This is the published version:

Doig, Brian, Williams, Julien, Wo, Lawrence and Pampaka, Maria 2006, Integrating errors into developmental assessment: 'time' for ages 8-13, in Proceedings of the 30th conference of the international group for the psychology of mathematics education, Charles University in Prague, Faculty of Education, Prague, Czech Republic, pp. 441-448.

Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30005949

Reproduced with the kind permission of the copyright owner.

Copyright: 2006, The Authors
INTEGRATING ERRORS INTO DEVELOPMENTAL ASSESSMENT: ‘TIME’ FOR AGES 8-13

Brian Doig
Deakin University

Julian Williams Lawrence Wo
University of Manchester

It is widely agreed that measurement is of paramount importance to students’ overall development in mathematics. This paper describes a developmental ‘map’ of students’ understanding and skills in measurement, focussed on the topic of Time, that integrates correct and incorrect student ideas. The map is based on a Rasch analysis of data from a large-scale UK national survey for standardising assessment for children from 5 to 14 years of age. It is demonstrated how a partial credit strategy enables a developmental map to be constructed to show students’ strengths and weaknesses in a meaningful and useful summative and formative manner. This map provides evidence, of both a summative and a formative nature, which may enable teachers to craft appropriate and successful learning experiences for children.

INTRODUCTION

The application of mathematics to daily life is firmly based on those aspects of mathematics commonly denoted as ‘measurement’ in primary school curriculum documents, a situation affirmed precisely thirty years ago in the National Council of Teachers of Mathematics annual year-book (Nelson & Reys, 1976). While there are some regional differences in nomenclature and definition, we argue that there is sufficient common ground for a description of the development of measurement understandings and skills in students to be of benefit to teachers and students. We will show that the large-scale collection of student responses to Time items enables a description of a hierarchy of student development, and that the use of a partial credit strategy enhances our understanding of how certain ‘errors’ may indicate progress towards understanding and eventual success.

The most common description of children’s development of measurement understandings and skills are to be found in curriculum documents, and their close relatives, text books. However, other education stakeholders have contributed support for this view of the development of measurement. The Cockcroft Report (Cockcroft, 1982) strongly endorsed the usefulness and necessity of learning to measure (see, for example, § 79 and § 269). In the United States, the late 1980s saw the National Council of Teachers of Mathematics (NCTM) publish its ‘Standards’ which included standards for measurement (National Council of Teachers of Mathematics, 1989). The measurement standards for grades K to 4 included awareness of the measurable attributes of objects, units of measurement, estimation of measurements, and using measurement in everyday contexts. The standards for
grades 5 to 8 further developed these topics to rates, indirect measurement, and derived units.

However, some mathematics education researchers have explored alternative ways of describing students’ mathematical development. Under the banner of Developmental Assessment, the work of Masters and Forster (Masters & Forster, 1996a, 1996b) has provided a great deal of encouragement to those seeking an alternative approach for thinking about development and assessment, and particularly, how assessment can inform curriculum (Doig & Lindsey, 2002). Similar research has yielded such maps, in such areas as mental computation (Callingham & Watson, 2004) and probability (Gagatsis, Kyriakides, & Panaoura, 2001).

In a recent study, drawing on aspects of Piaget’s work on the development of measurement ideas, Bladen and her colleagues (2000) describe a developmental ‘map’ that showed discrepancies between curriculum and children’s development. The authors claim that their ‘findings may serve as a warning to organizations that develop content standards to reconsider expectations of students … [and that] These data reveal a wide discrepancy between educational expectations and development of the concept of measurement’ (p. 11). Further, the authors raise the question of ‘how can these results be used to help teachers and students meet the state and national standards … ?’ (p. 11). This question was also raised by Williams and Ryan (2000) who examined the responses of a large number of students to national test items.

Diagnostic, or formative, information in most developmental maps is implied, and is carried by the notion of more is better: that is to say, development is shown through descriptions or examples of more achievement (usually as more correct responses) whereas we argue that a richer, more useful map, in O’Connor’s (1992) phrase, includes “accounting for errors” (p. 20) rather than simply counting errors (italics in the original). Thus, in order to explore students’ measurement development better, we set out to describe how student responses from a large-scale assessment program could be analysed, for some aspects of measurement, to produce a ‘map’ of typical development, one that would include errors and mis-understandings. Such a developmental map would describe increasing ability, and importantly, it would also describe typical mis-understandings and errors that accompany this development (Williams & Ryan, 2000). This ‘accompanying’ is a critical point, as it illustrates that development includes, for some students at least, new mis-understandings or the persistence of earlier problems. As Williams and Ryan (2000) suggest, the errors illustrated in the developmental map “provide a concrete reference point for teachers to engage with research findings and conceptual frameworks in the literature that would otherwise remain obscure and arcane” (p. 67). Further, they argue that “[the] analysis and interpretation [of mis-understandings and errors] may mediate between the research community and the teaching profession” (p. 68).
INSTRUMENTS AND DATA SOURCES

As part of a continuing program to develop formative assessments that also serve summative purposes (see, for example, Ryan, Williams, & Doig, 1998), a new set of age-standardised diagnostic assessment materials for ages 5-14 were developed by the Mathematics Assessment for Learning and Teaching (MALT) team from the University of Manchester in collaboration with the publishers Hodder Murray (see, Williams, Wo, & Lewis, 2005, for further details). The MALT items draw on content from the UK National Curriculum, and the measurement items assessed aspects of time, length, capacity and weight, measures of measures, conversion of units, problem solving, and estimation and rounding.

A nationally representative sample was drawn from schools in England and Wales, and students’ responses to this large set of mathematics items were collected. A total of over 14 000 students took part.

The items were placed into papers (30 mark points for Reception to Year 2 and 45 mark points for Years 3 to 9) with no common items between papers. Vertical test equating was conducted using a sub-set of students who sat two papers, and by using data from a pre-test in which students sat approximately half a paper from the following year. Validation included analyses at both the development and main test stages that suggested that the construct, scale and the vertical equating was safe (Ryan & Williams, 2005) although vertical equating becomes less reliable with changes to curriculum content in subsequent years.

The complete set of MALT measurement items can be accessed at www.education.man.ac.uk/lta/pme30/measurement.

ANALYSIS AND RESULTS

Like most of the research cited above, responses to the MALT items were analysed using Item Response Theory (IRT) (Bond & Fox, 2001; Rasch, 1960). Whilst there are several benefits of using this form of analysis, a particularly useful aspect is the nexus between a student’s ability, as measured by the assessment items, and the difficulty of correctly answering these items. In brief, this is that for any given level of student ability, it is possible to identify the likelihood of that student responding correctly to any item. Importantly, this means that the greater the difference between student ability and item difficulty the more likely the student will answer correctly and vice versa. In addition to this formative affordance, we may interpret positive differences between positions on the IRT scale as indications of progress or development. While the basis for connecting the difficulty of assessment items and student achievement is probabilistic, the relationship is of critical importance in diagnostic interpretations of a student’s assessment performance.

In a simple Rasch analysis items are taken to have two values: correct or incorrect. While this provides useful summative information about student achievement and understanding, responses that are incorrect are all treated as being equal, that is, they are the same. However, as an examination of the students’ responses shows, a range
of responses exists, many responses presenting possible insights into the different ways students think about the item to which they respond.

In the eight-year-old MALT test, item 30, shown in Table 1, children had to read an analogue clock and select the equivalent digital time (2:45). This generated three error-responses, of two qualitatively different kinds: the first pair refer to problems in reading a clock-face, while the last describes a difficulty in correctly interpreting the hour hand (before or after the hour). It will emerge that the latter is an error made by much less able students than the former two, which are quite close in developmental terms.

Table 1: Item codes and interpretations for a MALT item 8 – 30

<table>
<thead>
<tr>
<th>Item</th>
<th>Response</th>
<th>Description of imputed student thinking</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:45 (correct)</td>
<td>Read o'clock time (a quarter to)</td>
<td></td>
</tr>
<tr>
<td>3:45</td>
<td>Confused hour on the clock with the next hour when reading a clock-face</td>
<td></td>
</tr>
<tr>
<td>9:15</td>
<td>Confused hour hand and minute hand when reading a clock-face</td>
<td></td>
</tr>
<tr>
<td>2:15</td>
<td>Interpreted quarter to as quarter past on reading a clock problem</td>
<td></td>
</tr>
</tbody>
</table>

Although a single item may not provide a key to unlocking every student’s misunderstandings, collectively, item responses can reveal markers of students’ mathematical development. Thus, across a range of items, the patterns of incorrect responses provide formative information that adds richness to the developmental map.

However, these descriptions of responses need to be placed on the developmental map at an appropriate place. In order to do this, an alternative analysis, using a Masters’ PCM (Partial Credit Model, Masters, 1982) was conducted with ConQuest (Wu, Adams, & Wilson, 1998). As its name suggests, in a Partial Credit analysis each response is viewed as a step towards a fully correct response to the item. This requires the ordering of the erroneous responses from a ‘least developed’ group to more highly developed groups, to ‘correct’. This was done by ordering the responses by the mean ability (obtained from a dichotomous Rasch analysis) of those students making the error: then ordering the error responses so that errors made by students with higher abilities are given a higher order. In the case of item 8 Q30 discussed above, involving the reading of a clock, there were four responses (three errors and one correct) associated with increasingly ‘able’ students (meaning simply students of increasing competence in measurement overall). Only in this ordering will the Rasch PCM be well-fitting with acceptable fit statistics, and thresholds correctly ordered as for this item (shown at those points marked with “A” in Fig 1.)
Figure 1: Developmental map for some Time item-responses, n=14420
(approximately 1500 per year group), age 8-13
INTERPRETATION

A selection of meaningful responses to eight items focussed on the topic of Time are presented in the developmental map in Figure 1 (four correct responses and twelve errors). The item response descriptions are numbered with their test number, question number, and their step level: thus, 12Q13.3 is the third step (and, in this case, is the correct response) of item 13 in the 12-year-old MALT test. Correct responses are in bold text. An examination of the map shows that there are several sub-texts that provide formative information. In order to see these sub-texts clearly, related items are labelled similarly alphabetically. Four sub-texts are described here:

**Sub-text A** deals with the ability to read an analogue clock. The conventions of the clock-face are not as simple to master as many adults believe. This difficulty persists across a range of abilities, shown by the responses 8Q30.1, 8Q30.2, 8Q30.3, and is not fully achieved until the response 8Q30.4 is given. Note that the chart on the left of the figure indicates that only a minority of the sample are expected to achieve this.

**Sub-text B** focuses on the conceptual aspect of the representation of time by numbers. The idea that numbers can be used in other than a decimal form is apparently a difficult concept for many students. As responses 13Q14.1, 12Q13.1, 11Q12c.1, and 8Q16.1 show, this causes trouble for students across a wide age range from 8 to 13 years.

**Sub-text C** describes student development in reading information from a distance-time graph. These items may reflect also student development in graphicacy as well as in concepts of time and distance. Notwithstanding this, responses 10Q32b.1, 10Q32b.2, and 10Q32b.3 (a correct response) are all at the more difficult end of the developmental scale. This information enables teachers to plan for successful learning by being aware of the likely difficulties students may encounter.

**Sub-text D** describes the development of student flexibility in dealing with time represented in a decimal format. While responses 10Q19.1 and 10Q19.2 describe difficulties, success is achieved for students of slightly higher ability, shown by the higher position on the scale of response 10Q32a.2. Interestingly, rounding down appears to be a more difficult undertaking than rounding up.

While these interpretations are informative, they are not exhaustive, even for this one topic: they serve, however, to illustrate the point that well-designed summative assessments can be made to provide formative information.

CONCLUSIONS

We have shown how the award of credit for increasingly sophisticated errors in large-scale assessment can enrich diagnostic assessment through a Time map which charts the full range of student responses. While the description offered here provides a picture of students’ development in the measurement topic of time, it must be remembered that students have been exposed to the UK curriculum, which, by its very nature, attempts to scaffold a particular developmental pathway. However, we recommend the methodology to those involved in large-scale assessment in other
countries, and although our argument has been largely technical, we hope that the
general principal of rewarding errors with credit may be a significant part of the work
of changing perspectives on teaching and learning.

References

Bladen, E., Wildish, K., & Cox, J. (2000). The development of linear measurement in
2 - 12.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in

Callingham, R., & Watson, J. (2004). A developmental scale of mental computation with

Teaching of Mathematics in Schools under the Chairmanship of Dr. W. H. Cockcroft.*
London: Her Majesty's Stationery Office.

Doig, B., & Lindsey, J. (2002). Summative and formative assessment: Creating a tool for
(Eds.), *Mathematics Education in the South Pacific: Proceedings of the 25th Annual
227-235). Auckland, New Zealand: Mathematics Education Research Group of
Australasia.

assessment on probabilities: A Rasch measurement model analysis*. Paper presented at
the Twenty-fifth Conference of the International Group for the Psychology of


Council for Educational Research.

174.

Standards for School Mathematics*. NCTM.


science research and the framing of assessment policy. In B. Gifford & M. C. O'Connor
(Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests
(Expanded ed.)*. Chicago: University of Chicago.
**Doig, Williams, Wo & Pampaka**


