Let’s stop complaining about low cognitive levels of testing in time limited examinations: The case for fair testing practices

G. Green
School of Chemistry
University of the Witwatersrand
Johannesburg, South Africa
e-mail: gail@aurum.chem.wits.ac.za

M. Rollnick
School of Education
University of the Witwatersrand
Johannesburg, South Africa

Abstract
This article describes the process of determining cognitive levels of testing in first and second year university chemistry examinations and emanates from a larger study in which explanations were sought for problems experienced by second year students in chemistry at the University of the Witwatersrand. A user friendly scheme for establishing cognitive levels of questions was developed and this was used to determine the cognitive level of questions in 25 different examination papers written over a period of three years. It was found that testing at both first and second year level was predominantly at level II (comprehension). Possible reasons for this finding are explored and justified.

INTRODUCTION
It is generally accepted that what is learned in any subject will become more difficult as one progresses in a course. Although at face value this notion of ‘difficulty’ appears unambiguous, it is in reality not easily defined since it relates to the nature of the content taught, to the context within which the course is taught as well as to the cognitive level at which questions are posed in formal tests and examinations. While students are coping with a course, the difficulty of the course is usually not an issue, but becomes one when students do not fare as well as anticipated. This has historically been the case in the School of Chemistry at the University of the Witwatersrand (Wits) where there has been concern about the poor performance of students in the second year chemistry major class compared to first year.

Besides the negative impact of this poor achievement on student throughput rates there is also a feeling that if students perceive the second year major course to be abnormally difficult, this might negatively impact on the numbers and retention rates of potential chemistry majors at Wits. This view is based on the assumption
that students’ expectancy of success and their subsequent achievement behaviour (including choice of courses) plays a vital role in the choice of a major (McKeachie, Pintrich, Lin, Smith and Sharma 1990). In the period 1992 to 1998 the number of chemistry majors at Wits declined rapidly and although the rate of decline slowed down from 1999 to 2005, the trend remained worrying. It has been well documented that this has also happened in other parts of the world (Brand 1995). In order to investigate possible reasons for poor student performance in second year chemistry a decision was taken in 1999 to investigate how students themselves experience second year major chemistry at Wits. One of the factors investigated at the time was the cognitive level of questioning in examinations in second compared with first year. Although the data presented in this article are related to the abovementioned study and consequently are not recent, there are good reasons for not regarding the study as dated. First, a superficial analysis of similar data from 2003 and 2004 shows that little has changed since 1999, besides which the focus in this article is not on presenting findings based on a specific study. It is rather an attempt to share some of the insights gained during the research process and to present what may prove to be a controversial opinion about the cognitive level at which university examinations should be set. In order to present our arguments logically, we deviate from the format of a traditional report of research since we wish to show that

- assigning questions to cognitive levels (based on Bloom’s taxonomy) is not easily accomplished without some insider knowledge of the teaching situation.

Thereafter, our data will be presented which will reveal that

- in spite of what one might expect, there is little difference in the cognitive levels of testing as one progresses from year to year in chemistry at Wits (Green 2001).

Finally, we wish to argue that the similarity in levels of testing between years is not unexpected since

- most chemistry courses at universities worldwide are presented within the context of normal chemistry education (Van Berkel, DeVos, Verdonk and Pilot 2000).
- timed tests are considered by an increasing number of educators to be inappropriate for testing at cognitive levels above the level of application of knowledge (Felder and Brent 2002; Airasian and Miranda 2002).

CHOOSING A FRAMEWORK FOR ANALYSIS OF COGNITIVE LEVELS

There are a multitude of approaches which can be adopted in analysis of the cognitive level of examination items, most of which are based on Bloom’s taxonomy – the well-known hierarchy of six different levels of cognitive
objectives (Bloom 1956). However, when it comes to practical application of Bloom's hierarchy the literature reveals that its complexity often makes it unsuited to the needs of particular individuals and simplification is necessary before it becomes user-friendly. A case in point is the extreme simplification described by McMillan and Lawson (2001) where only two main categories were used by secondary science teachers to separate the cognitive level of assessments, that is, recall and higher order thinking. A more useful model for tertiary studies, as used by researchers like Zoller, Dori and Lubesky (2002), and Felder and Brent (1999) is shown in Figure 1 where Bloom's levels IV to VI collectively are considered to comprise the higher order thinking abilities.

If however the main focus of a study, as in our case, is on identification of higher order cognitive skills it is possible to use a system of only three levels, that is, level I (knowledge gained by rote learning), level II (comprehension and application, the traditional levels II and III) and level III (analysis, synthesis and evaluation, the higher order skills) (Farmer and Farrell 1980; Crossley 1980a). This simplification is justified in view of that fact that some, including Felder and Brent (1999) consider Bloom's levels II and III to be roughly equivalent in terms of demand and will be the model of cognitive levels used in the rest of this article. Zoller, Lubezky, Nakhleh, Tesser and Dori (1995) have shown that students fare better at algorithmic application than at conceptual (comprehension) questions which demonstrates not only that it is possible to solve problems without understanding, but also suggests that Bloom's levels II and III are very close in terms of cognitive demand. In our study our first task was therefore to assess the cognitive demands imposed by questions in examinations at Wits.
OVERCOMING PROBLEMS IN ASSIGNING TEST ITEMS TO COGNITIVE LEVELS

Although we had not anticipated it, we found that we experienced problems in drawing a distinction between examination questions at our levels II and III (higher order). There appears to be general agreement (Farmer and Farrell 1980; Crossley 1980b; Zoller et al. 1995) that a crucial characteristic separating lower order cognitive skills (LOCS) from higher order cognitive skills (HOCS) is the element of novelty to the respondent. The explanation which follows, clarifies the idea of ‘novelty’ according to Farmer and Farrell (1980).

If a test item requires application of appropriate generalisations and/or principles to a familiar situation then this item would obviously fall into level II, whereas applications in an unfamiliar setting would presumably raise it to level III. However, careful thought shows that questions which are easily identified as ‘type’ problems, even if they are worded differently to examples already encountered by students, should also be classified at level II, since a different context alone does not suffice in elevating a question to a higher cognitive level. This means that although it is relatively easy to categorise level I questions, other questions cannot by their wording automatically be assigned cognitive levels without due consideration being given to the teaching context. It is the task of the analyst to evaluate each question against a particular setting in order to decide whether it is indeed novel and unfamiliar, that is, requires mental restructuring (level III) or whether it is merely a particular type of problem in a new setting which would automatically assign it to level II. Bloom himself acknowledged the difficulty in devising level III test items which are ‘new yet real’, in which a student will apply an appropriate abstraction without having been prompted about which abstraction to use. A very simple example which illustrates what is meant by this is the following question: A student wants to cut a circle with radius 25 cm from a square of paper. Find the smallest area of square paper that could be used to produce the circle.

The reason why this problem requires HOCS is because before any action is possible the student has to come to the realization that the radius is exactly half the length of the side of the square. This operation involves mental restructuring of the problem and establishing links between the dimensions of a circle and a square. Another example taken from Wits first-year tutorials (Chemistry Dept 1998) follows:

Mercury pollution has serious consequences for health since mercury, a cumulative poison, eventually results in irreversible brain damage. A sample from a lake of surface area 100 km² and average depth 5.0 m was found to contain 0.50 \( \text{g Hg cm}^{-2} \). What is the total mass (kg) of Hg in the lake?

For an expert the problem is clearly of the type requiring unit conversions for its solution. However, before a student can commence with meaningful unit conversions, mental restructuring of the problem is required to visualize that the
mercury concentration is given for only a very small fraction of the content of the lake. Our experience from tutorials is that students have great difficulty in identifying the essence of the problem which means that they cannot solve it.

**CRITICAL THINKING VS PUZZLE-SOLVING**

Although he prefers to use the term critical thinking instead of HOCS, Hobden (2002) has come to the realisation that there is generally a narrow understanding of what it means to solve problems. Students are often given a selection of different problems to work through for practice. Hobden (2002) argues that application of routine or normal strategies to find solutions, even to new questions, does not constitute problem solving at all, since there is no reflecting, imagining, planning, categorising or reasoning involved. His definition of problem solving can be extrapolated to our problem of distinguishing between our cognitive levels II and III. It becomes clear that in true problem solving as in higher order questioning there has to be recourse to some form of critical thinking which is an essential element in working out a strategy to overcome the barrier between the existing and goal situations.

Hobden's views seem to coincide with those who argue that the difficulty of a subject is inevitably influenced by the nature of the teaching, whether it is in the context of Normal Science Education (NSE) or critical science (Van Berkel, De Vos, Verdonk and Pilot 2000). These ideas originate from Kuhn (1970) who claimed that NSE initiates and prepares students for the handling of normal science problems ‘that is for the activity of puzzle-solving as set within the current paradigm or disciplinary matrix, which is all that future normal scientists need in order to function successfully’ (Kuhn 1970). Pintrich's (2002) suggestion that metacognitive knowledge would be an appropriate category in a new taxonomy of educational objectives ties in well with the NSE view of science as puzzle-solving since the activity of metacognition requires knowledge of general strategies for learning, thinking and problem solving which are those associated with successful puzzle-solving.

One of the undisputed aims of a university education is to encourage critical thinking and it has been proved that under controlled research conditions it is possible to implement modest forms of critical science education (Van Hoewe-Brouwer 1996, quoted in Van Berkel et al. 2000; Zoller et al. 1995). However, despite much criticism, Kuhn (1970) continues to dispute the possibility of being able to train students for the practice of revolutionary or critical science and strongly contends that the puzzle-solving practices of normal science are indeed sufficient for aspiring chemists.

**THE PROBLEMS OF HIERARCHICAL TAXONOMIES**

Although all the revised taxonomies and strategies have inherent imperfections, the
‘untouched’ version of Bloom’s taxonomy is not above criticism either. Due to the hierarchical nature of Bloom’s taxonomy, a higher ability is usually assumed to include the lower level abilities. This is summarised below:

<table>
<thead>
<tr>
<th>Ability Category</th>
<th>Abilities Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>II</td>
<td>I and II</td>
</tr>
<tr>
<td>III</td>
<td>I, II and III</td>
</tr>
<tr>
<td>IV (above III)</td>
<td>I, II, III and IV</td>
</tr>
</tbody>
</table>

It is possible to argue against both the cumulative as well as the hierarchical view of Bloom’s taxonomy. For example, that it is possible to perform an evaluation (at Bloom’s level VI) without calling on level V abilities (i.e., synthesis). In short, performance at a higher level does not necessarily imply that all lower level abilities are included as well. The exception is that some knowledge (level I) is always necessary for performance of all operations at higher levels.

Another problem with Bloom’s model is that the presentation of cognitive objectives as a hierarchy unwittingly creates the perception that performance at the higher levels implies involvement with tasks of increasing difficulty. Felder (personal communication) argues that this is not valid since young children who cannot analyse nor synthesise can be taught to evaluate a number of options and then justify their choice (a cognitive ability at Bloom’s level VI). Another factor to consider is that even when students are thought to be operating at the lowest cognitive level and are faced with a large quantity of content to memorise, then higher order cognitive strategies like clumping of concepts to form new ideas, might come into play. Because of all these complexities, no assumptions can be made about the cumulative nature of any cognitive hierarchy, although a measure of rote learning is common to all other levels.

Once we had recognised that the normal activity in chemistry, including examinations was essentially puzzle-solving in the realm of NSE (which would per definition be confined to levels I, II and III) this made it possible for us to map a workable strategy for differentiating between questions. From the perspective of NSE levels II and III could be called routine puzzle-solving and more advanced puzzle-solving respectively.

We have stated that it is the task of the analyst to evaluate questions against the teaching that has gone before a decision can be reached concerning the cognitive level of a particular question. It is clear that analysis by an outsider, however familiar with Bloom’s taxonomy and the subject matter being examined, is insufficient. This flies in the face of traditional research which emphasises the objectivity of the researcher. The simplicity of the model we used in practice is represented schematically in figure 2.

Having set out our criteria for assigning questions to cognitive levels, a discussion of our data and findings follows.
Let's stop complaining about low cognitive levels of testing in time limited examinations . . .

Figure 2: Guidelines for classification into cognitive levels

DATA SOURCE

In our study, examination papers from the period 1996 to 1998 were selected for analysis. In the case of first year courses (which will be identified as 1A – 1C), selected papers (written either in November or June) were randomly chosen for analysis while papers from the six modules comprising second year chemistry at Wits (analytical, inorganic, physical, organic and materials chemistry as well as spectroscopy) were analysed. In the case of one of the first year courses (1C), there was an unexpected variation in level I testing between the chosen papers which compelled scrutiny of an additional paper (written in 1998) in an attempt to establish what the norm might be. This inconsistency in level I testing can possibly be understood in view of the finding by Mumba (2000) that lecturers themselves are unclear about cognitive demands imposed by questions.

DATA ANALYSIS

Level I questions were fairly easy to identify from the way they were phrased or alternatively by the fact that identical tutorial questions existed. This meant that some seemingly complex questions were classified by us as answerable by rote, because of our view that it is possible to learn virtually anything using memorisation techniques. This exercise of matching examination with tutorial questions was intended to ensure a reasonable degree of validity in the classification and it can be argued that this can be considered a virtually foolproof way of conducting a simple classification according to our simplified taxonomy.
Notwithstanding our confidence in the routine, a sample of papers was subjected to peer review to validate the method, producing an agreement of above 95 per cent.

A real weakness of the method is that it does not take individual differences into account since the assumption is that all students would experience questions at the same level. Each paper was analysed in its own right and compared only with what was actually taught in a particular course and how it was taught.

RESULTS OF ANALYSIS

The results of the analysis of first-year papers are in Table 1.

Table 1: Assessment of cognitive abilities required in first year examinations of different courses

<table>
<thead>
<tr>
<th>Paper</th>
<th>Course 1A 1997 (%)</th>
<th>Course 1A 1998 (%)</th>
<th>Course 1B 1997 (%)</th>
<th>Course 1B 1998 (%)</th>
<th>Course 1C 1996 (%)</th>
<th>Course 1C 1997 (%)</th>
<th>Course 1C 1998 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks allocated at Level I (knowledge)</td>
<td>1</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>30</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Marks allocated at Level II (comprehension/application)</td>
<td>87</td>
<td>77</td>
<td>88</td>
<td>79</td>
<td>60</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Marks allocated at Level III (HOCS)</td>
<td>12</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (n = total marks)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

What is immediately obvious from the data presented above in Table 1 is the consistency in level II testing. It is also clear that more level I testing occurs in course 1C than in the other first-year courses.

The results of the analysis of the 18 second-year examination papers placed under scrutiny have been condensed in Table 2. It must be noted that this is a summative table which masks a lot of important detail, for example the fact that there was a higher proportion of level I testing some modules compared with the others.

In common with first-year examinations, at second-year testing at level II predominates. What may be surprising to some is the relatively low percentage of the total marks assigned to testing at higher cognitive levels. Certain questions tend to reappear at regular intervals in certain subjects. It is interesting to speculate that, had questions reappearing from previous examinations been taken into consideration in determining cognitive levels, the proportion of level I and II questions might have been even higher.

G. Green and M. Rollnick
Table 2: Assessment of cognitive abilities required in answering second year examinations

<table>
<thead>
<tr>
<th>Level</th>
<th>1996 (%)</th>
<th>1997 (%)</th>
<th>1998 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level I (knowledge)</td>
<td>19</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Level II (comprehension/application)</td>
<td>79</td>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>Level III (HOCS)</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Total marks (n=total marks)</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

INSIGHTS AND PERSPECTIVES REGARDING RESTRICTIONS IMPOSED BY TIME-LIMITED TESTS

There are many ways in which the findings from the data can be interpreted. However, in the context of this article the preponderance of level II testing (summarised in Tables 1 and 2) may be considered by some to be disappointing in view of the commonly held idea that questioning at university level should occur mainly at the higher cognitive levels. However we wish to argue against this perspective, and present a counter argument that in fact the predominance of level II testing is neither unexpected nor unsatisfactory in view of the fact that the testing occurs in fixed time situations.

The teaching experience of Felder and Brent (1999) enables them to make the observation that, although it is considered pedagogically desirable to address all levels (albeit not necessarily sequentially) in every course, in practice undergraduate education deals almost exclusively with our levels I to II. It can be supposed that they are referring specifically to examinations, since it is not true for a course like chemistry where higher cognitive skills are required for some embedded tasks, like writing laboratory reports.

We wish to argue that one of the reasons for testing almost exclusively at lower cognitive levels lies in the innate characteristics of problems requiring HOCS which make them unsuited for limited time examinations. The argument is based on the fact that time constraints, the inherently stressful nature of the testing situation itself and the well documented phenomenon of test anxiety (Hembree 1988) make it unlikely that any but the best students would be able to answer questions in these categories in the time allowed. Critical thinking can take time!

This view is supported by many, including Felder and Brent (2002) who propose that tests which purport to show whether students can think independently (i.e., exhibit HOCS) are usually only advantageous to good puzzle-solvers even though other students may understand the course material just as well.

Lecturers should also accept restrictions imposed by time-limited examinations on possibilities for assessing performance at higher cognitive levels. It is possible to set fair questions involving the puzzle-solving which is the dominant feature of
NSE, which can nonetheless not be answered by rote. The reason for avoiding or minimising recourse to memorisation is that rote learning is not congruous with the deep learning which is supposed to take place at university, but even this deeper type of learning is enhanced through practice (Felder and Brent 2002). Even though our study shows the bulk of questioning occurs at level II, students still struggle to cope. This raises an interesting question about whether the number and/or complexity of the second year chemistry concepts themselves should be taken into account if possible explanations are to be sought for the failure rate. This relates to the nature of the material which we mentioned in the introduction to this article. Ramsden (1992) suggests that much of student learning might not be about understanding chemistry at all, but about adapting to the requirements of the assessment. If students perceive that rote learning techniques will bring success, then these will be the methods that they will use. There is also evidence from other studies that students may first make sure they understand concepts and then use rote learning strategies as a means of answering examination questions faster than they would be able to without memorisation (Kember 2000). Elbow (1986) explains that tests are at the heart of the dual, conflicting functions of a lecturer who acts simultaneously as gatekeeper and coach. As gatekeeper a lecturer is required to set appropriately high standards in tests. As coach, she has to do everything in her power to help them meet and surpass the test standards. The cognitive level of test questions is influenced by the instructional process (coaching) which has preceded the test (Airasian and Miranda 2002). A lecturer might have used certain level III questions in teaching. If students were exposed a second time to a similar level III question, because of its ‘familiarity’ this would relegate the question to level II on the second exposure. Only an insider would be able to put this all together and decide whether any question can, due to a particular teaching approach, be identified as a ‘type’ (i.e., routine) question or whether it indeed requires HOCS. Any researcher looking into cognitive demand should have some knowledge of examples students have been exposed to which would make it possible to avoid confusing ‘practised’ complexity (at level II) with true novelty (level III).

Even if students are unused to operating on higher cognitive levels they should be presented with opportunities to practice these skills outside of the examination situation. Any skill (including a higher level cognitive ability) is developed through practice and feedback (Felder and Brent 2002).

CONCLUSION

This report has described some of the problems associated with linking examination questions to specific cognitive abilities and, it is hoped, has provided some food for thought on the level of questioning appropriate in university examinations. We have shown that only an insider can meaningfully assign particular questions to cognitive levels. In addition our data support the finding of others (Zoller et al. 2002) that there appears to be little difference in the cognitive
levels of testing as one progresses from year to year in chemistry. We have proposed that this is not an undesirable situation in view of the fact that chemistry at Wits is presented within the context of normal chemistry education (Van Berkel et al. 2000) as well as the fact that timed tests by their nature are an inappropriate medium for testing HOCS (Felder and Brent 2002; Airasian and Miranda 2002). The imperative for any chemistry department wishing to improve HOCS would be to identify strategies which would ensure that outside of the examination situation their students are able to practice tasks requiring higher order cognitive abilities.

REFERENCES


