Perspectives in Education 2014: 32(1) http://www.perspectives-in-education.com
ISSN 0258-2236
© 2014 University of the Free State

A model for assessment: integrating external monitoring with classroom-based practice

Caroline Long, Tim Dunne & Gabriel Mokoena

The rationale for the introduction of standards in the United States in the late 1980s was that the quality of education would improve. Assessment instruments in the form of written tests were constructed in order to perform a monitoring function. The introduction of standards and the associated monitoring have been replicated in South Africa. It was intended that these elements would result in a more equitable education across the country. In neither of these countries has this process borne the results expected. The lack of substantive progress may be due to the fact that socio-economic disadvantage and, therefore, opportunity to learn, has not been addressed. It may also be that systemic-type assessments have little meaning for the teachers, and bear little relation to classroom assessments; the perceived lack of relevance resulting in a lack of commitment to the process.

Our education system is in need of serious deliberations about the broad curriculum goals relevant to society, a classroom environment that is conducive to learning and an assessment model that supports instruction. Within the assessment model we propose an instrument design that supports instruction. The assessment model includes a monitoring component, a formative component and a professional development component. We propose an assessment process where general trends can be reported for systemic purposes but also that the results of the individual learner progress obtained through both a monitoring and a formative component, are to be suitable for engagement by teachers. We honour the central teacher role in communicating both the emerging teaching successes and the currently troublesome areas of classroom learning challenges.

Keywords: Assessment, monitoring, formative assessment, Rasch measurement model, standards, mathematics education

Caroline Long
Centre for Evaluation and Assessment, University of Pretoria
E-mail: caroline.long@up.ac.za
Telephone: 012 420 4175

Tim Dunne
Department of Statistical Sciences, University of Cape Town
E-mail: tim.dunne@uct.ac.za
Telephone: 021 650 3643

Gabriel Mokoena
Centre for Evaluation and Assessment, University of Pretoria
E-mail: gabriel.mokoena@up.ac.za

Introduction

The aim of assessment has altered over the centuries from being an instrument only of discrimination and selection to being the means of supporting the potential development of all learners (Gipps, 1994; Black, 1998). The philosophy underpinning the second aim is that all individuals, personally and collectively, make a contribution to society by expressing their individual qualities, and through this contribution fulfil their own and some of their community's needs.

The standards movement in the late 1980s grew from the frustration of the fragmented schooling system in the United States and the desire to engage learners in diverse educational experiences (Wolk, 2012). This desire precipitated a change in educational focus from the ranking of students to the requirement that each student be educated to their potential. It was hoped through stipulating clear standards, providing the opportunity to learn and then assessing whether these standards had been attained, that the quality of education would be improved.

Many hold the belief that assessment embodies great potential for positive change in schools (Black & Wiliam, 1998; Schafer, 2002; Stiggins, 2002; Wiliam, 2009; Matters, 2009). However, where the policy makers and professional educators do not understand the proper use of assessment, it is possible that the 'assessment instruments and procedures that [have been] created for public accountability are doing more harm than good for certain students – they are causing students to learn less, not more' (Stiggins, 2002: 19).

For assessment to live up to its promise, we need to create a productive assessment environment where the emphasis is no longer on accountability in the sense of reward or punishment, but on the central role of assessment to alert the teacher, the learners themselves, the district officials and provincial departments to areas of need. In order to achieve this environment that is conducive to learning, the essential requirement for the assessment programme is to inform instruction.

Here we distinguish, with Schafer (2002), between curriculum and instruction. We note that the curriculum, the set of standards ideally constructed by a committee of experienced and expert teachers does not of itself educate the students. It is through instruction that the learner capacities are developed. The teacher therefore has to interpret the curriculum statements into instructional components through which learners gain proficiency. Schafer (2002) makes the distinction between summative assessment, which is the actualisation of the curriculum (for accountability objectives), and formative assessment which plays a central role in instruction by iterative feedback and informed adjustments. Assessment products and processes of high quality are demanded for both these types of assessment. They each perform specific functions at their respective sites, namely affirmation that the curriculum requirement has been met in the case of summative assessment, and that instruction has been informed and adapted as a result of formative assessment.

For assessment to perform its critical role in education, there is a prior requirement of a model of knowledge development (curriculum for the particular subject domain) and of cognitive development (understanding how learners progress to higher levels of proficiency) (Black, 1998; Long, 2011). This curriculum knowledge has to be made explicit to teachers in the form of curriculum statements of assessment standards. When external assessment instruments and test items, on which their students have been tested, are withheld, teachers need to be sure that their knowledge of the curriculum content domain is adequate or complete when compared with the 'actualised content and performance standards' envisaged by the authorities (Schafer, 2002: 88). A further critical requirement is to articulate classroom assessment products with external assessment instruments primarily for the purpose of informing teachers about the external assessments to which the schools and teachers may be held accountable. This requirement is premised on the notion that the external tests are carefully constructed from the perspective of both concept development and cognitive development.

While much of the literature to which we refer has been from the United States experience, we note that many of those trends and current concerns also arise in the South African educational context.

The South African educational context

South African schools range from highly resourced[i] private schools, where the aspirations of learners and teachers are to achieve world class standards, and moderately resourced rural schools where there are few resources of the type recognised as quality education. The hope and promise for an improved system of education post-1994 is yet to be fulfilled for the majority of learners.

The design and implementation of successive curricula were intended to lead to an equitable and high standard of education for all learners. Each new curriculum was to provide better guidance, within each knowledge domain, that made explicit requirements for an acknowledged proficiency. These standards implied in the curriculum and operationalised in the assessment tasks both at the classroom level and in external assessment programmes, such as provincial and national testing, were to be both the means and the ends for the improvement of teaching and learning in our schools. For the past 15 years testing at provincial, district and national levels, with a peppering of large-scale international and regional assessments, have been put in place on the implicit understanding that the more we test, the more likely it will be that the education system would improve. The reality is that there has been little substantial improvement in the areas of greatest need.

i. Andrich (2009) uses the terms 'highly resourced' and 'moderately resourced' to refer to the resources that enable the school to engage with curriculum matters, for example, highly experienced teachers or schools where time is made available for these purposes.

Failure of standards and monitoring

The most obvious reason for a lack of progress in the United States has been an array of socio-economic factors, that is, the 'the persistent discrimination that condemns low-income learners to mediocre education' (Wolk, 2012: 1). In areas of dire poverty, the efforts to turn schools around have been counter-productive, and the 'assessment instruments and procedures created for public accountability' are possibly doing more harm than good (Stiggins, 2002: 19).

A second related reason may be that teachers faced with socio-economic disadvantage lack the agency that would enable them to understand, interpret and uplift the learners, schools and communities. The sentiment expressed here aligns with the work of Batra (2009) whose vision for teacher education in India is that teacher agency would be the vehicle to transform that society to a true democracy. The extent of such a training programme is indeed vast, covering many components from understanding socio-educational issues to an in-depth mastery of associated subject knowledge.

A third reason may be that the current organisation and planning within American (and our own) public schools is not conducive to teaching and learning. The current arrangement, where time is the constant and learning the variable,[ii] inevitably leads to the situation where a large proportion of the class is left behind (Wolk, 2012). Here the aim of creating a productive assessment environment that accommodates differences among learners may generate attention to the needs of learners at their current phases of development (Stiggins, 2002).

A fourth reason points to the limitations of external assessment. The kind of instruction required to engender deep understanding of a particular knowledge domain may not be the instruction that leads to success on systemic tests. Bennett and Gitomer (2009) warn that the preparation for systemic-type tests may work against the attainment of deep domain knowledge. The tendency to use external or systemic assessment to hold teachers accountable points to a lack of understanding of the purpose of systemic-type assessment, and to misconceptions of their power.

The failure of decades of curricular outlining and the associated disappointment of repeated test administrations has led to the view that a debate and deliberation about the aims of education is necessary. At the *micro* level, the classroom arrangements have to be re-envisaged as sites of learning; at the *macro* level the systemic assessments that are put in place require strong links to instruction. But perhaps also it is necessary at the *meso* level to reconsider the overarching vision and purpose of education in a radical reconstruction. For descriptions of micro, macro and meso levels, see Thijs and van den Akker (2009).

[ii] We understand the importance of engaging with the critical concepts in the knowledge domain. This engagement demands more than covering the elements presented in curriculum documents.

Debate and deliberation

The two aims for assessment, ranking individuals on current proficiency and informed nurturing of the potential of each individual, align somewhat with the two conflicting tendencies in education described by Robert Young (1990), drawing on the work of Habermas. The first tendency is the assertion of a mode of education which 'seeks to meet the more urgent economic and political needs of the nation in its contemporary situation' (Young, 1990: 48). This mode calls for the strict marshalling of the knowledge and skills through defining a curriculum of a narrow or specific rather than a broad and general kind, and defining stringent criteria for passing into a new affirmed status, such as passing the matriculation certificate. This qualified status describes one as able to make a contribution to the society whose current conditions provide the frame of reference for those criteria. The second mode stresses 'the emancipation of the individual, and through the universalization of that emancipation, the development of autonomy-promoting social institutions, nationally and internationally' (Young, 1990: 48).

The vision we hold for an education system is that the three major goals, identified by Biesta (2009), that of imparting skills and knowledge, of socialisation (imparting and sharing cultural and societal values), and of individuation (the encouraging and support of the unique path of the individual learner), are all given substantial weight. The argument here, though not elaborated, is that it is in the context of the quest that each person has for a fulfilled and enriched life, that the motivation for learning skills and for gaining knowledge will be achieved. Here we must envision a developing world where the individual's roles and the institutions they constitute are continuously unfolding, determined not by past educational tradition alone, but also by engaging anew with the traditions of value that support an unfolding vision of a society surpassing an existing one (Osberg, Biesta & Celliers2008).

While the strong knowledge traditions that have gone before us must be transmitted, the way in which the knowledge and skills are reconstituted may be surprising and exciting. It is therefore important to cultivate an educational environment where the potential of each learner is regarded as unlimited and where the particular direction taken by a learner cannot be entirely pre-empted. Likewise, the specific direction of the teacher with her class must embody open-ended options serving the goals of transmission and empowerment, but not losing sight of the individual. While the core curriculum may be defined, within this conceptual space there should be some leeway for the creativity of teachers and learners to mould their individual and collective directions. The mitigating factor against a too narrowly defined curriculum is that the opening up of the global stage inevitably lights up the imagination of different individuals in unimagined ways. The exploring of that global environment will certainly demand that the core skills be mastered, but may also create a new vision of the future.

Ironically, it may be just this attention to the individuals themselves, the acknowledgement of the role to be played in the community, rather than the current preoccupation with knowledge and skills, that motivate the learners to master the relevant knowledge and skills and, surely, the same principle would apply to teachers. We note here that it is the child who learns; the role of the teacher is to motivate that learning.

Improved assessment

Given the backdrop painted here, it is imperative that we consider the questions: What do we want children to learn? How can the curriculum be planned to take them there? Are teachers equipped to implement the curriculum through designing instructional sequences[iii] and assessment instruments? How may learners be motivated to learn?

The answers to these questions are not obvious and will not be elaborated here. In summary, however, we propose a deep understanding of what has to be taught, the maximum development of each individual, extended assessment options with clear and appropriate standards of validity and reliability, designed to serve instruction, and attention to the critical notions of affirmation, motivation and the maintenance of self-worth (Stiggins, 2002: 31).

We present a strategy for providing common information to both the classroom teacher and also the external supporting or monitoring agency within which both the successful areas and the problem areas may be noted. Here we focus on skills and knowledge acquisition within a knowledge framework, and provide an example of mathematics. We place this strategy within a larger model which envisages the interaction of the monitoring function, a formative function within the classroom and a collaborative professional enrichment process. We note the imperative that testing practices and the statistical procedures associated with testing are applied in a transparent way. This transparency principle of good scientific practice also makes explicit the limitations of assessment procedures. Those of us tasked with the development of tests and the subsequent analyses need to retain the awareness that behind all the numbers are people who have the right to know the limitations of the numbers that are assigned to them. The onus is also on the researchers in this field to be receptive to the challenges to current assessment practices, and, where we have insights, to communicate them in such a way that they are understood and open to critique.

It transpires that one of the most fundamental challenges is to convey the truth that test instruments can only be regarded as measurements of performances or competences when several strict technical criteria have been satisfied in the design stages and verified in the primary analysis of performance data, prior to any measurement based inferences. This truth has profound consequences for assessment in design and practice.

[iii] This requirement is ambitious, but it is only when the teachers have control of the curriculum and the classroom that optimal leaning will take place.

Assessment resources

In support of the second, more democratic aim of education, that is, enhancing the contribution of each individual, scientific and technical resources are being developed in many research centres across the world (see Griffin, 2009, Bennett & Gitomer, 2009). These resources envisage for every discipline the provision of valid assessment instruments which are appropriately targeted for the group to be tested. In this way valid discriminations emerge between performances. These discriminations may also admit social purposes such as verifying attainment of pass and distinction criteria. Given proper subject expert inputs, and the use of important new theory for design and analysis of tests that have measurement-like properties, it is now possible to obtain precise numerical descriptions (measurements) of student performance and progress in a particular subject area, hence providing maximum information for efficient planning of educational programmes that are aligned to instruction (see Griffin, 2007, 2009; Long, 2011). This process, we believe, will contribute somewhat to the correction and amelioration of the vast discrepancies we find in the South African learning environment, through setting achievable goals that enable the teacher to extend each learner incrementally in the direction of greater proficiency.

Researchers and practitioners in education will recognise the importance of deep reflection into the principle features of a knowledge domain, for example, the exploration of the distinctive nature of mathematics (as presented in the work of Dantzig, 2007, and others). We should ask: “What are the knowledge and skills inherent in this domain that are critical to learners firstly for engagement with their everyday lives, and secondly, for the subsequent trajectories of those students?” Such trajectories may require that some learners are suitably equipped to enter advanced mathematics courses. Similar imperatives emerge in other disciplines.

Following the important phase of reflection, the scientific and technical skill of constructing items which operationalize aspects of the knowledge domain is critical. Here it is important to note that the curriculum as currently exemplified in curriculum documents is not sufficient for deciding on test items. A deeper engagement with curriculum is required to determine something of a plausible line of development within the knowledge domain. This view of domain development necessarily includes reflection on which underlying insights and skills are critical for learner progress in the domain. In this context, items and assessment questions are constructed so as to embody the domain, and extract evidence of both learner competence and current learner inadequacies.

Then, because the knowledge of the researchers and test designers does not guarantee that they are beyond making errors or sub-optimal inferences, a piloting phase is required to establish whether or not the current form of the test as a whole appears pitched at the correct level for its stated objectives. This question always involves examining whether or not the constituent items are functioning as expected by the designers. The design imperative is to obtain and report optimum information for the stakeholders involved - the schools and their teachers, the associated education department and the district officials. This phase of the test design is critical, and in this phase the input of teacher specialists can and should provide essential insights which impact on the test design. The function of a test includes not only the designation of content areas mastered, but also the identification of the elements that learners currently find obscure or challenging.

Underpinning this test design process is the notion that we are attempting to 'measure' proficiency within a field of knowledge, for example, mathematics. Applying the rigour of the classical theory of measurement to social sciences, and education, and in the construction of valid instruments, is not only in the interest of good science but also in the interest of social responsibility (Andrich & Marais, 2008). The scientific demand is not simple nor is it easily attained. Especially where assessment is used to identify individual differences or to order students, the notion of a more scientific approach to measurement is critical. As Thorndike noted in 1904 (in Wright, 1997), even the assessment of spelling raises issues of comparability and fairness in that there is no unit of spelling proficiency.

Where as in the physical sciences characteristics of objects such as length, mass, time and their many derivatives are tightly specifiable, in the social and educational sciences the characteristics of proficiency in a domain are much less tractable. To presume to make measurements in educational, psychological and sociological domains, there must first be an explication of how the domain is defined and understood. It is necessary to have an explicit context so that the validity of any instrument can be adjudicated as its fitness for a stated purpose in that context, by any appropriate peer community. Then the question of the extent to which an instrument achieves measurement criteria leads to insights into its reliability and utility.

In this quest to provide valid, reliable and useful instruments a resource of note is Rasch measurement theory (RMT) in which the mathematical model is underpinned by the invariance of measurement . The Rasch model requires that several measurement criteria are met in the data. Where these criteria are not satisfied, the particular items, and the associated learner responses, must be investigated further, along with a re-examination of the construct underpinning the test instrument. For details on this process, see Rasch (1960/80), Wright and Stone (1979), Andrich and Marais (2008), and Dunne, Long, Craig and Venter (2012).

A study in mathematics education

A research study applying the principles and processes of RMT, conducted in mathematics education at the Senior Phase and the Further Education and Training (FET) Phase, is currently in progress. Aspects of the study are presented here merely for the purpose of demonstrating the particular assessment model that is being advocated, and not for reporting on the study concerned.

The external priority for this project was to monitor learner progress over successive years. In order for this monitoring process to be valid and reliable, and to inform teaching and learning, several pre-conditions had to be in place.

The first requirement in this study was to find a degree of consensus within the reference group, constituted to include representative mathematics teachers and assessment specialists, about the critical and central concepts within the particular grade for which the assessment was intended. An investigation was conducted of the current and pending curricula, and in addition, a comparison was made between the local curriculum and the curriculum frameworks underpinning the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Australian Curriculum (see discussion document, in process).

In addition to the content to be explicitly included in the instrument, discussions were held concerning the response processes required for the various content areas at distinct levels of complexity. The common approach to identifying the response processes or cognitive domains is to use variations of Bloom's taxonomy (1956). Here, we found this approach inadequate for identifying the nuance of what was required for mathematics. For example, the levels for algebra response processes, while increasing in complexity, were defined differently from the geometry questions which, while requiring core mathematical reasoning skills, also demand an understanding of reasoning and justification.

It is at this stage of prospective item design and analysis that the involvement of teachers is necessary to establish common ground about what constitutes critical mathematics knowledge for the assessment context. Black (1998) avers that assessment should be aligned with instruction. To ignore this requirement renders the instruments somewhat insensitive to the context, and at worst the information obtained may be irrelevant. This type of teacher involvement can enrich the meaningfulness of instrument use and assessment outcomes.

It is important to note here that the Rasch model has specific requirements for the vindication of a claim of measurement-like properties within an instrument. By applying the model to a set of data, anomalies within the data will be highlighted. These anomalies are then checked by the subject experts for explanations. The source of any problems may be in the items themselves, in the learners, or in an unforeseen extreme difficulty with an aspect of the construct being tested (Smith & Plackner, 2009). This scrutiny of items and test occurs at both the pilot phase of testing and after the administration of the final instrument.

Person-Item map

An output feature of the Rasch measurement model is that both the item difficulty and the learner proficiency are aligned on the same scale. As one may observe (see Figure 1, Learner-Item map), the item difficulties located on the right side of the figure with their mean location set at zero, range from about –3 logits (relatively easy) to +2 logits (relatively difficult). For the technical details of how the items are calibrated see Wright and Stone (1979), Andrich (1988) and Dunne *et al.* (2012). The learners are implicitly depicted on the left side of the figure, where each cross represents two learners. Here the learner locations are fairly well spread across the items at the lower end of the scale.

The nature of the alignment of learners and items on the same scale is probabilistic. Learners located opposite a particular item are interpreted to have a 50% probability of correctly answering that item. This subset of learners share a common probability higher than 50% of being correct on each item below their aligned items, and less than 50% for any item above their aligned items. The probabilities of correct responses shared by the aligned group increase with the item distance below (and are easier than) the aligned items, and decrease with the item distance above (and are more difficult) than the aligned items. These model probabilities follow patterns determined by the logistic function.

```
        PERSONS [locations=estimated proficiency]          ITEMS [locations=estimated difficulty]
                                      |N24 N25
                                      |
                                      |
                                      |N23 D46
                                      |
                                      |
                                      |N22
 1                                    |
                                      |
                                      |N21
                                     X|
                                     X|N20 D15
                                     X|D14
                                     X|D11 D12 D13
                                     X|
                                    XX|N19
                                    XX|N18
                                    XX|N17
0                                  XXX|D08 D09 D10
                                  XXXX|D07 N16 33
                                   XXX|N14 N15
                                  XXXX|N13
                                XXXXXX|
                                 XXXXX|N12
                                XXXXXX|N11 N10
                              XXXXXXXX|
                            XXXXXXXXXX|N09
                              XXXXXXXX|D06
                            XXXXXXXXXX|
-1                             XXXXXXX|N07 N08
                              XXXXXXXX|D05
                               XXXXXXX|N06
                              XXXXXXXX|D04 N05
                               XXXXXXX|
                             XXXXXXXXX|D03 N04
                               XXXXXXX|
                              XXXXXXXX|
                                   XXX|N03
                                  XXXX|N02 D02
                                XXXXXX|
-2                                 XXX|
                                    XX|
                                   XXX|
                                    XX|
                                    XX|
                                     X|
                                      |
                                      |N01
                                      |
                                      |D01
-3
Each 'X' represents   2 cases
```

Figure 1: Person-Item map of learners' proficiency and item difficulty on a common scale

We see from the vertical scale that there are several items in this test above 0.5 logits, for which no learners are estimated to have a 50% chance of answering correctly. In fact the higher the item is located up the scale, the smaller the chance of even the highest achieving learners answering correctly. Although this model is probabilistic, and there is always some notional small chance of answering a difficult item correctly, the empirical information provided here is that the items above 2.0 logits are just too difficult to provide useful information for discriminations within this cohort of learners. For a next cycle some of these items would be adapted or removed, as they provide insufficient information to distinguish between the learners in this cohort. The same items might well be suitable for other cohorts.

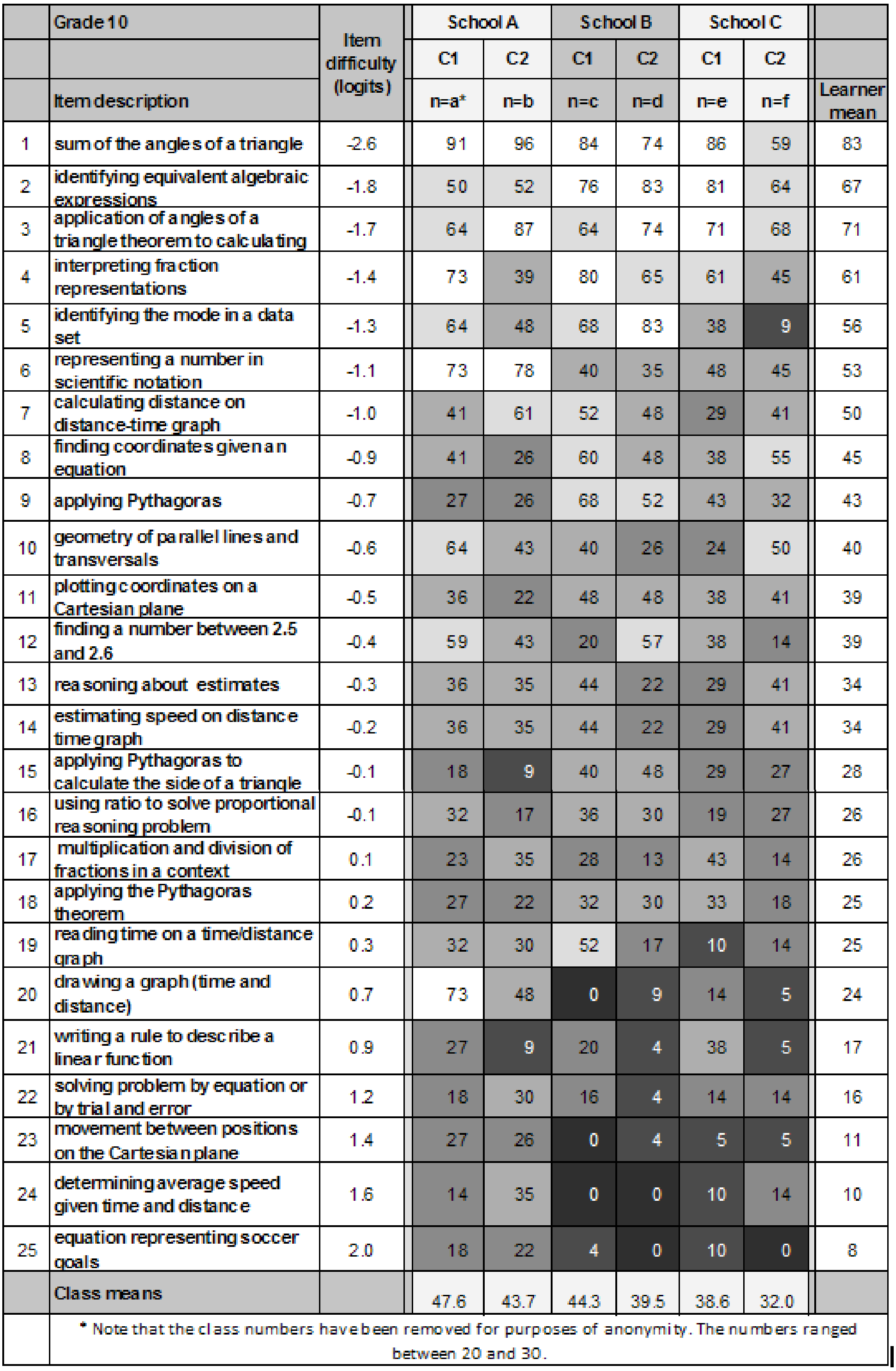
The importance of aligning the array of item difficulty locations with the array of learner proficiency locations is that when this alignment has been achieved, then optimal levels of discriminatory information may be obtained about learner competences. The instrument of this study clearly needs refinement before any future administrations to the cohorts of interest here. This outcome, diagnosing potential improvements towards increased fitness for purpose, we interpret as an advantage and not a deficiency of the test design process.

An item–class percentage correct chart

An outcome of interest as a result of the analysis was the construction of a matrix depicting the relationship between the items and the schools with their individual classes. An explanation follows, showing three schools, School A, School B and School C, each with two classes (C1 and C2) (see Table 1). The item numbers are listed in Column 1; the constructs tested are described in Column 2. The first item (labelled Item 1) involves knowing the *sum of the angles of a triangle*. The second item (Item 2) involves *identifying equivalent algebraic expressions*. In Column 3 the overall item difficulty (as estimated by the Rasch model) is reported. Note that the first item shown has a difficulty of –2.6 which denotes relatively easy in its context, and the second item has a difficulty location of –1.8, which while relatively easy compared with the complete set of items, is somewhat more difficult than the first item.

In the cells (for item by class) the percentage of learners with correct responses is shown. Note that these results have been adapted somewhat for illustrative purposes. The numbers in the classes were around 25. Focussing only on the first two rows of results, we observe a pattern in School A. While over 90% (about 23) of the learners have the simple geometry item correct, only around 50% (12) have the algebra question correct. Then, in School B, Class 1, some 84% (21) answered correctly for the geometry item, and 76% (19) for the algebra item. In School B, Class 2, the percentage correct for each of these items is reversed, a relative 74% (18) of learners answered the geometry item correctly, and 83% (81) answered the algebra item correctly. In School C, Class C1 appears more proficient than Class C2 when judging only on this information. The last column on the right reports the overall percentage of the whole cohort of this study who answered the item correctly. We may compare the individual class percentages with the overall percentage for each item as a reference point.

Table 1: Item by class matrix of percentage correct correct



	Grade 10	Item difficulty (logits)	School A		School B		School C		
			C1	C2	C1	C2	C1	C2	
	Item description		n=a*	n=b	n=c	n=d	n=e	n=f	Learner mean
1	sum of the angles of a triangle	-2.6	91	96	84	74	86	59	83
2	identifying equivalent algebraic expressions	-1.8	50	52	76	83	81	64	67
3	application of angles of a triangle theorem to calculating	-1.7	64	87	64	74	71	68	71
4	interpreting fraction representations	-1.4	73	39	80	65	61	45	61
5	identifying the mode in a data set	-1.3	64	48	68	83	38	9	56
6	representing a number in scientific notation	-1.1	73	78	40	35	48	45	53
7	calculating distance on distance-time graph	-1.0	41	61	52	48	29	41	50
8	finding coordinates given an equation	-0.9	41	26	60	48	38	55	45
9	applying Pythagoras	-0.7	27	26	68	52	43	32	43
10	geometry of parallel lines and transversals	-0.6	64	43	40	26	24	50	40
11	plotting coordinates on a Cartesian plane	-0.5	36	22	48	48	38	41	39
12	finding a number between 2.5 and 2.6	-0.4	59	43	20	57	38	14	39
13	reasoning about estimates	-0.3	36	35	44	22	29	41	34
14	estimating speed on distance time graph	-0.2	36	35	44	22	29	41	34
15	applying Pythagoras to calculate the side of a triangle	-0.1	18	9	40	48	29	27	28
16	using ratio to solve proportional reasoning problem	-0.1	32	17	36	30	19	27	26
17	multiplication and division of fractions in a context	0.1	23	35	28	13	43	14	26
18	applying the Pythagoras theorem	0.2	27	22	32	30	33	18	25
19	reading time on a time/distance graph	0.3	32	30	52	17	10	14	25
20	drawing a graph (time and distance)	0.7	73	48	0	9	14	5	24
21	writing a rule to describe a linear function	0.9	27	9	20	4	38	5	17
22	solving problem by equation or by trial and error	1.2	18	30	16	4	14	14	16
23	movement between positions on the Cartesian plane	1.4	27	26	0	4	5	5	11
24	determining average speed given time and distance	1.6	14	35	0	0	10	14	10
25	equation representing soccer goals	2.0	18	22	4	0	10	0	8
	Class means		47.6	43.7	44.3	39.5	38.6	32.0	

* Note that the class numbers have been removed for purposes of anonymity. The numbers ranged between 20 and 30.

Of course, percentages should not be reported uncritically, especially when the entire reference group is less than 100 in number, as in each of these classes and schools. The point of their use is to allow bridging between new messages that can be discerned from the table and familiar data descriptions and summaries. We note that the point of the table is not to infer any ranking of classrooms, but to explore a new form of engagement with the data in order to derive classroom-based insights into learning and teaching.

Some 25 items from the study (an arbitrary selection for purposes of demonstration) are presented in Table 1. The item by class cells are colour coded according to the following descriptive sequence: At least 70% (unshaded); from 50% to 69.9% (light grey); from 30% to 49.9% (medium grey); from 10% to 29.9% (dark grey); and 0% to 9.9% (darker grey).

The analysis of the table will not be presented here. However, we observe that one may identify similarities and differences across the 3 schools, across their 6 classes, and across pairs of classes within schools. It is also interesting to note that while an overall mean percentage is reported for each class, it is very difficult to categorically assert the mathematical superiority of any class, as their strengths vary from item cluster to item cluster. Another point of caution is that, when constructing a test with relatively few items, say 50, but attempting to cover an entire curriculum, the consequence of the design is that one has very little information within each particular item alone. What is more pertinent is for researchers together with teachers to use such a matrix to identify points of interest on which to have further discussions. These discussions may focus upon learner needs or instrument improvements.

We are able to establish proficiency zones by drawing on the alignment of persons and items on the same scale (see Person-Item map, Figure 1), and from an analysis of arbitrary but convenient more homogeneous sections of the test, for example establishing bands of one logit width, from –3.00 to –2.00 (very easy), from –2.00 to –1.00 (easy) and so on (see Table 2). We note that the item locations are continuous on the scale and that the use of integers to make label distinctions between the proficiency zones is merely for purposes of analysis, rather than inherently appropriate.

In general it may be noted here that the message from the data on learner performance is that Level 1 items require recall and some procedural knowledge, whereas Level 5 items require reasoning about the co-variation of variables. This message is independent of the criteria by which they were selected for inclusion. Hence it is also the justification for making a mathematical statement, based on knowledge of mathematical properties of items. Given the current emphasis in the curriculum documents, it is possible that the Level 5 skills do not receive enough attention in most classrooms.

Table 2: Summary skills audit by current proficiency level from high to low

Proficiency zones	*Location*	*Knowledge and skills*	*Performance description*

5 (greater difficulty)	[1.00 to 3.00)	knowledge of rational numbers;
		scientific notation;
		co-variation of variables;
		time and distance;
		reasoning and justifying
			The cluster of items at this level requires critical mathematical skills such as mathematical reasoning. There are pockets where schools have shown some proficiency, for example, School B, on formulating solutions for a problem, reasoning about statistical concepts and determining average speed

4 (difficult)	[0.00 to 1.00)	estimation skills;
		scientific notation;
		recognising the rule determining a linear function;
		understanding a time-distance graph
			This cluster of items proved to be very difficult for this cohort. There are pockets of relative success, for example, where, for School A, 70% of learners answered the distance-time graph question correctly. (See table 1.)

3b (moderately difficult)	[-0.50 to 0.00)	ordering rational numbers; applying ratio concepts to proportional reasoning problems; estimating speed on a distance-time graph; knowledge of basic geometry (angles of a triangle); applying Pythagoras' theorem	The most difficult items at this level are knowledge of a rational number and of Pythagoras' theorem. Pockets of difficulty were experienced with determining the next number in a geometric sequence
3a (moderately easy)	[-1.00 to -0.50)	knowledge of decimal fractions; understanding the Cartesian plane; recognising patterns within a linear function; applying Pythagoras' theorem	There appear to be difficulties with decimal fractions and with geometry applications

2 (easy)	[-2.00 to -1.00)	knowledge of place value; fraction notation and scientific notation; elementary statistics concepts	In some classes it appears that the topic of elementary statistics has not been covered. Scientific notation requires attention in some cases
1 (very easy)	(-3.00 to -2.00)	the sum of angles of a triangle, and applying this knowledge; naming the hypotenuse; identifying algebraic expressions of the same value	Proficiency across classes ranges from above 90% for an easy item to a relatively challenging item for one of the classes, around 50%

The implication for teaching is that suitable levels of current proficiency can be interpreted for clusters of learners whose proficiencies on the test appear to be aligned with a corresponding item level as identified in this study (see the Person-Item map).

In this analysis we invoke the concept of the zone of proximal development (ZPD) (Vygotsky, 1962), the cognitive space in which learners may receive optimal tuition but where the concepts are perhaps too difficult for learners to manage on their own. It is in this zone, where concepts appear currently neither so difficult as to be intimidating nor so easy as to be boring, that tuition may achieve its best results. This analysis by learner group does not necessarily imply group teaching, although it may be educationally efficient to design different educational experiences for each of the groups at some points along their learning trajectories.

Approaches to teaching are not discussed here, except to acknowledge that information of this nature elicited from the test may already be known to the proficient teacher. However we aver that with carefully crafted tests, insights of this sort may generate further discussion about the mathematics involved at different proficiency zones. An additional feature of this output is that the instrument, while providing some provisional information, may be revised and improved so as to inform teaching in more pertinent ways.

We also concede here that the information obtained from a test is directly related to the extent and quality of the theoretical work that preceded the construction of the test. In the case of this test with about 50 items in total, and with the requirement to completely cover the curriculum, the information may not provide the precision necessary to inform mathematics teaching of more specific curriculum areas. Nevertheless, this type of analysis may be the starting point for more rigorous research of the type required to understand more deeply the complexities of learning and teaching mathematics.

Extending the study

Against the background of the study briefly described above and other studies of this nature, as in Griffin (2007, 2009) and Long (2011), we ask the question "How may a nationwide assessment programme be envisaged, and constructed, such that it involves all the important role players, particularly teachers, in the design of a continuous range of mathematics tests covering Grades R through 12, accessible for regular use by teachers in their own classrooms?"

In envisaging such a programme we are warned against the unintended consequences of accountability systems where, according to Bennett and Gitomer (2009: 46):

> *... the end goal for too many teachers, students and school administrators has become improving performance on the accountability assessment without enough attention being paid to whether students actually learn the deeper curriculum standards those tests are intended to represent.*

An assessment programme designed to avoid a narrow and self-defeating goal should involve three central components, namely an accountability component, where information is fed back into the official channels responsible for the conduct and servicing of education; a professional development component, where the indicated areas of teaching and learning needs are catered for; and a formative development component, where teachers work closely and reflectively within their classrooms and professional learning communities (see Bennett & Gitomer, 2009). These components are conceived as having a common conceptual base that considers the curriculum requirements and the general findings that have emerged regarding effective teaching in particular subject domains.

The programme is based on the following principles:

- The development of proficiency of all learners from their current levels to greater levels of proficiency is paramount. Here we note that the learners currently situated higher on the proficiency scale deserve the educational experiences which challenge them to even greater proficiency, just as the learners currently situated on lower levels of proficiency require attention to assist them to move progressively to higher levels. We note also that varieties of mathematical proficiency may be required for particular vocations, and for particular students, and hence should be considered and explicit decisions recorded.

- A second principle is that teachers are intrinsic to any development plan that could serve the needs of the educational environment. This phase of the programme would involve both a professional development component in the content and pedagogy of mathematics at the grades in which the teachers are working, and a specific component on the construction of items whose objective is to assess the identified critical concepts. A pilot version of this phase has been conducted in collaboration with the mathematics development organisation COUNT with teachers in one district in KwaZulu-Natal (see COUNT-ZENEX Report, 2011).
- A third principle is the necessity of a formative aspect to the plan, in which teachers trial the formative assessment products, reflect and report on the current suitability of the items as well as the proficiency of the learners within the particular cluster of concepts of interest. Also intrinsic to this phase is the teacher reflection on the construct(s) in question, and further research into the particular area of mathematics.
- The principle of inclusiveness with respect to classroom teachers embraces the fact that the current mathematics classrooms are far from optimal, but that support and direction will be necessary. The importance of quality assurance at each site and at each phase of the programme is critical and can be articulated through suitable collaborative feedback loops. This component may be seen as an extension of professional development.

A model, in some respects similar to this proposed model, has been envisaged by Bennett and Gitomer (2009) to improve current systemic assessment programmes in the United States. An idea of relevance here is that, while the putative standards may be specified in curriculum documents, they cannot easily be translated into learner skills and knowledge, unless these same standards are sufficiently specific and extensive within the minds and practices of the teacher cohort.

Conclusion

The crisis in education and of the youth is not restricted to the South African context but is rather a worldwide phenomenon. This ferment may signal that critical elements of the society are debated and deliberated for the benefit of the youth and society as a whole. Young (1990) avers that for the purposes of education, including both transmitting the knowledge, skills and culture from previous generations, and reconstituting of society through vision and creativity, some vision of society to support our ideals is required. This vision is of necessity a collaborative enterprise that involves engagement at all levels of the society, in particular the educational institutions and the youth.

In this paper we acknowledge the important function of carefully crafted assessment. We also acknowledge, with Stiggins (2002), that authority covering assessment for accountability purposes in the hands of people unschooled in assessment practices may indeed cause harm to education systems. We have sought to demonstrate that powerful instruments with a diagnostic value for the classroom are possible. Every such instrument is limited but nonetheless can address specific current learning needs. Systemic assessments can have the same qualities, if both suitable care and planning are given to diagnostic objectives. Moreover we have argued that a culture of classroom assessment can be constructed and that it will support learning development. These initiatives can be configured in ways that also inform and enrich the same processes that systemic assessments serve.

Our assumption is that learning, however it is structured, relies crucially on the classroom. In that context, the teacher is a principal agent, and the role of assessment is to discern and direct current learning strengths and needs. Systemic efforts to enrich and support classroom assessment, conducted and led by a suitably purposeful group of specialists and teachers, will enrich a primary locus at which learning is meant to take place, namely the interaction of teacher and learner.

A shift is required in our beliefs about assessment and its role in achieving effective schools. The belief that reward and punishment are motivating factors has little traction for the teachers or the youth in the 21st century. The requirement rather is to focus on a wide array of assessment types that require students to present their work in the form of portfolios, projects and presentations. The redirecting of funds to classroom support is required, as well as attention to assessment that informs instruction and that affirms the self-worth of individuals which, in doing so encourages learning, is required. In addition, involving students in their own assessment provides a sense of agency over their own learning (Wolk, 2012).

In order for our teachers and educational institutions to play a critical role, the imperatives of education, learning progress and classroom assessment of current learning needs, in contrast with imperatives supporting accountability through systemic assessment, are to be kept in a moving equilibrium. We propose that, by including teachers in the cycles of an assessment programme, rather than making them and their learners merely the objects of such an exercise, we position the teachers in their rightful place as agents. They are not only transmitting the knowledge, culture and skills of previous generations, but are also agents of the reconstitution of society through a creative engagement with knowledge, culture and skills by the learners themselves.

References

Andrich D 1988. *Rasch models for measurement*. Newbury Park: SAGE Publications.

Andrich D 1989. Distinctions between assumptions and requirements in measurement in the social sciences. In JA Keats, R Taft, RA Heath & SH Lovibond (eds.). *Mathematical and theoretical systems*. North Holland: Elsevier Science Publishers.

Andrich D 2009. *Review of the curriculum framework for curriculum, assessment and reporting purposes in Western Australian schools with particular reference to years kindergarten to Year 10*. Perth: University of Western Australia.

Andrich D & Marais I 2008. *Introductory course notes: Instrument design with Rasch, IRT and data analysis*. Perth: University of Western Australia.

Batra P 2009. Teacher empowerment: The education entitlement – social transformation traverse. *Education Dialogue*, 6(2): 121-156 .

Bennett RE & Gitomer DH 2009. Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional development. In C Wyatt-Smith & JJ Cumming (eds.), *Educational assessment in the 21st century*. Dordrecht: Springer.

Biesta G 2009. Good education: What it is and why we need it. *Inaugural lecture*. Stirling: The Stirling Institute of Education.

Black PJ 1998. *Testing: Friend or foe*. London: Falmer Press.

Black, P. & Wiliam, D. (1998). *Assessment and Classroom Learning*. Assessment in Education: Principles, Policy & Practice, 5:1, 7-73.

Bloom BS (ed.) 1956. *Taxonomy of educational objectives*. New York: David McKay.

COUNT-ZENEX Report 2011. COUNT: Johannesburg.

Dantzig T 2007. *Number: The language of science*. London: Plume.

Dunne T, Long C, Craig T & Venter E 2012. Meeting the requirements of both classroom-based and systemic assessment: The potential of Rasch measurement theory, *Pythagoras*, 33(3), 1-16.

Gipps C 1994. *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Griffin P 2007. The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, 33: 87-99.

Griffin P 2009. Teachers' use of assessment data. In C Wyatt-Smith & JJ Cumming (eds.), *Educational assessment in the 21st century: Connecting theory and practice*. Dordrecht: Springer.

Long MC 2011. The mathematical, cognitive and didactic elements of the multiplicative conceptual field investigated within a Rasch assessment and measurement framework. Unpublished PhD thesis. University of Cape Town, Cape Town.

Matters G 2009. A problematic leap in the use of test data: From performance to inference. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century*. Dordrecht: Springer.

Osberg D, Biesta G & Celliers P 2008. From representation to emergence: Complexity's challenge to the epistemology of schooling. *Educational Philosophy and Theory,* 40(1): 213 – 227.

Rasch G 1960/1980. *Probabilistic models for some intelligence and attainment tests*. (Expanded edition with foreword and afterword by BD Wright). Chicago: University of Chicago Press.

Schafer W 2002. How can assessment contribute to an educational utopia? In R Lissitz & W Schafer. *Assessment in educational reform: Both means and end*. Boston: Allyn & Bacon.

Smith RM & Plackner C 2009. The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, 10(4): 424-437.

Stiggins R 2002. Where is our assessment future and how can we get there from here? In R Lissitz & W Schafer. *Assessment in educational reform: Both means and end*. Boston: Allyn & Bacon.

Thijs A & van den Akker J 2009. *Curriculum-in-development*. Enschede, Netherlands: SLO.

Vygotsky L S 1962. *Thought and language*. Cambridge, MA: MIT Press.

Wiliam D 2009. Assessment for learning: What, why and how? Institute of Education, University of London.

Wolk RA 2012. Common core vs. common sense. *Education Week*, 32(13): 35-40.

Wright BD 1997. A history of social science measurement. *Educational Measurement: Issues and Practice*, Winter: 33-52.

Wright BD & Stone MH 1979. *Best test design*. Chicago: MESA Press.

Young R 1990. *Critical education: Habermas and our children's future*. New York: Teachers College Press.