Assessment formats: do they make a difference?

Eleni Danili and Norman Reid

Centre for Science Education, University of Glasgow, Glasgow, G12 8QQ, UK
e-mail: N.Reid@mis.gla.ac.uk

Received 23 August 2005, accepted 13 October 2005

Abstract: This study has explored the relationships between the results of various formats of paper-and-pencil classroom assessments in five classroom chemistry tests. The formats of assessment that have been used were: multiple choice, short answer, and structural communication grid. The study was conducted in Greece with the participation of first year upper secondary public school pupils (Lykeio, Grade 10, age 15-16). The correlations between the different formats of assessment tended to be between 0.30 and 0.71. This is a wide range but even the highest value is well short of 1.0. This suggests that the best student found by one method is not necessarily the best student by another method. This raises questions about the validity of the formats of the assessment and what different formats of assessment are testing. [*Chem. Educ. Res. Pract.*, 2005, **6** (4), 204-212]

Key Words: Assessment formats, correlation, paper-and-pencil classroom chemistry assessment.

Introduction

Assessments play an important role in the teaching and learning process, and for specific uses. For individuals, assessments, particularly public examinations, profoundly affect life chances, not just in the first years after leaving school, but many years later. As Boud (1995. p. 35) stated "*the effects of bad practice are far more potent than they are for any aspect of teaching. Students can, with difficulty, escape from the effects of poor teaching, they cannot (by definition, if they want to graduate) escape the effects of poor assessment. Assessment acts as a mechanism to control students that is far more pervasive and insidious than most staff would be prepared to acknowledge*".

Indeed, to evaluate someone and make decision for his/her career and future is not an easy task to do. It is a very difficult one and carries with it awesome responsibility. Therefore, some authors characterizations for assessment were: "*both time consuming and potentially dangerous*" (Johnstone, undated, p. 2); "*a serious and often tragic enterprise*" (Ramsden, 2003, p. 13); "*nightmares*" (Race, 1995, p. 61).

Assessment can take many forms, and ideally, performance should be unrelated to the mode in which a test is administered. However, can this statement be true in a real situation? Evidence from research shows the effects of assessment task format on student achievement (e.g. Caygill and Eley, 2001). Moreover, Friel and Johnstone (1978a) showed that, if the same area of learning is assessed by normal, open-ended methods and also assessed by objective, fixed-response methods, two orders of merit are generated for a given group of students. Ideally, if a test is reliable (in a test, re-test sense) and same knowledge and understanding is being assessed, the two orders of merit should be actually very similar for the same sample of students. The best student by one method should be the best by another method, and so on down the line. In that case, if rank-order correlation is worked out between the two orders of merit, this should be 1.0, a perfect match in order. A complete reversal of the order would

give a value of -1.0 and a completely random pair of orders would give values tending to zero (Johnstone and Ambusaidi, 2000). Their research found that the rank-order correlation was about 0.6. This suggests that the two rank-orders of merit have in common only about 36% of the variance [$(0.6)^2 = 0.36$ that is 36%]. This suggests that the two orders of merit have some similarity, yet are by no means well matched (Johnstone and Ambusaidi, 2000).

In classroom assessment, the study of Yuh-Yin and I-Fen (2000) with science tests found that the correlation between multiple-choice items and short-answer question was 0.68 in one topic (solutions) and 0.77 in another topic (momentum), while the correlations between the same formats of assessment but in different content areas were smaller. Thus, for multiple-choice solution and multiple-choice momentum, the correlation was found to be 0.47, and for short-answer solution and short-answer momentum it was found 0.66. Moreover, correlations between multiple-choice items and short-answers questions with performance-based assessment were smaller even in the same area of content (0.48 and 0.46 respectively). It was assumed that, within the same content area, multiple-choice and short-answer tests measured similar cognitive components, while performance-based assessment emphasized different cognitive dimensions.

It is certain that no one method of assessment is adequate for testing a course. A battery of test methods is required to allow for a fair measure of our students' attainments (Balla and Boyle, 1994) and "*to cater for the range of student abilities, of testable objectives and student maturity*" (Johnstone, 2003). Indeed, Race (2003) argued that, "*the greater the diversity in the methods of assessment, the fairer assessment is to students*". However, in recent years there has been a temptation to adopt objective testing to cope with the rise in student population (Johnstone, 2003). This situation creates problems, because "*to conduct all assessment by this method is not advisable. The most intellectually mature students generally hate objective testing because they need room to expand and show their independence of thought*" (Johnstone, 2003). Moreover, each one of the formats of assessment can be claimed to disadvantage those students who do not give of their best in the particular circumstances in which it is used. Therefore, diversifying assessment so that students experience a range of assessment methods balances out the situation, and increases the chance that they will be able to demonstrate their best performance in at least some of the formats.

Design of the project

The aim of this study was to examine the correlation between the results of different paper-and-pencil formats of classroom assessment. The aim was to build on previous work by extending the range of formats and using a set of assessment across the work of many months of school study.

The work was conducted in Greece with the participation of first year upper secondary public school pupils (Lykeio, Grade 10, age 15-16) during March-April 2002 and during the school year September 2002 to May 2003. Five chemistry tests were designed and each chemistry test assessed pupils by a range of question formats asking about the same knowledge and understanding in the same topic.

There were difficulties and restrictions in relations to the format of questions that the researcher wanted to apply. For example, the researcher wanted to try not only Structural Communication Grid questions which allow for pattern seeking but also Structural Communication Grid questions which look for sequencing and even for a kind of 'objective essay' (Johnstone, 2003). However, the teachers objected to these questions because they thought pupils were not familiar with them and this might have caused problems.

The most common pencil-and-paper formats of assessment used in educational practice in Greece are:

Open-ended questions (OE)

Close-response question or objective tests such as:

Multiple-choice Questions (MC)

True-false,

Matching questions,

Identifying reasons to support assertions

Filling in blanks to complete statements

Grid questions (SCG)

Short answer question (SA)

Solving problems (mainly of algorithmic type)

The chemistry topics that were tested followed the timetable and the syllabus of the Greek schools. The tests were constructed after looking at the study questions with the Greek Chemistry textbook (Liodakis, 1999) the Standard Grade Chemistry book (Renfrew, 1995) and the textbook by Moore et al. (1998) in order to develop questions in formats and styles appropriate for the pupils. Thus, the tests were based on:

Test 1: Atomic structure, classification of matter, solubility

Test 2: The periodic table and chemical bonds

Test 3: Mole concept

Test 4: Acids, alkalis, pH, neutralisation

Test 5: Solutions

The researcher contacted several teachers of different schools and explained to them the purpose of the project. Schools were selected on the basis of teachers being willing to assist in what was a large project. Table 1 shows the range of question formats that have been used in each test of the project, the number of schools as well as the number of pupils who have been involved in the project.

Table 1: Type of formats of assessment used in each chemistry test and number of pupils have been involved

Chemistry Test	Type of formats	Number of schools	Number of pupils
Test 1	MC-SA	8	288
Test 2	SA-SCG	4	185
Test 3	SA-SCG	3	146
Test 4	MC-SCG-SA	7	321
Test 5	MC-SCG-SA	2	64

MC: Multiple-choice
SA: Short-answer (open-ended)
SCG: Structural communication grid

Only 64 pupils sat the Test 5. One of the reasons for that was that the test was given to the teachers towards the end of the school year and, at that time, the pupils usually are very busy with other activities. Thus, many hours of teaching are lost and teachers are mainly concerned to finish the teaching units and they are not willing to spend time to evaluate and assess the results of their teaching.

Weighting of marks was carefully decided to reflect the demand level of questions. For every section of each test, raw scores were converted into a percentage and these were combined to give the total mark for each pupil. Converting the raw scores to a common scale

makes it easier for the reader to compare mean scores between different tests and see patterns that might emerge from the study, though it does not alter the statistical results.

All the test papers were marked by the class teachers who were familiar with both the course and what was a reasonable standard. The test papers were then re-marked by the first author to ensure that standards were maintained. In fact, because there were strict marking schemes, marks were rarely modified on re-marking. It is easy to show that the inter-marker reliability will be very high given a tight marking scheme (typically over 0.95).

The tests were set to match the styles and standards typical in Greece and were based on the actual tests normally used. Mean marks did vary considerably, reflecting what normally happens in schools in Greece where all pupils have to take chemistry and a sizeable minority do *not* wish to take chemistry. In this way, the study reflected closely the reality of what actually happens. There is no reason to suppose that tests of different difficulty will lead to unreliability on its own in that the spread of marks in all tests suggests a good discrimination. Discrimination indices are quoted for the multiple-choice questions in the Appendix.

The aim was to make the various formats of the test in any content area as similar as possible. This is not easy and, in reality, tests seek to test *samples* of work in a content area. Nonetheless, the attempt was made as strictly as was possible.

Description of each chemistry test and statistical results

For each chemistry test, descriptive statistics and correlations between the different formats of questions were calculated. Both Pearson coefficient and Spearman's rho correlation between the formats of questions were calculated and were found to give similar values. However, because the distributions were frequently observed to deviate from the normal distribution, it was decided that the Spearman's rho coefficient was more appropriate and this is used in all subsequent discussion. The tests are shown in the Appendix.

Test 1: Multiple-choice vs. short-answer format

Test 1 was based on the introductory chapter of the Greek chemistry textbook. The content areas that it tested were atomic structure, classification of matter, and solubility. It consisted of two sections:

Section 1: 12 Multiple-Choice questions 12 marks,
Section 2: 5 Short-Answered questions 14 marks.

Section 1 had multiple-choice questions, which mainly require students to recognize or identify knowledge. In section 2 there were short answer questions covering the same thematic area. However, the demands on students were more than simply recognition and memorisation. The short-answer questions varied considerably. For example some required students to recall and define knowledge, others required to solve a numerical problem, which requires only a small numbers of steps but deep understanding of the concept involved, or to interpret a graph. Table 2 shows the descriptive statistics for test 1 and the Spearman's rho correlation between MC section and SA section. As can be seen from the table, the SA test was more difficult than the MC test. The Spearman's rho correlation between the MC and SA scores was found to be 0.71 (significant at the 0.01 level - 1-tailed). This correlation is the highest found in the whole study.

Table 2: Descriptive statistics of Test 1

Test 1	N	Minim.	Maxim.	Mean	S.D.
MC	288	17	100	64.3	20.4
SA	288	0	100	53.5	25.6

Spearman's rho between MC and SA = 0.71
significant at 0.01 level (1-tailed)

Test 2: Short-answer vs. structure communication grid

Test 2 was based on the periodic table and bonding theory chapters of the Greek chemistry textbook. It included two sections.

Section 1: 3 Short-answer questions 10 marks
Section 2: 1 Structural communication grid question 10 marks

In order to answer the test, pupils were allowed to have the periodic table in front of them. Thus, in both sections the questions require no recalling of the scientific facts. All questions require an understanding of taught concepts (the periodic table, the properties of the element and the concept of bonding theory), and an ability to interpret the presented information and to apply it. However, short-answer questions require pupils to use their language skills more, since the questions ask pupils to give explanations, e.g. for properties of compounds or for similarities of elements. Table 3 shows the descriptive statistics for the test 2. As can be seen for the table the SCG test was more difficult than the SA test and the Spearman's rho correlation between SA section and SCG section was found to be 0.38, which is relatively low.

Table 3: Statistics of the Test 2

Test 2	N	Minim.	Maxim.	Mean	S.D.
SA	185	0	100	52.2	30.7
SCG	185	0	100	36.7	25.4

Spearman's rho correlation between SA and SCG = 0.38
significant at the 0.01 level (1-tailed)

Test 3: Short-answer vs. structure communication grid

Test 3 was based on the mole concept and Avogadro's Law. It was a short test and included two sections.

Section 1: 2 Short-answer questions 10 marks
Section 2: 1 Structural communication grid question 10 marks

The questions require retrieval of declarative knowledge and procedural knowledge, as well as numerical problem solving ability (of the algorithmic type) in both formats of assessment

Table 4: Statistics of the Test 3

Test 3	N	Minim.	Maxim.	Mean	S.D.
SA	146	0	100	60.9	37.3
SCG	146	0	100	67.7	36.7

Spearman's rho correlation between SA and SCG = 0.55
significant at the 0.01 level (1-tailed)

Test 4 :Multiple-choice vs. grid vs. short-answer

Test 4 was based on the content area of acids; bases; oxides and neutralisation reactions. It had three sections:

Section 1: 13 Multiple-choice (MC) questions 13 marks.
Section 2: 2 Structural communication grid (SCG) questions 12 marks.
Section 3: 3 Short-answer (SA) questions 14 marks.

Mainly, the questions asked students to recall, define, recognise and apply knowledge.

Table 5: Statistics of Test 4

Test Format	N	Minim.	Maxim.	Mean	S.D.
MC	321	8	100	52.9	19.8
SCG	321	0	100	35.2	23.3
SA	321	0	100	34.5	31.4

Spearman's rho correlations between:
MC and SCG = 0.64
MC and SA = 0.64
SA and SCG = 0.66
all significant at the 0.01 level (1-tailed)

Test 5: Multiple-choice vs. grid vs. short-answer

Test 5 was based on the content area of solutions. It had three sections:

Section 1: 5 Multiple-choice questions 5 marks
Section 2: 1 Structural communication grid question 5 marks
Section 3: 3 Short-answer questions 5 marks

The questions were developed mainly from the chemistry book of Moore et al. (1999), in which the assessment questions test understanding and applying chemical concepts. Thus, the answers to the test did not require much memorisation and recall of chemical concepts but the ability to interpret the given information, and understanding of the concept of concentration in solutions, and how it changes when water is added to the solution or water is evaporated from the solution. It required arithmetic skills for answering the open-ended questions.

Table 6: Statistics of Test 5

Test 5	N	Minim.	Maxim.	Mean	S. D.
MC	64	0	100	67.5	28.6
SCG	64	0	100	68.6	26.4
SA	64	0	100	50.8	35.6

Spearman's rho correlation between:
MC and SCG = 0.46
MC and SA = 0.49
SA and SCG = 0.30
all significant at the 0.01 level (1-tailed)

It was expected that very high correlations between different formats of assessment in this test would be found because it was testing the same narrow area of understanding (concentration of a solution, and how the concentration changes by mixing two solutions or diluting a solution). Thus, it is surprising that the correlations were fairly low.

Comparison between Test 1 and 2 for the same group of pupils

Some pupils sat two tests (1, 2). This gave an opportunity to explore the correlations across content areas for the same and for different formats of questions for this group of pupils. Table 7 shows a correlation matrix between the different formats of questions in each test and for the three formats.

Table 7: Correlations across content areas in Test 1, 2

	Test 1		**Test 2**	
	MC	**SA**	**SA**	**SCG**
MC Test 1	1.00	0.54**	0.38**	0.26*
SA Test 1	0.54**	1.00	0.58**	0.42**
SA Test 2	0.38**	0.58**	1.00	0.54**
SCG Test 2	0.26*	0.42**	0.54**	1.00

** Correlation is significant at the 0.01 level.

Table 7 shows that the correlations between the different formats of assessment in the *same* content area both have a value of 0.54 (short answer – multiple choice; short answer structural communication grid).

The correlation between the short answer formats of assessment in *different* content areas is 0.58, this latter figure suggesting a reasonable reliability of testing.

Multiple choice in test 1 correlated with structural communication grid in test 2, with a value of 0.26 while short answers in test 1 correlated with structural communication grid in test 2, with a value of 0.42.

In one study a correlation of 0.6 was found between multiple choice and short answer question in the same content area. When the multiple-choice questions were re-marked using partial credit marking, the correlation rose to 0.9 (Johnstone and Ambusaidi, 2001). This suggests that test reliability itself is not the problem but the method of marking may be.

There are several factors that might explain the low correlations:

(a) Test reliability – probably not a major factor;
(b) Content area – pupils may perform differently with different material;
(c) Test structure – in table 7 three structures are used.

In multiple-choice tests, pupils tend to eliminate two responses and make a decision between the remaining two, the test being a test of recognition. In structural communication grids, pupils look at each box in turn and decide whether it contains a possible answer, this involves recognition, although thought may be required (in that the number of possible answers is not known). In short answers pupils have to interpret the question and generate answers by recall or thought. In essence, the three types of test are probing different skills and this probably explains the lack of perfect correlation. However, it does raise issues about what skills are being measured – psychological or knowledge of chemistry.

Conclusions

From all the tests it is clear that pupils' performances in MC section were higher than SCG items and SA items. In addition, in this study, the correlation values ranged from 0.30 to 0.71. The higher values tended to occur when *different formats* tested the *same content area* but none of these approached 1.0. We expected higher correlations between MC and SCG

questions because both are objective tests. However, the correlations are fairly low in these comparisons. Thus, differences in correlation did not simply arise because some questions require writing while others are objective (in the sense that only a number or letter is required or a box has to be ticked).

These findings are consistent with those of Friel and Johnstone (1978a), Yuh-Yin (2000) and Badger (1990). This suggests that the best student found by one method is not necessarily the best student by another method. If the two formats of assessment were simply testing the same content, then very high correlation would be expected. This also raises questions about the validity of the formats of the assessment. The main question is, what are the different formats testing? Are the different formats testing different abilities and skills, which involve different cognitive factors? Are the different formats testing chemistry or cognition? Are the ways in which the questions are presented having an impact of the pupils' performance (e.g. the use of pictures, or diagrams)? Thus the fundamental issues arising from the study are:

1. Are the different formats of questions testing different abilities or just different themes in a discipline? Probably both?
2. Is any particular format of assessment more valid than others?
3. Are the different formats related to differences between students in one or more psychological traits?
4. It might be reasonable to suppose that the use of multiple formats of assessment tests students more fairly than the use of a single format but on what basis can this be justified?

Clearly, assessment formats do make a difference to student performance. These results led to another study that sought to explore some of the psychological factors that might account for these difference. The findings of that study will be discussed in a future paper.

References

Balla J. and Boyle P., (1994), Assessment of student performance: a framework of improving practice, *Assessment and Evaluation in Higher Education*, **19**, 17-28.

Boud D., (1995), Assessment and learning: contradictory or complementary? In P. Knight (Ed.), *Assessment for learning in higher education*, London: Kogan Page Ltd.

Caygill R. and Eley L., (2001), *Evidence about the effects of assessment task format on student achievement*, Annual Conference of the British Educational Research Association.

Friel S. and Johnstone A.H., (1978), Scoring systems which allow for partial knowledge, *Journal of Chemical Education*, **55**, 717-719.

Johnstone A.H., (2003), LTSN Physical sciences practice guide: effective practice in objective assessment, Hull, LTSN.

Johnstone A.H., (undated), Unpublished lectures, unpublished manuscript, Centre for Science Education, University of Glasgow.

Johnstone, A.H. and Ambusaidi, A., (2000), Fixed Response: what are we testing? *Chemistry Education Research and Practice*, **1**, 323-328.

Liodakis S., Gakis D., Theodoropoulos D., Theodoropoulos P. and Kallis A., (1999), Chemistry A' Lyceum Athens: Organismos Ekdoseon Didaktikon Biblion.

Moore J., Stanitski C., Wood J., Kotz J. and Joesten, M., (1998), *The chemical world concepts and applications* (2nd ed.), New York, Saunder College.

Race P., (1995), What has assessment done for us and to us? In P. Knight (Ed.), *Assessment for learning in higher education*, London, Kogan Page Ltd.

Race P., (2003), *Designing assessment to improve Physical Sciences learning-exams*, http://www.physsci.ltsn.ac.uk/Publications/PracticeGuide/guide4.pdf [2003, 6/10/2003].

Ramsden P., (2003), *Learning to teach in higher education* (2nd ed.), London, Routledge Falmer.

Renfrew R. and Conquest N., (1995), *Standard Grade Chemistry*, London, Hodder and Stoughton.

Yuh-Yin W. and, I-Fen G., (2000), *Classroom assessment forms and their relations with cognitive components. An example from Taiwan,* Paper presented at the Annual meeting of American Educational Research Association, New Orleans, Louisiana.

The Appendixes associated with this paper can be found as separate PDF files at http://www.rsc.org/Education/CERP/issues/2005_4/index.asp