagnostic error analysis methodologies in the context of e-assessment in primary-level mathematics.

Damian Murchan[1], Elizabeth Oldham[2]

[1]School of Education, Trinity College, Dublin
[2]School of Mathematics, Trinity College, Dublin

Paper Presented at the
Annual Meeting of the Association for Educational Assessment - Europe,
Belfast, November 2011.

Direct correspondence to Damian Murchan, School of Education, 3091 Arts Building, Trinity College, Dublin 2, Ireland. Email: damian.murchan@tcd.ie

Introduction

This paper explores the potential of a Computer-Based Assessment (CBA) to provide useful diagnostic information in relation to children's performance on topics contained in the Irish primary mathematics curriculum (Government of Ireland, 1999). The study is set in the dual contexts of persistent concerns about mathematics achievement in Ireland and the particular potential of technology to support teaching, assessment and learning and formative assessment especially (Kingston & Nash, 2011). Concerns about student's mathematical attainment in Ireland reflect similar concerns internationally. Reports and commentary highlight weak performance in State Certificate examinations at both Junior Cycle (end of compulsory education) and Senior Cycle (end of upper secondary). Weak performance in some of the mathematical areas measured as part of international assessments lends further evidence of difficulty (Eivers and Clerkin, 2012; Perkins, Shiel, Merriman, Cosgrove and Moran, 2013). These and other concerns about mathematics performance levels have led to a range of policy responses.

Revised curricula and associated examinations have been introduced in post-primary schools. Known as Project Maths, these curricula emphasise broad conceptual understanding coupled with procedural fluency to enable students confidently and competently investigate real-world problems and contexts, while encouraging students to engage enthusiastically with mathematics (Jeffes, Jones, Cunningham et al, 2013). The curricular change has been accompanied with extensive professional development for teachers and concomitant development of instructional supports and material by the State and by private publishing organisations.

Recognising the importance of antecedent teaching and learning at primary level, a "bridging framework" has been developed to map the content of upper primary education onto the new mathematics curriculum at post-primary level (Cosgrove, Perkins, Shiel, Fish & McGuinness, 2012). The challenges evident at the post-primary level are reflected also in primary education. Whereas a revised curriculum was published in 1999, national monitoring suggests that students experience particular difficulty in relation to measures, general application of mathematical concepts and problem-solving in comparison with other areas of the curriculum (Eivers, Close, Shiel, Millar, Clerkin, Gilleece & Kiniry, 2010). A similar survey identified a zero average gain in mathematics over the period 1999-2004, with reports of up to one-fifth of pupils achieving below the standard expected at 4th Class (Surgenor, Shiel, Close and Millar, 2006). These and other data suggest the need to continually monitor progress of primary students' mathematical achievement and to identify what underpins it in terms of conceptual understanding of mathematics.

Purpose of paper

The study reported here reflects recommendations by Cosgrove et al. (2012) for schools to promote assessment for learning, strategies to foster problem-solving and greater use of aggregated data to identify strengths and weaknesses in mathematics. The paper explores one method of addressing challenges in mathematics achievement at primary level using the affordances associated with a computer-based assessment (CBA) of mathematics achievement. Specifically, the paper aims to:

1. Investigate the challenges and possible solutions associated with automated scoring of student responses, especially in the context of open-format items.

2. Illustrate the essential elements of the CBA, especially in relation to providing more granulated diagnostic information about children's performance on primary mathematics content and skills.
3. Appraise the efficacy of analysing examinee responses in relation to common errors made by students and of providing resultant test-level feedback to teachers.

Programme overview

The assessment was designed to facilitate ease of administration and scoring, while providing relevant information to teachers about the performance of students on different skill areas in mathematics and the types of errors being commonly made by students. Students enrolled in 3rd Class (grade) in participating schools accessed testlets online while in school, assessing their achievement in three content areas, Fractions, Money, Lines and Angles. Software captured student responses that were scored and analysed by the programme, with subsequent return of relevant information to teachers.

Student responses were analysed using traditional test theory resulting in the reporting of a range of descriptive scores based on individual items, topics, total test and mathematical skill, at the individual student and class levels. Given the restrictions associated with dichotomous (1,0) scoring of many CBAs, a broader three-category polytomous scoring model was employed. This facilitated the additional award of partial credit (0.5) for responses that display considerable understanding of the concept but where a small slip or error is judged to have occurred.

In addition to the generation of more traditional scores, analysis of student responses provided information about the type of common mathematical error evident in student responses. Identifying student errors in performance, is one of the approaches to diagnostic assessment identified by Bejar (1984) and Nitko & Brookhart (2007). The assessment reported here was informed by long-established research into the identification and classification of errors in mathematics as one way to facilitate better understanding of student conceptions about mathematics and therefore effect improvement (e.g. Brueckner & Elwell, 1932; Grossnickle, 1935; Lankford, 1974; Newman, 1977; Clements, 1980). Other research notes that errors in children's mathematics performance often result from the application of what are, to children, sensible conceptions or rules about mathematics rather than idiosyncratic random mistakes (Ginsberg, 1977; Confrey, 1990; Smith, diSessa & Roschelle, 1993), thus positioning the present research within a more constructivist understanding of error analysis.

Ten categories of errors were initially developed for the assessment by Burke (2011), drawing in part on diagnostic interview approaches championed by Lankford (1974), Newman (1977), Casey (1978), Clements (1980) and White (2005). Further refinement resulted in the specification of eight errors as outlined in Table 1. Categories in the model are not meant to be hierarchical, but reflect different classes of errors identifiable in students' responses.

SEE TABLE 1

The scoring and categorisation of student responses involved an initial blend of manual and automated manipulation initially, leading to final automated processing.

Given a desired emphasis on including both open-response and closed response items in the assessment, the potential number and variety of individual student responses is considerable. This presented a challenge to designing scoring algorithms that could mimic human raters who can easily discount misspellings and other structural elements of a response irrelevant to the mathematical concepts and skills being tested.

Of interest also was the extent to which particular errors load on items designed to assess three different process areas in mathematics: Understanding and recalling; Using procedures; Reasoning and problem solving. Various taxonomies exist for categorising process dimensions of mathematics learning. These include, for example, six categories in the Irish Primary Curriculum, five in the NCTM 2000 standards, five in the National Research Council framework, three in TIMSS 2011 and three in PISA 2012. Where differences occur, it is typically in the degree of specificity across a range of similar processes.

Methods

The paper describes a small-scale study illustrating the essential elements of a CBA, especially in relation to providing more granulated diagnostic information about students' performance on primary mathematics. We report on the extent to which it was possible to automate the scoring of a mathematics CBA comprising a range of item formats and appraise the efficacy of analysing examinee responses in relation to common errors made by students in the context of different process dimensions of mathematics.

The assessment was administered to 53 students located in three different 3rd classes in three participating schools during the 2010-11 school year. Items aligned with the mathematical content and skills contained in the national mathematics curriculum in Ireland were included, covering three topic areas: Money, Lines and Angles and Fractions. Detail about the specific objectives assessed in the CBA is provided in Table 2.

SEE TABLE 2

Student achievement was assessed using twenty three items, five measuring understanding and recalling, eight measuring use of procedures and ten measuring reasoning and problem solving. A range of item types were used. Scoring software provided a range of results in relation to items, topics, skill areas and total test. Conventional scoring facilitated the provision of traditional forms of feedback whereas error analysis provided teachers with details about the most common sources of errors detected in student responses.

Descriptive commentary and analysis is provided in the next section charting the challenges and solutions associated with capturing data from free-response items and embedding error analysis techniques in automated scoring at the item and overall test level. In addition, analysis of the frequencies of errors allocated to different process dimensions highlights the link between item conceptualisation and student responses.

Development and results

Capturing and scoring free responses

Seventeen of the twenty three items on the assessment required students to provide answers based on supply-type formats. The remainder of the test consisted of multiple-choice, figural drawing and multiple-answer formats in keeping with Scalise and Gifford (2006). Figure 1 presents three items illustrative of the free response tasks required of students in the CBA. Student responses were scored automatically and the discussion below highlights the robustness of the scoring to non-standard responses by students.

SEE FIGURE 1

Item A illustrates a procedures item that emphasised the renaming of cent as euro and recording using the decimal point. The task for students included the addition of cents, conversion of the answer to euro and the insertion of this answer, including the decimal point in the answer box. Students were encouraged to complete the calculation on paper first and then transfer their answer to the computer via the keyboard. Given polytomous scoring of each item on a scale 0, 0.5, 1, the *difficulty* of each item *i* (p_i) can be represented as the average score on the item across examinees, though this technical definition of item difficulty seems to contradict a layperson's use of the term difficulty (Crocker and Algina, 1986). As such, *p-values* are frequently used to summarise the extent to which items are successfully answered by students. The p-value associated with Item A for the sample of students was .71. Given that the focus in the objective underpinning the item was on renaming as opposed to calculating, some leeway was granted to students in relation to the calculation. Therefore, students who recorded answers such as 3.98c, 3.93c, 3.95 and 3.90 were automatically awarded partial credit by the scoring system. In addition, an error by one student associated with inadvertently hitting the space bar on the keyboard resulted in an answer of 3 .98 (space between the 3 and decimal point) as opposed to 3.98. This error, unrelated to the student's mathematical knowledge of place value in converting currency units, was accommodated within the scoring and full marks for the item were awarded.

Item B explores students' capacity to calculate a fraction of a set using concrete materials. Using the proxy materials of clicking on boxes, students were expected to click on any 10 boxes, though the majority clicked on a uniform set of contiguous boxes. The low p-value of .30 highlights confusion amongst students with many clicking on 5 boxes, ignoring the calculation required in relation to 5/8 of 16, with an associated error category of 4 (method execution, see Table 1) However, the error could be as a result of failure to understand the concept. Partial credit was automatically awarded to a student who clicked on nine boxes, assuming a correct calculation of ten but error in clicking/counting while one student opted for a different pattern of ten clicked boxes and was awarded full marks for the item.

The final illustrative item (C with p-value of .89) required students to respond by typing in a word. The obvious limitation with this approach is the possible confounding influence of misspellings on answers. To counter this, broad spelling filters were applied to ensure that answers that reasonably approximated the correct response of *horizontal* were not penalised. Accordingly, the following answers, provided by children, represent some of the incorrectly spelled responses recorded as mathematically correct by the scoring algorithm: *horizantal, horizontil, horizontel,horazotal, horizontal, hoaazal, horzantal, horistel, horazantl, howerzuntl, HOIZONTIL, HORISANTIL*. This illustrates the potential for programming the software to infer mathematical intention from free responses provided by students and to embed this in the scoring.

Given the capacity of the CBA to efficiently capture student responses in a way that addresses some of the negative affordances affecting free-response and CBA free response in particular, we look now at the use of these data to profile the errors being made by students.

Coding errors at the item level

Drawing inferences about patterns of student errors from total test scores requires careful prior embedding of error analysis techniques at the item level. This was achieved in the study by allocating student responses at the item level into one of a small number of item-specific *Response Codes* that highlighted several predicted errors. Initial identification of item response codes was informed by relevant literature (e.g. Cockburn and Littler, 2008; Haylock, 2006; Hansen et al., 2005; Lesh and Zawojewski, 2007; Orton and Frobisher, 1996; Verschaffel et al., 2007). Figure 2 and Table 3 illustrate an item on the topic of money along with the associated scoring rubric.

SEE FIGURE 2

The mathematical task for students involved subtraction. Roughwork paper was available to students, if needed, and they were required to enter the answer in the box on screen using the keyboard. Given the open nature of the item, a range of responses was predicted, as illustrated in Table 3. An asterisk is used to denote a small range of digits. For example €*.80 could be 1.80, 2.80, 4.80 etc, signifying an error in calculation.

SEE TABLE 3

Student responses to the item were expected to fall into one of seven categories and Figure 3 illustrates some student responses captured on the roughwork sheets.

SEE FIGURE 3

The response of €3.90 (example A in Figure 3) reflects correct understanding that the solution to the problem is to be found by subtraction, but possible carelessness in calculation or in transferring the answer to the computer. Similarly, although the student response 380 (example B) is technically incorrect given the directions implied by the answer box, the response does show understanding of the basic mathematical task implicit in the item. Both responses, therefore, suggest possible carelessness (slip) as the cause of the error, with the consequent award of partial credit (0.5) and allocation to item response codes 2 and 3 respectively.

A student response of 6.60 (example C) or even 6.70 indicates inability to convert the word problem into the appropriate mathematical task, where the student probably added instead of subtracted and is thus allocated to item response code 4 with a score of 0. Difficulties with renaming in subtraction are suggested by responses of 4.60, 4.20 (example D), while other errors such as answers in the range 300 to 500 or 3.00 to 5.00 suggest that the student knows the transformation of the word problem into a subtraction algorithm but makes errors of procedure in the subtraction.

Linking item response codes to overall Error Categories

As outlined earlier, the *response codes* at the item level, illustrated in the section above, are linked to a set of broader, over-arching *error categories* applicable to the mathematics curriculum as a whole. The filtering of student responses initially into different predetermined categories at the item level facilitated a second-order analysis that related individual item response codes to the broader set of eight error categories. This provides an overall profile of responses distinctly different from summed results or averages based on more traditional scoring. Table 4 presents a summary of the allocation of student incorrect item responses to the broad error categories. Data are presented in relation to the total test (23 items) and in relation to three subsets of the tests, namely, the skill areas of *knowledge and understanding*, *using procedures* and *reasoning and problem-solving.*

SEE FIGURE 4

In completing the assessments, students could return to items and change answers if they wished before submitting the test. The analysis here focuses on the final response made by students to an item. Accordingly, the software captured 1081 responses from students in 3rd Class. Of these 1081 responses, correct responses accounted for 619 (61%), leaving 462 incorrect responses, distributed across error categories as shown in percentage form in Table 4. Of the erroneous responses overall on the test, 40% were classed as relating to concept understanding. Of the erroneous responses to the five items specifically measuring understanding and recalling, 70% were classed as relating to concept understanding. Similarly, 57% of the erroneous responses to the eight items measuring Using procedures related to students' concept understanding.

The data highlight a loading of errors overall on conceptual understanding of the mathematics included in the tests. Four out of every ten errors recorded were linked to errors categorised as lack of basic understanding of concepts underpinning the items and the test as a whole. The second most frequent error (13%) observed on the test related to students' difficulties in executing the procedures, such as calculations, required to solve the tasks on the test. Students inability to choose the correct approach to follow in solving tasks represented one in every ten errors recorded overall, approximately the same number as careless mistakes made by students (9%). In 8% of cases, students seemed to look at the items but opted not to answer (Viewed but unattempted) whereas one in twenty errors was due to students not having reached the item. Eight percent of the responses fell outside the parameters of the automated analytic rubrics being employed and were, therefore, categorised as *No conclusion.*

Within the three mathematical skill levels, seven out of every ten errors made on items that were designed to measure basic mathematical understanding were, in fact, assigned in the marking to the conceptual understanding category, thus confirming an expected consistency in the coding system. For items measuring understanding and recalling, 14% of the occasions where the correct solution was not provided was due to the student not having recorded an answer despite viewing the item. In 16% of occasions of incorrect solution, it was because the student did not reach the item. The pattern of errors on items measuring students' use of mathematical procedures is more varied. There were eight items designed to assess students' facility with mathematical procedures and the average percent correct on those items was 54%. Where the correct solution was not provided by students, just fewer than six out of ten cases related to lack of conceptual understanding, 25% related to observable difficulties in executing the procedures, 13% due to carelessness and 5% due to the items being viewed but not completed. The last procedural item on the test was located mid way through the assessment, so it is not surprising that all students reached these items, as indicated in Table 4. Overall, the pattern for procedural items suggests that the diagnostic model is better oriented towards diagnosing difficulties with concept understanding than procedure. Possible explanations include lower levels of complexity in the items measuring conceptual understanding or fewer opportunities for diverse slips in those items. Further refinement of the model is warranted in this area.

A broad pattern emerges also in relation to items measuring reasoning and problem solving. One in five of the errors recorded related to students' inability to convert the mathematical prompt into the appropriate mathematical structure required to solve the task (category 2: strategy selection). One in five also related to concept understanding (category 1), with a total of 17% due to the related areas of setup of the mathematical procedure/calculation and execution of that procedure (categories 3 and 4). Seventeen percent of responses recorded by students were outside the automated parameters of the scoring algorithms and were, therefore, categorised as *No conclusion.* Six percent of items were not reached, perhaps not surprising as four of the reasoning items were amongst the final items on the assessment.

Overall, the diagnostic data lend some support to the potential to accurately capture and allocate errors through automated processing. There is evidence that the errors expected in items designed to assess different mathematical knowledge and processes occurred and were correctly categorised, though there is scope and challenge to further refine this process.

Discussion and conclusions

Development of the assessment reported here reflects advantages attributed to CBAs in terms of convenience, accelerated and objective scoring, timely feedback and potential positive influences on student motivation (Johnson and Green, 2006; Van der Kleij et al., 2011). Its design sought to maximise the affordances associated with CBAs (Bearne and Kress, 2001), for example through attractive, user-friendly interface, opportunity to skip on in the test and opportunity to review items retrospectively before submitting the final test, in keeping with Pommerich and Burden (2000), Wang et al (2007), Russell et al. (2003) and others. Heeding cautionary research by Choi and Tinkler (2002) and Kingston (2009), students were provided with the opportunity to complete roughwork on paper also. Test development and administration logistics were also informed by research on CBA item design (McKee and Levinson, 1990; Scalise and Gifford, 2006; Doukas and Andreatos, 2007), automated scoring (Bennett and Bejar, 1998) and the somewhat more intractable issues of unreliability of the hardware, software and internet connections (Higgins, Russell and Hoffman 2005; Kingston, 2009).

The findings of this study inform the literature in relation to CBAs in education, offering an expanded perspective on application and design of tests, scoring and interpretation. As a result of increased policy emphasis on levels of mathematical achievement and mathematical literacy worldwide there is greater emphasis on enhancing student learning in mathematics (Department of Education and Skills, 2011; Martio, 2009; McDonagh & Quinlan, 2012; Mullis et al, 2012; OECD, 2010). Raising achievement meaningfully requires improvements across the mathematical achievement range to narrow the gap between high and low performers, as evident in some countries. Accordingly, initiatives to understand better and address student difficulties in mathematics are warranted. The present study is an effort to harness emerging digital assessment techniques alongside more traditional error analysis methods to shed new light on student achievement and thereby facilitate improvement.

Results of the research confirm the viability of incorporating open, polytomously scored items in a CBA in keeping with the emphasis on complex scoring and validity espoused by Williamson, Bejar and Mislevy (2006), Feng et al (2009) and Eggen and Lampe (2011). Students had little difficulty in combining the use of roughwork paper with answering items on computer, supporting the case for using more authentic items that reflect the typical mathematical tasks encountered by students in school Once captured by the software, complex data were available for analysis and the scoring algorithms proved robust to variation in the ways in which responses were initially entered by students. It was possible to process, score, analyse and categorise student responses automatically and draw inferences in relation to many, but not all of those responses. Student responses were found to group in predictable clusters and even though some differences existed in answers, it was possible to allocate most answers to a relatively small number of response codes per item.

One of the main challenges found was in unambiguously converting the individual item response codes to the more generic error categories used to combine data and derive inferences across items and topics. In many cases, this conversion was unproblematic. However, the errors overall loaded heavily on the concept understanding category, representing four out of every ten instances. Whereas this might be expected to some lesser extent from the research of Newman (1977), Clements (1980) and Watson (1980), better explication of more specific dimensions within concept understanding might help reveal more nuanced interpretations and prove more useful to teachers. Having said that, it may be helpful for teachers to know that the majority of errors being made by a child or group are traceable to fundamental problems with their conceptual understanding rather than other features of mathematical knowledge and process. The high loading of errors on concept understanding for those items specifically designed to test understanding lends support to the design and approach undertaken with the CBA. The more mixed pattern of error loadings for items designed to test procedural knowledge invites further research given the greater loading expected on the method execution category. A more even distribution of errors across the reasoning/problem-solving items perhaps reflects the greater complexity of the tasks for students, who come to such items with a varying range of knowledge, skills and conceptions about mathematics. For some students who do not solve such items, the process highlights a host of challenges for students from lack of original understanding of the topic to many possible incorrect and obscure interpretations of the task.

As a marrying of traditional error analysis in mathematics with novel digital assessment technology in schools, the findings offer support for further research in the area. The specific nature of the items and the extent to which they unidimensionally focus on any one mathematical skill area is worth exploring, though any absolute designation of an item as eliciting one and only one skill requirement from a student is open to question. Different students attempting the same item may not bring the same skill to bear on the problem. For example, a problem-solving item may be novel to one student and require application of complex skills, whereas for another who has seen or tried a similar problem previously, the skill may reflect more the student's memory of the correct operations to apply rather than genuine on-the-spot reasoning.

Overall, the study and results are important in addressing concerns about persistent underperformance in mathematics by exploring novel approaches. Teachers should be ideally placed to engage in diagnosis and prognosis in relation to student difficulties in mathematics. However, a combination of teachers' own possibly limited subject matter knowledge and pedagogical content knowledge (Surgenor, Shiel, Close

& Millar, 2006, Delaney, 2010) and available time can attenuate such potential. Carefully constructed CBAs designed to supplement teachers' judgements can assist teachers in this important diagnostic and formative task. One of the driving forces behind CBAs is the promise of more efficient assessment, recording and reporting (Cook and Jenkins, 2010), thus leaving more time for teachers and students to act on the basis of the feedback provided. This suggests a challenging and positive role for test developers in broadening their approach to designing, scoring and reporting on mathematics performance.

This small-scale study offers a vision for an alternative way to gather and analyse student responses to mathematics tasks. A host of researchable issues remain in realising this vision: designing items to best measure different skill areas in mathematics; identifying and coding different erroneous responses at the item level; mapping item errors onto more holistic error categories and maximising internal coherence within the categories. Further research is also suggested in relation to how such information can be best understood and used by teachers and students. This suggests the need to offer teachers the opportunity to work with the approach and learn about its use, guided by appropriate professional development. We invite other researchers to engage with the growing literature on cognitive diagnostic assessment and re-examine the potential of computer-based error analysis assessment in mathematics as a modern tool available in the cause of improved mathematical performance amongst primary students.

References

Bearne, E., & Kress, G. (2001). Editorial. *Reading, Literacy, and Language, 35*(3), 89-93

Bennett, R. E., & Bejar, I. I. 1998. Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bejar, I. I. 1984. Educational diagnostic Assessment. *Journal of Educational Measurement, 21*(2), 175-189.

Brueckner, L. J., & Elwell, M. 1932. Reliability of diagnosis of error in multiplication of fractions. *Journal of Educational Research, 26*(3), 175-185.

Burke, E. 2011. *A framework for diagnosing students' difficulties within the Irish primary mathematics curriculum.* University of Dublin, Trinity College: Unpublished Master's Thesis.

Casey, D. P. 1978. Failing students: a strategy of error analysis. In P. Costello (Ed). *Aspect of motivation, Melbourne, Mathematical Association of Victoria*, 295-306.

Choi, S. W., & Tinkler, T. 2002. *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. http://www.ncme.org

Clements, K.A. 1980. Analysing children's errors on written mathematical tasks. *Educational Studies in Mathematics 11*, 1-21.

Cockburn, A. D. & Littler, G. (Eds.) 2008. *Mathematical misconceptions. A guide for primary teachers.* Los Angeles: Sage.

Confrey, J. 1990. A review of the research on student conceptions in mathematics, science and programming. *Review of Research in Education 16*, 3-56.

Cook, J., & Jenkins, V. 2010. *Getting started with e-assessment. Project Report.* Bath, UK: University of Bath.

Cosgrove, J., Perkins, R., Shiel, G., Fish, R., & McGuinness, L. (2012). *Teaching and learning in Project Maths: Insights from teachers who participated in PISA 2012.* Dublin: Educational Research Centre.

Crocker, L, and Algina, J. 1986. *Introduction to classical and modern test theory.* NY: Hold, Rinehart and Winston.

Delaney, S. 2010. *Knowing what counts. Irish primary teachers' mathematical knowledge for teaching.* Dublin: Marino Institute of Education and Department of Education and Science. Available at hppt://www.mie.ie

Department of Education and Skills. 2011. *The national strategy to improve literacy and numeracy among children and young people 2011-2020.* Dublin: Author.

Doukas, N., & Andreatos, A. 2007. Advancing electronic assessment. *International Journal of Computers, Communication and Control (IJCCC), 2*(1), 56-65. Available at http://journal.univagora.ro/

Egen, T.H.M and Lampe, T.M. 2011. Comparison of the reliability of scoring methods of multiple-response items, matching items and sequence items. *CADMO XIX*(2) 85-104.

Eivers, E. and Clerkin, A. 2012. *PIRSL and TIMSS 2011. Reading, mathematics and science outcomes for Ireland.* Dublin: Educational Research Centre.

Eivers, E., Close, S., Shiel, G., Millar, D., Clerkin, A., Gilleece, L., and Kiniry, J. 2010. *The 2009 national assessments of mathematics and english reading.* Dublin: The stationery office.

Feng, M., Heffernan, N.T., & Koedinger, K.R. 2009. Addressing the assessment challenge in an Online System that tutors as it assesses. In *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI journal)*. 19(3), 243-266.

Ginsberg, H. 1977. *Children's arithmetic: How they learn it and how you teach it.* Austin, TX: Pro-Ed.

Government of Ireland, 1999. *Primary school curriculum.* Dublin: The Stationery Office.

Grossnickle, F. E. 1935. Reliability of diagnosis of certain types of errors in long division with a one finger divisor. *Journal of Experimental Education, 4*(1), 7-16.

Hansen, A., Drews, D., Dudgeon, J., Lawton, F. & Surtees, L. 2005. *Children's errors in mathematics. Understanding common misconceptions in primary schools.* Exeter, UK: Learning Matters Ltd.

Haylock, D. 2006. *Mathematics explained for primary teachers.* Third Edition. London: Sage.

Higgins, J., Russell, M., & Hoffmann, T. 2005. Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning and Assessment, 3*(4). Available from http://www.jtla.org

Jeffes, J., Jones, E., Wilson, M., Lamont, E., Straw, S., Wheater, R. and Dawson, A. 2013. *Research into the impact of Project Maths on student achievement, learning and motivation: final report.* Slough: NFER.

Johnson, M. & Green, S. 2006. On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning and Assessment, 4*(5). Available from http://www.jtla.org

Kingston, N. M. 2009. Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22-37.

Kingston, N. M., and Nash, B. 2011. Formative assessment: A meta-analysis and a call for research. *Educational Measurement Issues and Practice, 30*(4), 28-37.

Lankford, F. G. 1974. What can a teacher learn about a pupil's thinking through oral interviews? Arithmetic Teacher, 21, 26-32.

Lesh, R. & Zawojewski, J. 2007. Problem solving and modeling. In. F. K. Lester, Jr. (Ed.). *Second handbook of research on mathematics teaching and learning, Volume 2*. pp. 763-804. Charlotte, NC: Information Age Publishing Inc / National Council of Teachers of Mathematics.

Martio, O. 2009. Long term effects in learning mathematics in Finland – Curriculum change and calculators. *The Teaching of Mathematics, 12*(2), 51-56.

McDonagh, S. and Quinlan, T. 2012. *Maths and national competitiveness: A discussion document*. Dublin: National Competitiveness Council

McKee, L. M., & Levinson, E. M. 1990. A review of the computerized version of the Self-Directed Search. *Career Development Quarterly, 38*(4), 325-333. Available at: http://web.ebscohost.com/ehost/detail?sid=50fde15f-4278-427a-be04-a01a1ddb9b09%40sessionmgr111&vid=2&hid=108&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=bth&AN=9605304043

Mullis, I., Martin, M., Foy, P., and Arora, A. 2012. *TIMSS2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Newman, M. A. 1977. An analysis of sixth-grade pupils' errors on written mathematical tasks. In M.A. Clements and J. Foyster (Eds.). *Research in mathematics education in Australia, Vol 1*, pp 239-258.

Nitko, A. J. & Brookhart, S. M. 2007. *Educational Assessment of Students 5th Edition*. Upper Saddle River, NJ: Pearson Education.

OECD. 2010. *The High Cost of Low Educational Performance: The Long-run Economic Impact of Improving PISA Outcomes*. PISA, OECD Publishing. doi: 10.1787/9789264077485-en

Orton, A. & Frobisher, L. 1996. *Insights into teaching mathematics*. London: continuum.

Perkins, R., Shiel, G., Merriman, B., Cosgrove, and Moran, G. 2013. *Learning for life: The achievements of 15-year-olds in Ireland on mathematics, reading literacy and science in PISA 2012*. Dublin: Educational Research Centre. Available at http://www.erc.ie

Pommerich, M. & Burden, T. 2000. From simulation to application: Examinees react to computerized testing. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2000.*

Russell, M., Goldberg, A. & O'Connor, K. 2003. *Computer-based testing and validity: A look back and into the future*. Boston College: Technology and assessment study collaborative. Available at http://www.bc.edu/research/intasc/PDF/ComputerBasedValidity.pdf

Scalise, K. & Gifford, B. 2006. Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning and Assessment, 4*(6). http://www.jtla.org

Smith, J. diSessa, A. & Roschelle, J. 1993. Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *Journal of the Learning Sciences, 3*(2), 115-163.

Surgenor, P., Shiel, G., Close, S. and Millar, D. 2006. *Counting on success. Mathematics achievement in Irish primary schools.* Dublin: The Stationery Office.

Van der Kleij, F. M., Timmers, C. F. & Eggen, T.J.H.M. 2011. The effectiveness of methods for providing written feedback through a computer based assessment for learning: a systematic review. *CADMO, 19*(1), 21-38.

Verschaffel, L., Greer, B., & DeCorte, E. 2007. Whole number concepts and operations. In. F. K. Lester, Jr. (Ed.). *Second handbook of research on mathematics teaching and learning, Volume 1*. pp. 557-628. Charlotte, NC: Information Age Publishing Inc / National Council of Teachers of Mathematics.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. 2007. A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219-238.

Watson, I. (1980). Investigating errors of beginning mathematicians. *Educational Studies in Mathematics, 11*(3), 319-329.

White, A. 2005. Active mathematics in classrooms: Finding out why children make mistakes – And then doing something to help them. *Square One, 15(4)*, December.

Williamson, D.M, Bejar, I.I., and Mislevy, R.J. 2006. Automated scoring of complex tasks in computer-based testing: An Introduction. In. D.M. Williamson, R.J. Mislevy, and I.I. Bejar. (Eds.). *Automated scoring of complex tasks in computer-based testing*. pp. 1-14. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Table 1
Categories of Student Mathematical Errors employed in the computer based assessment

	Category	Description
1	Concept understanding	Understanding of the key elements or principles governing a mathematical domain and recognising how the knowledge, elements or principles interrelate.
2	Strategy selection	Choice made by learner in converting semantic or symbolic prompt into appropriate mathematical structure. Frequently applies to converting problem prompts into mathematical/computational expression.
3	Setup of procedure	Difficulties in effecting appropriate spatial layout of mathematical symbols required for successful computation/operation.
4	Method execution	Difficulties in executing the procedures known to be required to solve the mathematical task; frequently but not restricted to use of algorithms.
5	Carelessness	Errors that occur apparently at random and unlikely to be repeated.
6	No conclusion	Impossible to estimate the specific nature or cause of error; includes items not attempted or not reached.
7	Viewed but not attempted / completed	The student viewed the item but did not provide any response. This can be interpreted as either (i) not attempting the item at all or (ii) not completing it to the stage of providing a response/solution.
8	Not reached	The student did not provide a response to the item because he/she did not view the item at all.

Table 2
Topics assessed in the computer based assessment.

Topic	Objective
Money	Rename amounts of euro or cents and record using symbols and decimal point. Solve and complete one-step problems involving the addition and subtraction of money.
Lines and angles	Identify, describe and classify vertical, horizontal and parallel lines. Classify angles as greater than, less than or equal to a right angle. Solve problems involving lines and angles.
Fractions	Compare and order fractions with appropriate denominators and position on the number line. Calculate a fraction of a set using concrete materials. Solve and complete practical tasks and problems involving fractions

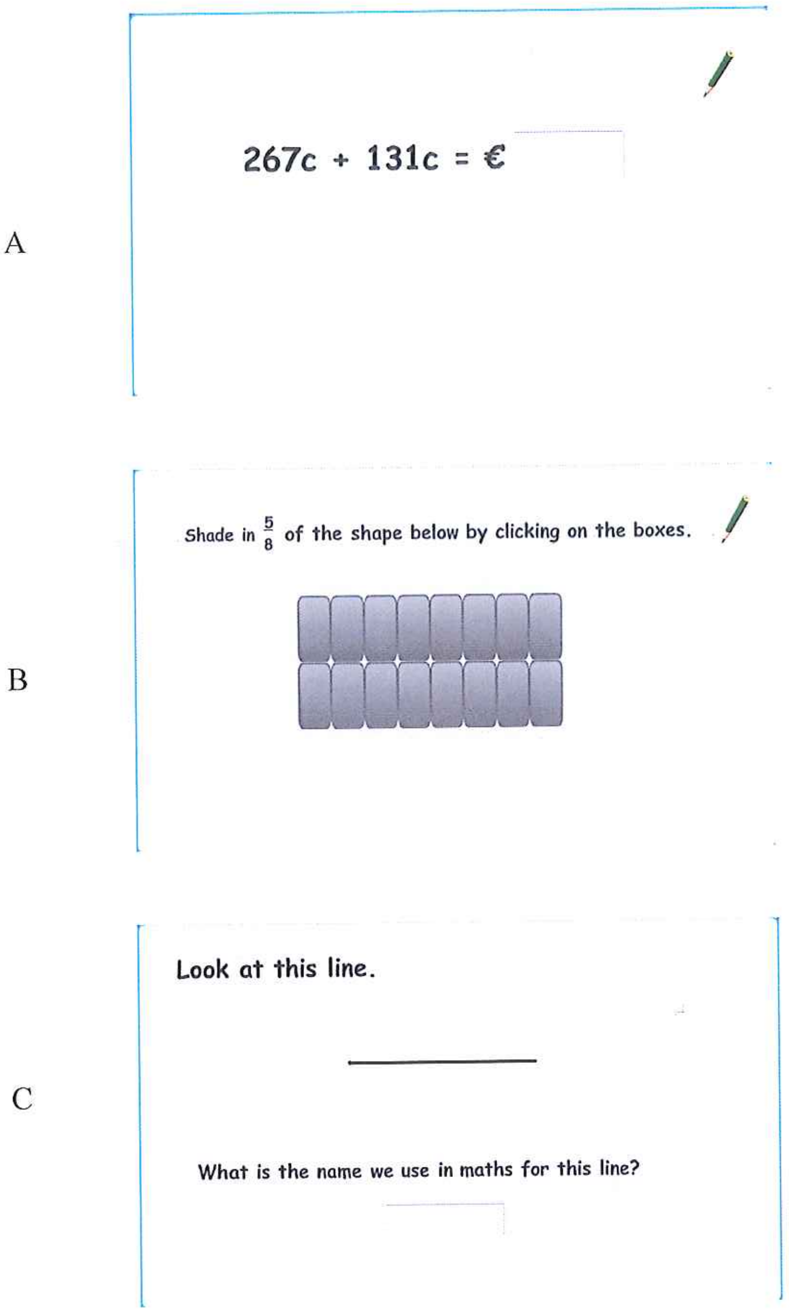
Table 3
Marksheet and Item Response Codes related to measures item in Figure 2.

Response Code	Description	Sample answer	Score
1	Correct	€3.80	1
2	Slip1: Minor error	*€*.80, 3.*0, €3.8**	0.5
3	Slip2: Omission of decimal	*380*	0.5
4	Unable to convert problem	€6.70, €6.60	0
5	Unable to rename (subtract)	4.60, 4.20, 420, 460	0
6	Error in method execution	300-500 or 3.00 – 5.00	0
7	No understanding of task	Other answers	0

Table 4
Distribution of incorrect student responses across Error Categories: Percentage of student responses in different categories. (Number of items in parentheses)

Error Category	Description	Total Test (23)[a]	Understanding & Recalling (5)	Using Procedures (8)	Reasoning & Problem Solving (10)
1	Concept understanding	40	70	57	20
2	Strategy selection	10	0	0	20
3	Setup of procedure	5	0	<1	10
4	Method execution	13	0	25	7
5	Carelessness	9	0	13	8
6	No conclusion	8	0	0	17
7	Viewed but unattempted	8	14	5	13
8	Not reached	5	16	0	6
		100	100	100	100

[a] Number of items



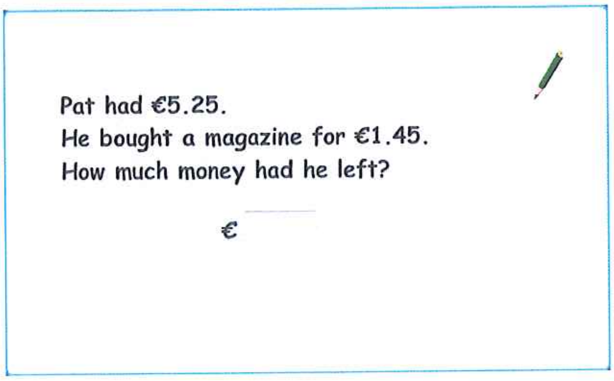
267c + 131c = €

Shade in $\frac{5}{8}$ of the shape below by clicking on the boxes.

Look at this line.

What is the name we use in maths for this line?

Figure 1 Items illustrating variety of response tasks for students



Pat had €5.25.
He bought a magazine for €1.45.
How much money had he left?

€

Figure 2 Sample measures item related to objective:
Solve and complete one-step problems and tasks involving the addition and subtraction of money.

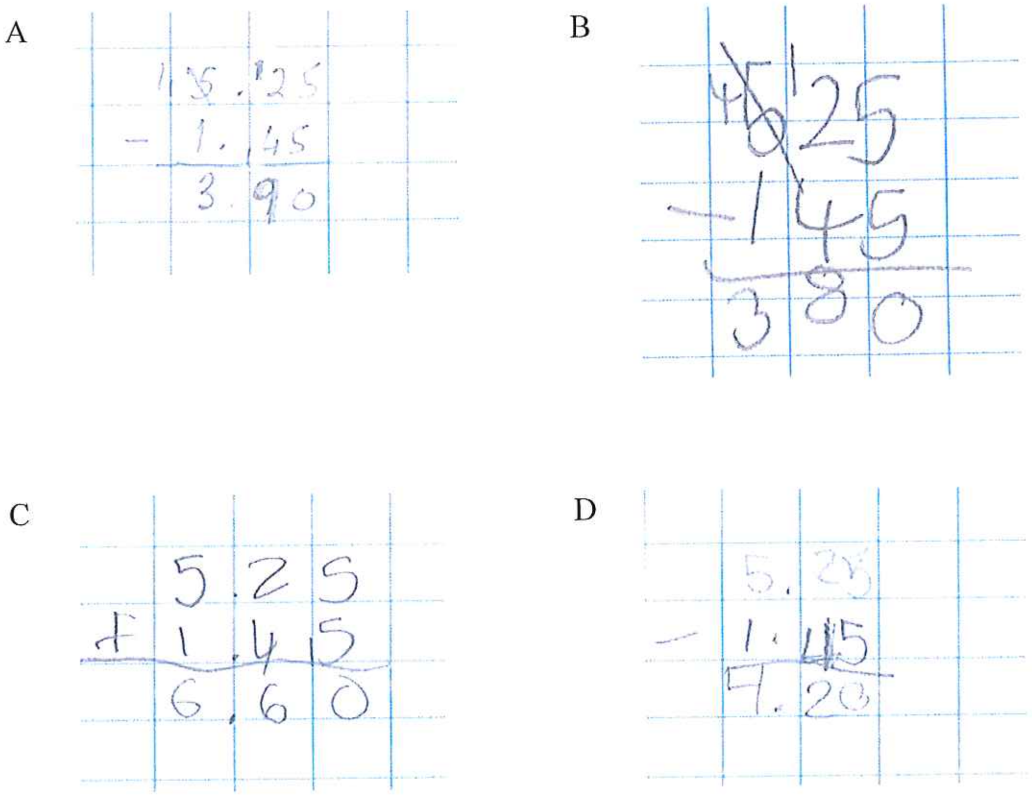


Figure 3 Selected student responses to measures item in Figure 2.