Ms Celeste Combrinck
Centre for Evaluation and Assessment, Faculty of Education at the University of Pretoria. Email: celeste.combrinck@up.ac.za
012 420 5680

Prof Vanessa Scherman
Department of Psychology of Education at the University of South Africa
012 429 4900
Email: scherv@unisa.ac.za

Prof David Maree
Department of Psychology at the University of Pretoria
012 420 2329
david.maree@up.ac.za

DOI: http://dx.doi.org/10.18820/2519593X/pie.v34i4.5
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
2016 34(4): 62-78

The use of Rasch competency bands for reporting criterion-referenced feedback and curriculum-standards attainment

Abstract

This study describes how criterion-referenced feedback was produced from English language, mathematics and natural sciences monitoring assessments. The assessments were designed for grades 8 to 11 to give an overall indication of curriculum-standards attained in a given subject over the course of a year (N=1113). The Rasch Item Map method was used to set cut-scores for the Rasch competency bands, after which subject specialists examined the items in each band. Based on the content and difficulty of the items, descriptions for the proficiency levels were generated. Learner reports described each individual's current proficiency level in a subject area as well as the subsequent level to be attained. This article shows how the Rasch Item Map method can be used to align assessments and curriculum-standards, which facilitates reporting learner performance in terms of criterion-referenced feedback and empowers learners, teachers and parents to focus on subject content and competencies.

Keywords: *Rasch competency bands, proficiency levels, criterion-referenced feedback, Rasch Item Map method, curriculum-standards, reporting learner performance*

1. Context

There is a growing realisation worldwide of the importance of learner content-focused feedback as opposed to only providing norm-referenced feedback (Bennett, Tognolini & Pickering, 2012). Norm-referenced feedback is based on numbers, means achieved as well as a comparison of a learner or groups of learners to others in the cohort(s). But with standards-referenced feedback, also known as criterion-referenced feedback, the spotlight is on informing learners about what they know and how they can increase their knowledge and skills (Bennet *et al.*, 2012, Long, Dunne & Mokoena, 2014). Standards-referenced feedback is crucial for meeting curriculum requirements as the criteria are based on learning standards set in curriculum documents (Great Schools Partnership, 2014). The alignment of assessments to curriculum-standards is linked to this type of feedback. Therefore, assessments would be designed to provide diagnostic information and feedback

in order to inform teaching and learning. This article specifically investigates methods using external and internal monitoring assessments for deriving standards-referenced feedback, which serves the purpose of developing accountability and a measure for comparability as well as presenting a method for deriving criterion-referenced feedback. The argument is made that by offering feedback that is content specific and linked to curriculum-standards, monitoring assessments, which comprise systemic evaluations, large scale studies and standardised tests in the schooling system, can serve the additional purpose of enhancing teaching and learning. Content-specific feedback is empowering to teachers and learners and when monitoring assessments are utilised, all stakeholders, including the participants, should benefit from the testing system (Long *et al.*, 2014).

Large scale assessments for norm and criterion-referenced feedback

Systemic and standardised tests are generally not utilised for diagnostic purposes, though such applications could contribute to an impact on the system they evaluate (Khosa, 2013). The main goal of most systemic testing systems is to serve as indicators of performance levels, mainly for norm-referenced comparison and for quality assurance in education (Jiao *et al.*, 2011; Osman *et al.*, 2008). Designing criterion-referenced tests can be time- and resource-intensive but has advantages that benefit those being tested and the systems in the long-term (Stone, Beltyukova & Fox, 2008). Such advantages include monitoring a system but also improving its functioning by identifying attained curriculum standards and pinpointing the next achievable levels. This approach taps into the notion of competency as moving along a continuum (Griffin, Gillis & Calvitto, 2007). Criterion-referenced results give feedback in terms of what skills and knowledge a person has gained, whereas norm-referenced feedback focuses on comparing a person to others in terms of achievement. The study described in this article has utilised criterion-referenced and norm-referenced feedback approaches, using the monitoring assessments to provide learners, teachers, parents, principals and funders with criterion-referenced and norm-referenced feedback. Norm-referenced feedback gave schools learner-level insight into where their school's achievement was located in relation to other schools. Criterion-referenced feedback gave insights into school-level proficiency levels and curriculum standards achieved by learners, in addition to offering subsequent target levels for teachers. During interactive workshops, teachers and subject specialists discussed ways in which the results could be used in developing/initiating interventions and thus enhancing classroom practice.

Learner-level feedback as an ethical application of large scale assessment

Large-scale assessments, such as standardised testing and systemic evaluations, are time consuming and expensive (Khosa, 2013; Popham, 1987). Time is taken from teaching and learning while results are generally used to broadly inform policy and not necessarily to give feedback to learners, teachers and parents. Providing learner-level feedback by using methods such as the Rasch Item Mapping method, is one way of assessing in a more ethical manner as participants can benefit more directly by receiving feedback which is explicit and useful due to its content-based nature (Long *et al.*, 2014). By providing criterion-referenced feedback, teachers can also benefit from the assessments as they become aware of specifically attained curriculum standards and how to structure their teaching to target standards which have not been fully achieved. Giving criterion-referenced feedback could also contribute to learner motivation (Boone, Staver & Yale, 2014). If learners realise that the testing does not influence their academic performance, they could be less motivated to write such assessments. In

contrast, if learners understand that they will receive usable feedback, which will also be disseminated to their parents and teachers, they could be more motivated to participate fully.

Identifying competency levels

The idea of using the Rasch Measurement Theory (RMT) to align assessments to curriculum standards is not a new idea. Ingebo (1989) discussed the use of RMT for alignment to curriculum-standards because the Rasch model creates an equal interval scale of curriculum tasks (items in tests) which can be used for a comparison to curriculum-standards.

> *As dichotomous items that are actually curriculum tasks are lined up and given values with respect to each other, these calibrations (values) are on an equal interval scale generated by the confluence of knowledge and position in the curriculum (Ingebo, 1989: 43).*

The use of the Rasch model to set standards and identify proficiency levels has been demonstrated by several studies (Boone *et al.*, 2014; Grosse & Wright, 1986; Long *et al.*, 2014; Shen, 2001; Stone, 2000). In fact, aligning item banks to curricula with equal intervals for every item was one of the main achievements of Benjamin Wright (Ingebo, 1987). Over the course of the last 30 years, much research on the use of RMT has been conducted in the United States, Great Britain, European countries and Southern Africa (Bond & Fox, 2015; Boone *et al.*, 2014; Wright & Grosse, 1993; Stelmack *et al.*, 2004; Wissing, 2013). Some of the recent literature on the use of RMT for setting standards and proficiency levels are discussed next.

Holster and Lake (2015) showed how items could be scaled with the use of the Wright map for diagnostic vocabulary tests and these results can be used to identify learners needing remedial intervention. Their study also demonstrated how identifying competency levels with the Rasch model are applicable for classroom use, curriculum planning and the refinement of vocabulary tests for placement purposes (Holster & Lake, 2015). In addition, Jiao *et al.* (2011) used the Mixed Rasch Model (multi-dimensional) to classify student performance in a simulation into proficiency levels by analysing item response patterns and the achievement represented on the latent trait by achievement. This resulted in high student classification accuracy and had the added advantage of assisting with the classification of borderline case or minimally competent students. This accurate classification was achieved by fitting the data to the Rasch model and using the intersecting points between adjacent distributions to distinguish varying proficiency levels. Similar studies, such as that of Jiao *et al.* (2011), are needed as the findings can be strengthened by using real data as opposed to simulated data. The study presented in this article is based on empirical data and addresses the limitation of using only simulated data to illustrate the usefulness of certain approaches.

Studies utilising the Rasch model to create competency bands are based on the theory that the item difficulty and person ability alignment reflects the complexity of the content and levels of proficiency in the content areas (Shen, 2001). Shen's study (2001) on medical licensing data compared the Angoff method, the Hofstee and the Rasch Item Map method. This study found that the Angoff method identified test subjects with expertise, whereas the Rasch Item Map method was more likely to identify those with fundamental knowledge. The Rasch and Hostee methods gave equivalent results. In the case of Shen's study, using the Rasch Item Map method made more sense, as fundamental knowledge was required to practise medicine and was more significant than specialist knowledge for those entering the field of medicine. The Rasch method also provided more criterion-referenced results whereas the other methods were more likely to yield norm-referenced results. The Rasch Item Map method

was also more time efficient, as reviewing maps versus reviewing individual items takes less time. Other methods may also lack content explanations of what a standard means, a crucial aspect of reporting learner and student performance levels (Shen, 2001). Other studies of a similar nature have been conducted with equivalent results, showing the potential advantages of using the Rasch Item Map method over that of traditional methods (Wang, 2003). Stone *et al.* (2008) demonstrated how the multifaceted Rasch model could be used to firstly identify minimal competence and then incorporate it into the model of standard setting, especially for criterion-referenced standards when assessments are scored with rubrics. They note that the rating scale used, the unique context of their study and their sample size all influenced the outcome, suggesting that more studies of this nature are required to evaluate the applicability of modelling minimal competence in other settings. Herrmann-Abell and DeBoer (2015) used RMT to map science items onto curriculum materials to compare, with precision, how ideas are taught and how learner understanding of these ideas progress. By aligning items to curriculum standards and progression of understanding, they were able to identify areas for improving teaching and learning in the subject areas. It should be noted that Rasch methods could be combined with other methods, as Bennett *et al.* (2012) did; the multi-stage Angoff procedure was used in conjunction with the Rasch theory to establish performance standards in curriculum-based exams.

The research and findings discussed in this article contribute to the body of knowledge by demonstrating how monitoring assessments can also be used for standards-referenced feedback. This article explains how to establish and define progression levels, how to structure feedback and why this is advantageous to learners and other stakeholders. The purpose of the study described here was to find ways to give teachers and learners useful feedback, more than just comparative information derived from external monitoring assessments. The aims included locating each learner on the subject area developmental continuum, defining what has been gained at each level as well as making explicit the subsequent level of development and curriculum-standard required.

2. Methods

The study

Seven independent high schools form part of an association that strives to give disadvantaged learners the opportunity to grow academically and socially. These seven schools have longer school days, smaller classes and Saturday classes to offer learners additional support. The schools have outside funding to maintain their intervention model of schooling. In order to monitor the progress of learners across schools, set an accountability system in place and to give feedback into the development of the schools academically, an educational research agency was approached to design and develop monitoring assessments. The findings discussed in this article are a component of that monitoring process.

Study group

One thousand one hundred and thirteen learners participated in the assessment study and the cohort for each grade differed in size (see table 1). There were 250 grade 8 learners, 251 grade 9 learners, 319 grade 10 learners and 293 grade 11 learners. More girls (77%) than boys (23%) participated in the study due to a girls-only school being included in the sample, in addition to the other schools having more female than male learners (a 60% girl: 40% boy composition). The sample sizes, overall and per grade, were judged to provide

adequate power, based on the fact that the whole group participated, with the independent schools being considered a population on their own due to their unique intervention model of schooling.

Table 1: Descriptive statistics of sample (percentage by column for grades)

Grade	Gender					
	Male		Female		Total	
Grade 8	52	20.4%	198	23.1%	250	22.5%
Grade 9	50	19.6%	201	23.4%	251	22.6%
Grade 10	87	34.1%	232	27.0%	319	28.7%
Grade 11	66	25.9%	227	26.5%	293	26.3%
Total	255	100.0%	858	100.0%	1113	100.0%

Assessment instruments

The curriculum standards for school subjects are contained in the National Curriculum and Assessment Policy Statement grades R-12, also known as CAPS (Department of Basic Education, 2011, 2012). The CAPS documents aim to set "minimum standards of knowledge and skills to be achieved at each grade" and this includes high, achievable curriculum standards (Department of Basic Education, 2016: 1). The documents also endeavour to show progression in content and context of each subject with development from the basic to the more complex skills and knowledge. In line with the curriculum standards as set by CAPS, subject specialists designed the assessment instruments for English language, mathematics and natural sciences for grades 8 to 11. The assessments were designed to cover the national curriculum topics that would be taught within a school year. The results provided an indication of knowledge and skills gained for a subject within a year and per curriculum area. Multiple-choice items (approximately 60% of a test) and constructed-response items were included in all the assessments. The average scores per school, per class and per learner were fed back into the system via school reports, workshops and data sets containing learner achievement. To give criterion-referenced feedback, a more qualitative approach was also sought to communicate the results, which led to the creation of content descriptive reports discussed in this article.

To determine the overall functioning of the monitoring assessments, the Rasch partial credit model was applied. Rasch Measurement Theory (RMT) holds up the ideal of measurement and aims to compare real item and person responses to this ideal. Where reality diverges from the ideal is where there is an indication for potential improvement (Herrmann-Abell & DeBoer, 2015; Linacre, 2011). The Winsteps 3.75.0 programme provides in-fit and outfit mean-square statistics (MNSQ) which reveal where items and persons fit and where there is misfit (Linacre, 2012). All statistics are reported in terms of log odds units and have a range of -5.00 to +5.00 with a mean set at 0.00 and a standard deviation of 1.00 (Bond & Fox, 2015; Boone *et al.*, 2014).

Table 2 shows the ranges for the Rasch item fit statistics for the three school subjects and for all four grades. Rasch mean square statistics have been found to remain relatively stable for

polytomous item type data and such statistics are relatively independent of sample size (Smith *et al.*, 2008). In-fit and outfit mean square statistics (MNSQ) should have an expected value of 1.0 and values which are above 2.0 are considered potentially problematic and noisy, while values above 3.0 degrade the measurement (three standard deviations above the mean). The mean ranges per subject were well within acceptable limits for the in-fit and outfit MNSQ statistics, with ranges between 0.98 and 1.01 across the three subjects. Maximum values of MNSQ values show that the mathematics grade 8 and grade 9 tests have outlying items with values above 2.00. The standard deviations are small and within expected limits. The standard error of the item mean lies well within the expected range of ± 0.33 logits, ranging from 0.11 to 0.20. The person separation index ranges were above 2.00, with the exception of one science test, which may require more items to separate low and high performers. For all other instruments, the values above 2.00 indicate that there was an appropriate spread of item difficulties. The item separation index ranges were above 3.00, demonstrating that there were enough learners answering the items to confirm the item difficulty hierarchy (Fisher, 2007; Linacre, 2011). Item reliability was also high, with most tests having reliabilities of .98. Overall, the assessment instruments functioned well and were deemed appropriate. It is also important to note that the assessments contained anchor items and cohorts were measured from 2012-2014 providing three points of measurement. The use of anchor items means that extrapolations across cohorts, with regard to progression, could be made.

Table 2: Range of Rasch item fit statistics for instruments from grade 8-11

	English Language		Mathematics		Natural Science	
	In-fit MNSQ	**Outfit MNSQ**	**In-fit MNSQ**	**Outfit MNSQ**	**In-fit MNSQ**	**Outfit MNSQ**
Mean	1.01 - 1.00	1.01 - 0.99	1.00 - 1.01	0.98 - 1.04	1.00 - 1.01	1.00 - 1.01
Standard Deviation	0.08 - 0.22	0.13 - 0.15	0.09 - 0.14	0.23 - 0.35	0.06 - 0.09	0.18 - 0.22
Maximum	1.16 - 1.27	1.25 - 1.67	1.26 - 1.47	1.69 -2.76	1.15 - 1.46	1.53 - 1.92
Minimum	0.80 - 0.86	0.59 - 0.73	0.77 - 0.84	0.45 - 0.68	0.53 - 0.72	0.04 - 0.64
Item separation index (reliability)	6.08 (.97) - 7.49 (.98)		7.19 (.98) -8.04 (.98)		5.36 (.97) - 8.01 (.98)	
Item S.E. of mean	0.12 – 0.14		0.16 – 0.20		0.11 – 0.13	
Person separation index (reliability)	2.03 (.81) - 2.41 (.85)		2.33 (.85) - 2.81 (.89)		1.88 (.78) - 2.63 (.87)	
Person S.E. of mean	0.03 – 0.05		0.04 – 0.05		0.03 – 0.04	

The instruments were designed to gauge curriculum knowledge as specified in the CAPS documents in a given academic year for a subject. As per good assessment design, a balance of items was designed for each topic area with easier and more difficult items for each topic. However, some topic areas consisted of items that are more difficult whereas other topics may have consisted of easier items. This is not only because of the nature of the topic but also considers that a balanced assessment includes easy, medium difficulty and difficult items. As

the assessment information is expansive, the grade 9 assessments are used for illustrative purposes and are shown in table 3.

Table 3: Number of items, mean %, comments for grade 9 assessments

Grade 9 Assessment topics	Number of items	Mean %	Comments
Mathematics			
Data handling & measurement	17	22.21	The *data handling and measurement* section contained the most difficult items, with the other sections being relatively easier. Overall the assessment was challenging for the learners, but item difficulty was set to curriculum standards and workshops were held with teachers to address gaps in learning.
Number, operations and relationships	11	35.65	
Patterns, functions and relationships	33	45.15	
Space and shape (geometry)	14	31.36	
Grand total	**75**		
English			
Reading for meaning (A)	24	61.42	*Poetry* was the most challenging section for learners with these items being considerably more difficult than items in other sections. The other sections had an even spread of difficulty from easier to more difficult items. Teachers said in workshops that poetry is more difficult for second language speakers due to the subtle nature of meaning in poetry.
Poetry	15	37.16	
Non-fiction text	16	54.23	
Visual literacy	8	57.27	
Reading for meaning (B)	16	64.17	
Grand total	**79**		
Natural science			
Earth and beyond	12	54.90	The topics were spread evenly in terms of difficulty. The *life and living* section had more difficult items than the other sections, followed by *matter and materials*. Teachers confirmed that these were sections which learners found more challenging in general and strategies for dealing with this was discussed in workshops.
Energy and change	14	54.43	
Life and living	30	33.72	
Matter and materials	31	44.41	
Grand total	**87**		

Procedure

The assessments were administered in November 2014, using standardised procedures and examination conditions. The educational research centre trained assessment administrators who implemented the testing process. Informed consent was obtained from parents for

learners in all grades and in addition, assent was obtained from learners older than 16 years. After administration of the assessments, specifically recruited and trained teachers conducted the scoring and moderation. Thereafter, the assessments were captured on item level and data cleaning, processing and analysis was conducted. The competency bands were set using procedures described in the next sections and thereafter subject specialists crafted the descriptions. Once descriptions were available, a report structure was created and reports generated using mail merge. The reports and data files were then sent to the schools to utilise.

3. Data analysis

Proficiency levels

The assessments used in this study included items based on a variety of rating scales, from dichotomous items, which had only right or wrong answers to items of five scores, with each increasing score denoting higher ability. Considering this, the Rasch partial credit model was used to analyse the results so that each item could be treated as an individual rating scale. Rasch-Thurstone thresholds set the 50% probability level (Linacre, 1998). These thresholds identify rating scaling positions on the latent variable at the precise point where observing each category is at the 50% probability level (Linacre, 2003, 2009). This dichotomises rating scales of items, so that that an item with, for example three scores (0, 1, 2), are dichotomised at the 0 and 1 as well as the 1 and 2 intersections (Linacre, 2003, 2009). This process simplifies the analysis and gives more precise locations as movement from one category to the next is compared per pair of ratings per item. Based on the 50% probability levels, the medians were calculated and used to create the number of levels that were decided upon *a priori*. For example, if five levels were hypothesised to exist in the assessment, the 50% probability levels were divided into five sections after sorting ascendingly and the median of each grouping was calculated which gave the logit points at which to group the person measures, from lower levels of proficiency to higher levels. Thereafter item maps, with the bands indicated, were generated and subject specialists were able to utilise the maps and align them with items to create descriptions per band or level. Figure 1 illustrates the Rasch Item Map method applied to the English language grade 8 results. Subject specialists agreed on three proficiency levels and cut-scores were set using the Rasch-Thurstone thresholds.

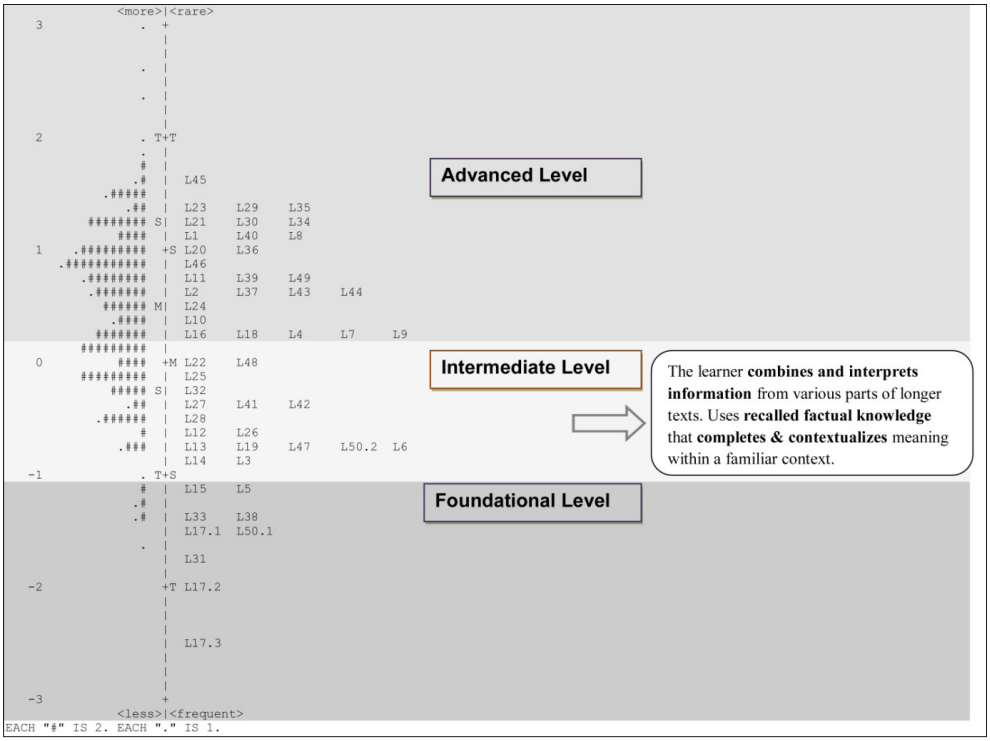

Figure 1: Item Map – Example of generating descriptions for grade 8 English language

Descriptions of different levels

Subject specialists for each subject were required to define the content areas of each band as indicated on the item maps after determining the points at which to cut the scores (see figure 1 as example). While it was decided beforehand to set five proficiency levels, examining the item maps showed that this was neither practical nor suitable for each subject. For English language specifically, the specialists determined that there were three proficiency levels when classifying the types of items and abilities associated with the items. Based on their recommendations, the levels were recalculated and item maps regenerated. The subject specialist re-examined the items and content in each level and after reviewing the items in each band, a description was generated per level. Other specialists reviewed and discussed the descriptions and once consensus was reached among the subject specialists, the descriptions were accepted and used for feedback. Table 4 shows the total number of levels per subject as well as examples of descriptions. Note that sections of the descriptions are shown as an example to illustrate the method used.

Table 4: Levels per subject and examples of level descriptions

Subject	Levels	Example
English language	Three levels: • Foundational • Intermediate • Advanced	Taken from the grade 8 intermediate description*: The learner combines and interprets information from various parts of longer texts using his/her recalled factual knowledge that completes and contextualises meaning within a familiar context.
Mathematics	Five levels: • Elementary • Intermediate • Adequate • Proficient • Advanced	Taken from the grade 9 proficient description*: Learners are able to write numbers in scientific notation and work with large numbers. A learner at this level can reason about decimal numbers and about rational and irrational numbers. The learner can use the operations on fractions in a context and work with the simple and compound interest formulae.
Natural science	Four levels: • Intermediate • Adequate • Proficient • Advanced	Taken from the grade 10 adequate description*: Learners at the adequate level are able to describe the mole as the SI unit for amount of substance. They can do basic stoichiometry calculations. They have some knowledge of ionisation energy. The learners are able to draw Lewis Dot Diagrams of elements. These learners can plot a heating curve for water.
*A section of the description for illustration purposes, descriptions per level were more extensive than illustrated here.		

Report design

This project served as an external monitoring system but in addition, results were also used to improve the teaching and learning within the schools. Teachers and principals attended interactive workshops in which the results were discussed. To give learners criterion-referenced feedback, a learner report format was devised in which the proficiency level with its description was shown as well as the subsequent level for which the learner should aim. Each learner received the criterion-referenced report to share with his/her parents. Teachers were provided with the learner reports, school reports as well as data sets containing the performance and proficiency levels of each learner per subject and per curriculum area. These results were used by schools to facilitate their intervention plan and to inform extra classes and Saturday classes so that learners could be assisted in content areas where developing proficiency was required. Table 5 below shows an excerpt from a learner report for grade 8 science. The table in the report first shows the level, in this case *adequate* and then offers a description of the curriculum-standard attained in the proficiency band. In the last column, the subsequent areas to attain are shown.

Table 5: Section of the learner report as an example from a grade 8 natural sciences section

Level	What you have learned	What comes next
3 Adequate	*You have gained knowledge about* matter, mixtures of elements *and* compounds. *You can recognise the* forces between particles *and discuss the* contraction of materials.	*You should focus on learning to* analyse chemical reactions *and* bonds *as well as identifying* which chemical test to use *in experiments.* *Learn how to apply your understanding of* density *in an investigation.*

4. Results and discussion

Proficiency levels across different cohorts

Figure 2 displays the three English language proficiency levels that could have been attained per grade based on the results. In grade 8, 41% of the learners were at the foundational level, 43% at the intermediate level and 16% had reached the advanced level. Considering the disadvantaged background of the learners and that they had only been attending school for a year, these results were to be expected. However, with each advancing year, fewer learners should fall in the foundational category with more learners moving to either the intermediate or advanced level. By grade 11, 36% of the learners could be classified as advanced, an improvement on the 16% at grade 8 level. Even though these are different cohorts, the selection criteria means that homogenous samples are selected yearly and some indication of improvement can be glimpsed from these results. The results indicate that English language proficiency increases with each advancing grade, possibly due to the increased exposure to English and the intervention model followed by the schools. Furthermore, as anchor items are included it is possible to make extrapolations across cohorts.

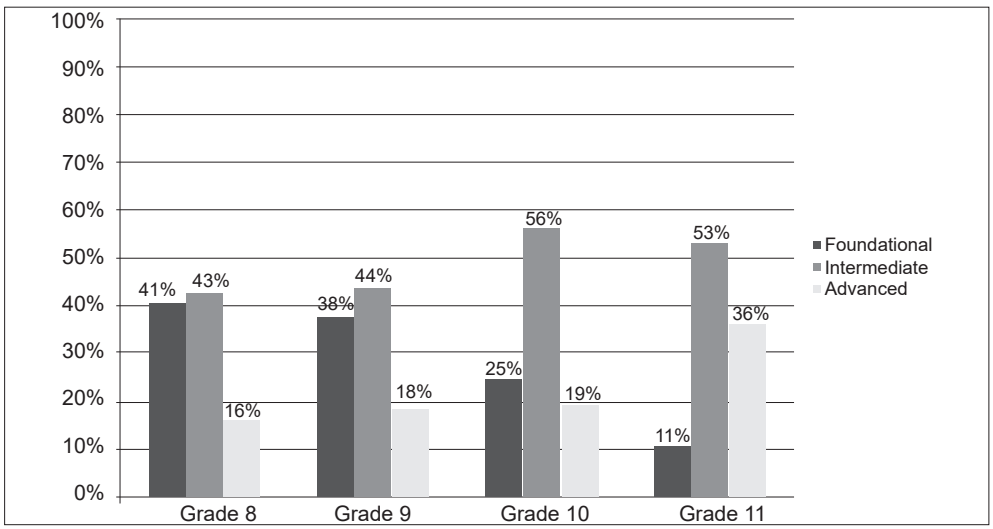

Figure 2: English language proficiency – percentage of learners in levels per grade

In figure 3, the results for the mathematics levels are presented. The results across the different grades remained mostly stable for grade 8, 9 and 10 with the majority of learners

falling into the adequate level. In grade 11, a movement towards the higher levels can be seen and more learners, 45% in total, are at the proficient level. It may take more years for the intervention school model to improve the mathematics ability of learners. Improved English language proficiency may also assist in improving performance in mathematics and science by improving the ability to read, comprehend and problem solve.

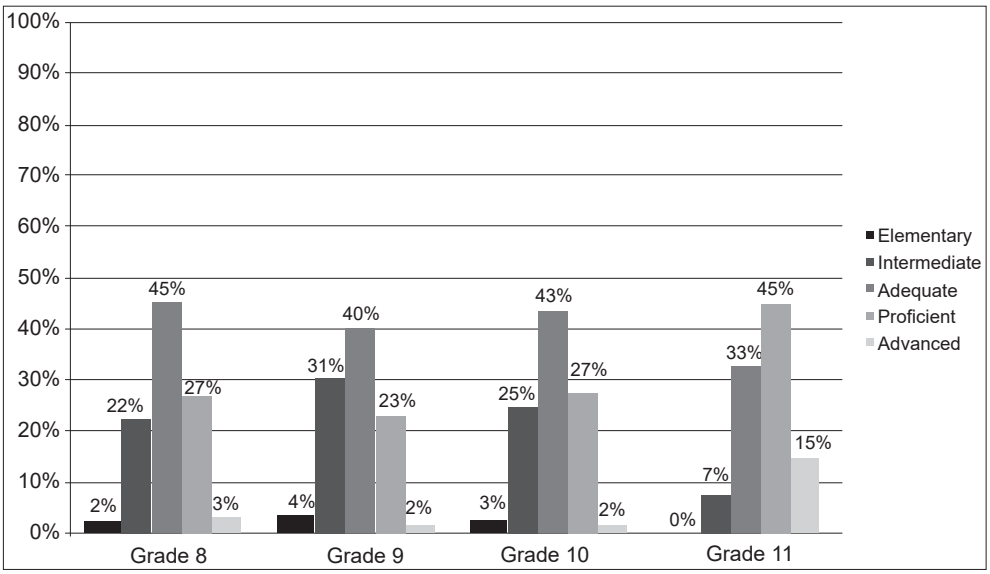

Figure 3: Mathematics proficiency – percentage of learners in levels per grade

Figure 4 illustrates the proficiency percentage levels of learners attained in the natural sciences assessments. Here a reversal of the pattern is observed, with more learners falling into the intermediate level (the least proficient level) with each advancing grade. Fewer learners seem to reach the more proficient levels in the higher grades, which may reflect the increasingly complex nature of the content and the challenges of attaining proficiency as difficulty in these content areas increases.

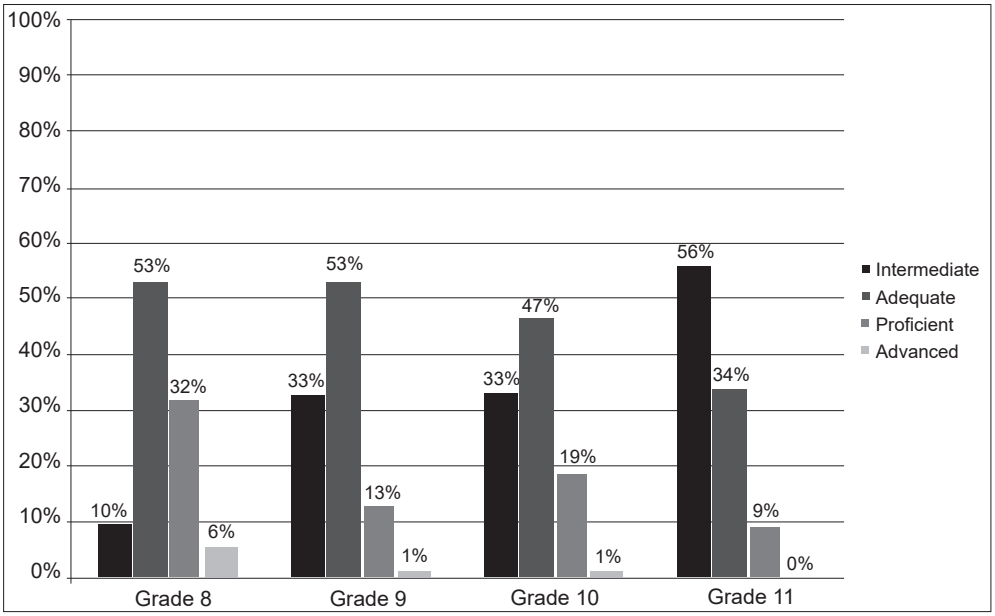

Figure 4: Science proficiency – percentage of learners in levels per grade

Teachers (N = 25) were asked how useful they found the reports to be. Eight per cent commented that they were *somewhat useful*, 56% reported that they were *very useful* and 36% verified that they were *entirely useful*. Comments sent by two of the schools are quoted below for illustration purposes.

School 1:

> *I love these, and will make sure that we communicate to teachers how to train parents how to read them. We are actively looking at how teachers can get a more accurate picture of every student for intervention.*

School 2:

> *Thank you for these comprehensive reports. They will be very helpful in driving our teaching.*

Figure 5 below illustrates the processes followed to design the learner feedback reports. The first step was to determine the cut-scores using the Rasch-Thurstone thresholds to calculate where the bands are located. Next, the subject specialists examined the items in each band and determined the skills and knowledge represented by each level. Subject specialists also influenced the setting of cut-scores when they found that the assessment bands were too broad or did not accurately reflect the proficiency levels. In such cases, subject specialist feedback facilitated recalculation of the bands. In the last stage, other subject specialists examined the descriptions and items and discussions assisted in reaching consensus on the descriptions, which resulted in the report format.

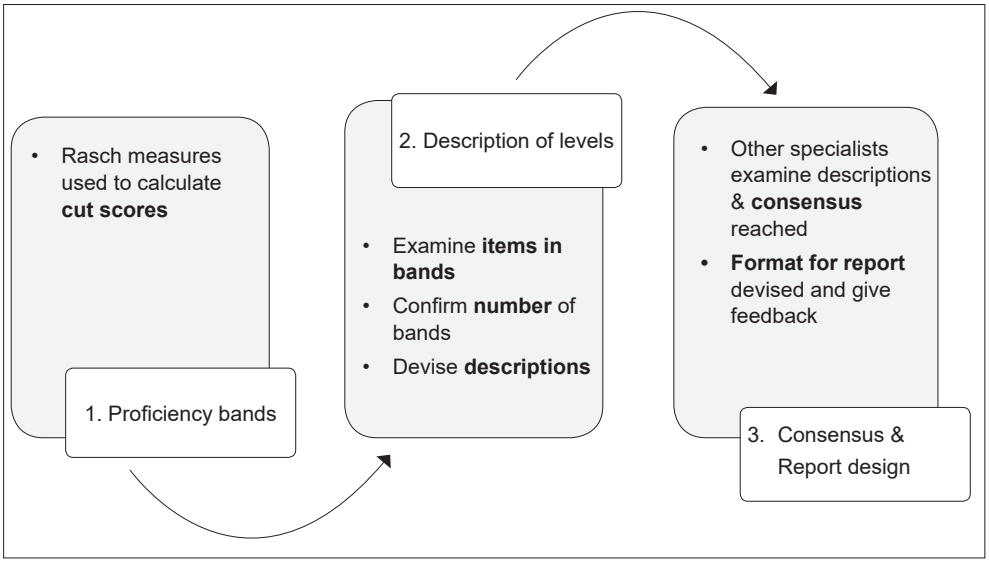

Figure 5: Process of creating criterion-referenced feedback

Implications for practice

The reporting of learner results should be moving past the "numbers only" era (de Vos & Belluigi, 2011; Green, 2002; Ottevanger, van den Akker & de Feiter; 2007; Popham, 1978). Increasingly there is a realisation that criterion-referenced results, based on curriculum-standards, are more aligned to the goals of educational assessment, which aims at mapping progression and outlining the subsequent developmental path. In addition, the equal interval measures provided by RMT also provide the opportunity to compare growth and development in subject areas (Ingebo, 1989).

This study has demonstrated the combination of using Rasch-Thurstone thresholds to set cut-scores for the proficiency levels and the value of subject specialists examining the items in the bands to create descriptions of the curriculum-standards represented by each band. The fact that not all subjects lend themselves to the same number of proficiency levels points to the fact that the quantitative numbers, the Rasch logits and the qualitative interpretations, by subject specialists, need to interact to inform the most appropriate definitions and explanations of what proficiency in a subject area means. Giving learners criterion-referenced feedback removes the focus from only aiming to attain a symbol and refocuses it on becoming more proficient in a learning area. This type of feedback is valuable in assisting teachers focus on specific subject areas in the curriculum, particularly where there are gaps in understanding so that learners can move more smoothly along the continuum of developing proficiency in a subject area. When teachers refocus their teaching, they are also likely to align their classroom assessments to the changes.

There are of course, limitations of implementation. Not all assessments have been designed to give criterion-referenced feedback. The process of setting cut-scores, creating descriptions for proficiency levels and validating the process requires expertise and resources to which teachers may not always have access. It should also be noted that in contexts where there are a greater number of learners in a class, it might also be more challenging to give each child

such detailed feedback. Where external assessments such as systemic and standardised tests are concerned, the focus may be on norm-referenced results for comparative purposes. This article suggests that even when benchmarking and comparing learners, schools or systems is the aim of the assessment, criterion-referenced reporting should be included as it can positively influence the learners and teachers and is specifically linked to curriculum-standards attainment. Large-scale assessments and systemic testing can be costly to implement and time-consuming for learners while disrupting teaching and school functioning. As a matter of social responsibility, the findings from such studies and assessments should be used to directly benefit the learners and criterion-referenced reporting is one way this can be accomplished.

While there are still challenges for reporting criterion-referenced results, the findings in this article suggests that progressively more emphasis should be placed on this type of feedback and less on norm-referenced feedback. The Rasch theory can effectively be used to set cut-scores to create proficiency bands and subject specialists should provide descriptions of each level and curriculum-standards represented in levels. The Rasch Item Map method can be used to align assessments and curriculum-standards by linking content to results. This results in diagnostic type feedback, which can be used by learners, parents and teachers to enhance teaching and learning.

Future research

It is suggested that continued research could expand the findings by having larger samples and conducting cross-validation studies using subsets of learners or test items (Jiao *et al.*, 2011). This would allow for examining the reproducibility of the competency levels in various sample sizes of learners and items across time. More stable parameters for classifying learners into competency bands could be identified, increasing the correct classification of persons. Ways in which to introduce criterion-referenced reporting in the school system should also be explored. Possibilities include assessments or items on a platform for teacher use which comprise proficiency bands for reporting purposes.

References

Bennett, J., Tognolini, J. & Pickering, S. 2012. Establishing and applying performance standards for curriculum-based examinations, Assessment. *Education: Principles, Policy & Practice,* 19(3), 321-339. https://doi.org/10.1080/0969594X.2011.614219

Bond, T.G. & Fox, C.M. 2015. *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Boone, W.J., Staver, J.R. & Yale, M.S. 2014. *Rasch analysis in the human sciences*. New York: Springer. https://doi.org/10.1007/978-94-007-6857-4

Department of Basic Education (DBE). 2011. *Curriculum and assessment policy statement (CAPS) grades 10-12: Physical sciences*. Pretoria: Government Printing Works.

Department of Basic Education (DBE). 2012. *National protocol for assessment in grades R-12.* Pretoria: Government Printing Works.

Department of Basic Education (DBE). 2016. *National curriculum statements (NCS) grades R-2*. Available at http://www.education.gov.za/Curriculum/NationalCurriculumStatementsGradesR-12.aspx [Accessed 20 October 2016].

de Vos, M. & Belluigi, D.Z. 2011. Formative assessment as mediation. *Perspectives in Education,* 29(2)*,* 39-47.

Fisher, W.P. 2007. Rating scale instrument quality criteria. *Rasch Measurement Transactions,* 2(1), 1095.

Great Schools Partnership. 2014. *Criterion referenced test*. Available at http://edglossary.org/criterion-referenced-test/ [Accessed 1 August 2016].

Green, S. 2002. Criterion referenced assessment as a guide to learning – the importance of progression and reliability. *Paper presented at the ASEESA conference, Johannesburg*. Available at http://www.cambridgeassessment.org.uk/images/109693-criterion-referenced-assessment-as-a-guide-to-learning-the-importance-of-progression-and-reliability.pdf [Accessed 1 August 2016].

Griffin, P., Gillis, S. & Calvitto, L. 2007. Standards-referenced assessment for vocational education and training in schools. *Australian Journal of Education,* 51(1)*,* 19-38. https://doi.org/10.1177/000494410705100103

Grosse, M. & Wright, B.D. 1986. Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Professions,* 9(3), 267-285. https://doi.org/10.1177/016327878600900301

Herrmann-Abell, C.F. & DeBoer, G.E. 2015. Using Rasch modeling to explore students' understanding of elementary school ideas about energy. *Paper presented at the NARST annual conference*, Chicago. Available at http://www.aaas.org/sites/default/files/CHA%26GDB-NARST%202015%20final.pdf [Accessed 1 August 2016].

Holster, T.A. & Lake, J.W. 2015. From raw scores to Rasch in the classroom. *Shiken,* 19(1), 32-41.

Ingebo, G. 1989. Educational research and Rasch measurement. *Rasch Measurement Transactions,* 3(1), 43-46.

Jiao, H., Lissitz, R.W., Macready, G., Wang, S. & Liang, S. 2011. Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling,* 53(4), 499-522.

Khosa, G. 2013. *Systemic school improvement interventions in South Africa: Some practical lessons from development practitioners*. Cape Town: JET Education Services.

Linacre, J.M. 1998. Thurstone thresholds and the Rasch Model. *Rasch Measurement Transactions,* 12(2), 634-635.

Linacre, J.M. 2003. Estimating 50% cumulative probability (Rasch-Thurstone) thresholds from Rasch-Andrich thresholds and vice-versa. *Rasch Measurement Transactions,* 16(3), 901.

Linacre, J.M. 2009. Dichotomizing rating scales and Rasch-Thurstone thresholds. *Rasch Measurement Transactions,* 23(3), 1228.

Linacre, J.M. 2011. *Winsteps® Rasch measurement computer program user's guide*. Beaverton, Oregon: Winsteps. Available at http://www.winsteps.com/manuals.htm [Accessed 1 August 2016].

Linacre, J.M. 2012. *Winsteps® Computer Software version 3.75.0*. Beaverton, Oregon: Winsteps.com.

Long, C., Dunne, T. & Mokoena, G. 2014. A model for assessment: Integrating external monitoring with classroom-based practice. *Perspectives in Education,* 32(1), 158-178.

Osman, S.A., Badaruzzaman, W.H.W., Hamid, R., Taib, K., Khalim, R., Hamzah, N. & Jaafar, O. 2008. Assessment on students' performance using Rasch model in reinforced concrete design course examination. *Recent Researches in Education,* 193-198.

Ottevanger, W., van den Akker, J.J.H. & de Feiter, L. 2007. *Developing science, mathematics, and ICT education in Sub-Saharan Africa: Patterns and promising practices*. South Africa: World Bank Publications. https://doi.org/10.1596/978-0-8213-7070-4

Popham, W.J. 1978. *Well crafted criterion-referenced Tests*. United States of America: Association for Supervision and Curriculum Development.

Shen, L. 2001. A comparison of Angoff and Rasch model based item map methods in standard setting. *Paper presented at the annual meeting of the American Educational Research Association*, Seattle. Available at http://eric.ed.gov/?id=ED452213 [Accessed 22 August 2016]

Smith, A.B., Rush, R., Fallowfield, L.J., Velikova, G., & Sharpe, M. 2008. Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology: Open Peer Review reports.* Available at http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-8-33 [1 August 2016].

Stelmack, J., Szlyk, J.P., Stelmack, T. & Babcock-Parziale, J. 2004. Use of Rasch person-item map in exploratory data analysis: A clinical perspective. *Journal of Rehabilitation Research & Development,* 41(2), 233-242. https://doi.org/10.1682/JRRD.2004.02.0233

Stone, G.E. 2000. A standard vision. *Popular Measurement: Journal of the Institute for Objective Measurement*, 4, 40-41.

Stone, G.E., Beltyukova, S. & Fox, C.M. 2008. Objective standard setting for judge-mediated examinations. *International Journal of Testing,* 8, 180-196. https://doi.org/10.1080/15305050802007083

Wang, N. 2003. Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement,* 40(3), 231-253. https://doi.org/10.1111/j.1745-3984.2003.tb01106.x

Wissing, M.P. 2013. *Well-being research in South Africa: 4 Cross-cultural advancements in positive psychology*. South Africa: Springer. https://doi.org/10.1007/978-94-007-6368-5

Wright, B.D. & Grosse, M. 1993. How to set standards. *Rasch Measurement Transactions,* 7(3), 315.