Applying CBA Technology to Traditional Error Analysis in Elementary Mathematics. Lessons from *Maths Assist.*

Damian Murchan[1], Elizabeth Oldham[2], Conor O'Sullivan[3]

[1]School of Education, Trinity College, Dublin
[2]School of Mathematics, Trinity College, Dublin
[3]Apierian e-Learning / Folens Group, Tallaght, Dublin

Paper Presented at the
Annual Meeting of the American Educational Research Association,
San Francisco, April 2013.

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

Abstract

The digital agenda in education has transformed practice with students increasingly being taught in a technology-rich environment. Alongside this rapid technical revolution in learning is a slower evolution in digital assessment. Teachers can harness the positive "affordances" of Computer-Based Assessment (CBA) to support student learning. Through analysis of test data and reflections by stakeholders, this paper investigates *Maths Assist*, a CBA where feedback to teachers and students specifically highlighted students' errors and misconceptions in elementary-level mathematics. The data suggest the potential to accurately capture and analyse errors through automated processing with evidence that student errors expected in relation to different mathematical skills were identified. The experiences of the test developers in refining the program are highlighted in the paper.

Introduction and Purpose of Study

Competence in mathematics has for a long time exercised the policy agenda within and across education systems. From the nascent international surveys of educational achievement to the present day, mathematics ranks amongst the "core" subjects in most education systems. Policymakers, politicians and the media routinely emphasise mathematics in their analysis of what is right, or wrong with educational practice. Recently, the term "PISA shock" is used to describe the emotional reaction by the policy community when students' performance in "core" areas such as mathematics as measured by the OECD survey is at variance with local expectations. After-effects of PISA shock manifest themselves in significant reform of curricula and overall educational structures and practices, as in the recent cases of Norway and Germany for example. In Ireland, concern with students' mathematics achievement (Eivers et al., 2010; Engineers Ireland, 2010) has prompted a national literacy and numeracy strategy (DES 2011) alongside a fundamental review of mathematics curricula at secondary level (Project Maths Development Team, nd).

What is clear from the vigorous policy debate around mathematics is that despite continual reform of curricula and teaching methods, many students still underperform in mathematics. Within the drive worldwide towards more broadly-based and skills-based curricula (Rychen & Salganik, 2003; Partnership for 21st Century Skills, nd; DES 2012), mathematics retains its central and separate place within curricula,

notwithstanding the policy rhetoric in relation to cross-subject skills such as critical thinking, problem solving, communication, collaboration, personal development, creativity etc.

Despite the range of techniques used by teachers in mediating mathematics curricula to students, many students fail to achieve. This suggests the need for additional approaches and techniques aimed at providing assistance to teachers and students. Two such approaches are the focus of this paper: Error Analysis and Computer Based Assessment (CBA). Neither of these is particularly new in isolation. Taken together, however, we believe that the incorporation of error analysis of students' mathematics responses within CBA represent the type of "affordances" associated with computers whereby certain activites and tasks are "made easily possible" by a medium such as a computer (Bearne & Kress, 2001: 90). The present study describes a novel assessment programme called *Maths Assist* that combines the diagnostic potential of error analysis with the administration and scoring potential of a CBA.

Error analysis in mathematics involves the identification and categorisation of specific types of errors committed by students when working mathematically, whether in class, on tests, or in conversation. It has as its premise the opportunity to use such information to better inform teachers, students and other stakeholders so that learning can be enhanced. However, for such an intuitively appealing approach, there are surprisingly few operational examples of error analysis in practice within education systems worldwide. We hypothesise that part of the reason for this centres on a number of related issues (i) the logistical complexity of conducting the one-to-one diagnostic interviews with learners typical of the approach, (ii) difficulty in convincing teachers that the process was worth the trouble, essentially a sense that the approach is not practical along with (iii) concerns about inferring thought processes from errors, especially written errors.

Here, then, is the second focus within this paper, drawing on the digital agenda in education. There has been a marked increase in the use of information and communications technology in teaching and learning, with students being increasingly taught in a technology-rich environment. Alongside the rapid digital revolution in learning is, however, a somewhat slower evolution in digital assessment. For assessment to be seen by teachers and students as relevant to the digitally-oriented pedagogical environment, greater alignment is required between stakeholder needs

and the nature of assessment. Whereas technological developments in the field of educational assessment are "inexorable and inevitable" (Bennett, 2002) as technology makes it possible to assess students in ways not previously envisaged, schools frequently operate in a misaligned state where learning is mediated through digital technologies but assessed using more traditional pencil-and-paper formats. The present study attempts to address this anomaly.

In this paper, we contend that the recent availability of highly flexible and sophisticated CBAs offer opportunity to reapply the principles and lessons of error analysis within classroom settings, in keeping with the early work of Drucker & McBride (1987) and more recent proposals from Buiffington & Clements (2011). One of the affordances of the computer is to remove some of the inefficiencies in procedures. The opportunities and challenges associated with the ease of testing and scoring by computers represents the second theoretical focus of the paper. Consequently, this paper explores the application of error analysis techniques in mathematics within the context of a CBA programme termed *Maths Assist*, designed for use with elementary grades and where feedback to teachers and students specifically highlighted students' mathematical errors and misconceptions. The paper investigates the implementation of the *Maths Assist* CBA across a broad range of mathematics topics with a sample of students in Ireland, drawing on test data, test developers' experiences and a limited selection of teacher perspectives. The main objectives underpinning the investigation are:

1. Illustrate an application of error analysis in mathematics mediated through a computer based assessment programme.
2. Evaluate the potential role of automated error analysis in providing credible feedback to teachers and students in relation to mathematics performance, particularly in relation to common sources of difficulty encountered by students across topics.
3. Explore the extent to which automated scoring algorithms could accurately capture student responses, particularly in the context of free-response (non-objective) items and across a range of mathematical topics

Theoretical framework

Error Analysis

This study draws on the long history of attempts to use information about student errors in elementary mathematics to diagnose difficulties. Early studies (Brueckner & Elwell, 1932; Grossnickle, 1935) typically focused on specific topic areas and emphasised the identification and classification of errors but often resulted in large and unwieldy numbers of categories. Subsequent studies focused on identifying smaller, more generic classifications of errors (Lankford, 1974; Newman, 1977; Casey 1978; Clements, 1980) to reveal more latent cognitive thought processes leading to incorrect item responses. This, along with other work by Confrey (1990), Borasi (1994), Smith, diSessa and Roschelle (1993) and Leighton and Gierl (2007) highlight the compatibility of error analysis techniques with constructivist ways of interpreting and promoting learning through assessment.

Conceptualisation and evidence for the efficacy of such approaches draw especially on work in Australia during the 1970s and 1980s by Anne Newman and others who embraced her diagnostic interviewing process. Whereas there was intense research activity during that particular period, the initiative never took hold to the extent that the positive findings suggested it should, despite some later investigations (Ellerton & Clements, 1996; Ayers, 2001, White 2005). One sustained implementation was initiated by the New South Wales Department of Education and Training in the form of its *Counting On* and *Counting On2007* programmes where the five "Newman prompts" centrally underpin the Department's advice to teachers (NSW, nd). An evaluation of the 2007 programme lends continuing support to the efficacy of the Newman procedure as a method to help determine why students make mistakes with written mathematics questions (White 2008).

Recent iterations of error analysis methodologies are consistent with the cognitive diagnostic assessment and classification approaches advocated by Leighton & Gierl, (2007) and Rupp, Templin & Henson (2010). These models seek to facilitate teachers in deriving more cognitively based inferences about what students can and cannot understand and do, especially as this relates to possible student misconceptions. The present study is framed within a more recent strand of research seeking to develop computer-based forms of cognitive diagnostic models as evident in the work of

Huebner (2010) and Buiffington & Clements (2011) that aims to draw on the affordances of e-assessment

Computer Based Assessment

One rationale for applying CBAs to cognitive diagnostic assessment models is to maximise the usability of the assessments for teachers. Whereas previous research attests to the efficacy and usefulness of error-analysis methods in mathematics, practical application has been relatively limited due to constraints implicit in the busy realities of classroom life and the logistical and professional challenges of scaling up Newman-like individual diagnostic interviews within the context of regular classroom life. The error analysis literature offers, therefore, as-yet unrealized potential for helping teachers better understand, and therefore, address students' difficulties in mathematics. The present study is designed to overcome the logistical, cumbersome obstacles to error analysis-oriented assessment in mathematics through building on the affordances implicit in CBA.

CBAs encompass a broad range of digital assessment techniques that are stored, delivered, answered and scored automatically using information and communication technologies (Cook and Jenkins, 2010). Once development efforts and costs are removed, the convenience and efficiency of CBAs have encouraged their implementation in a wide range of settings. The literature highlights a range of advantages associated with CBAs (Johnson & Green, 2006). Challenges remain in relation to development issues, item design, student interface, student familiarity, roughwork, scoring and the somewhat more intractable issues of unreliability of the hardware, software and internet connections (Bennett and Bejar, 1998; Choi and Tinkler, 2002; Higgins, Russell and Hoffman 2005; Scalise and Gifford, 2006; Kingston, 2009; Eggen & Lampe, 2011).

Whereas an alluring benefit of CBA is the provision of quick feedback to stakeholders, we argue that the nature and quality of the feedback is of greater importance. Many CBAs provide number correct feedback to students, or variants of number correct that depend on subsequent transformation and scaling of data based on a range of item response models. Moving from these approaches to embracing different forms of feedback requires both change in test developers' habits and the support of users, particularly teachers. Maximising the diagnostic information from student responses is a key goal and can be achieved by analysing the examinee's

responses to each item and drawing inferences across items about more generalised common mathematical errors and conceptions that may underpin and undermine student learning in mathematics.

Incorporating error analysis techniques within CBAs requires certain assumptions. Much of the evidence for error analysis derives from work based largely on individual diagnostic interviews conducted with students. However, inferring thought processes solely from errors on written tests is not unproblematic (Burke, 2011) and caution is required in interpreting results carefully.

Towards a combined model

The benefits of error analysis and CBA were combined in the development of an assessment programme known as *Maths Assist*. This programme, aimed at the middle elementary grades, employs software to allocate student responses to a small number of over-arching categories that describe the mathematical action and thought processes implicit in the observed response.

The CBA was designed in the context of the national Primary Mathematics Curriculum (Government of Ireland, 1999), using test blueprints that reflected the objectives of the curriculum, including three different forms of mathematical skills. Fifteen of the topic areas from the mathematics curriculum were selected and short, focused diagnostic tests were developed to measure student achievement on and misconceptions about the topics. Participating students accessed the diagnostic tests online, from their classrooms.

Student responses were automatically captured by software and analysed using classical test theory in addition to automated error-analysis procedures. The results of the analyses were returned to teachers by email. Processing and scoring of the tests yielded three different forms of data. Simple polytomous raw scores were estimated for each student based on their performance on each item where the response to an item was judged as correct (1 point), partially correct (0.5) or incorrect (0). As such, this facilitates the compilation of item difficulty and discrimination statistics and the generation of item-level, topic-level and test-level feedback to teachers. Overall, this polytomous score is used to generate item and test-level data about individuals and the class as a whole across all the items and for specific subsets of items in relation to content and skill level.

Separate processing of the data used students' individual item-level responses to categorise a response into one of a small number of item-specific response options or outcomes, termed an *Item Code* (IC). Examples include whether the response reflected a careless slip, partial completion of a required method to find a solution, inability to convert and mathematically frame a word problem, error in calculation, unable to rename in subtraction, confusion of two concepts etc. Initial ideas for the breadth and nature of the possible response codes were drawn from the literature while implementation of the diagnostic testing programme over time offers empirical evidence for these responses and for the identification of new response codes.

The *Item Codes* are linked to a set of broader, over-arching *Error Codes* (EC) applicable to the mathematics curriculum as a whole. This requires a judgement about the best match or fit of the examinee's (incorrect) *Item Code* with a list of seven Error Codes. Drawing on the basic approach implicit in Newman's (1977) five categories of response (or error), our model includes seven possible categories into which a student's incorrect response is classified. These categories transcend topic areas in mathematics and provide an overall picture of the outcomes of students' engagement with test items linked, in part, to the typical errors that students make in solving test items. The Error Codes included in *Maths Assist* are presented in Table 1. These codes form the basis for deriving student profiles related to errors or misconceptions evident in their performance as measured by the CBA. Categories in the model are not meant to be hierarchical, reflecting, rather, different classes of errors identifiable in students' responses. Further details on the codes are available in Burke (2011).

Table 1
Global Error Codes used in categorising student responses on *Maths Assist*

Code	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	Method execution	Difficulties in executing the procedures known to be required to solve the mathematical task; frequently but not restricted to use of algorithms.
4	Carelessness	Errors that occur apparently at random and unlikely to be repeated.
5	No conclusion	Impossible to estimate the specific nature or cause of error; includes items not attempted or not reached.
6	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.
7	Not reached	The student did not provide a response to the item because he/she did not view the item at all.

Accordingly, over a series of items or the entire test, individual item responses are categorised first into *Item Codes* and then into global *Error Codes* thus facilitating compilation of a pattern of overall errors and misconceptions. Scoring rules are applied to identify items that proved particularly challenging for each class or group of students; other scoring rules and parameters govern the identification and reporting of students' achievement and progress on mathematical content and skill areas. Once diagnostic data are extracted from student responses, a feedback pack is generated and returned electronically to the teachers, to aid reflection by teachers and students. The pack contains information about the dominant errors being made by the class and by individual students at a particular time. This provides a manageable task for the teacher in addressing a limited number of dominant and recurring errors within a class.

Research Framework and Methods

Data for the paper draw from an exploratory diagnostic CBA with elementary students in Ireland. The research, in keeping with Tasakkori and Teddlie (2003), is underpinned by a pragmatist philosophical paradigm that accepts as reality the different ways in which students and indeed teachers perceive mathematics and is thus consistent with the view of student conceptions as "children's beliefs, theories,

meanings and explanations" and of student errors as frequently representing overgeneralizations on the part of the students, rather than idiosyncratic random mistakes (Confrey, 1990; Smith, diSessa & Roschelle, 1993).

Theory underpins the conceptual basis to this investigation. Research has suggested a place in education for digital resources and for the analysis of students' errors in mathematics, accumulated theory that can be bridged by applying error analysis to CBAs. The investigative processes, encapsulated in a number of specific research objectives highlighted earlier, reflect relevant understanding of epistemology, theoretical perspective, methodology and methods and the vigorous debate about the relative merits of positivist, phenomenological and pragmatist research traditions (Cresswell, 2005; Robson, 2011). As a consequence of a pragmatic research orientation adopted, the tendency to take sides in the "paradigm wars" with one of either quantitative or qualitative approaches to data gathering was eschewed in the design of the study, in keeping with the freedom and the multiple opportunities espoused by Johnson and Onwuegbuzie (2004: 14). As such, the research can be characterised as a pragmatic, mixed methods exploration of a developmental project generating primarily quantitative with some limited qualitative data to inform the research objectives in a way that "seeks to combine both quantitative and qualitative traditions on the basis that research issues in education are often so complex that the insights of both approaches are required if we are to gain a good understanding" Newby (2010, p. 92).

Sources and Data

The present study was conducted in the context of the national mathematics curriculum in the Republic of Ireland. The methodological framework drew on two sources of data, comensurate with a mixed methods design.

The majority of the data are quantitative test data generated by the CBA programme itself. The CBA was administered to 175 third and fourth grade students in five schools during the course of the 2011-12 school year, as summarised in Table 2. Participating schools were characterised by small or medium student enrolments, school sizes consistent with the majority of schools in the country. As such, a number of the students were located in multi-grade classes, sitting alongside peers from lower or higher grade levels where the teacher simultaneously teaches a number of grade

levels, typically two or three. As part of the study, separate responses were sought from participating teachers using a questionnaire format employing scaled likert items and more open response formats. This instrument explored teachers' perspectives on (i) the accessibility and user friendliness of the CBAs, (ii) the usefulness of the CBA overall, (iii) the usefulness of the feedback and examples of ways in which they used the feedback, and (iv) the extent to which students enjoyed and engaged with the assessment process.

Table 2
Participating schools and grades

School	School Size	# Teachers	# Students 3rd Grade	# Students 4th Grade
1	Medium	3	44	33
2	Small	2	8	9
3	Small	1	7	6
4	Small	2	26	18
5	Small	1	12	12
Total		9	97	78

A series of short, unspeeded CBAs were accessed by students over the internet at times chosen by the 9 participating teachers to suit their own groups. Five major mathematical topics were included: *Number, Algebra, Shape and space, Measures* and *Data* and each of these was subdivided into a smaller number of sub-topics as presented in Table 3. Participating teachers had the freedom throughout the year to administer the assessments as single testlets or in sets of 2 or 3 topics and to administer as few or as many as they wished.

Test specifications incorporated three types of mathematical skills: *Conceptual Understanding; Using Procedures;* and *Reasoning and Problem-Solving*. Each CBA typically included 6-12 items and students entered their answers on computer using the keyboard or mouse. A range of open and closed item types were employed, drawing, in part, on the taxonomy by Scalise and Gifford (2006) with emphasis on free-response items where students used the keyboard to type in numbers or other data into answer boxes on screen, along with other open formats. Student responses were automatically captured by software and analysed using classical test theory in addition

to automated error-analysis procedures. The results of the analysis were returned to teachers by email.

Table 3
Topics assessed in CBA: Grades 3 and 4

Mathematical Area	**Topic**	**Grade 3 #objectives**	**G3 # items**	**Grade 4 # objectives**	**G4 # Items**
Number	Place value	3	8	2	7
	Addition / subtraction	2	5	1	3
	Multiplication	5	12	4	12
	Division	3	10	4	8
	Fractions	4	9	5	9
	Decimals	4	9	6	9
Algebra	Number sentences	2	6	2	8
	Number patterns & sequences	4	9	2	6
Shape & space	2-D shapes	5	10	4	10
	Lines and angles	4	8	4	8
	Symmetry	2	6	2	7
Measures	Time	3	6	4	7
	Money	2	6	3	7
	Weight	4	6	3	8
Data	Representing & interpreting data	5	8	3	8
TOTAL		*52*	*118*	*51*	*117*

Student test responses captured by the software were aggregated and analysed using descriptive statistics and procedures associated with classical test theory. Polytomous item scores were extracted for each student and these were used in deriving aggregate scores that facilitated the provision of traditional forms of feedback to teachers. Item Codes and overall Error Codes were also generated for each student. The error codes were used to identify the most common sources of errors detected in students' responses over collections of three topics. To this end, students were required to take tests on any three topics following from which processing was initiated and feedback subsequently provided.

In the context of this paper, analytic methods consist of descriptive statistics of test scores and associated distributions of errors in relation to the test as a whole and sub-scores based on the three mathematical skills. Reliability of the tests was estimated using Cronbach's Alpha. Data from the teacher questionnaires were analysed using numeric and qualitative approaches. Given the very small number of questionnaires returned (4), partly a function of late distribution at the end of the school year, the questionnaires are used merely to offer additional observation and commentary in relation to teacher perspectives.

Results

Applying scoring and analytic rubrics

The majority of assessment items required students to provide rather than select answers, as illustrated in Figure 1. Student responses were scored automatically and the scoring algorithms proved robust to idiosyncratic responses by students.

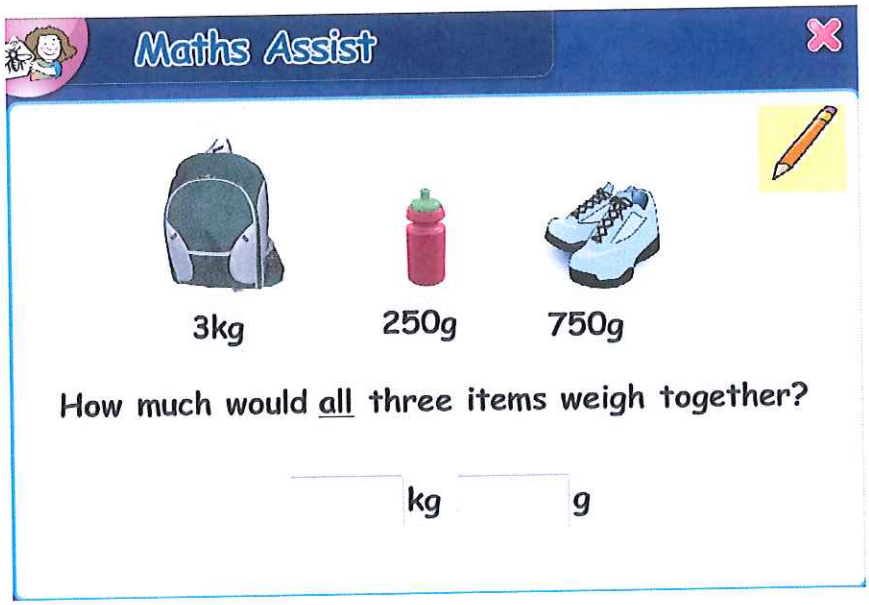

Figure 1 Sample measures item related to 3rd Grade objective: *solve and complete practical tasks and problems involving the addition and subtraction of units of weight (kg and g).*

The mathematical task for students in this item involved the addition of weights and appropriate conversion of grammes to kg. Roughwork paper was available to

students who were required to enter the answer in the boxes on screen using the keyboard. Given the open nature of the item, a range of responses was expected, as highlighted in Table 4.

Table 4
Expected student responses, associated Item Code, polytomous scoring and Error Code related to Figure 1.

Item Code	Description	Sample answer	Score	Error Code[1]
1	Correct	4kg	1	
2	Slip: Minor error	4000kg	0.5	4
3	Incorrect strategy selected	2kg, 2000g, 3250, 3750	0	2
4	Incorrect setup	1003, 1030, 1300	0	3
5	Unable to carry in addition	3900	0	3
6	Error in the procedure	9103, 1070	0	3
7	No conclusion	Other response	0	5

[1] See error codes in Table 1

Responses to the item were expected to fall into one of seven item response categories (item codes). For example, a response of 4000kg (Item Code 2) reflects correct understanding of 1000g in a kg, but possible carelessness in converting the answer back into appropriate units, with the consequent award of partial credit (0.5) and a designation of "*carelessness*" in the overall error codes underpinning the diagnostic model (Table 1). A student response of 2 kg or 2000g suggests choice of an incorrect operation (subtraction instead of addition), thus receiving a score of zero and suggesting, in broader model terms, a strategic error (Error Code 2). Item Code 4 also suggests incorrect setup of the calculation, leading to answers such as 1003, 1030 and 1300 (with consequent designation as Error Code 3), as highlighted by the student roughwork answers in Figure 2, drawn from an administration of the item with a previous cohort.

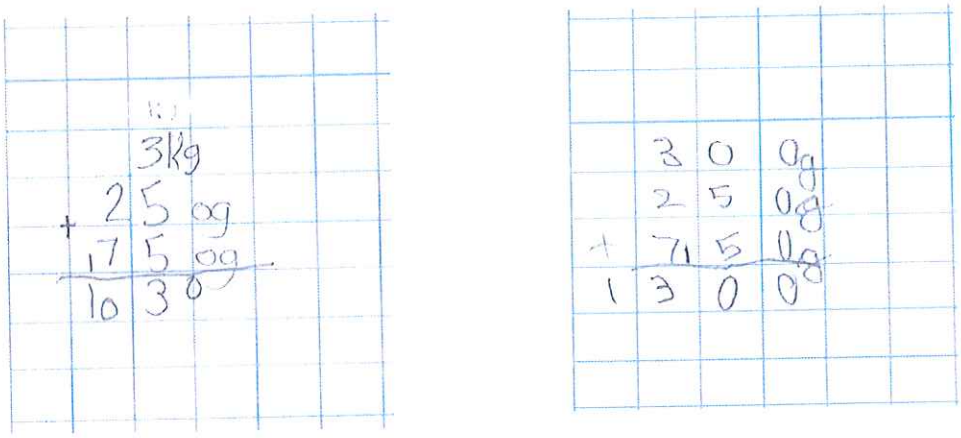

Figure 2 Selected illustrative student responses to weight item in Figure 1.

The automated process underlying the above illustrates the technical capacity of the software to infer mathematical intention from free responses provided by students and embed this in the scoring.

Another example is drawn from the topic Lines and Angles (Figure 3). This example illustrates the capacity of the software to accommodate the inevitable and mathematically-irrelevant mis-spellings in relation to the correct answer (acute).

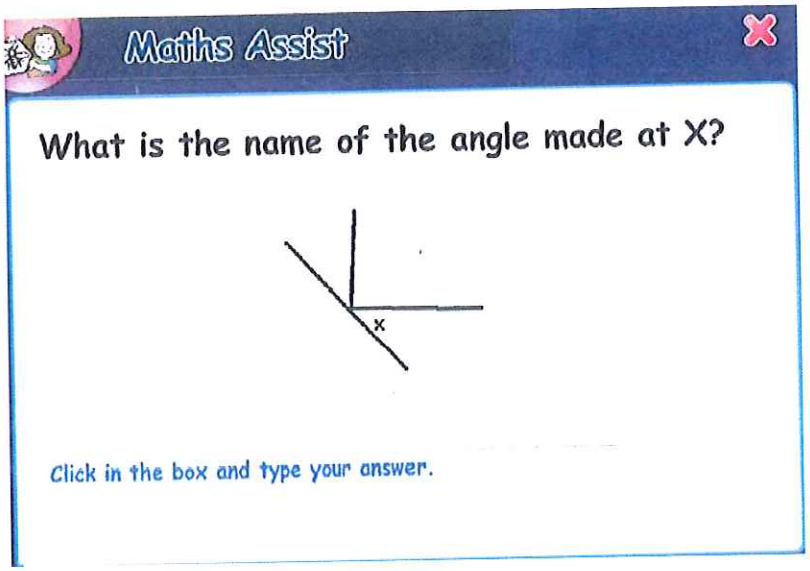

Figure 3 Sample Shape & Space item related to 4th Grade objective: *Draw, discuss and describe intersecting lines and their angles*

The objective called for students to "draw, discuss and describe intersecting lines and their angles". Part of the interpretation of this objective within the curriculum involves students identifying acute, obtuse and right angles, the approach taken in

Figure 3. Responses to the item were expected to align with one of 5 item codes, as indicated in Table 5.

Table 5
Expected student responses, associated Item Code, polytomous scoring and Error Code related to Figure 3.

Item Code	Description	Sample answer	Score	Error Code[1]
1	Correct	Acute (or variants of spelling)	1	
2	Slip: None allowed	-	0.5	4
3	Misread diagram	Straight angle, right angle	0	1
4	Incorrect strategy: answered in degrees	Child answered using degrees	0	2
5	No understanding of acute angles	Any other answer	0	1

[1] See error codes in Table 1

As an item designed around measuring relatively simple conceptual understanding, it was expected that incorrect responses would largely reflect this skill dimension, as represented in Item Codes 3 and 5. Item trialling had indicated some instances of students attempting to respond in terms of degrees, and this is reflected in Item Code 4, with a corresponding Error Code designation of strategy selection, though of course this might also be considered a careless mistake (Error Code 4).

Overall results

Teachers had full flexibility in administering as many topics/testlets as they wished to their students and there was considerable variation in test response patterns as a result. All data and corresponding statistics presented in the paper, therefore, are on the basis of the tests taken by students, where the numbers of students taking any particular test varied. Overall statistics indicate a test programme of moderate difficulty (full scale p-values .59 and .62 in both grades, as indicated in Table 6.

Table 6
Descriptive statistics on tests: Overall and by Skill level. Data on % scale 1-100.

	Grade 3		Grade 4	
	Mean Score	**Standard Dev**	**Mean Score**	**Standard Dev**
Total test (All items taken)	59	18	62	19
Conceptual understanding	62	18	61	26
Using procedures	54	27	68	21
Reasoning & problem solving	56	23	56	25

The mean score on the test across all students and items in relation to Grade 3 was 59%, with a standard deviation of 18%, with broadly similar results for 4th Grade. Data reported are in relation to raw scores as the data scores were not normalised. Therefore, caution is required in over interpreting mean differences in skill level p-values as presented in Table 6. The raw score data show that items measuring students' conceptual understanding were relatively easier that items measuring procedures and reasoning in 3rd Grade, with a somewhat different pattern in 4th where procedural items were easier.

Table 7 presents summary statistics for both grades by mathematical topic. Again, caution is urged in relation to the non-standardised nature of the student scores. The results highlight a range of student performance on the different topics, bearing in mind that not all students took tests on all topics. From the data we see that students performed well on some topics: division, number patterns, lines and angles, weight and data in 3rd; place value, number sentences, lines and angles, and time in 4th Grade. More difficult topics were evident also: fractions in both grades, decimals and symmetry in 3rd, for example.

Table 7
Descriptive statistics on tests by Topic. Data on % scale 1-100.

Mathematical Area	Topic	Grade 3		Grade 4	
		Mean score	Standard Dev	Mean score	Standard Dev
Number	Place value	57	14	88	14
	Addition & subtraction	63	25	54	28
	Multiplication	58	22	69	24
	Division	71	22	52	25
	Fractions	49	30	51	30
	Decimals	54	17	63	24
Algebra	Number sentences	75	23	88	28
	Number patterns & sequences	77	17	93	08
Shape & space	2-D shapes	66	19	69	14
	Lines and angles	70	17	82	19
	Symmetry	50	23	79	26
Measures	Time	60	16	83	22
	Money	68	25	63	21
	Weight	83	19	50	25
Data	Representing & interpreting data	77	14	75	23

Reliability of the tests

Students took different patterns of topics, so efforts to calculate overall reliability estimates for the two consolidated tests are problematic. Given the topic-by-topic nature of the test administration, individual topic reliability is more relevant. As the tests are not designed as speeded tests, with no suggested time limits for teachers to apply (as verified in data presented later in Tables 9 and 10), we felt that internal consistency reliability estimates were appropriate. Cronbach's Alpha was used to estimate the extent to which the items in a topic assess homogeneous content and skills and indices are presented in Table 8 for Grades 3 and 4.

Table 8
Cronbach's Alpha reliability estimates for topic tests

Mathematical Area	Topic	Grade 3		Grade 4	
		Alpha	# Items	Alpha	# Items
Number	Place value	.35	6	.30	5
	Addition & subtraction	.56	5	.16	3
	Multiplication	.75	11	.84	12
	Division	.69	10	.72	8
	Fractions	.79	9	.80	9
	Decimals	.50	9	.65	9
Algebra	Number sentences	.51	6	.93	8
	Number patterns & sequences	.45	8	.15	3
Shape & space	2-D shapes	.65	10	.20	10
	Lines and angles	.30	7	.58	8
	Symmetry	.39	6	.81	7
Measures	Time	-.06	6	.7	7
	Money	.68	5	.62	6
	Weight	.48	5	.67	8
Data	Representing & interpreting data	-.11	7	.77	8

The tests were aligned with the topics and objectives in the national mathematics curriculum, with the consequence that there was not uniform length to the tests and, as highlighted earlier in Table 3, numbers of objectives and items varied from topic to topic. This has implications for the reliability of the scales, as demonstrated in the data in Table 8. In estimating reliability, some items were deleted from the calculations due to zero variance observed in the data. Many of the alpha estimates are encouragingly high (Grade 3 fractions, multiplication, division, money; Grade 4 number sentences, multiplication, symmetry, fractions, data etc). Others are significantly lower (for example place value in both grades), with two particular challenges in relation to time and data in Grade 3. These latter data are in need of further analysis in relation to the negative covariances between items within the

relatively short scales (6 and 7 items respectively), though the small numbers of students responding to these topics may be relevant also.

What we can draw from these reliability data lends some support to the potential to develop high-quality, consistently functioning items within a diagnostic CBA. Though a crude measure, the average of the Grade 3 reliability estimates (omitting time and data) is .55, with five of the estimates greater than 0.6. The average in 4th Grade is .59, with 10 estimates greater than 0.6 and seven greater than 0.7. Given the role of each CBA as part of a larger and continuous testing programme accessed by students, it can be argued that the high levels of reliability normally expected in the case of one-shot summative standardised tests may not be as necessary, in keeping with the analyses of Linn & Miller (2005) and Nitko & Brookhart (2007). The latter's suggestion that "*the more important and the less reversible is the decision about an individual based on the assessment instrument, the higher the reliability should be*" (p. 81) is noteworthy. The *Maths Assist* programme is a process whereby teachers receive continual data in relation to their students, thereby reflecting a very different instrument from the more traditional types of standardised multiple-choice achievement tests.

Use of the Error Codes

Application of the error coding to student responses facilitated compilation of data reflecting the distribution of incorrect student responses across the designated error codes. Scoring software captured all data inputted by students during the testing session, including information about the number of attempts made at an item, skipped items, items to which students returned etc. These data provide a rich array of information in relation to student behaviour on the tests. The filtering of student responses into different predetermined categories provides an overall profile of responses distinctly different to summed results or averages based on dichotomously or polytomously scored procedures. Tables 9 and 10 presents a summary of the allocation of student incorrect responses to a range of error categories for Grades 3 and 4 respectively. These data reflect students' performance (i) across all items taken and (ii) across the three subsets of the tests, namely, the skill areas of *knowledge and understanding, using procedures* and *reasoning and problem-solving.*

Table 9
Distribution of incorrect student responses across error categories: (Percentages of student responses in different categories). Grade 3

Error Code	Description	Total Test	Understanding & Recalling	Using Procedures	Reasoning & Problem Solving
1	Concept understanding	47	62	35	33
2	Strategy selection	6	7	5	16
3	Method execution	14	6	25	17
4	Carelessness	6	4	10	4
5	No conclusion	11	<1	20	20
6	Viewed but unattempted	13	20	4	9
7	Not reached	<1	<1	0	<1

Overall, the Grade 3 results show moderate success for students on the items on the test as a whole and for the three skill areas. Of total student responses across all items, 61% indicated no error, with corresponding figures of 64, 59 and 58 percent respectively for the three skill areas. With 61% of all the student responses on the tests correct, it is in relation to the 39% incorrect responses remaining that the data in Table 9 offer insight. Of those incorrect responses, the distributions across the seven Error Codes reveal interesting patterns. On the total test, just under half of the errors loaded on concept understanding, with lesser loadings on strategy selection, method execution and carelessness. In 11% of cases, the automated software was unable to assign student responses to any of the codes, so "*no conclusion*" was drawn. It was possible to quantify that in 13% of instances where students did not get the correct solution, they had viewed items, but did not provide an answer and only in a handful of cases did the student not reach all the items (Error Code 7).

Of the erroneous responses to the 46 items specifically measuring understanding and recalling, 62% were identified as relating to concept understanding. Of the errors made on the 38 items designed to assess students' capacity in using procedures, one quarter were allocated to the method execution error code, with over one third (35%) relating to concept understanding. Errors were distributed widely across reasoning and problem solving items also, with 16% allocated to strategy selection, and again one third relating to concept understanding.

What these data indicate is that errors are distributed across different categories and that some expected patterns are evident, for example, the heavy loading of

concept understanding errors on items designed to measure students' understanding and recalling and a non-trivial level of carelessness by students, especially where they were required to follow pre-learned mathematical procedures (20%). Other patterns seem less clear, though the variety of knowledge and skills required by students when solving problems may well be reflected in the broad distribution of errors in that skill area.

In Grade 4, a slightly higher percentage of responses were correct overall (64%), with values of 69, 64 and 59 percent for the three skill areas respectively. Table 10 focuses on the remaining 36% of the responses that were incorrect. Similar patterns emerged in Grade 4, though some differences were evident also. Errors relating to concept understanding and method execution dominate across the total test. The pattern on items measuring understanding and recalling are broadly similar to that in Grade 3, with similarities also in items measuring using procedures, though with more evidence of carelessness. Method execution (37%) appeared as a more dominant error category in reasoning items at 4th Grade. Overall, errors attributable to (avoidable) carelessness are clearly evident in the data. At both grade levels, there are quite a number of errors that evaded definitive categorisation by the software and that were, consequently, designated *No conclusion*. Clearly this is an area for future development and refinement.

Table 10
Distribution of incorrect student responses across error categories: (Percentages of student responses in different categories). Grade 4

Error Code	Description	Total Test	Understanding & Recalling	Using Procedures	Reasoning & Problem Solving
1	Concept understanding	34	65	38	12
2	Strategy selection	10	6	6	14
3	Method execution	28	13	28	37
4	Carelessness	11	10	19	6
5	No conclusion	12	2	2	23
6	Viewed but unattempted	6	3	6	8
7	Not reached	<1	<1	1	<1

Tables 11 and 12 show in a more granulated way for separate topics the patterns of student responses as categorised using the error coding. In contrast to Tables 9 and 10, the following tables indicate within the tables the percentages of responses that were correct (Error Code 0).

Table 11
Distribution of student responses across Error Codes. Percentages of student responses in different categories, by Strand Unit, Grade 3.

	Error Code[a]	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number	Place value	57	8	7	3	0	1	25	0
	Addition, subtraction	60	0	1	9	7	22	0	0
	Multiplication	58	16	3	10	0	5	8	0
	Division	71	10	3	15	0	1	1	0
	Fractions	48	29	4	6	1	7	5	0
	Decimals	52	16	4	5	5	13	4	0
Algebra	Number sentences	74	4	9	8	1	0	4	0
	Number patterns & sequences	76	11	5	6	2	0	0	0
Shape & space	2-D shapes	66	28	2	2	1	0	1	0
	Lines and angles	70	22	8	0	0	0	0	0
	Symmetry	48	48	0	0	3	0	1	0
Measures	Time	53	6	5	8	14	15	0	0
	Money	63	5	2	19	11	0	0	0
	Weight	82	3	1	3	1	10	0	0
Data	Representing & interpreting data	76	11	1	10	2	0	0	1

Error Code[a]
0 =No error - correct response
1 = concept understanding
2 = strategy selection
3 = method execution
4 = carelessness
5 = no conclusion
6 = viewed but unattempted
7 = not reached

Table 12
Distribution of student responses across Error Codes. (Percentages of student responses in different categories). By Strand Unit, Grade 4

	Error Code[a]	0	1	2	3	4	5	6	7
Number	Place value	87	4	5	1	1	2	0	0
	Addition, subtraction	52	0	3	0	4	38	3	0
	Multiplication	67	5	1	12	6	5	3	0
	Division	46	10	8	18	12	4	1	0
	Fractions	50	20	8	9	2	7	3	0
	Decimals	61	21	1	12	3	1	2	0
Algebra	Number sentences	88	6	6	0	0	0	0	0
	Number patterns & sequences	89	1	0	1	9	0	0	0
Shape & space	2-D shapes	69	30	0	0	1	0	0	1
	Lines and angles	80	12	4	0	5	0	0	0
	Symmetry	77	18	0	0	5	0	0	0
Measures	Time	83	6	0	2	0	4	4	1
	Money	63	5	15	10	1	3	3	0
	Weight	50	11	0	29	1	6	3	0
Data	Representing & interpreting data	75	3	0	13	1	0	6	1

Error Code[a]
0 = No error - correct response
1 = concept understanding
2 = strategy selection
3 = method execution
4 = carelessness
5 = no conclusion
6 = viewed but unattempted
7 = not reached

Conclusions and Significance of the Research

Overall, the analyses lend some support to the potential to accurately capture and allocate errors through automated processing using CBAs. Undoubtedly the complexity of the error coding described in this study would significantly overburden even the most enthusiastic of teachers, a factor that we believe is related to the relative underuse of error analysis approaches in teaching and learning within school systems. This paper offers a glimpse of how such a complicated process might be achievable

through application of CBA. There is some evidence also that the errors expected in items designed to assess different mathematical skills did, in fact, materialise and were correctly categorised, though there is scope and challenge to further refine this process.

Finally, although this is a small sample of students and teachers, there is evidence from some of the participants of enthusiasm about the CBA and its impact on their own planning and teaching and on their students. A small number of the teachers returned brief questionnaires and their responses indicated satisfaction with the nature of the CBA and its ease of navigation for students: *"children really enjoyed the idea of doing the tests online. Very enthusiastic compared to doing written test."* Of course, the potential biasing impact of the Hawthorne effect needs to inform interpretation of findings in this and similar studies of innovative practices, especially those that involve computers. Teachers highlighted the value to them of receiving information in relation to the most prevalent type of error for individual students and for the class as a whole. Illustrative of the challenges still to be resolved in this program and with online tests in general is the comment from a teacher in a small rural school that the *"internet connection was poor and this delayed us a lot."* This frustration with the internet connectivity in her school has the potential, over time, to erode enthusiasm for the CBA itself, exemplified by her observation that the CBA was *"all in all a fabulous programme children enjoyed and benefited hugely from participation."*

The bulk of the error analysis literature in mathematics predates the technological revolution that has embraced education recently. That research endorses the use of diagnostic approaches, sometimes associated with interviews, to explore the specific conceptions and misconceptions that guide student actions in mathematics. This message is, however, juxtaposed with reservations about the heavy workload implications of a granulated diagnostic process. CBAs may help resolve this tension and the present paper outlines how technology may offer a proxy vehicle for such detailed processes. The paper describes and critically examines the potential of a CBA to build on and finally realise the opportunities inherent in error analysis of students' mathematical performance in elementary level, thus addressing mathematical underachievement as one of the major policy imperatives in education systems worldwide.

References

Ananiadou, K. and M. Claro (2009), *21st Century Skills and Competences for New Millennium Learners in OECD Countries*, OECD Education Working Papers, No. 41, OECD Publishing.
doi: 10.1787/218525261154

Bennett, R. J. (2002) Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *Journal of Technology, Learning and Assessment, 1*(1). Available from http://www.jtla.org

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.

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

Borasi, R. (1994). Capitalising on errors as "springboards for inquiry," A teaching experiment. *Journal for Research in Mathematics Education, 25(2), 166-208.*

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

Buffington, P. & Clements, M. (2011). Eliciting mathematics misconceptsion (EM2): a cognitive diagnostic assessment system. Grant description, available online. http://ies.ed.gov/funding/grantsearch/details.asp?ID=1083

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 computerbased assessment in the K-12 setting. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2002.*

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

Dede, C. (2010). Comparing Frameworks for 21st Century Skills. In J. Bellanca & R. Brandt (Eds.), *21st Century Skills* (pp. 50-75). Bloomington: Solution Tree Press.

DES / Department of Education and Skills (2011). Literacy and numeracy for learning and life: The national strategy to improve literacy and numeracy among children and young people 2011-2020. Dublin: DES.

Drucker, H., McBride, S., & Wilbur, C. (1987). Using a computer-based error analysis approach to improve basic subtraction skills in the third grade. *Journal of Educational Research, 80*(6), 363-365.

Eivers, E., Close, S., Shiel, G., Millar, A., Clerkin, A., Gilleece, L, & Kiniry, J. (2010). *The 2009 national assessments of mathematics and English reading.* Dublin: Educational Research Centre.

Engineers Ireland (2010). *Report of task force on education of mathematics and science at second level.* Available online: http://www.engineersireland.ie/public/20100211-mathematics_and_science_at_second_level.pdf

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., and Jenkins, V. (2010). *Getting started with e-assessment. Project Report.* Bath, UK: University of Bath.

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/

Eggen, T. J.H., & Lampe, T. T.M. (2011). Comparison of the reliability of scoring methods of multiple-response items, matching items, and sequencing items. *CADMO, 19*(2), 85-105.

Government of Ireland (1999). *Primary School Curriculum, Mathematics.* 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.

Higgins, J., Russell, M., and 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

Heubner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical assessment, research & evaluation,* 15(3). Available online: http://pareonline.net/pdf/v15n3.pdf

Johnson, M. & Greene, 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 at http://escholarship.bc.edu/jtla

Johnson, R. B. & Onwuegbuzie, A. J. (2004). A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

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

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

Leighton, J. P., and Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education. Theory and applications.* Cambridge: Cambridge University Press.

Linn, R.L, & Miller, D. (2005). *Measurement and Assessment in Teaching, 9th Edition*, Upper Saddle River, NJ: Pearson Prentice Hall.

Lubienski, S. (2011). Mathematics education and reform in Ireland: An outsider's analysis of Project Maths. *Irish Mathematics Society Bulletin, 67* (Summer), 27-55. Available at http://www.maths.tcd.ie/pub/ims/bulletin/

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

Newby, P. (2010). *Research methods for education.* Harlow, Essex: Pearson Education Limited.

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.

NSW Department of Education and Training (nd). *Newmans's prompts. Finding out why students make mistakes.* http://www.curriculumsupport.education.nsw.gov.au/secondary/mathematics/numeracy/countingon/index.htm

OECD (May 2000). *Definition and selection of competencies: theoretical and conceptual foundations (DeSeCo). Background paper* Neuchatel: DeSeCo Secretariat.

Partnership for 21st Century Skills. (nd). *Framework for 21st Century Learning.* Available at http://www.p21.org/storage/documents/1.__p21_framework_2-pager.pdf

Perkins, R., Cosgrove, J., Moran, G., & Shiel, G. (2012). *PISA 2009: Results for Ireland and changes since 2000.* Dublin: Educational Research Centre. Available at 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.*

Project Maths Development Team (nd). *Overview of Project Maths*. Available at http://www.projectmaths.ie/overview.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods and applications*. New York: NY: Guilford Press.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education 10*(3), 279–94.

Rychen, D. and Salganik, L (2003). Highlights from the OECD project Definition and Selection Competencies: theoretical and conceptual foundations (DeSeCo). Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April 21-25, 2003. *ERIC document ED476359.*

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.

Tashakkori, A. & Teddlie, C. (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks: Sage.

White, A. (2008). *Counting on. Evaluation of the impact of the Counting on 2007 programme*. New South Wales: State of New South Wales Department of Education and Training. Available at http://www.curriculumsupport.education.nsw.gov.au/secondary/mathematics/assets/pdf/counting_on/co_eval_2007.pdf