Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory

Authors:
Tim Dunne[1]
Caroline Long[2]
Tracy Craig[3]
Elsie Venter[4]

Affiliations:
[1]Department of Statistical Sciences, University of Cape Town, South Africa

[2]Centre for Evaluation and Assessment, University of Pretoria, South Africa

[3]Academic Support Programme for Engineering in Cape Town, University of Cape Town, South Africa

[4]Independent researcher

Correspondence to:
Tracy Craig

Email address:
tracy.craig@uct.ac.za

Postal address:
ASPECT Office, EM319.1 Electrical and Mechanical Engineering Building, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

Dates:
Received: 18 July 2011
Accepted: 07 Oct. 2012
Published: 21 Nov. 2012

How to cite this article:
Dunne, T., Long, C., Craig, T., & Venter, E. (2012). Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory. *Pythagoras, 33*(3), Art. #19, 16 pages. http://dx.doi.org/10.4102/pythagoras.v33i3.19

The challenges inherent in assessing mathematical proficiency depend on a number of factors, amongst which are an explicit view of what constitutes mathematical proficiency, an understanding of how children learn and the purpose and function of teaching. All of these factors impact on the choice of approach to assessment. In this article we distinguish between two broad types of assessment, classroom-based and systemic assessment. We argue that the process of assessment informed by Rasch measurement theory (RMT) can potentially support the demands of both classroom-based and systemic assessment, particularly if a developmental approach to learning is adopted, and an underlying model of developing mathematical proficiency is explicit in the assessment instruments and their supporting material. An example of a mathematics instrument and its analysis which illustrates this approach, is presented. We note that the role of assessment in the 21st century is potentially powerful. This influential role can only be justified if the assessments are of high quality and can be selected to match suitable moments in learning progress and the teaching process. Users of assessment data must have sufficient knowledge and insight to interpret the resulting numbers validly, and have sufficient discernment to make considered educational inferences from the data for teaching and learning responses.

Introduction

The assessment of mathematical proficiency is a complex task. The particular challenges inherent in this process depend on a number of factors, including the definition of what constitutes mathematical proficiency, an understanding of how children learn and the approach adopted as to the purpose and function of teaching. Besides these central questions in mathematics education, there are important questions to consider about the relationship between classroom-based assessment and systemic assessment types. Whilst there is potential for positive information exchange between these two types of assessment, more often there is an unnecessary conflict or simply a lack of constructive communication. Classroom teachers are at times perplexed by the outcomes of systemic assessment, confused about what action to take as a result of the reported outcomes and, in the worst-case scenario, demoralised. The quest for positive information exchange demands that questions about quality at both classroom and systemic sites are addressed (see also Wyatt-Smith & Gunn, 2009, p. 83).

In this article, we differentiate explicitly between the two broad types of assessment: classroom-based and systemic (or external) assessment.[1] The rationale for assessing, the demands of the stakeholders, the forms of the assessment instruments and the types of data produced can and do vary substantially. Having briefly explored the differences between the two assessment types, we discuss the broad distinction between two approaches to learning and teaching, one that may be termed *a developmental approach* and one that may be termed *a deficit approach* (Griffin, 2009). The particular approach adopted within a context will inevitably impact on the choice of and reasons for assessment.

If systemic assessment is to be useful within the classroom the results need to be interpreted by teachers and found applicable in the classroom context. Underlying this requirement of applicability is the presence of a model of developing mathematical proficiency that includes both plausible conceptual development (from the mathematical perspective) and cognitive development (from the learner perspective). A model such as envisaged here should be somewhat loosely configured and address common issues so that it does not exclude different approaches to mathematical teaching and learning (see Usiskin, 2007). Such a degree of coherence (from broad consensus towards a developmental model, to a working curriculum document that outlines the broad ideas, to a more specified curriculum at school level and a school programme

1.For detailed descriptions of assessment types and a coherent framework, see Black (1998).

providing more detail[2]) is at present a legitimate dream to work towards. Also envisaged in the dream is the idea that *professional development*, *accountability testing* and *formative classroom experience* are integrated around core core aspects of the discipline (Bennett & Gitomer, 2009). The theoretical insights informing a developmental model and the elaborated assessment programme are not the immediate concern of this article. We propose merely to show how applying Rasch measurement theory (RMT) may support such a project.

An essential part of that support is the facility of the Rasch model to yield measurement-like differences and changes. These quantities can enrich the evidence accessible from classroom-based assessment and satisfy the expectations of external stakeholders, in particular if one takes a developmental approach to learning (Griffin, 2009; Van Wyk & Andrich, 2006).

An example is presented which illustrates the intervention potential of an assessment programme that adheres to RMT and within which the Rasch model is applied. We advocate that this model should be seriously considered for inclusion in the approach to national systemic and external assessment programmes, in particular for mathematics.

In essence, we explore the question: What model of assessment may support teaching and learning in the classroom, and in addition enable broad-based monitoring of learning progression within districts and provinces? Reciprocally: How might systemic assessments not only serve their intrinsic purposes to inform decision-makers about performance levels in broad strokes, but simultaneously inform and enrich teaching and learning within the variety of classroom level challenges into which these single instruments intrude?

Classroom and systemic assessment

The important distinctions between and commonalities within classroom-based assessment and systemic assessment types are discussed below. In addition, the complexities involved in reporting results at an individual level and monitoring change over time are noted.

Classroom-based assessment

The teacher in the classroom is concerned with the learning processes and development of the learners in her class. Successful assessment is often of a formative nature and can emerge as continuous assessment, which helps to direct learning and teaching; the summative component, recording marks for the purpose of reporting, also plays a role.[3] The rationale for a teacher to run assessment exercises is to determine whole-class and, particularly, individual levels of current development, to diagnose current obstacles to learning progress, and to provide subsequent targeted scaffolding to appropriate classroom segments. In the best scenario, the forms of evidence used in classroom assessment may vary, from projects requiring extended planning to quick quizzes. Such variety embraces different learning styles and different facets of mathematical proficiency and adheres to cognitive science principles (Bennett & Gitomer, 2009, p. 49).

The stakeholders in classroom-based assessment are the teacher and the learners. The data sets produced by the classroom assessment exercises are not necessarily designed to be expressly meaningful to anyone outside the classroom, although inevitably and importantly teachers within a school community may share ideas and discuss assessments and their results. The particularity and the immediacy of a test or assessment give it currency in the classroom context and for the classroom processes, at a specific period in time.

We may note that in any classroom test or assessment, the teacher is generally concerned with a current spectrum of learner skills and needs in the class, which invariably differs from the spectrum that confronts the educational decision-maker at a district or provincial level. The learners in a particular class may have test performances that are on average well above or well below the average performance associated with all learners of the corresponding grade in an entire school district or province, in the same or an equivalent test. Moreover, the variation of individual test performances within any particular classroom will generally be substantially less than the overall variation in performances on the same instrument across the school district or province.

Systemic assessment

Whilst classroom assessment is generally fine-grained and topic specific, external systemic is generally broadly banded, and attempts to 'cover the curriculum'. From the perspective of the education departments, and in some cases other stakeholders such as funders of programmes, major purposes of systemic assessment are to assess the current performance and variability within a particular cohort of learners, according to some sort of external benchmark of desired proficiency, and to monitor progress, also according to some external standards for change and performance improvements over time. Overall averages (or percentage scores) and the associated pass rates (learner percentages at or above a specified pass criterion) may be deemed particular elements of interest, but their meaningfulness nonetheless has to be argued and established in a suitable robust exposition. These outcomes should be interpreted in relation to other assessment types, for example classroom-based assessment (see Andrich, 2009).

For systemic and external assessments, the sheer extent of the testing programmes, and the development time period and financial constraints, may impact resources and available turn-around time for testing, scoring and data capture. Systemic test designers may thus be obliged to limit the types of items to multiple choice or short-answer responses, to limit testing time to (say) a single period of a school day and, in consequence, to limit the maximum number of items that can reasonably be attempted.

2.See Thijs and Van den Akker (2009) for descriptions of curricula at the macro, meso and micro levels.

3.We consider the terms formative and summative assessment not as referring to discrete entities, but as depicting points on a continuum. Assessment moments may have elements of both kinds.

In an ideal situation, a systemic assessment is designed to produce, from a single short dipstick event, performance data about the current health of educational systems, which is meaningful to stakeholders, district officials and state educational bodies. This body of data and its interpretation may result in decisions requiring or offering intervention or other monitoring functions.

The Department of Basic Education 2009 review claims that 'externally set assessments are the *only credible method* of "demonstrating either to parents or the state whether learning is happening or not, or to what extent"' (Dada et al., 2009, p. 36, citing Chisholm et al., 2000, p. 48, [*emphasis added*]). We contest that claim and, with Andrich (2009), maintain the view that the results of external assessment must be considered in conjunction with classroom assessment, rather than alone. In fact, one may argue that to invoke only external test results to convince stakeholders whether or not learning is happening at the individual or class level, and even perhaps at the grade level in a school, amounts to dereliction of duty and is a dangerous, unethical practice. The claim (of invoking only externally set assessment) itself is unethical, however, if it does not sufficiently address the complex issues of causation that lurk within the extensive variation of student performance on the test.

Similar critique of inordinate emphasis on systemic tests, offered by Bennett and Gitomer (2009), rests primarily on two counts: firstly that systemic testing has unintended detrimental consequences for teachers, learners, administrators and policy makers, and secondly that this type of assessment generally offers limited educational value, as the assessment instrument is usually comprised largely of multiple choice or short answer questions (p. 45).

A systemic assessment may in its totality give a valid overview of system-wide performance on the test instrument (through its constituent items) for the part of the subject and grade curriculum or domain which actually appears within a finite test. Possibly, by astute design and professional concurrence, the test may satisfy further criteria, so as to be viewed as a valid assessment of the whole curriculum at a system-wide level. The attainment of such all-encompassing curriculum validity would, however, require a complete revision of the current systemic test design, as noted and proposed by Bennett and Gitomer (2009).

Whatever the virtues of a systemic test instrument, it simply cannot give the same level of precise inference about the performance of the individual, class or grade within a school as it does for aggregations at district or province levels. This comment applies even to the highly informative instrument[4] we analyse further in this article. For that reason, any interpretation of classroom or grade performance data for a school has to be tempered with a deeper contextual understanding of those units of aggregation, for example, the particular class and the particular grade, and the school in its context and the history of its learners.

4.We distinguish here between a highly informative instrument and an instrument which through rigorous analysis and revision may be regarded as valid and fit for purpose.

TABLE 1: Performance categories associated with percentage attained.

%	Performance categories
0–29	Not achieved
30–39	Elementary achievement
40–49	Moderate achievement
50–59	Adequate achievement
60–69	Substantial achievement
70–79	Meritorious achievement
80–100	Outstanding achievement

Reporting at an individual level

A fairly recent expectation is that the results of systemic assessment be made available to parents. This new access to information may be well intentioned, but the form of the information is problematic, precisely because the data from a single and necessarily limited instrument are so fragmentary and imprecise. Systemic assessment is generally not fine-grained enough to report to teachers, or parents, the results of individual learners, as if these single test performance results, ascertained from an instrument of about an hour's duration, are on their own an adequate summative insight into a year's progress in the classroom.

Even bland descriptions limited to only pass versus fail criteria for a systemic test should be supported by some vigorous and robust debate amongst curriculum specialists, and result in an explicit consensus, before such pass or fail designations of test performance outcomes are communicated. These discussions may be most productive if they occur before a test is finalised for use, and again after the tentative results are available, with explicit minutes recorded at both stages.

In some systemic tests administered under the auspices of the Department of Education (2005), designations of performance categories are assigned to the percentage of maximum scores attained on the instruments, as in Table 1.

On the basis of the systemic test score alone, a learner or parent is given a qualitative description that, however well intentioned, is simply arbitrary, invalid and possibly fraudulent, until other evidence justifies the descriptions offered. It is arguable that such descriptions are generally damaging, but especially when test design has not been informed at all by any criteria for item construction and selection that might relate to either the cut-points and the preferred 10% intervals or the adjectives chosen.

Table 1 indicates instead a tortured avoidance of any verbal signals that learning is in distress, and of any recognition that some children are at precarious risk in the subject.

When systemic tests are designed[5], there do not appear to be any explicit conditions or attempts made to warrant such achievement categorisations. Their valid use would suggest explicit design and the selection and inclusion of items precisely for the vindication of such verbal descriptions. For a 40-item test, the seven performance designators seem to imply a hierarchy of items, comprising 12 simple items

5.These divisions may be the intentions of the test designers. In practice this balance is difficult to achieve.

that all basically competent learners should have mastered, four items specifically associated with each of the five 10% interval categories, and eight items that address curriculum elements at the highest levels of cognitive demand for the particular grade cohort of the test.

Such a table as a criterion-referenced end-product, using cut-points and descriptions, may be a laudable goal, but we suggest it cannot be achieved in the time horizons of planning and test design currently applicable in national and provincial tests. If we are correct, then it becomes important to redesign the systemic test construction agendas and timelines to ensure that the criterion-referenced outcomes are validly constructed within the instruments.

In general, current systemic tests include items that explore elements of the curriculum that warrant particular attention. For each of these elements, a test instrument will indicate how many learners exhibit the desired mastery in the associated responses. Thus the instrument can validly diagnose a series of current particular needs or inadequacies. Summarising the item performance of a class of moderate size in a grade will give an indication of those curriculum elements of which those learners as a group do not yet have mastery.

These indications, inferred from items that have elicited evidence of low proficiency, do not however identify what factors are contributing to the performances, whether good or poor. The class item scores report states, rather than relationships or processes. They may tell us where a problem is to be found, but not why it arose and what may be necessary to address it effectively.

Reporting change

Any objective or intention to use systemic test performances to report on change between years, and possibly on trends over time, will involve an enormous amount of preparatory work to ensure the test performances for the various time periods are truly comparable. There needs to be demonstrable evidence that the associated tests are effectively equivalent. Where it is not possible to use the same instrument on two separate occasions, construction of equivalence is difficult and must be undertaken rigorously. Where the same instrument is used within too short a time frame, the problem of response dependence[6] and appropriate targeting has to be addressed.

Such preparatory work will involve subject and teaching expertise, but must necessarily impact on test construction and assessment. Without this work, and associated extensive piloting of all the test items or instruments in question as well as linking and equating processes, any apparent comparisons of individual test performances to measure change over time must be regarded as moot. It is safer to regard them as invalid until an equivalence relationship between performances over time or across tests has been explicitly argued and demonstrated.

6.Statistical techniques to resolve or account for issues of item dependence across replications of a single instrument for a particular cohort of learners are possible, and even necessary, to ensure validity of results (see Andrich & Marais, 2012).

A major purpose of equivalent tests is to legitimate comparisons. We may wish to examine progress within an individual over time, or to contrast the competencies elicited from two distinct cohorts of learners. Whilst such comparisons may and should admit and use profound qualitative insights and inferences, there is often an intention to seek numerical evidence to bolster those conclusions, and to argue their consequentiality. For that reason, *inter alia*, it will be of interest to obtain measurement-like outcomes of test instruments, in order to allow use of appropriate numerical differences and perhaps numerical ratios.

Inferences about systemic and classroom testing

We argue that systemic testing is valuable as an external assessment technique at broader levels of aggregation, such as district or province, but is substantially less valuable where aggregation is narrower, such as at class and school level.

A well-functioning system of external assessment would involve teachers in the development of the test instruments. It would also feed the results and analysis back into constructive professional development, intended ultimately to impact on classroom practice. In reality, the current design cycles of systemic testing and most external assessments, with or without envisaged professional development support, are too short. The cycles do not encourage adequate engagement with teachers at either the design or analysis stages.

Whilst engagement with teachers may not be a sufficient condition in itself to ensure subsequent effects in the classroom, it is certainly critical that assessment results make sense to teachers, and that the credibility and relevance of the outcomes are pursued. A systemic testing model proposed by Bennett and Gitomer (2009), provides an alternative model which avoids many of the pitfalls mentioned previously. This model includes three intersecting phases: an accountability phase, a formative phase and a professional development phase, wherein the engagement of teachers is a critical feature of the process.

Summation and comparison

Designers of any test instrument face the challenge of informing both the classroom and external stakeholders. We argue that, alongside this, traditional instruments assume validity of arithmetical functions, such as summation and comparison, which are not necessarily grounded in sound statistical theory.

We note that every assessment instrument will involve the summation of item scores. The validity of adding these scores underpins all assessment practice. Our current conventions of practice assume this operation is reasonable in every test, even though we may in contrast alert learners to the errors of adding apples to pears or grapes to watermelons. The unique role of Rasch measurement models in confirming the admissibility of summing test item scores to obtain a test-performance indicator, and in supporting interpretations of test results, will be outlined shortly.

Comparison of assessment performances using numerical differences requires that there is some common scale against which the two sets of performances can be authentically captured as numbers of a common kind. Then we may compare by subtraction. In effect we mimic the way we compare 23 apples with 26 pears by obtaining a distinct currency values for each individual fruit of each set, and then use additions and a subtraction. We must assure ourselves that we can discern differences by use of a common inherent unit.

Rasch approaches also allow evidence of change to emerge from the differences observed between two testing contexts whose comparability has been carefully constructed. The potential of the Rasch model to support use of information-enriched assessment for constructive classroom intervention in order to bring about changes in learning will be described shortly.

Here we argue simply that educational objectives of assessing performance and monitoring for numerical evidence of change must rely on the admissibility of summing item scores and of subtracting test scores. Authorities need to explicitly establish and not simply assume that the conditions for using arithmetic operations are inherently defensible parts of the assessment instruments and their processes.

Deficit versus developmental approaches to learning and teaching

Griffin (2009) makes a distinction between deficit and developmental learning approaches. A deficit approach may 'focus on the things that people cannot do and hence develop a "fix-it approach" to education, and thereby focus on and emphasise "cures" for learning deficits' (p. 187). The deficit approach is common practice where systemic assessment design processes take place in a short time period within the school year, with less than optimum engagement with any teachers and schools, and constrained to the use of a single instrument for a limited extent of class time.

These practices are followed by a period of data scoring and capture, an extensive analysis being performed on the data, and some form of particular aggregated data provided to the schools, many months after the assessment was designed, and of no possible diagnostic value for the same classrooms from which the data arose.

Invariably, the media are informed of the 'research' and information such as '*x*% of learners in Grade *z* cannot handle concept *y*', thereby exemplifying a deficit approach. The Grade *z* teachers then, possibly as a result of a circular letter informing them that only *x*% of their learners have mastered concept *y*, change their teaching plans and focus an inordinate amount of energy on teaching concept *y*. The mathematical concept *y* may not singly be the problem, but may indicate a constellation of concepts that have not yet been mastered (see Long, 2011; Long, Wendt & Dunne, 2011). To focus on concept *y* without understanding the bigger picture may in many cases be counterproductive. Certainly opting for such *post hoc* 'teaching to the test' is something of a backward move, unless of course the 'the test is worth teaching to' (Bennett & Gitomer, 2009).

A developmental approach builds on and scaffolds from the existing knowledge base of individual learners. This approach, advocated by Steinbring (1998), requires that a teacher be attuned to the learner's current understanding and hence current location on a developmental path. The teacher has to be able to diagnose and analyse the various students' current constructions of mathematical knowledge within a curriculum. Then she has to compare these constructions with the mathematics knowledge required (informal assessment), and to adjust her teaching accordingly so as to facilitate the transition (Steinbring, 1998). This process happens against the background of a carefully constructed sequence of learning experiences, exhibiting a suitable sequence of logical and evolving mathematical concepts and theorems that are to be learnt.

The developmental approach resonates with the work of Vergnaud (1988). He emphasises the important link between learners' current intuitive knowledge and the targetted more formal knowledge, and where the teacher's role assures scaffolding of the formal knowledge. The perceived 'errors' highlighted in a deficit model become the stepping stones to greater understanding and the construction of generalisable mathematical concepts.[7]

Something of a paradigm shift is required in order to focus on a developmental trajectory which takes into account the network nature of mathematical concepts and considers that learners may learn different concepts at different rates and in different sequences. This shift may obviate a learning approach where the focus is only on those mathematical objects and skills which cannot yet be exhibited fluently. What is required is an assessment instrument which can more reliably inform teachers of the locations of learners along an intended trajectory of development. Such an assessment instrument may also more reliably inform the education departments, and stakeholders such as funders, of the current learning requirements of particular cohorts of learners, at least in the associated curriculum elements, through an explicit sequential rationale.

When a test is well-designed for its purpose of distinguishing between different levels of learner performance, then we may simply order individual learner performances from lowest to highest, and order test items by their observed levels of difficulty. By partitioning learner scores into a range of ordinal categories, and similarly defining ranges of item difficulties, we may ascertain associations between these groupings that suggest educationally meaningful sequences within teaching and learning. Such a device is produced by

7.The answers to constructed response items in a systemic test set are often found to be partly correct, thus supporting Vergnaud's (1988) notion of 'concepts-in-action'. The transition from localised concepts-in-action to formal and generalisable concepts is the challenge of mathematics education.

the Rasch measurement model and can be easily incorporated into systemic-testing design, so as to permit the provision of supplementary diagnostic information about items, *inter alia* for later communication after assessment results have been analysed.

Rasch measurement theory

We argue that the assessment opportunities provided by the application of the Rasch measurement model can resolve the potential conflicts between the contrasted viewpoints discussed above: classroom based assessment and systemic type assessment, and a developmental model and a deficit model. A well-designed assessment instrument, or sets of instruments, can provide detailed information on the individual development of each learner as well as simultaneously informing external stakeholders on the educational health of an education system. The requirements of the Rasch model resonate with the requirements of good educational practice.[8]

Rasch measurement theory is explained in a number of publications (Andrich, 1988; Rasch, 1960/1980; Wilson, 2005; Wright & Stone, 1979, 1999). A comprehensive application of the Rasch model to a mathematical area, the multiplicative conceptual field, can be found in Long (2011), and an application pertaining to language assessment in Griffin (2007, 2009). In this article, the purpose is merely to illustrate the application of the Rasch model in one systemic test, through stipulating the requirements of the Rasch measurement model and through depicting the outcomes in a form that has the potential to inform both stakeholders and teachers, and to mitigate the misunderstandings that may arise when only aggregated data is used.

Whilst the example test was designed for a systemic application, it exhibits features which suggest areas of improvement in a subsequent design. The choice of setting happens to be mathematical, but the methodology is not tied to any single discipline.

The Rasch measurement model is based on a requirement that measurement in the social sciences should aspire to the rigour that has been the hallmark of measurement in the physical sciences (Wright, 1997). A great deal of qualitative and theoretical work is required in order to construct a valid measurement instrument, as in the physical sciences (such as the thermometer, ruler, scale, or clock). In the natural sciences measurement devices are designed for specific contexts. Though notions such as length, mass and time have universal application, the selection of the specific instrument by which we choose to measure those characteristics is necessarily determined in part by the context in which we seek to comprehend and measure levels of extent and variation in extent.

8.The model was developed by Georg Rasch in the 1950s in order to solve an educational dilemma: that of measuring reading progress over time with different tests (Rasch, 1960/1980). Equating and linking of tests over time, initiated in the 1950s, are examples of the immense power of the Rasch model.

In the social sciences, including education, the first step required towards measurement-like observations is to make explicit the construct to be tested. The operationalisation of the construct as various items, indicative of various levels of proficiency, makes up a test instrument whose overall purpose is to approximate measurement of a characteristic or an ability of persons. This ability is assumed to be plausibly described by a location on a continuum, rather than merely by membership of a discrete ordered category.

In designing a test instrument we are obliged to consider and specify both the construct of interest that we seek to measure, and the context or type of context within which the instrument is intended to be applicable. Having identified and described the construct of interest and designed a plausible test instrument, as a collection of items selected with an educational context in mind, the next step is administering the test to the intended study group (see Wright & Stone, 1979).

There is no requirement that the items are all of equal or equivalent difficulty. They will generally be a collection with elements at various levels of difficulty.

Of particular importance at this step is that the test instrument has been properly targeted to the cohort to be tested. This objective is a substantial challenge, because it involves hazarding judgements about how the overall study group will respond to the items, both individually and collectively, but doing so prior to having any corroborative information. In an educational context this challenge requires subject expertise, teaching experience and pedagogical insights into the learning journeys particular to the subject. It will preferably include cycles of engagement about suitability of items and item structures with specialists in the field, namely specialist teachers.

Appropriate targeting is the requirement that the instrument will be able to distinguish effectively between various levels of performance across the spectrum of learner achievement arising in a specific context of assessment. We may be particularly interested in distinguishing between overall performances at precisely those levels most frequently observed in the assessments. The adequacy of our targeting will contribute to the precision which an assessment instrument can achieve, and hence validly discern differences in ability, in the specified context.

We note that targeting a test instrument, in order to maximise its prospective use for distinguishing between performances in a specified study group, is a completely different issue from using the test as an instrument to decide which learners have performances exhibiting a desired level of competence in the curriculum of the test subject. We may order a set of student performances from best to worst, regardless of what subsequent judgment we may care to make about which of them attain a pass or attain a distinction on the basis of the ordered test scores.

This difference between distinguishing and deciding arises from the contrasting nature of norm-based (internal ordering) and criterion-based (external notions of pass and distinction) inferences, Rasch models can admit the strengths of both norm and criterion approaches. More will be said on the matter of criterion-referencing in the illustrative example.

The Rasch model and its item requirements[9]

The essential idea underpinning the Rasch model for measuring ability by performance on a test, is that the whole test comprises a coherent set of appropriate items. Each item is conceptually relevant to the purpose of the test: it consistently gives partial information about the ability which we seek to measure (justifying a possible inclusion of the item), it enriches the information provided by all the other items collectively (contradicting possible redundancy and exclusion of the item), and it is substantially free of characteristics which might obscure the information obtainable from the instrument (contributing to the precision, rather than to uncertainty, of the instrument, and being free of bias).

Dichotomous items

In educational settings, the Rasch model is a refutable hypothesis that measurement of an ability is being approximated by the test instrument outcomes in a specified context. It postulates that the ability level of a particular person can be represented by a single number β_n. In its simplest form for dichotomous items (with outcomes success or failure, scored as 1 and 0), the model assumes that single numbers δ_i represent the difficulty levels of the items.

Each outcome of an interaction between a person and an item is uncertain, but has a probability governed only by these two characteristics, that is by (β_n, δ_i). The Rasch model avers that the arrays of numbers β_n and δ_i are on the same linear scale, so that all differences between arbitrary pairs of these numbers such as $(\beta_n - \delta_i)$, and hence also $(\beta_n - \beta_m)$ and $(\delta_i - \delta_j)$, are meaningful. Through these differences we may not only assign probabilities to item outcomes, but may also measure the contrasts between ability levels of persons and the contrasts between difficulty levels of items, and offer stochastic interpretations of those contrasts.

The probability of any learner answering any item correctly is a function of only the difference between the locations of ability of the specified learner and the difficulty of the particular item $(\beta_n - \delta_i)$. The model demands that no other person factor or item factor or other consideration intrudes into the probability of success on the item, and that the net joint interaction effect of the person ability and item difficulty is dominated entirely by that difference.

The logistic function with parameters (β_n, δ_i), expresses the probability of a person n with ability β_n responding successfully on a dichotomous item i, with two ordered categories, lower and upper designated as 0 and 1, in the equation:

$$P\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} = \frac{e^{x(\beta_n - \delta_i)}}{\lambda_{ni}} \qquad \text{[Eqn 1]}$$

Here P is the probability, X_{ni} is the item score variable allocated to a response of person n on dichotomous item i, the number x is an observed score value (either 0 or 1), where β_n is the ability of person n and δ_i is the difficulty or location of item i. Note that Equation 1 does not require any restrictions on either of the real numbers β_n or δ_i, but it does require that the two values can be subtracted. The function of the denominator λ_{ni} in Equation 1 is simply to ensure the (two) probabilities for the dichotomous item sum to 1.

The relationship of item to learner is such that if a learner labelled n is at the same location on the scale as an item labelled i, then $\beta_n = \delta_i$ or $(\beta_n - \delta_i) = 0$. In consequence the two probabilities for the ordered categories are equal. Then substituting this zero difference for the bracketed terms into Equation 1 implies that the learner of any ability level will always have a 50% chance of achieving a correct response to any dichotomous item with a difficulty level equal to his or her ability level. If an item difficulty is above the ability location of any learner, then the learner has a less than 50% chance of achieving a correct response on that item, but if the item is located lower on the scale than the person location, the learner would have a greater than 50% chance of achieving a correct response.

The graph of Equation 1 for a specified value of δ_i is obtained by setting the probability on the vertical axis, and person parameter β_n on the horizontal axis (see Figure 1). The result is a symmetric s-shaped ogive curve, with a midpoint at $(\delta_i, 0.5)$. This curve is termed the item characteristic curve. The ascending curve (from low on the left to high on the right of the figure) indicates the probability of obtaining a correct response. The descending curve (from high left to low right) gives the complementary probability of obtaining an incorrect response.

Equation 1 suggests that if we consider the subset of all persons whose common ability is precisely β_n, then each of them will always have exactly the same probability of obtaining a score one, at each and every item whose difficulty is given by a

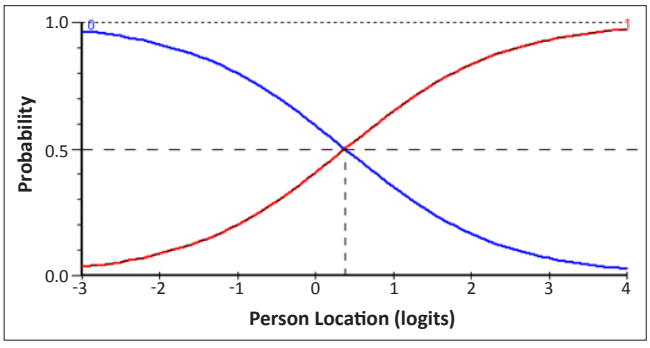

FIGURE 1: Item characteristic curve and complement: probability of 0/1 responses.

9.This section may be omitted on first reading, but readers are encouraged to become familiar with the underlying mathematical logic of the Rasch model.

specific value δ_i. Similarly, these persons of common ability will all have the same probability of a zero score, for all items at the specified difficulty level. Moreover, this equivalence of probabilities will continue but with a revised common probability value, at any new set of items which are all at a distinct but common difficulty location δ_j, where $\delta_j \neq \delta_i$.

Equation 1 is a stringent requirement, but it is exactly as required for a dichotomous test item to be validly considered as unbiased and equally fair to all persons who take the test. It may appear to be only one equation in this format, but each version comprises two probability statements (for the values $x = 0$ and $x = 1$).Then there are $2 \times N \times K$ equations summarised within Equation 1 as Rasch models require that same stochastic structure for all possible $N \times K$ combinations of N persons in a study group, each interacting with each of K (dichotomous) items in a test instrument.

Multiple choice items

Items offering a multiple choice amongst a closed set of response options are handled in the same way as dichotomous items. Some minor adaptations allow the analysis of test data to address the extent to which preference for the various false distractor items may exhibit patterns that vary over the ability range of the persons taking the test.

Polytomous items

Modifications of Equation 1 allow the probability relationships to be extended to polytomous test items that permit maximum score categories higher than one, for partially or completely correct responses. For polytomous items we permit each item response to be recorded as an ordinal category indicated as a single number within the set 0, 1, … m, where $m > 1$.

It is important to note that we are making an ordinal set of categories, recorded primarily by numbers. Rules for allocating these number labels will be set out in a scoring memo for the polytomous item. Because we assume expert construction of each item and its scoring memo, we expect that higher item scores will be associated with higher abilities β_n, and conversely that lower scores will be associated with lower abilities β_n.

We are only saying the labelled categories 0, 1, … m are distinct and uniquely ordered. We are not saying that unit differences between the scores x and $x + 1$ are the same, regardless of x. We are not considering any ratios to be valid. Here 2 is more than 1 but is not two times 1. Likewise 3 is higher than 2 and 1, but is not 3 times 1, nor 2 plus 1.

This initial ordinal structure is therefore distinct from using the category labels x as marks. But we may go on to assume the labels to be marks, and also allow addition of these marks across all items. Then, for any particular item, as the marks x increase, we will expect higher total performance scores in general, and specifically, higher averaged total scores at each new higher observed label x.

Simultaneously, but distinctly, we also assume that higher levels of person ability β_n will be associated with both higher item-score labels and marks x, for each polytomous item, and hence also with higher test total performance scores. These addition strategies are perfectly plausible and coherent, and have been common practice perhaps for many decades. But the issue of the conditions under which they can be defended as modes of obtaining objective and meaningful totals must still be addressed.

The levels of person ability can range over the entire set of real numbers $(-\infty, \infty)$. A consequence of the ordering of our categories in any polytomous item is that we also expect that each such item partitions the full ability range into a sequence of $(m + 1)$ consecutive disjoint intervals, over which the corresponding most likely item category label or score will be 0, 1, 2, … m and in that ordering.

If we wish to make inferences about the relative abilities of individual persons the Rasch measurement model is the only route by which to do so. All other models permit only vague general statements about the distribution of abilities for unspecified persons.

Software packages to perform Rasch analysis through stages of model checking, diagnostic processes and estimation procedures are available on the internet, and from development laboratories. This study made particular use of the RUMM2030 suite of programs. In the reported data (see Table 2), the five polytomous items are represented by the average thresholds.

The Rasch model and consequences for test design

Good test design seeks to have every item satisfying the design criteria outlined above. What Rasch methodology offers is the possibility of checking each of those item requirements, their collective functioning, and the various independence requirements. Constructing a valid instrument will require some arduous tasks at item level. When the item and independence requirements are each found to be reasonably satisfied by the test item data, the astonishing power of the Rasch model is harnessed.

Statistical theory guarantees us that under these required conditions we can not only find a valid estimate of ability for each learner, but that for any person, the sum of his or her item scores is the key element in estimating that ability, and that all other detailed information from the data is neither needed nor helpful in the estimation process. We note that this sufficiency does not imply the total performance score itself is a suitable measure of the ability, but that the person ability measure is a mathematical function involving only that person's total score.

The same statistical theory also guarantees a similar result for items: counting how many of the N persons have been assigned into each of the $(m + 1)$ score categories of an item, that is finding that item's score frequencies, is sufficient to

obtain valid estimates for both the m thresholds of that item and for its average level of difficulty. No other information from the data is required, and no other information from the data set could possibly improve the estimation process. Again this sufficiency of the $(m + 1)$ category frequencies for the m threshold estimates does not imply the frequencies themselves are suitable measures for the thresholds, but rather that threshold estimates are simply a mathematical function involving only those frequencies, whilst the person estimates are determined by the array of total scores.

These two types of simple estimation structures are extraordinary. These simplicities do not hold for any other model than the Rasch measurement model. The Rasch model is essentially an hypothesis that an ability is measurable, indirectly, from test instrument data in a specified context.

If the observed data do not fit the requirements of the Rasch model, then these measurement-like advantages, however desirable, do not arise. In consequence there is no way to coherently provide any statistical inferences relating to individual people or specific items, other than by frequency tables. Any long-term intention to make statistical comparisons between or within cohorts over time is irrevocably undermined.

When the data fits a Rasch model, suitable transformation of the raw total scores for persons and raw frequencies of score categories of each item will enable calculation of estimates for both learner ability parameters and all item thresholds and average difficulty levels. All these estimates may then be legitimately represented and located on the same scale or linear dimension. All differences obtained from any pair of these $N + M$ estimates have an explicit stochastic interpretation.

The estimated item difficulties are calibrated to have a mean of zero[10], and then the relative difficulties of the items are located accordingly. Thereafter the learner proficiencies are estimated in relation to the corresponding learner performance on each of the items. Figure 2 (in the illustrative example) displays a summary of item difficulty and person ability estimates in the same diagram. On the right side, all the items from the test instrument are located at their levels of relative difficulty. On the left side, all the learners are located at their individual levels of proficiency on the same vertical axis. Each learner is however only shown in the figure as hidden amongst the collective contributors to the cross (×) symbols at the particular interval in which their estimates appear. Note that the display gives valid insights into the test performance, but that no notions of fail, pass or distinction have been specified.

The Rasch measurement model suggests an assessment system which provides statistically sound data and analysis which can inform classroom teaching as well as external stakeholders in a contextually meaningful way. We support our argument with an example drawn from recent practice in secondary school assessment.

An illustrative example

A test instrument (K = 40 items) was designed for the purposes of measuring learner proficiency on Grade 8 mathematics. The test, as is common practice, combined several mathematical strands, such as data and probability, geometry, algebra, and number. The test was administered over a cohort of Grade 8 learners (N = 49 104) in one South African province. The study data was analysed applying the Rasch model, for the purposes of confirming appropriate difficulty level of the instrument as a whole for the learners and to identify and describe learner ability in relation to the test items (Long & Venter, 2009).

The mean of all item locations is set at zero as a standard reference point in the Rasch measurement model[11]. The item difficulties are estimated and located on the scale. The learner ability values are then estimated. The learner proficiency estimates are located on the same scale in relation to the items. For the purposes of this analysis the scale was divided into bins of equal width. The left hand side of Figure 2 is a simplified histogram for the estimated ability values[12]. The chosen scale is the log-odds or logit scale, derived from using the logarithm of odds (the ratio $O = \frac{\Pr(X=1)}{\Pr(X=0)}$). Within this scale all the parameter estimates satisfy the required measurement-like properties, and have consistent stochastic interpretations.

We note that Figure 2 immediately provides decision-makers with an extensive but quick diagnostic summary of which items can be correctly answered by at least half (50%) of the learners at a set of specified ability levels, and which items are correctly answered by fewer than half of the tested persons at specified ability levels. The diagram provides a label in which the item number in the test is specified, and the item content is easily obtained by reference to that label.

Here visual inspection of the proficiency histogram will suggest that the person (ability) mean is below the zero item mean, being located at approximately -1.0 logits. This negative location indicates that the test instrument is not appropriately targeted for the tested Grade 8 group as a whole. In consequence, somewhat less than optimum information for distinguishing between performance abilities on this test is obtainable for this cohort on this test. This graphical feature of the output indicates that the test could be improved to better match the variation in the study group. The data suggest that for this study group, more items of below the current average difficulty would improve the

10.There is a technical reason for setting the item mean equal to zero. A simple explanation is that there needs to be one arbitrary origin for all item difficulties because the data can only inform us about differences between item parameters in Equation 1, hence differences between person and item parameters.

11.The software, RUMM2030 (Andrich, Sheridan & Luo, 2011), a programme designed to support the features and requirements of the Rasch measurement model, has been applied here.

12.The terms 'ability' and 'proficiency' are both used to describe the location of persons. Proficiency is the preferred term as it denotes a current state rather than an innate condition.

TABLE 2: Items ordered from difficult to easy, with item location, standard error, item type, domain and item description.

Item	Location	SE	Item type	Domain	Item description
I36	2.79	0.26	Poly	Data	Finds the mean of a data set
I40	2.42	0.23	Poly	Geometry	Calculates the coordinates reflected about the x-axis
I38	2.14	0.21	Poly	Number	Calculates rate problem
I37	1.81	0.19	Poly	Geometry	Calculates volume of a cylinder
I39	1.09	0.15	Poly	Number	Determines the exchange rate
I10	0.99	0.14	MC	Number	Calculates percentage increase
I23	0.98	0.14	MC	Geometry	Finds surface area of a prism
I26	0.75	0.14	MC	Geometry	Applies Pythagoras' theorem
I34	0.70	0.13	MC	Data	Calculates total number in a stem and leaf plot
I35	0.55	0.13	MC	Data	Finds the mode of a data set
I33	0.46	0.13	MC	Data	Finds the median of a data set
I06	**0.45**	**0.13**	**MC**	**Algebra**	**Manipulates algebraic fractions**
I03	0.42	0.13	MC	Geometry	Estimates length measure in centimetres
I08	0.12	0.12	MC	Number	Calculates fractions
I21	**0.08**	**0.12**	**MC**	**Algebra**	**Substitution of variables**
I15	**0.06**	**0.12**	**MC**	**Algebra**	**Addition and subtraction of algebraic terms**
I13	0.03	0.12	MC	Geometry	Calculates angles of a triangle
I17	0.00	0.12	MC	Geometry	Applies horizontal translation
I19	**-0.01**	**0.12**	**MC**	**Algebra**	**Solves problem applying multiplicative reasoning**
I32	-0.08	0.12	MC	Data	finds the range of a data set
I30	-0.08	0.12	MC	Data	Calculates theoretical probability
I25	-0.11	0.12	MC	Number	Applies knowledge of integers and square roots
I02	-0.20	0.12	MC	Number	Finds temperature difference, represents with integers
I31	-0.20	0.11	MC	Geometry	Determines exterior angle
I28	**-0.23**	**0.12**	**MC**	**Algebra**	**Reasons about the square root of algebraic expression**
I12	**-0.25**	**0.11**	**MC**	**Algebra**	**Solves a linear equation**
I01	-0.34	0.12	MC	Number	Identifies an irrational number
I11	-0.52	0.12	MC	Geometry	Reflects shape about the x-axis
I05	-0.60	0.11	MC	Geometry	Identifies coordinates of a linear function
I16	-0.69	0.11	MC	Number	Calculates fractions of time
I09	-0.70	0.11	MC	Number	Orders integers
I20	**-0.70**	**0.11**	**MC**	**Algebra**	**Calculates arithmetical sequence**
I22	-0.88	0.11	MC	Geometry	Identifies faces of a solid object
I07	-0.88	0.11	MC	Geometry	Knowledge of angles of a quadrilateral
I24	-1.03	0.11	MC	Data	Reads a pie chart
I27	**-1.04**	**0.11**	**MC**	**Algebra**	**Converts additive problem into algebraic expression**
I18	-1.05	0.11	MC	Number	Understands multiplication before addition convention
I04	**-1.38**	**0.11**	**MC**	**Algebra**	**Recognises and predicts patterns**
I14	-1.74	0.12	MC	Geometry	Identifies the net of a solid object
I29	-3.14	0.17	MC	Data	Interprets a bar chart

power of the test to distinguish between proficiencies at the lower segment of the person range, where most of the study group are located.

Augmenting the instrument with new items in the targeted range might make the instrument appear easier in the sense of possibly improved performances for all learners who performed well enough on the new items. That artefact of apparently increased scores and likely increased percentages, necessary in seeking better power to make finer comparisons between learner performances in the mid-range, will usually require a revised view of any corresponding criterion-referenced judgments such as pass-fail or distinction-pass applicable in a revised instrument.

These revisions require precisely that same expert judgment which we hope originally contributes to the design of every systemic test, and to its educational interpretation, being exercised by the inclusion of new items and the interpretation of their consequences.

For learners clustered around the person mean, there are some items (below them) which are relatively easy, some items for which according to the model learners in this cluster have a 50% chance of answering correctly, but most items in the test (above them) are relatively difficult for this cluster of learners (fewer than 50% of them will answer correctly on any of the highest sets of items).

Table 2 presents the same items from most to least difficult vertically down a table with brief descriptions of the $K = 40$ items in the associated levels. The easiest items therefore address the interpretation of a bar chart (I29) and the identification of a net (I14). The items, calculating rate (I38), coordinate geometry (I40) and calculating the mean (I36) emerge as the most difficult.

For ease of analysis, some equally spaced levels, also denoted as proficiency zones, have been superimposed on the person-item map (see Figure 2). Items I15, I13, I17 and I19 are of average difficulty and therefore aligned with the item mean

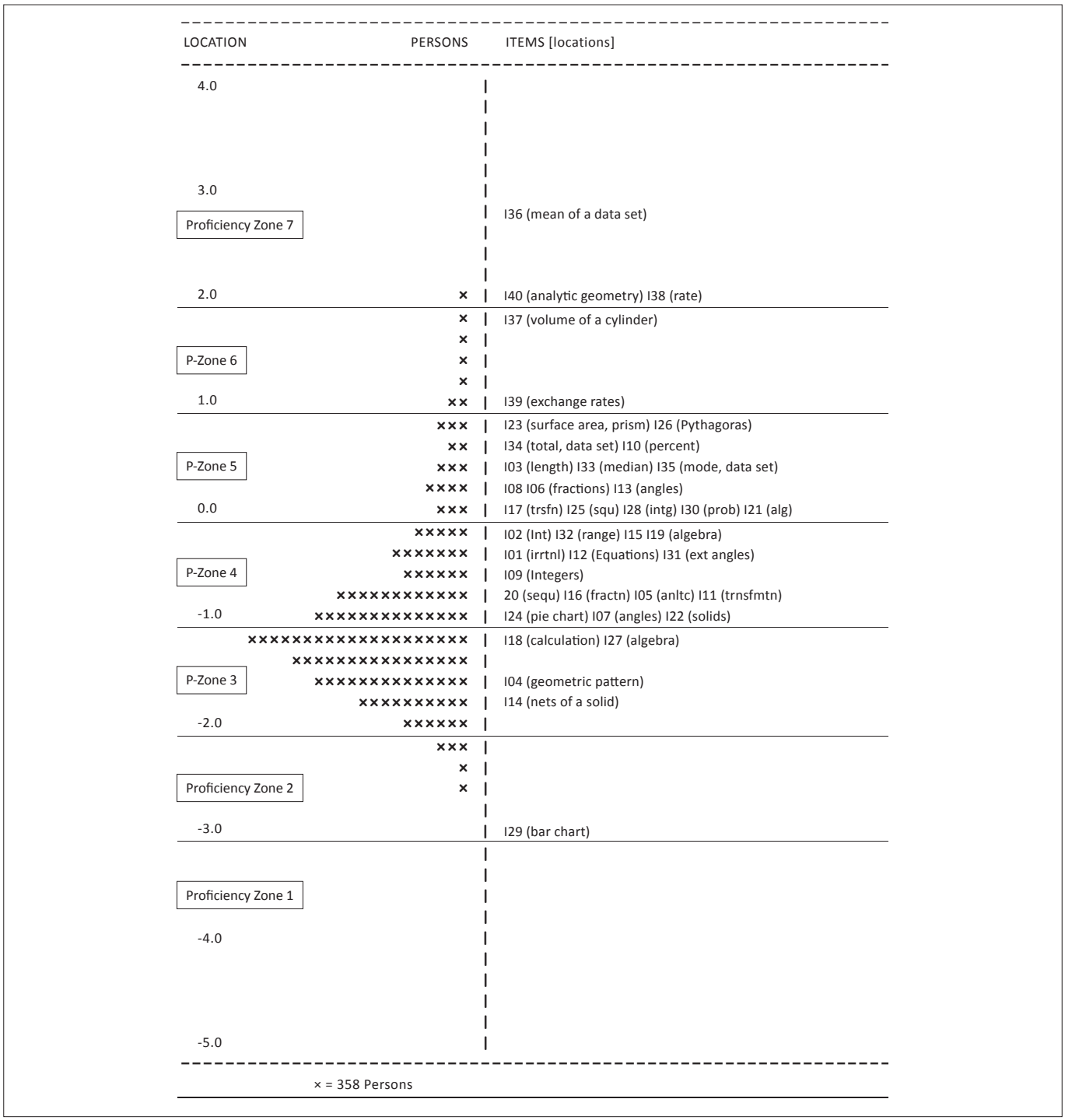

FIGURE 2: Person-Item map approximating person proficiency and item difficulty on a common scale.

set at zero (see logits -0.01 to +0.06, see also Table 2). I29 and I14 are the easiest items, located at the lower end of the scale (logits -3.14, and -1.74), with I38, I40 and I36 the most difficult items, located at the top end of the vertical scale (logits 2.14–2.79).

For the few learners at proficiency zones 8, 9 and above (not shown in Figure 2 due to the scale chosen), there are no items which challenge their mathematical proficiency. For learners at proficiency zones 6 and 7, there are five items located at a matching level, Items I39, I37, I38, I40 and I36, and for which the learners have around a 50% chance of being correct.

Analysis of relative locations of learner proficiency and item difficulty in separate individual construct strands (for example Algebra) allows stakeholders in both classroom-based and systemic assessment to further research and provide some appropriate intervention. For example, lesson sequences may be developed which attend to the increasing algebraic demands and the associated cognitive skills proximate to current levels of interpreted ability.

Retrospectively, according to the model:

> when the amount [*extent*] of latent trait possessed by the candidate was equal to the amount [*extent*] needed to demonstrate the

> criterion behavior, the probability that the person could demonstrate the behavior [*in this instance*] was 0.50. This [*criterion*] was an important idea in defining a person's [*current*] ability, but it was crucial to the assessment being used to improve learning, identify appropriate teaching resources and to develop current policy. (Griffin, 2007, p. 90)

Systemic assessments and classroom intervention strategies

We now make an educational assumption. We allow that the changing proficiencies between learners mapped against the static display of item difficulty as we move up Figure 2, will be very similar to the progression of proficiency on the corresponding curriculum elements particular to an individual learner. We assume that the learner is increasingly engaged in the teaching and learning classroom on tasks related to the test material and over time becomes better able to tackle items of greater difficulty up the vertical sequence. This assumption is debatable, since there is not necessarily only one pathway to mathematical sophistication in any grade. However its utility is that it allows us to interpret the static Figure 2 (with item descriptions) as part of a developmental model.

For each set of learners clustered at a level in Figure 2, we have some idea of the types of items which the cluster can currently manage (i.e. for which they have at least 50% chance of success). We also have some idea of the types of items just some small distance above the current cluster level, and hence located in what may be called the zone of proximal development (Vygotsky, 1962) for that cluster of persons:

> The idea of ordering criteria and locating the criterion where the probability of success for each person is 0.50 can be linked to Vygotsky's research which was driven by questions about the development of human beings and the role that formal education plays in the process. The challenge for educators was to identify students' emerging skills and provide the right support at the right time at the right level. It was in this context that Vygotsky's construct of Zone of Proximal Development (ZPD) – the zone in which an individual is able to achieve more with assistance than he or she can manage alone – was conceptualized. (Griffin, 2007, p. 90)

By specifying an assumed zone of proximal development for each cluster level, the teacher uses the test information to make teaching efforts more efficient. In this structure the teacher imposes temporary clusters within the class so as to more easily divide teaching efforts and time between groups with similar current needs, as reflected by the tested subject proficiency. For example, learners located in proficiency zone 3 have four items located within a similar zone. For these learners the model probability is 0.5 or 50% for answering correctly. For learners in proficiency zone 2, these same four items will be more difficult in general.

From a conceptual development perspective, we see in Table 2, where the algebra items are in bold, that they are spread nicely over almost the whole range of item difficulties, and well aligned with learner proficiencies, therefore giving a fair reflection of learner proficiencies in algebra. See Items I04, I027, I020, I012, I019, I015, I015, I021, I006, arranged from least difficult algebra item (logit -1.74) to most difficult algebraic fractions item (logit 0.45).

The potential is there, in the case of this systemic assessment, of identifying a hierarchy of competences within algebra through which learners could be guided in the small setting of a single classroom. The hierarchy of competences evident in Table 2, was derived from the responses collected from a very large sample of learners and not just from one classroom. This hierarchy could reflect increasing challenge in mastery of algebra as generally experienced by learners of that age. The development of a sequence of items, aligned with the theory of emerging proficiency in algebra, has the potential therefore to empower the researcher or professional teacher communities to structure learning opportunities in an informed manner, mapped to the needs of clusters of learners in her class whose proficiency has been mapped onto the same scale.

The efficacy of the instrument depends on the theoretical work that has informed the instrument and that also informs the analysis and the inferences to be made from the analysis. But given high quality theoretical work underpinning test construction and rigour in the refining of the instrument, we propose that the application of Rasch measurement theory may provide the means for meeting the needs of both the teacher or learners and the stakeholders interested in outcomes of large-scale assessment.

Complementary strategies

The advantage of identifying and targeting current need groups, emerging even from a non-optimal systemic test as reported here, arises if the results are known quickly. In large and complex educational structures where quick turn-around from data to results at a learner level is not easily achieved, it may be useful to consider an alternative complementary assessment strategy beyond systemic testing.

An external resource of a large collection of items, sufficient for several tests at any parts of the likely person ability range, along with associated already prepared diagnostic information, can be marshalled, and made available for devolved use by schools, grade leaders and teachers. There may be a need to provide facilitative scoring arrangements (e.g. electronic marking and outputs as provided for the example test in this article) so that the richness of the assessment resource feeds timeously into teaching. Given suitable systemic test and scoring resources, it will then be feasible for any classroom to be focused upon its own current needs, across all the very diverse ranges of classroom proficiency and school contexts.

Making this option for selection and downloading of items feasible will require prior resource implications and processes. Many proposed items will need to be submitted, cleared for use, piloted and, where necessary, adapted. There

will be some attrition due to unsuitable proposals, and some necessity to ensure breadth of cover for the resource. All items will require grading and diagnostic ancillary information. The associated collaborations will generate teacher collegiality and contribute to professional development of classroom diagnostic skills and intervention initiatives.

In this scenario, district and provincial decision-makers can usefully supplement external systemic-test results apparently signalling classrooms in current distress, with detailed analysis of the assessment initiatives and interventive strategies currently explored, or not yet explored, in those environments. Thus any systemic need to address incompetence or inexperience in the classroom can be informed in part by systemic tests, and give rise to other information or information processes that will be fairer to all teachers, affirming the dedicated and competent and alerting to incompetence or neglect.

Why Rasch

The importance of requiring data to fit Rasch models is that fitting the model guarantees that scores arising from items which independently obey Equation 1, may always be summed together. These person totals and category counts will always permit separate estimation of each of the N person ability parameters and each of M item difficulty parameters.

Only Rasch models have this property of guaranteeing the summation process to obtain a valid overall test score. All other methods (whether based on so-called traditional test theory or on so-called 2-parameter and 3-parameter structures for item responses) simply assume the summation is valid, even if there is demonstrable evidence that test items scores do not behave additively. In other words, all other models for summing of test item scores into a collective indicator will only assume the internal consistency within and between item scores as an incontestable truth, whereas the Rasch model allows the data to signal when such summation is dubious or false.

This issue of permissible summation is not simply a mathematical nicety. It is an ethical imperative. If we claim we have an instrument that consistently accumulates scores from appropriate component parts, we are obliged to assess the extent to which both the accumulation and the behaviour of the parts are confirmed by the evidence in the data.

We note that there is no requirement that the persons interacting with the items of an instrument are a random sample of any kind. The persons are simply part of the context, and not representative of any group other than themselves. We seek to make inferences about the relative abilities of any and all the persons tested.

Similarly, the items are not intended as a random sample from possible items. We seek to make valid inferences about the manner in which the selected items collectively discriminate between the persons who are the source the data.

Where is the catch?

In practice the validity of the output and analysis on which Figure 2 and Table 2 are based, is conditional on the adequacy of the fit of the test data to the Rasch model requirements. Checking the requirements of the model is an extensive and difficult task, precisely because this particular model embodies all the many requirements that permit measurement-like estimates. All these requirements should be checked. It may transpire that several iterations of design, analysis and identification of problems are required, before an instrument is deemed to be satisfactory for its intended measurement purposes. The checking of the fit is sketched here so as to obviate any impression that displays like Figure 2 are simply routine outputs of a test instrument and software which can be accepted without justification and analysis.

The checking of model fit is the first of a set of cyclical processes, the purpose of which is to understand the data and where necessary to improve the functioning of the instrument. Here we distinguish between items that fit the model, items which are under-discriminating (often when learners are simply guessing), and over-discriminating items arising from item response dependence (e.g. where a correct response on a previous item increases the probability of a correct response on a current item).

A further possible violation of requirements to be considered when applying RMT is differential functioning of an item across distinct learner groups. For example, boys at an ascertained proficiency level may perform much better than girls at the same level on a particular item that involves bicycle gears. Checking for these group differences is important in the interest of assuring fairness of all items for all groups. Strategies for diminishing the effects of differential item functioning are to be found in the literature (Andrich & Hagquist, 2012; Andrich & Marais, 2012).

The Rasch model is essentially a single complex hypothesis built from several requirements about a context, about a test instrument and its constituent items, and about the way in which the context and instrument interact to produce special forms of measurement-like data. The whole purpose of the Rasch model might be characterised as seeking to make valid inferences at the level of an individual person and to avoid being limited only to inferences about the patterns within a totality of persons in a given context. It is inevitable that in demanding so much more detailed utility of an instrument of any kind, there will be more stringent properties required within its construction. In addition, we will require detailed description of the contexts within which such an instrument can be validly used.

Here we will take care to specify all the major requirements, and indicate some of the ways in which each of those requirements may be invalidated by evidence. Note that a single invalidation of any one requirement may be sufficient to send a test instrument back to a revised design stage, the beginning of a new cycle of iteration towards a data set with a validated Rasch measurement model.

One such context may be the mathematical abilities of learners in a specified grade in all schools of a province. A test instrument is constructed with the purpose to measure the abilities of all the learners in the context, with sufficient precision. It will be impossible for the test instrument to yield exact measures, because it is composed of discrete item scores, subject to uncertainty. However we all recognise there is a point at which non-exact measures may be subject to such high levels of uncertainty that their utility is lost. In consequence all parameter estimates should be reported with an associated standard error of measurement, or by confidence intervals, as well as by point estimates. We may note that increasing numbers of persons will imply reduced standard errors for item parameters, and increasing numbers of items will imply reduced standard errors for person parameters.

The test instrument and its items are expected to explore and reflect an underlying single dimension, rather than more than one dimension. One may argue that the complexity of mathematics implies more than one dimension. Detailed discussion on the topic of unidimensionality may be found in Andrich (2006). Here we note that unidimensionality implies all aspects of the test 'pulling in the same direction'. Undue language difficulty for example, would be an example of an unwanted dimension.

On this single dimension we hypothesise that it is possible to meaningfully locate all N person abilities at particular numbers on a number line. We require that this arrangement must operate in such a way that all comparisons between person abilities would be consistently represented on the number line. We require that all K item average difficulties and all M item difficulty thresholds can be similarly organised on a single dimension, and that all comparisons of item parameters are consistently preserved. In addition, we require that the same straight line be used for both person and item arrangements, and that the two arrangements can be interwoven so that all differences of the type $(\beta_n - \delta_i)$ will also be consistently preserved.

Further, attention must be given to any extreme scores for persons and items. No test can usefully deal with estimating abilities for persons who score either 0% or 100% correct, except when further new assumptions are justified, or when new relevant information becomes available from beyond the current data set. Items on which 0% or 100% of persons are correct, tell us nothing about the distinct person abilities. These item data cannot contribute to a Rasch model for distinguishing either between persons, between 0% items or between 100% items, and are therefore eliminated from the analysis.

Some violations of the required independence may arise only from specific persons or specific items. For each item and for each person we may calculate the corresponding Item fit and Person fit statistics. The values obtained for these statistics assess evidence for dependencies between item responses for any particular person, and dependencies between person responses for any particular item. The statistics identify items or persons for whom the interaction data does not conform to the required Rasch expectations.

After identifying anomalous persons and anomalous items, the test designers have to explore what can be learnt from those elements. For the instrument, this process may involve changing or even dropping any anomalous item(s). The wording, structure and content of the item(s) will guide this choice. In general the final form of every item should enrich the collective power of the test instrument to distinguish between various persons on the basis of their ability alone.

For the specified context, finding that any particular subset of persons responds anomalously, often warrants exploring their removal from the analysis. If a person's item responses are random or incoherent, they do not address the construct which the items are intended to embody. Given that the vast majority of other learners are responding appropriately, we may eliminate the anomalous learners, precisely because their data are not contributing to an understanding of the relative difficulty of the items. In fact, including their anomalous item data will obscure the patterns in the data, and hence affect both the estimates obtained for the other learners and the estimates for the item parameters.

We may eliminate such data, but must record the elimination and its rationale. This strategy still preserves a diagnostic value, for example identifying students who simply randomly guess for all or part of the instrument may have value for educational interventions.

Only one ability-difficulty dimension is the intended construct of interest. However, it may be the case that an instrument taps into several dimensions, all inter-related in some way. Checking an instrument involves exploring if there is a suggestion that more than one dimension emerges from the data (Andrich & Marais, 2011).

Having ascertained that the data largely manifest as a single scale for the person performances and the item difficulties, we check if each of the items suitably contributes to our objective of a measurement process. This process is lengthy and detailed (Andrich & Marais, 2011). It is also complicated, especially when by construction we seek to have an instrument with substantive validity, and that validity requires distinct aspects of the single dimension to be included. For example, we may in a mathematics test require items that tap into algebra, arithmetic, geometry and data handling.

The data should be scrutinised for violations of the homogeneity of the learner responses over any features other than ability itself. Comparisons of the graphs produced by the Rasch analysis software for two or more groups may assist in determining whether various explanatory variables or factors give evidence for differences between groups.

Specifically we may check whether or not evidence exists for suspecting any items to be under-discriminating (as when learners are guessing rather than engaging with items), or over-discriminating (as when an item requires pre-knowledge or a threshold concept).

Discussion

The example provided serves to illustrate the potential of an application of the Rasch model to an assessment instrument should the requirements be met. The potential of such an assessment model with its subsequent analysis is dependent on the quality of the instrument, and therefore on the prior theoretical work that has preceded the development and selection of items. Whilst in this example some worthwhile information is available for the stakeholders to observe, the potential for a more nuanced instrument may be envisaged. We note that the Rasch model is used routinely in TIMSS (Trends in Mathematics and Science Study) and PISA (Programme for International Student Assessment) to scale item difficulties and proficiency scores (see Wendt, Bos & Goy, 2011).

Given a well-targeted test instrument, informed by adequate theoretical investigation within the substantive discipline of the test, there is the potential for informing both the stakeholders and the educational officials. Well-targeted instruments may also require some type of pilot testing or external benchmarking. As it transpired, this well-intentioned test did not match the target population very well. Inferences can be explored to improve this aspect of the test instrument. Nonetheless, diagnostics relevant to the teaching of the material relating directly and indirectly to the test are readily available from the design work on the construction of the test. The design work permits the explicit statements in Table 2, and the ordering of items from the data, to suggest sequences of teaching and learning. It is readily conceded that further iterations with some altered or replaced items may produce revised Table 2 summaries that will conceivably be mildly or radically improved in usefulness.

One may ask whether the information presented in this analysis is not already known to the stakeholders and education officials. We recognise the test design as somewhat typical of assessment instruments expected by current systemic assessment programmes; they should 'cover the curriculum'. The issues may be well known, but the problem of coherence within such a test when analysed from a developmental learning approach is less explicitly recognised.

By its generality of coverage, the systemic instrument provides only scant or generic developmental information to the teacher. Perhaps it is time for cycles of systemic assessment of a more focused and limited nature, for example, an instrument with a focus only on algebra where the skills and concepts may be operationalised in a set of items requiring increasingly complex and critical skills that elaborate on the key areas identified in the literature. Associated specific developmental elements can be marshalled at the design stage, and modified in terms of the emerging patterns of the applied test context, to inform more specific target interventions for algebra in the classroom.

Conclusions

Any approach to mathematics assessment almost certainly follows a predicated view of teaching and learning, which in turn rests on an understanding of the central features of mathematics. The implicit beliefs underpinning current assessment practice may benefit from debate and explicit acknowledgement of any underpinning philosophy. For example, what view of learning and what view of evidence underpins the claim that 'external' assessment is the only credible method of demonstrating that learning is happening in schools (Dada et al., 2009)?

The recommendations resulting from the Department of Education review (Dada et al., 2009) are that continuous and broad-based assessment is limited and that external assessment at Grades 3, 6 and 9 be enshrined in policy. Given that this policy decision has been adopted, it is critical that the external assessments work in conjunction with classroom assessment. The relevant grade teachers, rather than being the objects of the testing policy, should be participants involved in the construction and analysis of tests. We aver that a collaborative strategy supporting regular use of formative assessments may impact more directly on their teaching, in ways that better address learner needs, and hence improve learning of the subject.

In answer to the question: What model of assessment may support teaching and learning in the classroom, and in addition enable broad-based monitoring of learning progression within districts and provinces?, we advocate an approach which takes seriously the critical elements of mathematics, in the formulation of a developmental trajectory.

Systemic provision of a large variety of test items and their diagnostic support material, together with informed and deliberative selection by committed teachers for classroom use, with facilities for electronic data capture and/or marking, are important strategies. Routine classroom tests drawn from such item bases can simultaneously support classroom innovations, whilst providing district structures with information about classroom efforts and needs. In such extended contexts, occasional systemic testing can be interpreted against a wider range of contextual information.

The role of assessment in the 21st century is 'extremely powerful' (Matters, 2009, p. 222). According to Matters, this role can only be justified on condition firstly that the assessment is 'of sufficient strength and quality to support its use', and secondly that the 'users of assessment data have sufficient experience and imagination to see beyond the numbers' (p. 222).

Assessment against this background of theoretical rigour fulfils a requirement of the Rasch measurement theory that the construct of interest be made explicit. The practical unfolding of the construct, in items that are realisations of the construct, is then formulated as a test instrument. The output from the Rasch model, provided the prior requirements are met, has the potential to inform current teaching practice, to orchestrate teacher insights into the challenges of their own classrooms and initiate two-way communication between classrooms and decision-makers.

Acknowledgements

Acknowledgement is due to the Schools Development Unit of the University of Cape Town for the development of the test instrument in collaboration with Caroline Long and Elsie Venter, and to the Western Cape Education Department for commissioning the work.

Competing interests

We declare that we have no financial or personal relationship(s) which might have inappropriately influenced our writing of this article.

Authors' contributions

T.D. (University of Cape Town) contributed to the conceptualisation of the article, and to the detailed explanations of the Rasch model. He wrote extensive sections of the article. He was also involved as an advisor to the original project where the data were collected. C.L. (University of Pretoria) was project leader for the original project where these data were collected. Together with E.V. (Independent Researcher) she was responsible for the initial analysis in the original project, and independently conducted the re-analysis using RUMM software for this article. She contributed to the conceptualisation of the article. She wrote the remaining elements of the article. T.C. (University of Cape Town) contributed to the conceptualisation of the article and thereafter assisted with critical revision of the manuscript. E.V. worked with C.L. on the pilot study analysis and the subsequent analysis of the data.

References

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, CA: Sage Publications.

Andrich, D. (2006). *On the fractal dimension of social measurements I.* Perth: Pearson Psychometric Laboratory, University of Western Australia.

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.A., & Hagquist, K. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioural Statistics, 37*(3), 387–416. http://dx.doi.org/10.3102/1076998611411913

Andrich, D., & Marais, I. (2011). *Introductory course notes: Instrument design with Rasch, IRT and data analysis*. Perth: University of Western Australia. PMCid:3217813

Andrich, D., & Marais, I. (2012). *Advanced course notes: Instrument design with Rasch, IRT and data analysis*. Perth: University of Western Australia.

Andrich, D., Sheridan, B., & Luo, G. (2011). *RUMM2030 software and manuals*. Perth: University of Western Australia. Available from http://www.rummlab.com.au/

Bennett, R.E., & Gitomer, G.H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional development. In C. Wyatt-Smith, & J.J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–62). Dordrecht: Springer. http://dx.doi.org/10.1007/978-1-4020-9964-9_3

Black, P.J. (1998). *Testing: Friend or foe*. London: Falmer Press.

Chisholm, L., Volmink, J., Ndhlovu, T., Potenza, E., Mahomed, H., Muller, et al. (2000). *A South African curriculum for the 21st century. Report of the Review Committee on Curriculum 2005*. Pretoria: DOE. Available from http://www.education.gov.za/LinkClick.aspx?fileticket=Y%2bNXTtMZkOg%3d&tabid=358&mid=1301

Dada, F., Dipholo, T., Hoadley, U., Khembo, E., Muller, S., & Volmink, J. (2009). *Report of the task team for the review of the implementation of the National Curriculum Statement*. Pretoria: DBE. Available from http://www.education.gov.za/LinkClick.aspx?fileticket=kYdmwOUHvps%3d&tabid=358&mid=1261

Department of Education. (2005). *The national protocol on assessment for schools in the General and Further Education and Training band (Grades R to 12)*. Pretoria: DOE.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, *33*, 87–99. http://dx.doi.org/10.1016/j.stueduc.2007.01.007

Griffin, P. (2009). Teachers' use of assessment data. In C. Wyatt-Smith, & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 183–208). Dordrecht: Springer. http://dx.doi.org/10.1007/978-1-4020-9964-9_10

Long, C. (2011). *Mathematical, cognitive and didactic elements of the multiplicative conceptual field investigated within a Rasch assessment and measurement framework*. Unpublished doctoral dissertation. University of Cape Town, Cape Town, South Africa. Available from http://web.up.ac.za/sitefiles/file/43/314/Long,_M__C__(2011)__The_mulitplcative_conceptual_field_investigated_within_a_Rasch_measurement_framework_.PDF

Long, C., & Venter, E. (2009). *Report on the Western Cape Grade 8 Systemic Assessment Project*. Pretoria: Centre for Evaluation and Assessment, University of Pretoria.

Long, C., Wendt, H., & Dunne, T. (2011). Applying Rasch measurement in mathematics education research: Steps towards a triangulated investigation into proficiency in the multiplicative conceptual field. *Educational Research and Evaluation*, *17*(5), 387–407. http://dx.doi.org/10.1080/13803611.2011.632661

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: Connecting theory and practice* (pp. 209–225). Dordrecht: Springer. http://dx.doi.org/10.1007/978-1-4020-9964-9_11

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

Steinbring, H. (1998). Elements of epistemological knowledge for mathematics teachers. *Journal of Mathematics Teacher Education, 1*, 157–189. http://dx.doi.org/10.1023/A:1009984621792

Thijs, A., & Van den Akker, J. (2009). *Curriculum in development*. Enschede: Netherlands Institute for Curriculum Development (SLO).

Usiskin, Z. (2007). Would national curriculum standards with teeth benefit U.S. students and teachers? *UCSMP Newsletter*, *37*, 5–7. Available from http://d75gtjwn62jkj.cloudfront.net/37.pdf

Van Wyk, J., & Andrich, D. (2006). A typology of polytomously scored items disclosed by the Rasch model: Implications for constructing a continuum of achievement. In D. Andrich, & G. Luo (Eds.), *Report no. 2 ARC linkage grant LP0454080: Maintaining invariant scales in state, national and international assessments* (n.p.). Perth: Murdoch University.

Vergnaud, G. (1988). Multiplicative structures. In J. Hiebert, & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 141–161). Hillsdale, NJ: National Council of Teachers of Mathematics.

Vygotsky, L.S. (1962). *Thought and language*. Cambridge, MA: MIT Press. http://dx.doi.org/10.1037/11193-000

Wendt, H., Bos, H., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, *17*(6), 419–446. http://dx.doi.org/10.1080/13803611.2011.634582

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. London: Lawrence Erlbaum.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45. http://dx.doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B.D., & Stone, M.H. (1979). The measurement model. In B.D. Wright, & M.H. Stone (Eds.), *Best test design* (pp. 1–17). Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.

Wyatt-Smith, C., & Gunn, S. (2009). Towards theorising assessment as critical inquiry. In C. Wyatt-Smith, & J.J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 83–102). Dordrecht: Springer. http://dx.doi.org/10.1007/978-1-4020-9964-9_5