Dr Edith R. Dempster
School of Education,
University of KwaZulu-Natal,
Pietermaritzburg, dempstere@
ukzn.ac.za

Dr Nicola F. Kirby
School of Education,
University of KwaZulu-Natal,
Pietermaritzburg

DOI: http://dx.doi.org/10.18820/2519593X/pie.v36i1.7
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
2018 36(1): 94-110
© UV/UFS

Inter-rater agreement in assigning cognitive demand to Life Sciences examination questions

Abstract

Taxonomies of cognitive demand are frequently used to ensure that assessment tasks include questions ranging from low to high cognitive demand. This paper investigates inter-rater agreement among four evaluators on the cognitive demand of the South African National Senior Certificate Life Sciences examinations after training, practice and revision. The taxonomy used was based on the cognitive dimension of the Revised Bloom's Taxonomy, with analysis, evaluation and synthesis combined into one category. Descriptors from the Revised Bloom's taxonomy were slightly modified to suit Life Sciences. Inter-rater agreement was poor to fair, but pairwise percent agreement did not reach acceptable levels. Evaluators found it difficult to assign cognitive demand to examination items, and constantly referred to the descriptors. We question the usefulness of a taxonomy of cognitive demand when individuals differ in their interpretations of the levels of cognitive demand. The results indicate that standards of Life Sciences examination papers cannot reliably be assessed by evaluating cognitive demand using Bloom's Taxonomy.

Keywords: *assessment; cognitive demand; inter-rater agreement; Bloom's taxonomy; reliability*

1. Introduction

Exit-level examinations are a form of summative assessment, the results of which serve several different purposes in an education system.

- Selection of students for further study or the world of work (Broadfoot, 2007).
- Monitoring what learning has been achieved (Burke, 2010; Harlen, 2012; Newton, 2007; Taras, 2005) and whether the goals or outcomes stated in a curriculum have been achieved. The monitoring purpose enables authorities to ensure accountability of schools and the educational system (Atkin & Black, 2003; Kellaghan & Greaney, 2004). It also promotes equality of provision of educational opportunities across the schooling system (DBE, 2015: 13).
- Examinations enable comparability of educational standards across different syllabuses, between years,

between subjects and between countries (Baird *et al*., 2000; Cresswell, 2000; Kellaghan & Greaney, 2004).

Early studies comparing school-leaving qualifications were generally qualitative, providing a subjective evaluation of the whole qualification (Eckstein & Noah, 1989; Kellaghan & Greaney, 2004). Their reliability is therefore questionable. Efforts to improve the reliability of cross-national studies are illustrated by a major study comparing the curriculum and assessment at exit level of four subjects in 16 countries (Ofqual, 2012). Expert judgement was used to evaluate the demand of comparable qualifications. As explained by Pollitt *et al.* (2007: 168),

> *There is no statistical indicator of demands, and no prospect of our developing objective scales for assessing them. Instead, we rely on the judgement of experienced professionals.*

The South African National Senior Certificate (NSC) has three important functions. It has a strong selection function as it determines access to different post-school qualifications or employment based on the results achieved. It also has a monitoring function and, thirdly, has a comparability function as educational standards can be compared before and after curriculum revision or benchmarked against similar qualifications in other countries.

Post-apartheid South Africa has seen several curriculum revisions, accompanied by public accusations of lowering of standards (see, for example, Pauw *et al*., 2012; Joseph, 2016). In 2016, the marks of 28 of the 58 NSC subjects written were adjusted upwards and 4 subjects' marks were adjusted downwards during the standardization process that follows the marking of examinations (Davis, 2017). Standardization is carried out by a committee appointed by South Africa's quality assurance body, Umalusi Council for Quality Assurance in General and Further Education and Training. It is intended to limit large fluctuations in results for each subject from year to year.

In an open letter to the CEO of Umalusi, opposition parliamentarian Davis (2017) accused Umalusi of making upward adjustments without evidence that examination papers in 28 subjects were more difficult, by which he meant more cognitively demanding, than in previous years. Davis recommended that standardization should begin by examining the cognitive demand of examination papers before students' mark distributions were interrogated. Davis' (2017) recommendation is premised on the reliability of evaluations of cognitive demand by external raters. This paper addresses the reliability of such evaluations of cognitive demand.

Umalusi has investigated the standards of examination papers before and after a curriculum revision in 2008 (Bolton, 2009a), and benchmarked the South African NSC against international qualifications such as the Cambridge A levels and the International Baccalaureate (Grussendorff *et al.*, 2010). Author (2011) describes an effort to compare the NSC examinations in biology with equivalent examinations of Ghana, Kenya and Zambia.

Between 2008 and 2015, Umalusi conducted annual external evaluations of the standards of NSC examination papers in the subjects with the highest enrolment in South Africa, including Life Sciences. These evaluations included analysis of the cognitive demand of the examination papers. Such evaluation reports contribute to standardization decisions that follow each year's examinations.

Umalusi's benchmarking and examination evaluation exercises have been rigorous, quantitative and systematic (see, for example, Bolton, 2009a; Grussendorff *et al.,* 2010). Teams of three to four analysts drawn from different sectors of the education system have conducted the evaluations. Examination questions have been analysed individually for cognitive demand. Nevertheless, even though team members were retained for successive studies, it was noticed that inter-rater agreement was frequently poor (Bolton, 2009b).

The present study was conducted after a curriculum change in 2014. The process provided an opportunity to formally evaluate inter-rater agreement on assignment of examination questions to categories of cognitive demand in NSC Life Sciences examinations. The research question addressed the extent to which team members agreed on the application of a taxonomy of cognitive demand after intensive practice and revision. The findings have implications for the reliability and credibility of decisions based on expert evaluation of the cognitive demand of examinations.

1.1 Cognitive demand in examination papers

Most examining bodies provide a weighting of the categories of cognitive demand to be expected in their examination papers. The most recent South African Life Sciences curriculum specifies that: 40% of the marks should assess knowledge; 25% understanding, 20% applying knowledge and 15% analysing, evaluating and synthesizing knowledge (DBE, 2011: 67). The curriculum does not provide explicit criteria for each category of cognitive demand although it does provide a list of helpful verbs for each category.

Pollitt *et al.* (2007: 169) define cognitive demand of an assessment item as *the cognitive mental processes that a typical student is assumed to have to carry out in order to complete the task set by the questions.* Elliot (2011: 11) defines cognitive demand more broadly as *the level of knowledge, skills and competence required by typical learners*.

Cognitive demand requires that examiners and evaluators of the examination questions predict what thinking processes a student will use to make sense of a question and construct a response to it. A considerable number of frameworks, taxonomies and models relating to cognitive processes involved in the activities of thinking and learning exist (Moseley *et al.,* 2005). The most popular taxonomy of cognitive demand for over 60 years has been that of Bloom *et al.* (1956).

Bloom’s Taxonomy has been criticised by some assessment specialists. Forty years ago, Wood (1977: 204) described the problem with Bloom’s Taxonomy as being that *too many people have accepted the Taxonomy uncritically. Knowledge, Comprehension, Application, Analysis, Evaluation and Synthesis are still bandied about as if they were eternal verities instead of being hypothetical constructs constantly in need of verification.* Wood (1977) stated that assigning weightings of cognitive demand to formal assessment tasks is not justified because it implies a precision that does not exist. He described the organisation of Bloom’s Taxonomy as *remarkably ad hoc and not grounded in any psychological principles other than that knowledge is straightforward and anything involving mental operations is more difficult* (1977: 205). Twenty-five years later Sugrue (2002) reiterated Wood’s criticisms of Bloom’s Taxonomy because it was developed before cognitive science had progressed.

A Revised Bloom’s Taxonomy was produced after many years of discussion (Anderson *et al.,* 2001). The Revised Bloom’s Taxonomy changed the noun forms of cognitive processes to verbs, and reversed the order of the last two levels of cognitive demand. It introduced a second dimension to accommodate different types of knowledge on which the cognitive operations were performed. The cognitive dimension is Remember, Understand, Apply, Analyze, Evaluate and Create. The knowledge dimension is Factual, Conceptual, Procedural and Metacognitive. Combining the two dimensions allows for 24 possible combinations of cognitive process with type of knowledge.

Bloom's Taxonomy is assumed hierarchical, with Knowledge being the least demanding and Evaluation the most cognitively demanding. The Revised Bloom's Taxonomy has Create as the most cognitively demanding cognitive skill. Thus, examination questions assigned to higher order cognitive skills are assumed more demanding than questions assigned to lower order cognitive skills (Wood, 1977).

However, reasoning may be easier than remembering for a large proportion of the student population. The following essay question in an IEB examination provides an example of a higher order question of low difficulty. It is an argumentative essay on a socio-scientific issue.

> *Do you think the South African natural environment will survive the human population increase in this country?*
>
> *Read the source material carefully and present a debated argument to illustrate your point of view.*
>
> *To answer this question, you are expected to:*
>
> - *Select relevant information from Sources A to G below. Do not attempt to use all the detail provided.*
> - *Integrate your own biological knowledge. However, do not write an essay based solely on your own knowledge.*
> - *Take a definite stand on the question and arrange the information to best develop your argument.*
> - *Write in a way that is scientifically appropriate and communicates your point of view clearly.*
>
> *Write an essay of not more than 1 to 2 pages to answer the question.*
>
> *(IEB Paper 2, 2014)*

Several short pieces of source material are provided, presenting different sides to the argument. The task requires students to decide on a standpoint, analyse the source material, evaluate which sources support the argument, and synthesize a coherent argument. It therefore incorporates three higher order cognitive skills as described in the original Bloom's Taxonomy and the Revised Bloom's Taxonomy. In practice, most students perform well on this type of question, because its structure is familiar to them. They experience this high order task as easy.

The Revised Bloom's Taxonomy has refined Bloom's types of cognitive demand by providing fairly detailed descriptors of each cognitive skill and knowledge type (Anderson *et al.,* 2001). Moseley *et al.* (2005) advocated the Revised Bloom's Taxonomy for helping teachers to align learning objectives, instruction and assessment.

Pollitt *et al.* (2007) cited the Revised Bloom's Taxonomy as a possible instrument for analysing the cognitive demand of examination papers, but were concerned that its main purpose was to evaluate the cognitive demand of educational objectives.

A comprehensive Ofqual (2014) study comparing Senior Secondary Assessment in high-achieving countries used an analytical scale of cognitive demand based on the work of

Edwards & Dall'Alba (1981). Pollitt *et al*. (2007) describe how they modified the Edwards and Dall'Alba analytical scale, trialled the new scale and further modified it based on the comments of analysts. The final instrument is known by the acronym CRAS, representing Complexity, Resources, Abstractness and Strategy (Crisp & Novakovic, 2009). Each assessment task is assigned a rating of 1 – 4 on each criterion, with 1 being lowest demand and 4 being highest demand. It is possible, although not advisable, to calculate a single index of demand for an examination paper (Pollitt *et al*., 2007).

Dempster (2012) compared the 2008 Cambridge A-level, the 2006 International Baccalaureate Organisation Higher Level (IBO HL) and the 2008 South African National Senior Certificate final examinations using a three-level Bloomian-type taxonomy, comprising Remember, Understand and Apply, and Reason and Synthesize. The results showed that the A-level had the highest proportion of marks allocated to Understand/Apply, and the lowest proportion of marks allocated to Remember of the three examinations. The IBO HL had a high proportion of marks allocated to Remember, similar to the examinations of the NSC. The results were counter-intuitive, because the content of the IBO HL curriculum was judged to have the greatest depth of the three curricula.

Dempster (2012) re-evaluated the examination papers using CRAS. The IBO HL emerged as having the highest CRAS score on each of the four parameters, followed by the Cambridge A-level, and then the NSC. CRAS has the advantage of taking into account the complexity, abstractness and technical nature of the subject matter, which is absent in Bloomian taxonomies. The essay question cited earlier would have received a low CRAS score, because the subject matter lacks complexity and abstractness, most of the resources are provided, and students are coached in the strategy for writing an argumentative essay.

1.2 Inter-rater reliability in using a taxonomy of cognitive demand

Crowe *et al.* (2008) developed the Blooming Biology Tool (BBT), based on the original Bloom's Taxonomy, specifically to assign a level of cognitive demand to questions on biology-related topics. Three lecturers teaching different biological subjects developed the BBT (Crowe *et al.,* 2008). After intensive practice, the three lecturers achieved agreement of at least two out of three evaluators in over 90% of the 500 questions analysed independently. A sample of 36 students, trained to use the BBT and assign a Bloomian ranking to each question in their assessments achieved inter-rater reliability >80% for 31 of 51 test questions.

High inter-rater reliability was claimed for a Bloomian analysis of five biology-related examinations, which had been criticised for over-emphasizing recall of facts (Zheng *et al*., 2008). Interestingly, Crowe, Wenderoth and Dirks conducted the Bloomian ratings of the examination items (Zheng *et al.,* 2008). The Blooming Biology Tool needs to demonstrate that its reliability extends beyond its originators.

Näsström and Henriksson (2008) compared the Revised Bloom's Taxonomy and Porter's taxonomy as applied to standards and assessment in a chemistry course. Inter-rater reliability for classification of standards was significantly better for the Revised Bloom's taxonomy than Porter's taxonomy. Their study was limited to two evaluators, both teaching on the same course.

Bloom *et al*. (1956) claimed to have achieved a high level of agreement when classifying thinking and learning outcomes through discussion, but Wood (1977) reported that teachers using Bloom's taxonomy to classify examination questions found it difficult to reach agreement on higher-order categories. Sugrue (2002) also stated that Bloom's taxonomy is unreliable, to the extent of it being impossible to achieve consistent application by different people. However, she does not provide evidence for her claims in her short paper.

Pollitt *et al.* (2007: 189) cite many studies (e.g. Greatorex *et al.,* 2002; Fearnley, 1999; Griffiths & McLone, 1979), which report that evaluators had difficulty interpreting the statement classifiers with regard to scales of cognitive demand, and differed in their assignment of scale values. They concluded that judgement is inherently comparative and only approximately quantitative, due to the problem of trying to pin down relative meanings of words.

1.3 The present study

South African NSC examinations are set by several examiners, moderated by internal and external moderators and finally evaluated by a team of experts. The expert evaluator team appointed by Umalusi conducts a post-examination analysis of the papers before they have been marked. Their task is to ensure compliance with the prescribed weighting of categories of cognitive demand, and to express an evidence-based judgement of the overall standard of the examination papers (Umalusi, 2014a). Examiners, moderators and evaluators follow the taxonomy of cognitive demand specified in the South African Life Sciences curriculum (DBE, 2011: 67).

Agreement among examiners, moderators and evaluators on the meaning of each category of cognitive demand is a pre-requisite for valid judgements. The present study was a by-product of a commissioned comparison of the cognitive demand of NSC Life Sciences examination papers before and after the implementation of a revised curriculum in 2014. The four-member evaluation team used the Revised Bloom's Taxonomy to write descriptors for each of the four prescribed categories of cognitive demand. The structure of the evaluation project provided an opportunity to assess inter-rater agreement in the interpretation of the taxonomy of cognitive demand. The results have implications for the validity of assigning weightings of cognitive demand in examinations, and the reliability of examiners', moderators' and evaluators' assignment of assessment items to levels of cognitive demand.

2. Method

All public schools write the National Senior Certificate set, marked and controlled by the Department of Basic Education (DBE). Some independent schools write exit-level examinations set, marked and controlled by the Independent Examinations Board (IEB).

Both examining bodies follow very similar curricula, and Umalusi assures their quality.

Candidates from both examining bodies write two Life Sciences examination papers in an examination session. Each examination paper contains questions requiring short answers (such as multiple choice, matching columns, giving definitions for terms, and supplying missing words in a short text) worth 50 marks, questions requiring longer answers of one or two sentences and interpretation of data (80 marks), and an essay question worth 20 marks. Candidates are required to answer all questions. The IEB examinations include a separate Practical examination, worth 50 marks.

After examinations have been marked, Umalusi standardizes the results to ensure that the frequency distribution of marks remains within the range of the previous three years' mark distribution. Decisions about mark adjustments are informed by reports from moderators and a team of expert evaluators, as well as the distribution of marks for the current year. The expert evaluators are tasked, amongst other things, with evaluating the cognitive demand of each examination paper.

The expert evaluation team in 2014 consisted of:

- a University academic experienced in Life Sciences teacher education and evaluation of examination papers (E - the team leader);
- a subject advisor in Life Sciences (J);
- an experienced Life Sciences teacher from a public school (L);
- an experienced Life Sciences teacher from an independent school (R).

Table 1: Taxonomy of cognitive demand used for expert evaluation of Life Sciences examinations (adapted from Anderson *et al*., 2001).

Type of cognitive demand	Descriptor
Recognise or recall information (K = Knowing science)	**Recall** from memory or **recognize** from material provided explicit information, details, facts, formulae, terms, definitions, procedures, representations.
Demonstrate understanding (U = Understanding science)	**Communicate understanding** of a Life Sciences concept, idea, explanation, model or theory, for example to: • **interpret**: change from one form of representation to another (e.g. pictures to words; words to pictures; numbers to words; words to numbers; pictures to numbers). • **exemplify**: find a specific example or illustration of a concept or principle. • **classify**: determine that something belongs to a category. • **summarize**: abstract a general theme or major points. • **infer**: draw a logical conclusion from presented information. • **compare**: detect similarities and differences between two objects or concepts. • **explain why**: create a cause-and-effect model of a system or concept.
Apply procedures, facts and concepts (A = Applying scientific knowledge)	• **Use**, **perform** or **follow** a basic/standard/routine procedure/ rule/method/operation. • **Use**/**apply understanding** of biology facts, concepts or details from a known context to an unfamiliar context.

Type of cognitive demand	Descriptor
Analyse, evaluate, create (AEC = Evaluating, analysing or synthesizing scientific knowledge)	• **Analyse** complex information adopting a variety of appropriate strategies to solve novel/ non-routine/complex/ open-ended problems. • Apply multi-step procedures. • **Evaluate** or make critical judgements (e.g., on qualities of accuracy, consistency, acceptability, desirability, worth or probability) using background knowledge of Biology. • Create a new product by integrating biological concepts, principles, ideas and information; make connections and relate parts of material, ideas, information or operations to one another and to an overall structure or purpose.

The taxonomy of cognitive demand used by the evaluation team, with descriptors, is shown in Table 1.

The evaluation team held a training session in October 2014 with intensive discussion of the taxonomy shown in Table 1. The team worked through three examination papers from previous years together, discussing each individual question and sub-question until consensus was reached on the category of cognitive demand. The team then separated and independently analysed a further nine papers from 2012, 2013 and exemplar papers for 2014.

The team leader collated results, and questions identified where inter-rater agreement was low. A second meeting was held early in November 2014 to re-visit the instrument and revise analyses until greater consensus was reached. This was the practice phase of the research project.

The final set of examination papers for 2014 was analysed by evaluators working independently. Two examination papers from each of the examining bodies were analysed, giving a total of four examination papers and 244 individual items. Each examination question was entered on a spreadsheet, and the team leader recorded the ratings of demand from the four evaluators. Table 2 is an extract from the spreadsheet to illustrate how the team leader recorded the ratings. We omitted the IEB Practical paper because it was not comparable with the DBE examinations. The independent analysis of the final 2014 examination papers provided an opportunity to evaluate inter-rater agreement on assigning questions to categories of cognitive demand after intensive training and practice.

Table 2: Sample of spreadsheet showing collated results for cognitive demand.

Item	Marks	Cognitive demand by evaluator			
		E	L	J	R
2.2.1	1	U	A	K	K
2.2.2	2	U	U	U	U
2.2.3	3	U	AEC	U	U

Each team member also completed a questionnaire asking them how confident they felt about their ratings, and to what extent they referred to the criteria for assigning cognitive demand.

2.1 Statistical analysis of inter-rater agreement

Inter-rater agreement is used extensively in tests of agreement in medical diagnoses and psychological assessment (Altman, 1991). Gwet's AC_1 (Gwet, 2011) was chosen as the most suitable coefficient of agreement for this study, because it can be applied to more than two raters, and is unaffected by ratings that are skewed towards the marginal, e.g., rating most questions as Knowledge. Ratings that are skewed towards the marginal interfere with the correction for chance agreement.

Interpretation of coefficients of agreement followed the recommendations of Fleiss *et al.* (2003), who suggested that coefficients >0,75 represent excellent agreement beyond chance, 0,4 to 0,75 represent fair to good agreement, and <0,4 poor agreement beyond chance.

Per cent agreement is the average pairwise per cent agreement for each item. The agreements among all possible pairs are calculated and averaged for each item. For example, item 2.2.3 in Table 2 resulted in three evaluators agreeing (E. J and R) and one evaluator (L) disagreeing. The pairwise percentage agreement is calculated for all possible pairs (E and L, E and J, E and R, L and J, L and R, J and R).

	E	L	J
L	0		
J	100	0	
R	100	0	100

The average pairwise percentage agreement is ((100 x 3) + (0 x 3))/6 = 50%. The percentage agreement is then averaged for all the items on an examination paper. Percentage agreements of 90% or greater are nearly always acceptable; 80% is acceptable in most situations, and 70% may be appropriate in some exploratory studies (Neuendorf, 2002). Percent agreement needs to be included in evaluation of inter-rater reliability, because it either supports or refutes the agreement coefficient.

3. Findings

3.1 Inter-rater agreement

Inter-rater agreement within the evaluation team for the 2014 final papers is presented in Table 4 and Figure 1. Table 4 shows, for example, that there were 59 individual items in the IEB Paper 1 examination. All four raters agreed on the cognitive demand of 18,6% of the items, while three agreed on 42,4% of the items. Two raters agreed and two differed in 20,3% of the items, while two agreed on one rating, and another two agreed on a different rating in a further 17% of the items. No raters agreed on 1,7% of the items. This explanation applies to all the papers analysed.

Table 4: Percentage of items at each level of agreement, and statistical test results for cognitive demand in four Life Sciences examination papers (n=4 evaluators)

	4 agree	3 agree	2 agree + (2 sets 2)	None agree	Gwet's AC_1	Pairwise Percent agreement
IEB Paper 1 (59 items)						
% agreement	18,6	42,4	20,3 + (17,0)	1,7	0,33 Poor	49
IEB Paper 2 (52 items)						
% agreement	42,3	38,5	11,5 + (7,7)	0	0,56 Fair	66
DBE Paper 1 (68 items)						
% agreement	35,3	39,7	25,0 + (5,9)	0	0,53 Fair	63
DBE Paper 2 (65 items)						
% agreement	33,8	29,2	16,9 + (20,0)	0	0,46 Fair	58

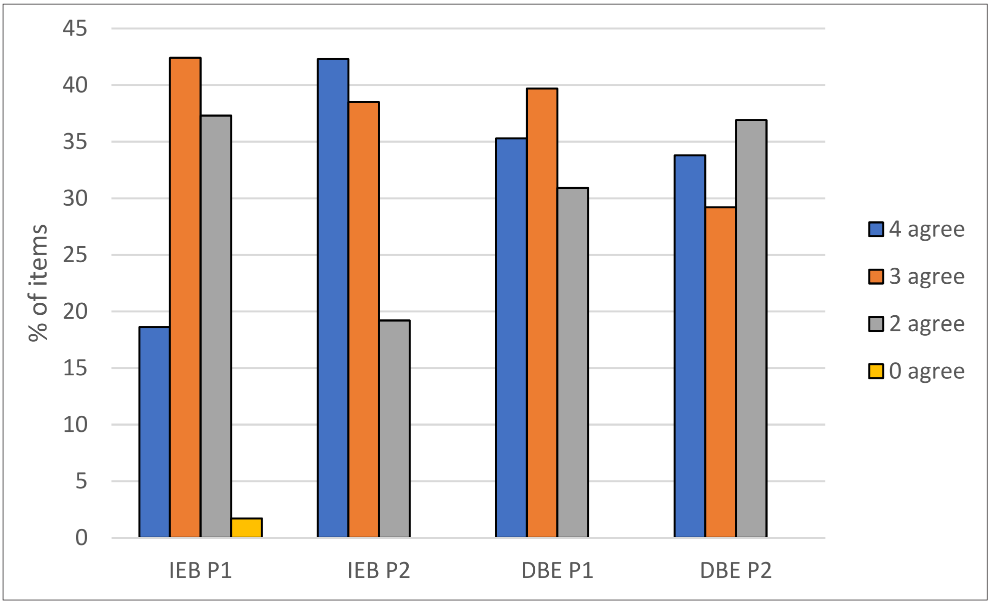

Figure 1: Percentage agreement among four evaluators on the cognitive demand of four examination papers.

The percentage of items on each paper on which three or four evaluators agreed was 61% for IEB P1, 63% for DBE P2, 75% for DBE P1 and 81% for IEB P2 (Figure 1). This is considerably lower than 91% achieved by at least two of three raters rating questions using the Blooming Biology Tool (Crowe *et al.,* 2008).

Gwet's AC_1 showed that IEB P2, DBE P1 and DBE P2 achieved "fair agreement", and IEB P1 achieved "poor agreement" among evaluators, based on the categories of agreement provided by Fleiss *et al.* (2003). None of the analyses approached 0,75 (excellent agreement).

Average pairwise percentage agreement was lowest for IEB P1, but all pairwise agreements were unacceptable according to Neuendorf (2002). Percentage agreement was congruent with coefficients of agreement.

In the questionnaires, all evaluators indicated that they were not particularly confident about their analyses of cognitive demand, and that they referred to the criteria listed in Table 1 all or most of the time.

3.2 Identifying outliers

Given the diverse professional experiences of the evaluation team, an analysis was conducted to determine whether any one team member could be identified as consistently different from the others. The mean number of items assigned to each level of demand was calculated for the four examination papers analysed. Results are shown in Figure 2.

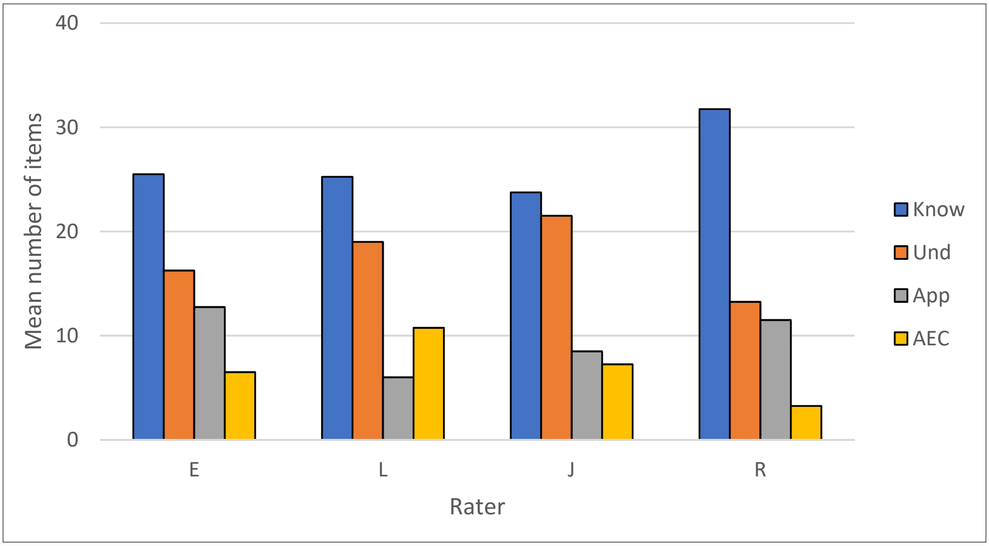

Figure 2: Mean number of items assigned to each cognitive category by four evaluators (n = 4 examination papers comprising 244 items) (Know = Know; Und = Understand; App = Apply; AEC = Analyse, Evaluate, Create)

Figure 2 shows variation among the evaluators in the mean number of items assigned to each category of cognitive demand for the four examination papers. Evaluator R stands out because she assigned more items to the cognitive category *know* and fewer items to *understand* and *AEC* than the other three evaluators. However, the Pearson chi-square statistic was not large enough to be statistically significant (p= 0,33). Cramer’s V (0,12) indicates a small effect size of evaluator on the overall variance.

In order to explore the effect of individual raters further, variation was analysed for each examination paper separately. Results are shown in Table 6. According to the Pearson chi-square statistic, evaluators’ results for IEB P1 and P2 were not sufficiently different to be statistically significant.

Table 6: Number of items assigned to each category of cognitive demand by examination paper and evaluator.

IEB P1 (59 items)					
Evaluator	Cognitive category				Statistical test results
	Know	Understand	Apply	AEC	χ^2 (9) =15,9; p =0,07; Cramer's *V*=0,15
E	17	22	12	8	
L	26	17	5	11	
J	11	28	9	11	
R	24	20	10	5	
Mean±SD	19,5±6,9	21,8±4,7	9,0±2,9	8,8±3,5	
IEB P2 (52 items)					
E	20	11	15	6	χ^2 (9) =12,2; p =0,21; Cramer's *V*=0,14
L	22	12	7	11	
J	22	16	9	5	
R	29	12	7	4	
Mean±SD	23,3±4,0	12,8±2,2	9,5±3,8	6,5±3,1	
DBE P1 (68 items)					
E	36	17	11	4	χ^2 (9) =18,6; p <0,05; Cramer's *V*=0,15
L	32	24	2	10	
J	35	20	7	6	
R	42	10	9	3	
Mean±SD	36,3±4,2	17,8±5,9	7,3±3,9	5,8±3,1	
DBE P2 (65 items)					
E	29	15	13	8	χ^2 (9) =20,7; p <0.05; Cramer's *V*=0,16
L	21	23	10	11	
J	27	22	9	7	
R	32	11	20	1	
Mean±SD	27,3±4,7	17,8±5,7	13,0±5,0	6,8±4,2	

Both DBE papers had sufficient variation among evaluators to achieve a significant Pearson chi-square statistic. Cramer's V shows a small, but significant effect size of evaluators on the chi-square. In both papers, evaluator R assigned more items to *know*, and fewer to *understand* and *AEC* than the other evaluators. In DBE P2, evaluator R assigned more items to *apply* than other evaluators did. Evaluator L stands out in that she assigned more items than other evaluators did to *AEC* in both DBE papers.

4. Discussion

The lack of agreement among four expert evaluators is clearly illustrated in the data presented here. Despite intensive training and practice with feedback, inter-rater reliability remained low. Our results support the view that identifying cognitive demand is subjective. Larger teams of raters may increase inter-rater reliability, because they dampen the effect of one outlier. Subjectivity is acknowledged in the Ofqual (2014) report, where it is accepted that expert judgement is rarely definitive, since it represents the views of a diverse group of experts who are influenced by their individual professional experiences.

The evaluators did not feel confident about their ratings, and referred constantly to the list of descriptors. While the evaluators clearly thought deeply about what cognitive processing students had to do to answer a question the questions did not neatly fit the descriptors.

The findings presented here raise questions about agreement among examiners, moderators and evaluators on what is meant by each cognitive category of the prescribed Bloom's Taxonomy. Alignment between the prescribed weighting and the actual weighting presents a challenge in the absence of common agreement on the meaning of categories of cognitive demand. Comparisons of the "standards" of successive examination papers is questionable when inter-rater reliability in judging cognitive demand is low, even after intensive practise. This raises doubts about the value of moderators and evaluators' reports to inform standardisation of marks, as recommended by Davis (2017).

Further complexity is added when one considers the multitude of factors influencing the demand of examination questions. High order questions that recur in successive papers lose their discriminatory power because of their familiarity to students. They do not necessarily indicate increasing or decreasing standards of examinations. Davis (2017) incorrectly assumes that high cognitive demand is indicative of examinations that are more difficult. As pointed out by Pollitt *et al.* (2007), difficulty is a measure of the performance of learners on an assessment task, which can only be reliably evaluated by studying mark distributions. It is distinct from cognitive demand, as illustrated by the high cognitive demand but easy essay question in the IEB examination papers.

Cognitive demand is influenced by the breadth and depth of the subject matter being assessed (Opposs & Moss, 2012). In this regard, the CRAS scale of demand has advantages over Bloomian taxonomies because it considers the abstractness of the subject matter being assessed, and the complexity of the operations required to complete the task. Inter-rater reliability of CRAS has not been assessed, but Crisp and Novacovic (2009) reported that raters found it particularly difficult to assign a level of abstractness to an item. Nevertheless, CRAS holds potential as an alternative taxonomy to Bloom's.

Wood (1977) and Sugrue (2002) recommend having no taxonomy at all, replacing it with performance objectives that inform structuring of examination questions. Kanjee and Moloi (2016), in consultation with teams of experts, developed a standards-based approach to reporting South African Grade 6 assessment results. The standards were based on performance level descriptors for four levels of performance and cut scores. The cut scores were determined by expert judgement of how many out of ten learners would be able to correctly answer each item on a national assessment (Kanjee & Moloi, 2016: 38).

Kanjee and Moloi's (2016) determination of cut scores may be difficult to apply in South Africa's diverse educational context. Experts' answers to the question "How many of 10 just proficient learners will get this item right?" will be influenced by their own home language and learning and teaching experiences. The panel of experts determining cut scores would need to be large and representative of the entire educational system. The present study illustrates how unreliable the subjective judgements of a panel of four experts can be.

The three specific aims of the South African Life Sciences curriculum, with their descriptors, provide a set of performance objectives (DBE 2011: 11-17), which could be developed into performance level descriptors. The standard of the examination would then reside in experts' prediction of what percentage of "just proficient" learners would answer each question correctly. However, such a judgement would be based on the current standard of education, which is below what is deemed appropriate (Umalusi, 2014a). Even experts will have difficulty conceptualising the performance of what a "just proficient" learner ought to be able to do, as compared with their lived reality. A standards-based approach for reporting National Senior Certificate results is currently unattainable.

5. Conclusion

This study illustrates the difficulties of achieving inter-rater agreement on the interpretation of levels of cognitive demand in examination questions. Despite training, practise and revision, evaluators still failed to achieve high levels of inter-rater reliability. We recommend that the standardisation committee uses the results of analyses of levels of cognitive demand with caution, and that examining bodies consider critically their practice of prescribing weighting of Bloomian levels of cognitive demand in examinations. We also urge examiners and moderators to consider alternatives to Bloom's taxonomy, particularly those that are informed by cognitive science.

6. Acknowledgements

The authors wish to thank the evaluators whose conscientious work enabled this analysis to be undertaken. Umalusi is thanked for initiating the project to evaluate consistency of standards in successive examinations. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number 104666).

References

Altman, D.G. 1991. *Practical statistics for medical research*. London: Chapman and Hall.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J. & Wittrock, M.C. (Eds.). 2001. *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* Abridged Edition. New York: Longman.

Atkin, P. & Black, J.M. 2003. *Inside science education reform*. Buckingham: Open University Press.

Baird, J-A., 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

Bloom, B.S. (Ed.), Englehart, M.D., Furst, E.J., Hill, W.H. & Krathwohl, D.R. 1956. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Bolton, H. 2009a. *From NATED 550 to the new national curriculum: Maintaining standards in 2008. Part 1: Overview report.* Pretoria: Umalusi.

Bolton, H. 2009b. *From NATED 550 to the new national curriculum: Maintaining standards in 2008. Part 3: Exam paper analysis.* Pretoria: Umalusi.

Broadfoot, P. 2007. *An introduction to assessment*. University of Virginia: Continuum.

Burke, K. 2010. *Balanced assessment. From formative to summative*. Bloomington: Solution Tree Press.

Cresswell, M.J. 2000. The role of public examinations in defining and monitoring standards. *Proceedings of the British Academy,* 102, 69–120.

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

Crowe, A., Dirks, C. & Wenderoth, M.P. 2008. Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education,* 7(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Davis, G. 2016. *Matric 2016: An open letter to Umalusi*. Available at https://www.da.org.za/2016/12/matric-2016-open-letter-umalusi/ [Accessed 12 Jan 2017]

Dempster, E.R. 2012. Describing cognitive demand of Biology examination papers: a comparison of two instruments, paper presented to *Standards in education and training: The challenge*, Muldersdrift, May 2012. DOI: 10.13140/RG.2.2.18477.28649

Department of Basic Education (DBE). 2011. *National curriculum and assessment policy statement: Life Sciences for the further education and training phase. Grades 10–12.* Pretoria: Department of Basic Education.

Department of Basic Education (DBE). 2015. *National senior certificate examination report 2015.* Pretoria: Department of Basic Education.

Eckstein, M.A. & Noah, H.J. 1989. Forms and functions of secondary school-leaving examinations. *Comparative Education Review,* 33, 295–316. https://doi.org/10.1086/446860

Edwards, J. & Dall'Alba, G. 1981. Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education,* 11, 158–170. https://doi.org/10.1007/BF02356779

Elliott, G. 2011. A guide to comparability terminology and methods. *Research Matters: A Cambridge Assessment Publication, Special Issue 2*. Cambridge: Cambridge Assessment.

Fearnley, A. 2000. A comparability study in GCSE mathematics. A review of the examination requirements and a report of the cross moderation exercise. A study based on the Summer 1998 examination and organised by AQA (NEAB) on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). *Techniques for monitoring comparability of examination standards.* London: QCA.

Fleiss, J.L., Levin, B. & Paik, M.C. 2003. *Statistical methods for rates and proportions,* 3rd ed. Hoboken: John Wiley and Sons. https://doi.org/10.1002/0471445428

Griffiths, H.B. & McLone, R.R. 1979. *Qualities cultivated in mathematics degree examination.* London: Social Science Research Council.

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: Higher Education South Africa/Umalusi.

Greatorex, J., Elliott, G. & Bell, J.F. 2002. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 Examination and organised by the Research and Evaluation Division, UCLES for OCR on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). 2007. *Techniques for monitoring comparability of examination standards*. London: QCA.

Gwet, K. 2011. *Handbook of inter-rater reliability. The definitive guide to measuring the extent of agreement among raters,* 4th ed. Gaithersburg: Advanced Analytics.

Harlen, W.D. 2012. Between assessment for formative and summative purposes. In J. Gardner (Ed.). *Assessment and learning* 2nd ed. London: Sage. https://doi.org/10.4135/9781446250808.n6

Joseph, N. 2016. *What matric results reveal about SA's school system.* Available at http://mg.co.za/article/2016-01-06-what-matric-results-reveal-about-sas-school-system [Accessed 11 February 2016].

Kanjee, A. & Moloi, Q. 2016. A standards-based approach for reporting assessment results in South Africa. *Perspectives in Education*, 34(4), 29–51. DOI: http://dx.doi.org/10.18820/2519593X/pie.v34i4.3.

Kellaghan, T. & Greaney, V. 2004. *Assessing student learning in Africa*. Washington DC: World Bank.

Moseley, D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. 2005. *Frameworks for thinking. A handbook for teaching and learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511489914

Näsström, G. & Henriksson, W. 2008. Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology,* 16 (6), 667–690.

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

Newton, P. 2007. Clarifying the purposes of educational assessment. *Assessment in Education,* 14(2), 149–170. https://doi.org/10.1080/09695940701478321

Ofqual. 2012. *International comparisons in senior secondary assessment: Full report.* Ofqual.

Opposs, D. & Mapp, L. 2012. *International comparisons in senior secondary assessments*. Available at http://www.ofqual.gov.uk/downloads/category [Accessed 20 April 2012].

Pauw, J., Dommisse, J. & van der Merwe, J. 2012. *1 in 6 matrics got less than 10% for maths*. Available at http://city-press.news24.com/SouthAfrica/News/1-in-6-matrics-got-less-than-10-for-maths-20120128 [Accessed 1 February 2017].

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

Sugrue, B. 2002. *Problems with Bloom's taxonomy*. Available at https://eppicinc.files.wordpress.com/2011/08/sugrue_bloom_critique_perfxprs.pdf [Accessed 1 February 2017].

Taras, M. 2005. Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies,* 53(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Umalusi. 2014a. *Consolidated post-exam analysis report 2014. Content subjects – IEB.* Available at www.umalusi.org.za/docs/assurance/2015/ieb.pdf [Accessed 15 Nov 2017].

Umalusi. 2014b. *Consolidated post-exam analysis report 2014. Content subjects – DBE*. Available at www.umalusi.org.za/docs/assurance/2015/dbe.pdf [Accessed 15 Nov 2017].

Wood, R. 1977. Multiple choice: A state of the art report. *Evaluation in Education,* 1, 191–280.

Zheng, A.Y., Lawhorn, J.K., Lumley, T. & Freeman, S. 2008. Application of Bloom's taxonomy debunks the MCAT Myth. *Science,* 319, 414–415. https://doi.org/10.1126/science.1147852