Time to make our mark

STEVE NEWSTEAD *believes psychology can improve the reliability and validity of student assessment.*

A COLLEAGUE recently suggested to me over coffee, only half in jest, that we should completely abandon the assessment of students. He pointed out that we have a good idea of what sort of degree class they will get before they even arrive, since we know that factors such as previous qualifications, age and gender all serve to predict how well students will perform, as does the university at which the degree is studied. If we can make a reasonably accurate prediction based on such factors, why bother to put students through all the stress of examinations and other types of assessment? Even more importantly (at least from his perspective), why devote so much staff time to setting and marking assessments when they add so little extra information?

Slightly to my embarrassment, he went on to cite some of my own research in support of this view. We know, for example, that marking of essays is not very reliable – the same piece of work can attract very different marks from different markers (Newstead & Dennis, 1994). We also know that biases may exist that favour one group of students over others (Bradley, 1984; Dennis *et al.*, 1996). And we know that students cheat, with more than half of them reporting having plagiarised material (Newstead *et al.*, 1996). If the assessment system is so flawed, why bother? Why not let staff get on with their research and students get on with the other benefits of a university education, such as learning about independent living, discussing the issues of the day, and having a good social life?

What is assessment for?

This basic question must be given some sort of answer before one can determine ways of improving methods of assessment. But it depends who you ask – presumably the lecturers who do the assessment have some sort of expertise here, but they give rather divergent answers. In a qualitative study Samelowicz and Bain (2002) found that at one end of the spectrum, lecturers perceived the purpose to be assessing students' ability to reproduce information; at the other end, the purpose was seen to be that of assessing the ability to integrate, transform and use information purposefully. It seems probable that the perceived purposes of assessment vary between different subjects and different types of knowledge, but at the very least it is clear that there is no single perceived purpose of assessment among those who do the assessment.

Nor is there a close match between lecturers' and their students' perceptions of assessment. Norton (1990) investigated staff and student perceptions of the key factors determining the mark awarded to a first-year psychology essay. Students tended to stress the amount and accuracy of the information in the essay, whereas lecturers stressed how the information was used to develop an argument. More recently, Maclellan (2001) has shown that lecturers emphasise the importance of assessment in motivating students and in diagnosing the effectiveness of teaching and learning. Students, however, tend to see the main purpose of assessment as that of ranking or grading their performance.

At a more conceptual level, there are different purported objectives of assessment. Some stress the formative aspect of assessment, in other words its effect in enhancing students' learning and development. Others stress the summative aspect, where assessment aims to achieve an overall summary (often a single number or grade) of a student's performance that can be compared with that of other students. It is possible to have assessments that fulfil both of these purposes simultaneously, but these are quite rare – not surprisingly given the conflicting aims of the two types of assessment.

In recent years there has been a move away from what is sometimes called the

WEBLINKS

Learning and Teaching Support Network:
www.ltsn.ac.uk/genericcentre/index.asp?id=16892

Teaching practice issue – assessment and examining: *ltsnpsy.york.ac.uk/ltsnasp/teachpracissuesl.asp?id=5*

measurement model, in which the aim is to place student performance on a scale and compare it with the performance of other students. More and more education experts are now emphasising the standards model of assessment, in which the purpose is primarily to ensure that a certain level of skill or competence has been reached. The measurement model usually leads to marking on a scale, as happens with degree classifications. The standards model usually leads to pass/fail marking and is often advocated by those wishing to move to a profile system, in which students are given a detailed breakdown of their performance rather than a single mark. The standards model is currently common practice in the assessment of students at master's level.

All of the above is little more than scene-setting for a discussion of how the assessment system can be improved. However, it does illustrate the complexity of the issue. Improvements can only be measured against the purposes of the assessment system; but if there is no real agreement about those purposes, the task becomes well-nigh impossible.

And there is another complicating factor. Biggs (e.g. 1993) has argued on many occasions that higher education is an integrated system. This means that a change in one part of the system is likely to have knock-on effects, sometimes unpredictable and often quite dramatic, elsewhere in the system. Thus changes in the assessment system might lead to changes in, for example, the motivations and perceptions of both students and staff. It also means that a change in one part of the assessment system is likely to have effects on other aspects of assessment. The following discussion of possible improvements needs to be considered against this background.

Improving reliability

It is widely agreed that the type of assessment we carry out, still dominated by essay-based exams, is unreliable. This was shown in a major study carried out by Hartog and Rhodes (1935) and has been confirmed many times since (Caryl, 1999; Dracup, 1997; Elander & Hardman 2002; Laming, 1990; Newstead & Dennis, 1994). The proposed solutions to this are many and varied but all bring problems of their own. Perhaps the most novel of these ideas is Laming's (1990) suggestion that each essay should be marked only in comparison to the essays read immediately before and after it, and that this relative assessment should be on a simple five-point scale (ranging from much better to much worse). Whenever I have discussed this with students they have thrown their hands up in horror. They point out that the mark they would obtain might depend on chance factors, since the same piece of work could get a high mark if surrounded by essays of poor quality and a low mark if surrounded by work of high quality. This is true, but Laming's method does have the distinct advantage of being within the known limits of human judgement, whereas marking on a 100-point scale is completely beyond our capabilities, as Miller (1956) pointed out many years ago.

Another suggestion to increase reliability is to abolish essays and to use only objective tests. This is essentially what has happened in the US, where graduate admissions staff rely more on the results of nationally delivered objective tests than they do on the grade point average awarded by the university where the student has studied. The objective test in question is the Graduate Record Examination (GRE), which tests verbal, numerical and logical reasoning ability, in addition to subject knowledge. Meta-analyses have shown that these tests provide a reasonable predictor of success in graduate school (Kuncel *et al.*, 2001). What is more, there is the potential to computer-generate at least some of these tests in such a way that each candidate is provided with a unique set of items that are adapted to the ability level of the candidate (Newstead *et al.*, 2002).

'The obvious problem…is that we do not know what a valid assessment is'

However, there are also potential disadvantages to such objective tests. Lecturing staff often point out that such tests do not measure all aspects of success at postgraduate level, for example writing skills and creativity. Nor is there any evidence of their ability to predict the success of those students who do not go on to graduate school, since such students do not take the GRE. There is also evidence that some sorts of objective tests might favour males over females (e.g. Murphy, 1982). While they probably have an important part to play in assessment, it seems unlikely that objective tests will take over from more traditional assessment methods in the foreseeable future.

Another suggestion for improving reliability is the use of criteria and marking schemes. The assumption is that these will reduce individual differences in marking and remove the more subjective aspects of assessment. The evidence that behavioural anchors can improve the reliability of ratings in general is not clear-cut (see, for example, Landy & Farr, 1980), and there is surprisingly little evidence on the use of criteria in the assessment of students' work. Hartog and Rhodes (1935) found that marking criteria reduced the variability of marks, but also led to an overall increase in marks. In the study that we carried out (Newstead & Dennis, 1994) we asked markers to score essays on five criteria as well as giving an overall mark, but these five marks were very similar to the overall mark and seemed to add very little further information. Elander and Hardman (2002) used a more comprehensive range of assessment criteria, but still found very high intercorrelations between these and the overall mark. Criteria may help to improve reliability, but the effect is unlikely to be great. In addition, it must be borne in mind that the use of marking criteria may bring about other changes, possibly altering students' approaches to their studies and making them more strategic in their learning.

One thing we can be sure of is that the more assessments are made, the greater the overall statistical reliability will be. It therefore seems reasonable to suggest that assessors should, where possible, deliver a large number of smaller assessments rather than a small number of very large ones.

Improving validity

The obvious problem in trying to improve validity is that we do not know what a valid assessment is, since there is no consensus on the purposes of assessment. However, one might expect that performance at degree level should predict future success on leaving university. It is widely accepted that university graduates earn more than their counterparts who do not enter higher education. Indeed, this has been one of the central arguments in the government's case

for introducing increased students' fees. But there is a flaw in this argument. It seems possible, even probable, that those who benefit from a university education would have out-earned their counterparts even if they had not gone to university. The students who attend university are among the most able and hence would have been expected to be more successful even without their university education.

Comparing graduates with non-graduates is a fairly crude way of determining validity. A slightly more refined measure might be the extent to which degree class predicts future success, though there appear to be relatively few studies that have looked at this. Brennan and McGeevor (1987) found that earnings did correlate with degree classification. However, their study took place several years ago and pre-dates the massive expansion of higher education in the 1990s.

There is a pretence that degree standards are equivalent

All in all, it is difficult to find any firm evidence that assessment at degree level is valid as measured by the later success of the graduates. If the aim of higher education is to improve employability, then one way of improving validity is presumably to make assessments more relevant to the world of work. Unfortunately, there is little consensus on what skills are required by employers and even less on how to measure these. Given the state of our knowledge, it is virtually impossible to make any sensible suggestions about how validity might be improved.

Improving standardisation

The output of the assessment system should in principle be relatively easy to interpret. The output is usually phrased in terms of degree class (first, upper second, and so on), and these grades therefore ought to have some standardised meaning. The truth is that they do not, since there are marked differences between institutions, between disciplines and over time (see Newstead, 2000, for a summary). There is a pretence that standards are equivalent, and external examiners are entrusted with ensuring that this is the case, but this is demonstrably ineffective.

If we seriously wish to ensure comparable standards then radical measures are required. It is, for example, well known that coursework tends to lead to higher marks than examinations. In a recent study by Bridges *et al.* (2002), this difference was present in all disciplines and as high as two thirds of a degree class in some. Hence there would need to be some attempt to standardise assessment methods. Of course, the problem here is that this would reduce the flexibility of staff to introduce new methods of assessment, or to tailor the assessments to the needs and aims of their own course.

I suspect, however, that we need to do much more than this if we are to have truly comparable degree classifications. One suggestion from within our own discipline is that of Howarth (1993), who believes we should have some kind of national examination. His ideas are not quite as radical as they seem at first sight, since he suggests that this national exam would count for only about one third of the final degree and would cover just the basic areas of psychology. The other two thirds would come from the project and from assessments geared specifically to the more specialist material covered by each university.

Whenever suggestions such as this are put forward there are howls of protest. It is claimed that it would reduce the freedom of universities to teach what they deem appropriate and to assess it as they see fit. While I have much sympathy with this viewpoint, the fact remains that without such drastic measures there will never be

References

Biggs, J. (1993). What do inventories of students' learning processes really measure? *British Journal of Educational Psychology, 63*, 3–19.

Bradley, C. (1984). Sex bias in the evaluation of students. *British Journal of Social Psychology, 23*, 147–153.

Brennan, J. & McGeevor, R. (1987). *Graduates at work: Degree courses and the labour market.* London: Jessica Kingsley.

Bridges, P., Cooper, A., Evanson, P., Haines, C., Jenkins, D., Scurry, D. *et al.* (2002). Coursework marks high, examination marks low: Discuss. *Assessment and Evaluation in Higher Education, 27*, 35–48.

Caryl, P.G. (1999). Psychology examiners re-examined: A 5-year perspective. *Studies in Higher Education, 24*, 61–74.

Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J., & Cohen, G.M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General, 126*, 393–413.

Dennis, I., Newstead, S.E. & Wright, D.E. (1996). A new approach to exploring biases in educational assessment. *British Journal of Psychology, 87*, 515–534.

Dracup, C. (1997). The reliability of marking on a psychology degree. *British Journal of Psychology, 88*, 691–708.

Elander, J. & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology, 93*, 303–328.

Hartog, P. & Rhodes, E.C. (1935). *An examination of examinations*. London: Macmillan.

Howarth, C.I. (1993). Assuring the quality of teaching in universities. *Reflections on Higher Education, 5*, 69–89.

Kuncel, N.R., Hezlett, S.A. & Ones, D.S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181.

Laming, D. (1990). The reliability of a certain university examination, compared with the precision of absolute judgements. *Quarterly Journal of Experimental Psychology, 42A*, 239–254.

Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

Maclellan, E. (2001). Assessment for learning: The different perceptions of tutors and students. *Assessment and Evaluation in Higher Education, 26*, 307–318.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Murphy, R.J.L. (1982). Sex differences in objective test-performance. *British Journal of Educational Psychology, 52*, 213–219.

Newstead, S.E. (2000). Silk purse or sow's ear: A psychological perspective on recent developments in higher education. *Psychology Teaching Review, 9*, 1–10. [A shorter version of this paper appears in *The Psychologist* (2000), *13*, 184–188]

Newstead, S.E., Bradon, P., Handley, S. Evans, J.StB.T. & Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S.H. Irvine & P.C. Kyllonen (Eds.) *Item generation and test development.* Mahwah, NJ: Lawrence Erlbaum.

Newstead, S.E. & Dennis, I. (1994). Examiners examined: The reliability of exam marking in psychology. *The Psychologist, 7*, 216–219.

Newstead, S.E., Franklyn-Stokes, B.A. & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology, 88*, 229–241.

Norton, L.S. (1990). Essay writing: What really counts? *Higher Education, 20*, 411–442.

Samelowicz, K. & Bain, J.D. (2002). Identifying academics' orientation to assessment practice. *Higher Education, 43*, 173–201.

true standardisation of degree classes. External examiners almost certainly cannot achieve this, however well they are trained. If we do not ensure some sort of comparability, then this is likely to be imposed on us by others. Profiles of performance may be introduced, or nationwide objective tests such as the GRE, or degree-level exam boards along the lines of the (currently much-maligned) A-level boards.

Some of these solutions may seem attractive, but they are unlikely to be universally welcomed, and could be highly problematic if introduced for political rather than academic reasons. It is not known what the side-effects of introducing a national exam would be, but it could well have a major impact on the attitudes and motivation of both staff and students.

Conclusion

It would be comforting if, as psychologists, we were able to point out solutions to the problems of student assessment. However, psychologists (myself included) seem to be rather better at pointing out problems than delivering solutions. What is more, there are, as we have seen, good reasons for the inability to suggest solutions, since there is no real consensus over the purposes of assessment, and since we know that higher education is an integrated system. It is possible to make some relatively minor suggestions for change that would seem to be fairly non-controversial, for example that there should be a reasonably large number of assessments and that these should not be all of the same type. But it is difficult to go beyond that. Other changes depend on what the purpose of assessment is perceived to be. If the aim is to test consolidation in memory, then we can, for example, suggest that a reasonable lapse should occur between presentation of material and testing (Conway *et al.*, 1997). But if the aim is to test skills acquisition, such a delay may be unnecessary.

'If we seriously wish to ensure comparable standards then radical measures are required'

It is rather a trite and unsatisfactory conclusion to reach, but we will not be able to make sensible suggestions until we have agreement on what we are supposed to be assessing and until we know what systemic effect any changes are likely to have. In the absence of these, we can do little more than tinker around the edges. But revolutions often have innocuous beginnings, and this is certainly an area where psychology has the potential to improve the higher education experience.

■ *Stephen E. Newstead is in the Department of Psychology, University of Plymouth. E-mail: snewstead@plym.ac.uk.*