Art. #1539, 11 pages, https://doi.org/10.15700/saje.v38n3a1539

Inter-rater agreement in assigning levels of difficulty to examination questions in Life Sciences

Edith R. Dempster and Nicki F. Kirby
School of Education, University of KwaZulu-Natal, Pietermaritzburg, South Africa
dempstere@ukzn.ac.za

Public perception of "declining standards" in school-leaving examinations often accompanies increases in pass rates in school-leaving examinations. "Declining standards" to the public means easier examination papers. The present study evaluates a South African attempt to estimate the level of difficulty, as distinct from cognitive demand, to exit-level examination papers in Life Sciences. A team of four expert raters assigned a level of difficulty ranging from 1 (easy) to 4 (very difficult). Invalid items were assigned a difficulty level of 0. The reference point was "the ideal average South African learner." Discussion and practice was conducted for 12 examination papers, followed by individual analysis of four examination papers. Inter-rater agreement for the final four papers was low. Raters assigned most items to difficulty levels 1 and 2, indicating that unreliability may be caused by the instrument having too many levels. Raters' predictions of levels of difficulty supported the actual mark distribution for private school candidates, but not for public school candidates. The "ideal average South African learner" is an unsuitable reference point in the unequal educational landscape of the public school system. We recommend that the instrument be modified by reducing the number of levels of difficulty and removing the hypothetical reference point.

Keywords: comparability; difficulty; examinations; inter-rater agreement; reliability; standards

Introduction

Exit-level examinations at the end of schooling play a powerful role in life opportunities for students, with the results determining whether a student qualifies for entrance to higher education, employment, or whether a qualification is accredited by other countries (Leyendecker, Ottevanger & Van den Akker, 2008). Comparing the standards of different qualifications or across years within the same examining body is usually a subjective judgement of the whole qualification made by expert analysts (Eckstein & Noah, 1989; Leyendecker et al., 2008). Umalusi Council for Quality Assurance in General and Further Education has developed an objective method for comparing the standards of examination papers in the National Senior Certificate (NSC) across years. The present study evaluates the reliability of Umalusi's method of comparing the level of difficulty of Life Sciences examinations in the NSC. It does so by analysing inter-rater agreement among four raters when they independently rated the levels of difficulty of individual examination questions. The findings have implications for the reliability of expert rating as a technique for comparing standards of examinations in other subjects and contexts.

Comparability of difficulty in examination papers has been an ongoing problem in large-scale, high-stakes examinations in other parts of the world (Coe, 2008; Crisp & Novaković, 2009). In South Africa and in Britain, rising pass rates lead to public accusations that "standards have fallen" and the examinations must be easier than previous examinations (see, for example, Davis, 2016; Jansen, 2017; Paton, 2011). Jansen (2017:para. 2) wrote in a newspaper article: "Passing Grade 12 in South Africa is actually quite easy, and it means very little. The standards are low and the marks are adjusted upwards for most subjects."

South Africa has experienced curriculum revisions and changes in the structure of the exit level examinations at the end of Grade 12 between 1994 and the present (Department of Basic Education [DBE], Republic of South Africa, 2014b). Improving education was a priority of the post-apartheid government, and the pass rate in the exit-level examinations became the primary indicator of how well that goal was being achieved (DBE, Republic of South Africa, 2014b). By 2013, when the pass rate peaked at 78.3%, public and professional concerns about the standard and quality of the examinations were raised (DBE, Republic of South Africa, 2014b).

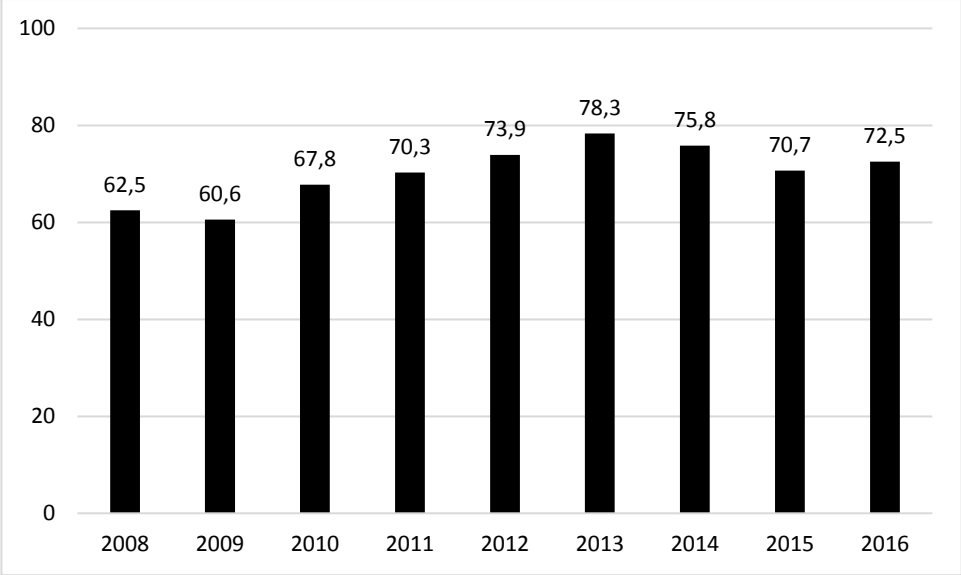

Figure 1 Pass rates (% of total candidates) for the NSC 2008–2016 (Figures from DBE, Republic of South Africa, 2016:36)

Given the high status of the exit-level examination results and the political imperative to improve the pass rate, it was important to demonstrate to the public and tertiary institutions that a rising pass rate is due to improved learning rather than lower standards of the examinations. Comparing national standards year-on-year is only valid post-2008, which was the first year in which all students wrote the same national examinations, based on the same curriculum, and received the National Senior Certificate (NSC). The minimum requirements for a pass in the NSC are 40% in three subjects, one of which must be Home Language, and 30% in a further three subjects. Students may fail one of the seven subjects. The Home Language may be any one of the eleven official languages of South Africa (DBE, Republic of South Africa, 2014b). Figure 1 shows the pass rates from 2008–2016.

Public criticism of the 18% increase in pass rates over the period 2009–2013 led to the Minister of Basic Education commissioning an independent task team to report on the standard and quality of the NSC. It released its report in June, 2014. One of its key findings was that the standard and quality of the NSC was improving (DBE, Republic of South Africa, 2014b), although there were still serious concerns about aspects of the examination process. The contrasting views of the ministerial task team and public critics point to different understandings of what we mean by "standards." The problem of how the British press and public understand "examination standards" is the subject of several research papers (see, for example, Baird, Cresswell & Newton, 2000; Coe, 2010). Baird et al. (2000) conclude that judging standards is a subjective process, influenced by the values of the person who makes the final decision.

The changing pass rates in South Africa over the period 2008–2016 have been influenced by circumstances. A new curriculum (the National Curriculum Statement, or NCS) was examined for the first time in 2008. Thus, increasing pass rates from 2009 onwards could be attributed to increasing familiarity with the NCS and the style of the examination papers, thereby decreasing the difficulty of the examinations. As from 2014, all examinations were set on a revised NCS, known as the Curriculum and Assessment Policy Statement (CAPS).

The South African exit-level examinations are rigorously monitored by Umalusi. To compensate for unanticipated variations in the mark distribution, a standardisation process sees the marks for the current year adjusted to match the average frequency distribution curve for the previous three to five years (Umalusi, 2016). In 2016, marks for 28 of the 58 subjects were adjusted upwards, and four subjects were adjusted downwards (Davis, 2016). Similar adjustments in 2015 were justified on the grounds that the examination papers were "demonstrably more difficult" than previous years' papers (Umalusi, 2015e). The large-scale upwards adjustments in 2016 were criticised by many commentators, including opposition politicians (Davis, 2016) and commentators (Jansen, 2017), on the grounds of declining standards of the examination papers.

Life Sciences in the NSC

Life Sciences has an enrolment of close to 300,000 students in the NSC. It was first examined in 2008. A second, revised Life Sciences curriculum was examined in 2011. The NCS-CAPS, first examined in 2014, was the third revision of the Life Sciences

curriculum since 2008. Notwithstanding the numerous curriculum changes, the proportion of candidates meeting the minimum pass requirement of 30% has remained relatively constant, as shown in Figure 2.

Figure 2 shows that examination of new curricula in 2008, 2011 and 2014 did not adversely affect the pass rates for those years. However, the official pass rate is released after standardisation has been applied. The final mark is a combination of examination marks, school-based assessment[i] marks, language compensation[ii] and standardisation (DBE, Republic of South Africa, 2014b). Figure 2 shows the raw pass rates for 2010–2013, which are the only available figures (DBE, Republic of South Africa, 2014b). Adjusted scores do not give a true indication of the difficulty of examinations, but the available raw scores indicate fluctuation in performance over the period 2010–2013. This could be ascribed to variation in the levels of difficulty of the examinations.

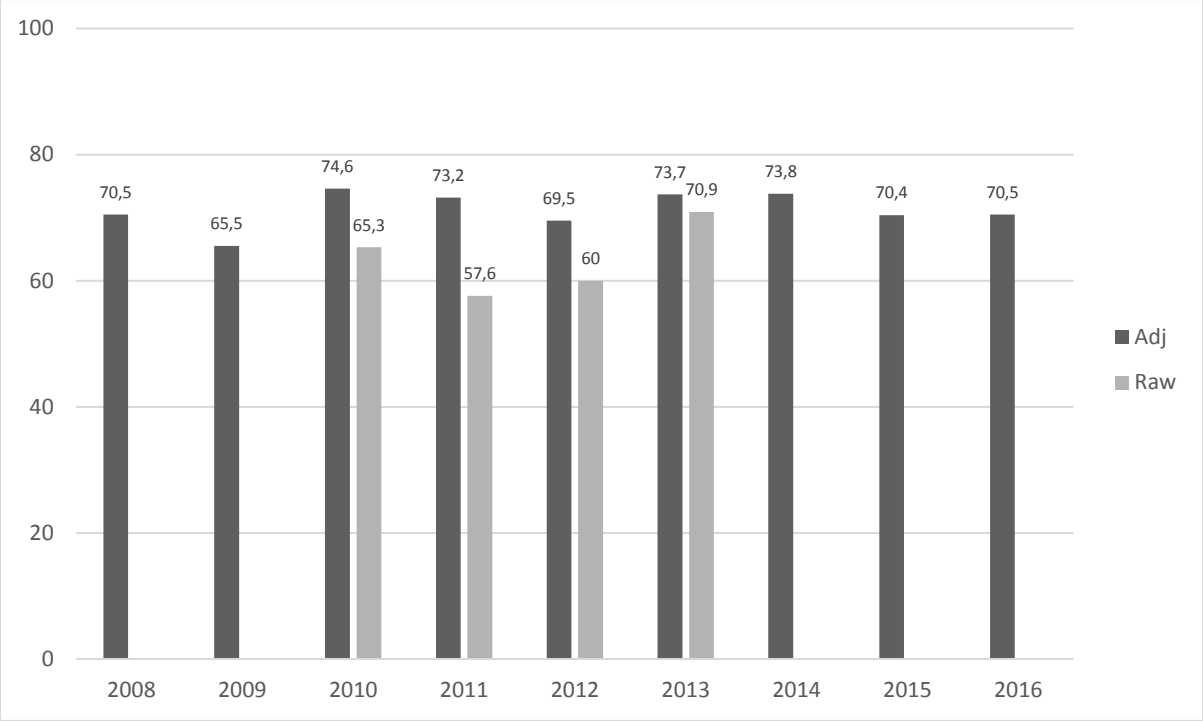

Figure 2 Adjusted pass rates (% of total candidates) for NSC Life Sciences 2008–2016 (DBE, Republic of South Africa, 2016:51; Umalusi 2015d:89). Raw pass rates for 2010–2013 only (DBE, Republic of South Africa, 2014b:167).

Several studies have compared the South African NSC curriculum and final examinations with other exit-level examining bodies. Umalusi, together with Higher Education South Africa (HESA), benchmarked the NSC against the Cambridge International Examinations, the International Baccalaureate and the Namibian Senior Secondary Certificate. NSC Life Sciences examinations were rated as less difficult than Cambridge AS and A-level, and International Baccalaureate Higher Level. However, they were judged to be more difficult than International Baccalaureate Standard Level, Cambridge International General Certificate for Secondary Education (IGCSE) and Namibian Higher Level and Ordinary Level (Grussendorff, Booyse & Burroughs, 2010).

The South African Department of Basic Education conducted benchmarking exercises with Cambridge International Examinations, the Scottish Qualification Authority (SQA) and the Board of Studies, New South Wales in 2010 and 2012 (DBE, Republic of South Africa, 2014b). In both years, external evaluators identified considerable problems with Life Sciences examination papers, in which the level of questioning was deemed too low, with too many closed questions. The SQA report in 2012 identified low cognitive challenge in the Life Sciences examinations, citing too many closed questions, very easy questions related to datasets and graphs, too few questions requiring knowledge of experimental procedure, insufficient development of scientific literacy, and short reading passages. All three examining bodies recommended that the examination papers ought to include critical thinking skills (DBE, Republic of South Africa, 2014b).

HESA evaluated the curriculum and examinations of 14 NSC subjects in 2012. The evaluators were subject specialists from 11 South African universities. The findings for Life Sciences were

that questions were mostly set at low cognitive levels, with very few higher order questions, which is considered to be essential for higher education (DBE, Republic of South Africa, 2014b).

Benchmarking studies therefore concur in expressing concern about the standard of Life Sciences examinations in the NSC. In the face of growing public criticism and poor external evaluations, it became imperative to track the level of difficulty of examination papers.

Distinguishing between Difficulty and Cognitive Demand

Difficulty is defined as "an empirical measure of how successful a group of students were on a question" as distinct from cognitive demand, which is defined as "the 'mostly' cognitive mental processes that a typical student is assumed to have to carry out in order to complete the task set by a question" (Pollitt, Ahmed & Crisp, 2007:169). Cognitive demand is described by a taxonomy such as Bloom's Taxonomy. Difficulty is derived from the ability of a student and the difficulty of the assessment task (Stiller, Hartmann, Mathesius, Straube, Tiemann, Nordmeier, Krüger & Upmeier zu Belzen, 2016). Although level of difficulty is affected by cognitive demand, it is possible for items to have low cognitive demand but high level of difficulty and vice versa. It is clear from Davis' (2016:para. 9) open letter to the Chief Executive Officer of Umalusi that he conflates cognitive demand with level of difficulty, by saying "… adjusting the raw mark upwards is justified if the exam paper was demonstrably more difficult (i.e. more cognitively demanding) than previous years."

The difference between cognitive demand and level of difficulty is illustrated by the following item from the Trends in International Mathematics and Science Study (TIMSS) 2003.

> The burning of fossil fuels has increased the carbon dioxide content of the atmosphere. What is a possible effect that the increased amount of carbon dioxide is likely to have on our planet?
> A. A warmer climate
> B. A cooler climate
> C. Lower relative humidity
> D. More ozone in the atmosphere

South African 8th Grade students who answered this question were divided into two groups, based on the apartheid classification of their schools:

1. Those who attended schools that were previously reserved for black African children (n = 1,019 students).
2. Those who attended schools that were previously reserved for Indian, coloured and white children (n = 212 students).

In Group 1, 19.6% of students selected the correct answer, while 45.8% of Group 2 students answered correctly. The question had the same cognitive demand for both groups, but was clearly more difficult for Group 1 students than Group 2 students (Dempster, 2007).

Factors Impacting Level of Difficulty

According to Pollitt et al. (2007), level of difficulty is most reliably estimated by analysing the scores obtained by students after completing an assessment task. Baird et al. (2000) list the challenges associated with comparing standards of examinations both before and after examination results are known. The difficulty of a question can vary for different cohorts of students, and over time (Crisp & Novaković, 2009). Coe (2010) lists many factors that could affect the difficulty of an examination, such as school type, quality of teaching, student motivation, gender, time devoted to the subject and level of interest in the subject. The diversity and inequity in South African schools creates an environment where "difficulty" is relative to a multiplicity of contextual factors.

Stiller et al. (2016) identified three features of multiple choice questions assessing scientific reasoning that increased item difficulty, namely: length of response options, use of specialist terms, and processing abstract concepts. One feature decreased item difficulty, viz.: processing data from tables. Stiller et al. (2016) recognised that processing abstract concepts is part of cognitive demand, which in turn contributes to item difficulty. Dempster and Reddy (2007) found that sentence complexity (number of words per sentence) was associated with poor performance of South African students answering multiple choice questions in the TIMSS 2003 study. The number of unfamiliar words (usually scientific terms) also contributed to difficulty, but its effect alone was not large enough to be significant.

Although no taxonomy of levels of difficulty exists, Leong (2006) proposed four locations in a test item where difficulty may reside. These were content difficulty, stimulus difficulty, task difficulty, and expected response difficulty. Task difficulty includes the cognitive demand of a test item, under the assumption that lower order cognitive processes are generally easier than higher order cognitive processes. Leong also identifies invalid moderators of difficulty, which impede or confound the measurement of a construct. Grammatical errors in the question, unclear mark allocation, and incongruence between mark scheme and question are a few examples of invalid questions.

Estimating Item Difficulty

Item response theory (IRT) has enabled a difficulty level to be assigned to test items by calibrating those items with a large number of students (Wauters, Desmet & Van den Noortgate, 2012). Coe (2008) used Rasch analysis to compare the levels of difficulty of different subjects in the General Certificate in Secondary Education (GCSE) examinations. Coe proposed that the comparisons are useful to identify an underlying construct, which he identified as "general academic ability." Coe found that GCSE subjects were not equivalent in terms of level of difficulty, but that a single trait, which he termed 'general achievement,' explained 83% of the observed variation in performance in 34 subjects. Coe (2010) proposed that comparability of examinations could be achieved by identifying the common construct revealed by performance in the examinations.

Wauters et al. (2012) claim that IRT is the most accurate measure of item difficulty. However, IRT-based calibration with a large sample size is not always possible in real assessment settings. They compared six alternative methods of estimating difficulty with IRT-calibrated results: *proportion of correct answers*, *learner feedback*, *expert rating*, *one-to-many comparison (learner)* and the *Elo rating system*. Not surprisingly, they found that the *proportion of correct answers* was the most closely related to IRT-calibrated difficulty estimates. *Expert rating* was the fourth most reliable measure of difficulty out of the six alternative methods.

The year 2014 marked a change in curriculum for the South African NSC, with the first examinations based on the NSC-CAPS curriculum. IRT-based calibration of test items was not possible because no previous examination questions based on the new curriculum were available. Umalusi's standardisation committee did not have historical norms on which to base its decisions. Umalusi therefore contracted teams of expert analysts to rate the level of difficulty of examination papers in a number of subjects before examination results were available. The method used matched Wauters et al.'s (2012) *expert rating* method of estimating difficulty.

Recognising that cognitive demand and level of difficulty are separate attributes of examination questions, Umalusi developed an instrument which requiring teams of expert raters to allocate each item on the examination papers to a type of cognitive demand, and separately to a level of difficulty (Umalusi, 2015b). Each team's report contributed to Umalusi's standardisation committee's decisions. This paper evaluates the reliability of expert rating of difficulty by analysing inter-rater agreement among four expert raters for Life Sciences. If raters achieve a high level of agreement, expert rating is a reliable method of estimating difficulty of examination questions. If raters do not achieve a high level of agreement, expert rating is unreliable. Reasons for unreliability must then be sought.

Method

In South Africa, two major examining bodies offer exit-level examinations in Life Sciences. Private schools are fee-paying schools and constitute a small proportion of the schools in the country. Most public schools are fee-free, although a few are semi-independent of the state and charge fees. Both private and public schooling systems follow the same curriculum. The private Independent Examinations Board (IEB) and the public Department of Basic Education (DBE) set and administer examinations independently, but all examination papers are quality controlled by Umalusi.

Candidates for both examining bodies write two theory examination papers in Life Sciences. The examination papers follow the same format, but each paper tests knowledge and skills related to different topics in the curriculum.

> Section A: Multiple choice and other questions requiring short answers (50 marks);
> Section B: A variety of questions requiring interpretation of diagrams or data, short written answers to specific questions, and graph-drawing (80 marks);
> Section C: Essay (20 marks).

Umalusi tasked a team of four expert raters with estimating the level of difficulty of the 2014 examination papers compared with the levels of difficulty of examination papers of the previous three years (Umalusi, 2015b). The team comprised one Life Sciences teacher from each of a Quintile 5 public school (rater TD) and an independent school (rater TI), and one subject advisor for Life Sciences in public schools (rater SA). The team leader, a university academic (rater UA), has considerable experience in evaluating the cognitive demand and level of difficulty of Life Sciences examination papers. Rater TD had also participated in several previous evaluations of the standard of Life Sciences examination papers. Raters SA and TI were new to the process. The team therefore represented a diversity of professional experience of the South African educational landscape.

The evaluation team was required to make a judgement on a scale of 0–4 of the level of difficulty of each item. An item was defined as the smallest unit of a question on each examination paper. The task required raters to estimate levels of difficulty for a student of average intelligence who was assumed to have studied the whole syllabus and been taught by a competent teacher (Umalusi, 2015b). Items that contained invalid sources of difficulty (Leong, 2006) were scored as 0. Level 1 in the scale of difficulty was "easy" for the average student to answer, 2 was "moderately challenging," 3 was "difficult" and 4 "very difficult." Level 4 items would discriminate the highest-achievers from other

students (Umalusi, 2015b). The task asked raters to consider four sources of difficulty proposed by Leong (2006), namely content difficulty, stimulus difficulty, task difficulty and expected response difficulty (Umalusi, 2015b).

The evaluation team met and discussed possible contributors to levels of difficulty. Reliance on specialist terminology, concepts that students traditionally find difficult, abstract concepts, and clarity of the wording of questions were identified as factors contributing to difficulty.

The team analysed three past papers together as a group in October 2014. Each item on each examination paper was assigned a level of difficulty after discussion. The team then separated, and analysed a further nine past papers independently. Individual analyses were collated by the team leader, who identified items where less than three raters agreed. At a second meeting in November 2014, the team revised their analyses until they reached closer consensus on levels of difficulty. The team worked well together, and discussions were conducted in a cordial manner.

The final four papers of December 2014 were analysed by each team member working independently. Results were used for statistical testing of inter-rater agreement after the intensive practice in October and November. Analysts also completed a questionnaire in which they were asked how difficult it was to assign a level of difficulty to a question, and whether they referred to criteria to decide.

Statistical Analysis

Measures of inter-rater agreement have been applied in various fields to assess inter-rater reliability (Fleiss, Levin & Paik, 2003). Gwet's Agreement Coefficient (AC1) can be applied to multiple raters and multiple-item responses on a nominal scale (Gwet, 2014) and was the most suitable coefficient of inter-rater agreement in the present study. Furthermore, Gwet's AC1 statistic is stable when ratings are skewed towards marginal response categories as was the case in this study (e.g., high frequencies of ratings for difficulty levels 1 and 2 category), which interferes with the correction for chance-agreement (Gwet, 2014).

Fleiss et al. (2003:604) suggest that for most purposes, values of the agreement coefficient greater than 0.75 represent excellent agreement beyond chance, values between 0.4 to 0.75 represent fair to good agreement, and values less than 0.4 suggest poor agreement beyond chance.

The average pairwise percent agreement for each item indicates the overall impact of chance agreement correction, and either supports or refutes a coefficient of agreement (Neuendorf, 2002). The agreement among all possible pairs is calculated and averaged for each item. For example, if two raters (UA and SA) agree on a level of difficulty and two raters (TD and TI) agree with each other, but disagree with UA and SA, the average pairwise percentage agreement is calculated for all possible pairs (UA & TD, UA & SA, UA & TI, TD & SA, TD & TI, SA & TI).

	UA	TD	SA
TD	0		
SA	100	0	
TI	0	100	0

The average pairwise percentage agreement is ((100 x 2) + (0 x 4))/6 = 33 percent. The percentage agreement is then averaged for all the items on an examination paper. Percentage agreements of 90% or greater are nearly always acceptable (Lombard, Snyder-Duch & Bracken, 2002); 80% is acceptable in most situations; and 70% may be appropriate in some exploratory studies (Neuendorf, 2002).

Results

Table 1 Inter-rater agreement on level of difficulty (including agreement on invalid questions) for four Life Sciences papers (n = 4 raters)

	4 agree	3 agree	2 agree + (2 sets 2)	None agree	Gwet's AC1	Pairwise percent agreement
IEB Paper 1 (59 items; 7 classified as invalid across the 4 raters)						
% of items	15.3%	39.0%	27.1% + (17.0%)	1.7%	0.34 Poor	45%
IEB Paper 2 (52 items; 6 classified as invalid across the 4 raters)						
% of items	23.1%	42.3%	21.2% +(11.5%)	1.9%	0.42 Fair	52%
DBE Paper 1 (68 items; 5 classified as invalid across the 4 raters)						
% of items	16.2%	51.5%	14.7% +(17.7%)	0%	0.41 Fair	51%
DBE Paper 2 (65 items; 8 classified as invalid across the 4 raters)						
% of items	6.2%	63.1%	20.0% +(10.8%)	0%	0.33 Poor	45%

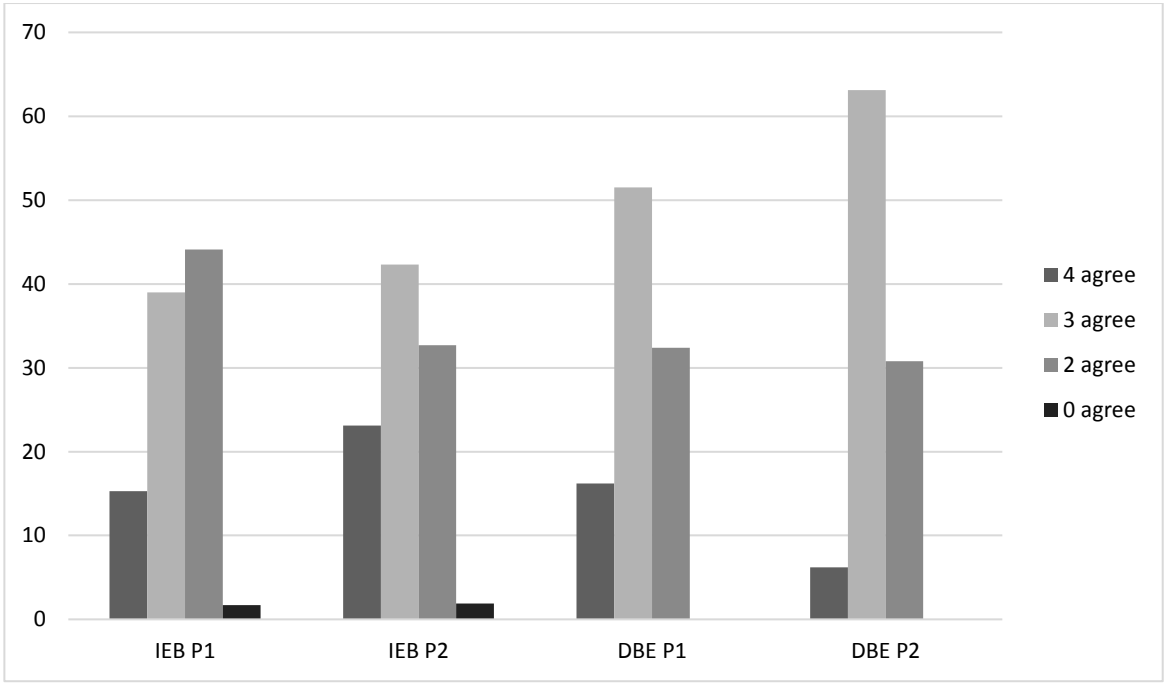

Figure 3 Percentage of items at each level of inter-rater agreement on difficulty

Table 1 summarises the level of agreement among the four raters for each examination paper. To enable comparison among the papers, levels of agreement are expressed as percentages of the total number of items on each examination paper. Thus, for IEB Paper 1, all four raters agreed on 15.3% of the 59 items and three raters agreed on 39% of the items. In 27.1% of the items, two raters agreed and two disagreed, while in a further 17% of items, two raters agreed on one level of difficulty, while the other two agreed on a different level of difficulty. All four raters disagreed in 1.7% of the items. This interpretation applies to all four examination papers.

Figure 3 shows the same information as Table 1. Table 1 and Figure 3 show that complete disagreement among the four raters rarely occurred, and that complete agreement among all four raters was low. In IEB P1, the highest percentage agreement was between two raters, while for the three remaining papers, the highest percentage agreement was among three raters. The percentage of items where three or four raters agreed was 69.3% for DBE P2, 67.7% for DBE P1, 65.4% for IEB P2 and 54.3% for IEB P1.

Using the guidelines developed by Fleiss et al. (2003), the coefficient of agreement (Gwet's AC1) showed that IEB P1 and DBE P2 had poor agreement beyond chance, and IEB P2 and DBE P1 were just above the boundary for fair agreement. Pairwise percent agreement was below 70% in all four papers, falling to 45% in IEB P1 and DBE P2. It did not achieve the 70% considered acceptable for exploratory studies in the social sciences (Neuendorf, 2002).

In the survey questionnaires, raters stated that they were confident of their ratings of levels of

difficulty, and rarely referred to criteria discussed in the training session. They acted intuitively, based on their knowledge of the curriculum and what types of questions students find easy, and which they find difficult. Apart from UA, other team members reported that they found it difficult to identify invalid questions.

The second part of the study attempted to understand whether poor inter-rater agreement could be ascribed to any one or two raters. The different professional experiences of the raters could have led to different perceptions of difficulty. Since the category "invalid question" was rarely used, its results are omitted from this analysis. For each rater, the number of items assigned to each level of difficulty was totalled for each of the four papers. The mean ±*SD* was then calculated for all four papers. Figure 4 shows the results.

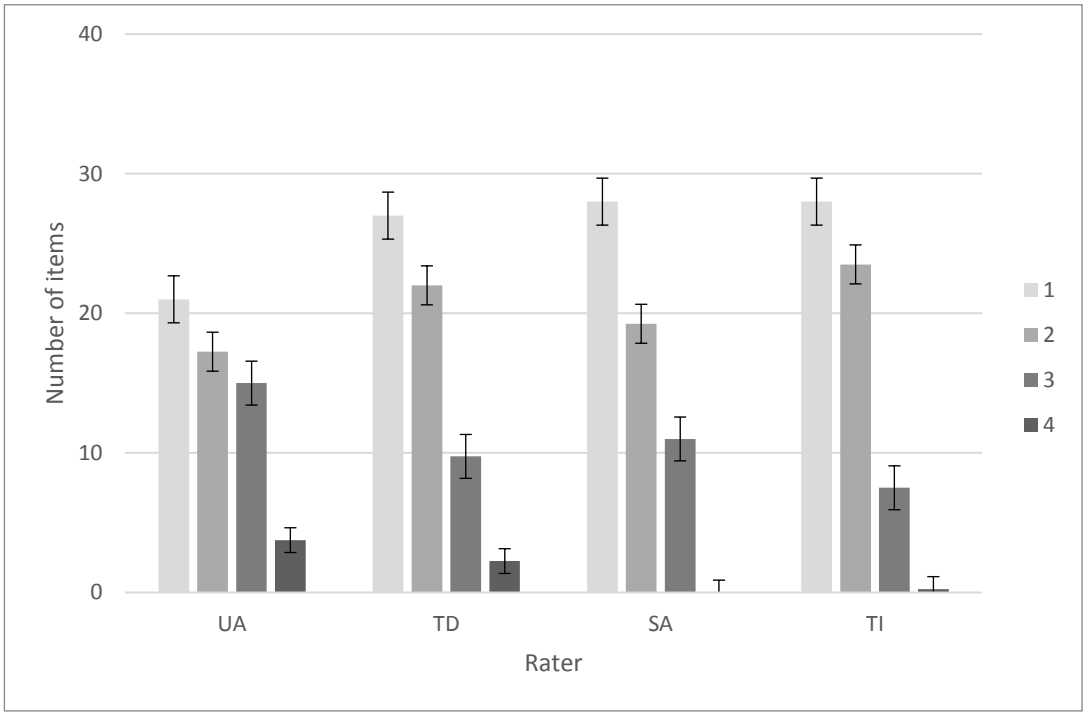

Figure 4 Mean ± *SD* number of items assigned to levels of difficulty 1–4 by each rater (n = 4 examination papers)

Figure 4 shows that all four raters rated more items as level 1 (easy) than any other category. The second-highest level was 2 (moderately challenging). Level 4 (very difficult) was rarely used by any of the raters. Rater UA rated more questions as *difficult* and *very difficult* than the other three raters. Rater TI rated more questions as *easy* or *moderately challenging* than all other raters. The small standard deviations indicate that there was little variation in each raters' ratings among the four papers.

A chi-squared test yielded a value of 15.61, which is not large enough to be significant (p = 0.16). Cramer's V measures the strength of association between two categorical variables, in this case, rater and use of levels of difficulty. Cramer's V confirmed the results of the chi-squared test and showed that the effect size was small (V = 0.154, p = 0.134).

Table 2 summarizes the percentage of items assigned to levels 1 and 2, 3 and 4 and invalid items by each rater. Table 2 points to a possible source of unreliability. Since most of the items were rated 1 or 2 by all four raters, unreliability could result from disagreement in distinguishing *easy* from *moderately challenging*. This could be a flaw in the task, which had too many levels of difficulty.

Table 2 Summary of mean percentage of items assigned to invalid, easier (1 & 2) and more difficult (3 & 4) levels

Rater	Invalid	Levels 1 + 2	Levels 3 + 4
UA	6.6%	62.3%	31.2%
TD	0%	80.4%	19.7%
SA	4.9%	77.0%	18.0%
TI	1.6%	85.2%	13.1%

Discussion

This study provides empirical evidence that, despite intensive practice and discussion, a team of four expert raters achieved low inter-rater agreement in evaluating the level of difficulty of Life Sciences examination papers using a 5-level rating. This is in agreement with Wauters et al. (2012), who found that expert rating was the fourth most accurate of six methods to estimate item difficulty in examinations. Our findings also support the view of Baird et al. (2000) that standard-setting is subjective and likely

influenced by the professional experience, values and subject competence of the standard-setters. Davis (2016) argues that the cognitive demand of South African examination papers (which he wrongly equates with level of difficulty) ought to be the starting point for standardisation, not the marks. The present study has shown that even with extensive practice and discussion, expert rating of level of difficulty in Life Sciences is unreliable.

Further exploration of the way in which individual raters differed in their analyses revealed that raters assigned many more items to levels 1 and 2, and few items to 3 and 4. The source of much disagreement therefore lies in the subtle distinction between *easy* and *moderately challenging*. The problem then lies with the instrument, which requires too fine a distinction to enable inter-rater reliability.

The extent to which the predicted levels of difficulty matched the marks obtained by students in the 2014 examinations indicates that factors unknown to expert raters affect the examination marks. Although the raters rated the examination questions as mostly easy or moderately challenging, changes to the structure of the 2014 examination papers led raters to suggest that students might experience the examinations as more difficult than in previous years (Umalusi, 2015b). The mark distribution showed that the opposite was true: the marks for 2014 were higher than previous years, and were lowered to match the three-year norm (Umalusi, 2014). The Umalusi media statement on the approval for release of NSC examination results for 2014 states the following:

> "… the learner performance in 2014 was the best in any previous year […]. A downward adjustment was therefore done."

Further support for the unreliability of raters' assessment of difficulty emerged in 2015, when the raters judged the examination papers for the DBE to be easier than 2014 (Umalusi, 2015a). The standardisation committee found that the marks were considerably lower than the historical norms, and the marks were adjusted upwards (Umalusi, 2015e). Clearly, raters' assessment of the level of difficulty of Life Sciences examinations does not match the results.

The effect of context on students' experience of the difficulty of an examination paper is illustrated by the mark distributions for DBE and IEB examinations. Both sets of examination papers were judged by all the raters in the present study to contain mostly easy and moderately challenging questions (Umalusi, 2015b, 2015c). After standardisation, 49.0% of the 5,177 IEB candidates achieved a final mark above 70% (IEB, 2014), while only 8.5% of the 284,298 DBE candidates achieved the same benchmark (DBE, Republic of South Africa, 2014a). Raters' estimations of the levels of difficulty of examination items as mostly easy or moderately challenging were supported by IEB schools' results, but not DBE schools' results.

This mismatch between raters' evaluations of examination papers and actual performance casts doubt on the reliability of expert raters' estimation of levels of difficulty of examinations in DBE schools. Items rated by expert raters to be easy or moderately challenging are experienced by most DBE learners as difficult or very difficult. Here we agree with Coe (2010) that difficulty of an examination is affected by many factors. In the South African context, quintile of the school, quality of teaching experienced, amount of time devoted to teaching the subject, and student motivation are likely to contribute to difficulty.

Unreliability can at least partly be ascribed to the instrument, which required fine distinctions between four levels of difficulty. In addition, raters were asked to make judgements for the "ideal average South African learner." Raters ascribed most items to difficulty levels 1 and 2, which may be correct for the ideal average IEB learner, but not for the current majority of DBE candidates. Inequality in the educational system makes it difficult to conceptualise the ideal average DBE learner.

Conclusion and Recommendations

This paper has shown that expert raters achieve low inter-rater agreement using a 5-level taxonomy of levels of difficulty. Raters rated most of the items on 2014 examination papers for IEB and DBE as easy or moderately challenging, more rarely as difficult and very few items as very difficult. Most of the unreliability could therefore be attributed to lack of agreement on the distinction between "easy" and "moderately challenging" items. We recommend the levels of difficulty should be reduced to "easy/moderately challenging" and "difficult/very difficult" to increase reliability. The category "invalid difficulty" should be retained to capture items that contain errors or that lack construct validity.

While the concept of the "ideal average South African learner" is a noble aspiration, it is difficult to conceptualise in the diverse South African educational landscape. This requirement of the task may have influenced raters to rate most items as easy or moderately challenging, but the reality for most learners is clearly quite different. If Umalusi wishes to continue estimating levels of difficulty of examinations before results are available, it ought to revise the standard against which items are to be evaluated. The professional experience of individual raters is likely to influence their concept of the "ideal average South African learner" since a teacher from a Quintile 5 school has a different experience of learners than a teacher from a Quintile 1 school. We therefore recommend removing the

hypothetical benchmark, given the diversity of school contexts in South Africa.

At present, low inter-rater reliability indicates that expert rater evaluations of levels of difficulty should be used with caution in standardising the NSC results. Reliability can be improved with a revised task instrument.

Acknowledgements

We wish to express our thanks to Umalusi for making this study possible, and to the four raters who spent many hours assigning levels of difficulty to items. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 104666) - https://doi.org/10.13039/501100001321.

Notes

i. School-based assessment contributes 25% of the final mark.
ii. Students who offer an African language as Home Language qualify for a 5% compensation on the mark they have obtained in any non-language subject (DBE, Republic of South Africa, 2014b).
iii. Published under a Creative Commons Attribution Licence.

References

Baird JA, Cresswell M & Newton P 2000. Would the real gold standard please step forward? *Research Papers in Education*, 15(2):213–229. https://doi.org/10.1080/026715200402506

Coe R 2008. Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education*, 34(5):609–636. https://doi.org/10.1080/03054980801970312

Coe R 2010. Understanding comparability of examination standards. *Research Papers in Education*, 25(3):271–284. https://doi.org/10.1080/02671522.2010.498143

Crisp V & Novaković N 2009. Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation & Research in Education*, 22(1):3–15. https://doi.org/10.1080/09500790902855776

Davis G 2016. Open letter to Umalusi. *Polotiki News*, 30 December. Available at https://polotiki.com/2016/12/30/open-letter-to-umalusi-gavin-davis/. Accessed 23 May 2017.

Dempster ER 2007. Textual strategies for answering multiple choice questions among South African learners: What can we learn from TIMSS 2003? *African Journal of Research in Mathematics, Science and Technology Education*, 11(1):47–60. https://doi.org/10.1080/10288457.2007.10740611

Dempster ER & Reddy V 2007. Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, 91(6):906–925. https://doi.org/10.1002/sce.20225

Department of Basic Education, Republic of South Africa 2014a. *National Senior Certificate examination 2014: Diagnostic report*. Pretoria: Author. Available at https://www.slideshare.net/NITheP/2014-nsc-diagnostic-report. Accessed 1 August 2018.

Department of Basic Education, Republic of South Africa 2014b. *Report of the Ministerial Committee to investigate the current promotion requirements and other related matters that impact on the standard of the National Senior Certificate*. Pretoria: Author. Available at https://www.researchgate.net/publication/303230414_Report_of_the_Ministerial_Committee_to_investigate_the_current_promotion_requirements_and_other_related_matters_that_impact_on_the_standard_of_the_National_Senior_Certificate. Accessed 13 June 2017.

Department of Basic Education, Republic of South Africa 2016. *National Senior Certificate examination report 2016*. Pretoria: Author. Available at https://www.education.gov.za/Portals/0/Documents/Reports/NSC%20EXAMINATION%20REPORT%202016.pdf?ver=2017-01-05-110635-443. Accessed 13 June 2017.

Eckstein MA & Noah HJ 1989. Forms and functions of secondary-school-leaving examinations. *Comparative Education Review*, 33(3):295–316. https://doi.org/10.1086/446860

Fleiss JL, Levin B & Paik MC 2003. *Statistical methods for rates and proportions* (3rd ed). Hoboken, NJ: John Wiley & Sons.

Grussendorff S, Booyse C & Burroughs E 2010. *Evaluating the South African National Senior Certificate in relation to selected international qualifications: A self-referencing exercise to determine the standing of the NSC* (Overview report). Pretoria, South Africa: Higher Education South Africa (HESA) & Umalusi. Available at http://www.umalusi.org.za/docs/research/2010/iq_overview_report.pdf. Accessed 2 August 2018.

Gwet KL 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed). Gaithersburg, MD: Advanced Analytics, LLC.

Independent Examinations Board 2014. *2014 NSC examination results*. Pretoria, South Africa: Author.

Jansen J 2017. Dear Grade 12 pupil: Let's talk about your matric pass. *Rand Daily Mail*, 5 January. Available at https://www.businesslive.co.za/rdm/lifestyle/2017-01-05-dear-grade-12-pupil-lets-talk-about-your-matric-pass. Accessed 23 May 2017.

Leong SC 2006. *On varying the difficulty of test items*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore, 21–26 May. Available at http://www.iaea.info/documents/paper_1162a1d9f3.pdf. Accessed 1 August 2018.

Leyendecker R, Ottevanger W & Van den Akker J 2008. *Curricula, examinations, and assessment in secondary education in sub-Saharan Africa* (World Bank Working Paper No. 128). Washington, DC: The World Bank. Available at http://siteresources.worldbank.org/INTAFRREGTOPSEIA/Resources/No.5Curricula.pdf. Accessed 3 October 2016.

Lombard M, Snyder-Duch J & Bracken CC 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Neuendorf KA 2002. *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Paton G 2011. Ofqual: Alarm over falling A-level and GCSE standards. *Telegraph*, 12 May. Available at https://www.telegraph.co.uk/education/secondaryeducation/8510925/Ofqual-alarm-over-falling-A-level-and-GCSE-standards.html. Accessed 8 May 2017.

Pollitt A, Ahmed A & Crisp V 2007. The demands of examination syllabuses and question papers. In P Newton, JA Baird, H Goldstein, H Patrick & P Tymms (eds). *Techniques for monitoring the comparability of examination standards*. London, England: Qualifications and Curriculum Authority.

Stiller J, Hartmann S, Mathesius S, Straube P, Tiemann R, Nordmeier V, Krüger D & Upmeier zu Belzen A 2016. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, 41(5):721–732. https://doi.org/10.1080/02602938.2016.1164830

Umalusi 2014. *Media statement: Approval decisions by Prof John Volmink (Chair of Umalusi Council) 30 December 2014*. Pretoria, South Africa: Author. Available at www.umalusi.org.za/docs/pr/2014/pr1230a.pdf. Accessed 2 June 2017.

Umalusi 2015a. *Comparison of the National Senior Certificate examinations administered by the Department of Basic Education: 2013 - 2015. Accounting, Business Studies, Economics, Geography, History, Life Sciences, Mathematics, Mathematical Literacy and Physical Sciences.* Pretoria, South Africa: Author. Available at https://www.umalusi.org.za/docs/assurance/2016/Composite-Report-DBE.pdf. Accessed 7 June 2017.

Umalusi 2015b. *Consolidated post-exam analysis report 2014: Content subjects - DBE*. Pretoria, South Africa: Author. Available at www.umalusi.org.za/docs/assurance/2015/dbe.pdf. Accessed 1 June 2017.

Umalusi 2015c. *Consolidated post-exam analysis report 2014. Content subjects - IEB.* Pretoria, South Africa: Author. Available at http://www.umalusi.org.za/docs/assurance/2015/ieb.pdf. Accessed 1 June 2017.

Umalusi 2015d. *Indicators Report 2008-2013: National Senior Certificate*. Pretoria, South Africa: Author. Available at http://www.umalusi.org.za/docs/reports/2015/Indicators-Report-2008-2013.pdf. Accessed 7 June 2017.

Umalusi 2015e. *Media statement: Approval decisions by Prof John Volmink (Chair of Umalusi Council) 30 December 2015*. Pretoria, South Africa: Author. Available at https://www.umalusi.org.za/docs/pr/2015/pr1230.pdf.pdf. Accessed 7 June 2017.

Umalusi 2016. *Requirements and specifications for standardization, statistical moderation and resulting* (Version 5). Pretoria, South Africa: Author. Available at https://www.umalusi.org.za/docs/reports/2016/REQUIREMENTS%20AND%20SPECIFICATIONS%20FOR%20THE%20STANDARDISATION%20STATISTICAL%20MODE.pdf. Accessed 7 June 2017.

Wauters K, Desmet P & Van den Noortgate P 2012. Item difficulty estimation: An auspicious collaboration between data and judgement. *Computers & Education*, 58(4):1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020