Wanted! Evidence based guidelines for unseen invigilated examinations

Ieva Stupans
Division of Health Sciences, University of South Australia, ieva.stupans@unisa.edu.au

Assessment has been the subject of vast amounts of literature in higher education for at least the past two decades. In undergraduate health science and science courses assessment of practical classes or clinical performance are quite common assessment components for students. Additionally, assessment components also frequently include unseen, invigilated, timed examinations. They are widely used to summarise what students know at certain times i.e. assessment of learning, in order to report achievement and progress. This is in spite of extensive literature around engaging students in assessment for learning through coursework assessments, particularly in the context of group work. This paper explores aspects of unseen invigilated examinations, such as their alignment with 'traditional' as opposed to 'alternative' assessments, the use of multiple choice questions, whether of a lower or higher cognitive level, the mix of multiple choice questions, short answer and essay questions used in papers, open book versus closed book papers and argues that there is a need for the development of evidence-based guidelines or principles which help guide and inform practice for the construction of unseen invigilated examinations.

Introduction

Current mainstream educational thinking is dominated by a constructivist view, which argues that deep learning occurs when a learner is actively engaged in learning activities and occurs where previous learning is linked with current i.e. constructivism focuses on knowledge construction, not knowledge reproduction. Constructivism contrasts to the direct instruction (instructivist or objectivist view) of education, which presumes knowledge exists independently of the student. Surface approaches to learning occur when there is a mere completion of the tasks at hand i.e. within the instructivist view (Ramsden, 1992).

We know that teaching, learning and assessment are linked. The role of coursework assessments in the constructivist view has been commented on extensively in the context of assessment *for* learning (Gibbs & Simpson, 2004-05). Students prefer coursework assessment and tend to gain higher marks from coursework assessments than they do from examinations; coursework marks appear to be a better predictor of long term learning of course content than are examinations; and lastly, the quality of student learning has also been shown to be higher in assignment-based courses (Gibbs & Simpson, 2004-05). Within mainstream constructivist thinking in the context of assessment *for* learning, there is extensive peer reviewed literature on the concept of teachers and students engaging in the assessment process as co-participants, the importance of feedback quality and timelines and topics such as peer assessment approaches and assessment of online contributions amongst others (Angelo and Cross 1993; Gibbs & Simpson, 2004-05; Harlen, 2005).

Coursework assessment is perceived to have greater relevance and utility for profession-related needs and lifelong learning. However, it has been acknowledged that work within this area is needed, in order to develop rigorous and enduring assessment standards, achievable within the constraints of conventional university practice (Williams, 2008).

The unseen invigilated exam

What has been commented on far less in the literature is 'best practice' or 'evidence-based practice' associated with unseen invigilated i.e. supervised (Merriam-Webster Online, 2008) examinations. Evidence-based practice identifies evidence that there may be for a practice and rates it according to how sound it may be. Its goal is to eliminate unsound practices. Table 1 presents an overview typology of examinations referred to in the literature. For the purposes of this paper, examinations refer to unseen assessments completed by students individually; their completion by students is invigilated. Arguments for or against timing of examinations are not explicitly referred to in this paper.

Table 1: Typology of Examinations

- Timed and untimed
- Composite or single question type - multiple choice questions, short answer questions, essay questions
- Open book, partial open book, 'cheat sheets' permitted and closed book

Assessment has been described as being for learning or as of learning through coursework type assessments or through examinations respectively. Assessment has also been described as 'traditional' or 'alternative', based on the learning domains of Bloom's taxonomy that the assessment best measures. 'Traditional' assessment measures learning at the lowest levels of Bloom's cognitive domain: knowledge and comprehension (Robles & Braathen, 2002). In contrast, 'alternative' assessment is positioned to measure learners' higher-level thinking skills of synthesis, analysis, and evaluation (Robles & Braathen, 2002). Within Bloom's taxonomy, educational objectives are ordered hierarchically; learning at higher levels depends on the attainment of the skills and abilities at the lower levels. It is important to note that assessment *of* learning is not mutually exclusive from alternative assessment, which measures higher level thinking skills although assessment of learning is frequently associated with question spotting, cramming and short-term knowledge retention (Williams, 2008). Therefore, the concept of structuring questions which allow students to demonstrate higher-level thinking skills is one which needs to be developed.

Alternative assessment may first be contextualized in real-world applications and second, involve students in problem-solving amongst a number of other criteria such as involving students in determining assessment criteria and focusing on collaborative skills as well as intellectual achievements in assessments (Williams, 2008). Unseen invigilated examinations may include problem solving questions (an example of alternative assessment), however it has been acknowledged, alongside the need for validity, there are consequent problems of task specification and consistency of marking (Maclellan, 2004). It should be acknowledged that issues around task specification and consistency of marking occur irrespective of whether assessment is regarded as being alternative or not.

Unseen invigilated examinations are a frequently used form of assessment, particularly in health disciplines and sciences. A recent audit from an Australian university (Taylor, 2006) indicated that approximately 50% of first year subjects had final examinations within their assessment profile, within sciences this was approximately 80%. A US study of pharmacy programs (Kirschenbaum, Brown, & Kalis, 2006) found that written assessments and/or

examinations were used in the case of 81% of assessments at Colleges and Schools of Pharmacy. Amongst academics in the sciences and health areas, unseen invigilated examinations are considered highly appropriate for assessment *of* learning, particularly with respect to certification or accreditation of learning by an external body, such as a health profession accreditation body which frequently use examinations as part of their processes. For example, FIP, the pharmacy arm of the World Health Organisation recommends 'a final examination should lead to the granting of a diploma or degree signifying either achievement of the academic requirement for recognition as a pharmacist' (FIP, 2000).

The views of sciences and health sciences academics supporting the continued use of unseen invigilated examinations are essentially that assessment measures for course work assessments have not yet developed the level of validity necessary to make appropriate a dependence on these measures for assessing student performance (Maclellan, 2004). Validity has been defined as a condition that exists when tests 'reflect achievement on the dimensions that the school wishes to evaluate' (Black & Duhon, 2003). There are disparate views presented in the literature as to whether validity can or cannot be achieved without reliability i.e. 'the extent to which a test, or any form of measurement, yields consistent results' (Bers & Smith, 1990; Moss, 1994; Black & Duhon, 2003).

It has been claimed that 'well-developed written examinations can provide a high level of validity and reliability in measurement of some types of learning' (James, McInnis, & Devlin, 2002). How do we then develop best practice for unseen invigilated timed examinations of an open or closed book type?

Extensive guidelines, based on personal viewpoint, for unseen 'tests' can be sourced from literature (Felder, 2002). It has been suggested in these guidelines that problems in tests should only cover content which has been covered, problems should not be overly tricky, with solutions that need to be worked out on the spur of the moment; examinations should not be so long that only the best students can finish in the allotted time; have excessively harsh grading, or inconsistent grading. Others have commented on the unfair examination with 'assessment by ambush' (McCoubrie, 2004) in which questions are chosen to discriminate between high and low achievers leading to omission of essential parts of the curriculum, because they are 'too easy'. Assessment design techniques such as 'blueprinting' are suggested (Crossley, Humphris, & Jolly, 2002). A test 'blueprint' defines and outlines the proportion of questions to be allocated to each content area and the cognitive knowledge levels of the questions.

In high stakes tests there is a premium on reliable marking. Large classes also present a challenge; there is reference to 'a growing reliance on exam-based assessment ……with an increased use of multiple-choice and short-answer or 'tick-a-box' questions' (James et al., 2002). This has the effect of reducing what is assessed to what can be readily and reliably marked (Harlen, 2005). Optical scanning of MCQ student answer sheets can be performed and with appropriate software, item analysis including percentage of students choosing the correct option and the biserial correlation for the item can be calculated, along with discriminator analysis and the KR20 for the test/exam overall (Kehoe, 1995). MCQ provide an ideal vehicle with which to assess 'body of knowledge'. Various texts are available which detail the features of 'good' MCQ (Kehoe, 1995; Case & Swanson, 2002), therefore questions can be well written, however one of the key questions is that of MCQ and their assessment beyond basic knowledge recall. Problem solving requires that students are able to apply, analyse and synthesise information. It is therefore appropriate to ask whether MCQ, which

test higher order cognitive thinking, can be written to test subject material and more importantly, used successfully in administered MCQ tests and exams. It has been suggested that writing of MCQ, which address higher order cognitive thinking rather than knowledge recall, is unlikely to be accidental (Stupans, 2006).

Should examiners construct and students sit, papers which include MCQ and short answer and essay type (extended answer) questions? The incontestable answer - assessment should align with the course objectives, but should both question styles be included in the paper? Results from a medical school examination in which students answered questions either by choosing the correct MCQ alternative or provided an explanation for the reasons for choosing the correct alternative, indicated that answers alone discriminate adequately among students with different levels of knowledge and ability (Schwartz & Loten, 1999) suggesting that by setting appropriate questions, the MCQ question style may be sufficient. Perusal of examination papers has also suggested that construction of essay questions which assess higher order cognitive skills is also not a 'simple task' (Palmer & Devitt, 2007). A study of composite reliability of the examination undergraduate clinical examinations – MCQ, extended matching questions, short-answer questions, essays, an objective structured clinical examination and history-taking long cases indicated that examination structure must be carefully planned and results combined with caution. The components testing different aspects of knowledge and clinical skills must be carefully balanced to ensure both content validity and parity between items and test length (Wass, McGibbon, & Van der Vleuten, 2001).

Clarity of questions in the unseen invigilated timed examination may be an issue. Studies of papers from high stakes MCQ medical examinations have indicated that imprecise terms are frequently used in papers and there is a wide range of interpretation amongst the examiners about their meanings (Holsgrove & Elzubeir, 1998). This has led to promulgation of guidelines for the construction of MCQ (Case & Swanson, 2002) (Lorusso, 2004); however the author of this paper was not able to locate an analysis of short answer or essay type question interpretations published in peer reviewed literature. The following refers to formulation of questions for online discussion, however can easily be used to describe questions in exam papers.

'One of my favorite movie scenes occurs in the Pink Panther Strikes Again. Peter Sellers, as Inspector Clouseau, is standing at the front desk of a hotel and sees a dog lying by the front door. In an exaggerated French accent, he asks the clerk, 'Does your dog bite?' The man answers, 'No.' Walking toward the door, Clouseau bends down to pet the dog; it growls and then bites him. Aghast, he exclaims, 'I thought that you said your dog does not bite!' The man responds, 'Oui, monsieur, but that is not my dog.

Obviously, Inspector Clouseau did not ask the right question.' (Toledo, 2006)

It has been argued that in the information age the closed book, invigilated final examination has become an 'anachronism' (Williams, 2006). This has been eloquently argued in the following

'Consider the following scenario, common to workplaces all over the world, each and every day.
Which is the more probable?
(a) Boss to employee. Look, we've got a real problem here … you've got an M.B.A. haven't you? Can you write me a report on this and email it to me by 9 a.m. tomorrow?

(b) Boss to employee. Look, we've got a real problem here … you've got an M.B.A. haven't you? Can you lock yourself away in that room, don't talk to anyone, don't browse the web or open any books and give me your answers to these multiple-choice questions in 3 hours' time?' (Williams, 2006).

If the concept of a closed book, invigilated final examination can no longer be argued for and this type of assessment has become an 'anachronism' (Williams, 2006), academics may also choose to use open book examinations, partial open book examinations i.e. an examination in which students have access to a limited number of resources or examinations that students sit having prepared a crib sheet – generally an A4 page of notes that students construct prior to the examination and then refer to during the examination.

Studies with crib (cheat) sheets are variously viewed (reviewed, Hamed, 2008). They have been suggested to have no effect on student performance for either higher order or lower order examination questions and to have no effect on student anxiety in some studies (Dickson & Miller, 2005), whereas other studies suggested their use increased students' deep learning and decreased student stress around unseen invigilated examinations (Erbe, 2007). The use of crib sheets has enhanced performance (Dickson & Bauer, 2008). Most importantly, students who expect to use crib sheets during testing rely on the sheet for information, suggesting that students do not learn the course material as well when they expect to use a crib sheet (Dickson & Bauer, 2008).

An appealing use of open book tests to support development of students' study skills i.e. assessment for learning, particularly those of weak students, has been described by Phillips (2006). In this case, the students' learning is motivated through a high level engagement strategy (Biggs, 1999) which results in students, particularly weak students, showing significant improvements in test scores on consecutive tests. This work has been extended through positive findings of the impact of training in open book test-taking strategies on student test performance in online, timed, open book tests (Rakes, 2008). These findings provide potentially useful strategies for scaffolding of students for assessment of higher level thinking skills in open book exams.

Encouraging academics to change their assessments from closed book, invigilated final examinations to open book, partial open book or crib sheet supported examinations requires an evidence based development of best practice principles for their use. The above discussion indicates that currently a range of views prevail as to the benefits and limitations of such assessments. First, it is important to note that this is before consideration of questions that may be set in such examination papers and the extent to which higher order cognitive skills are required to be demonstrated in order that credit is given for student answers. Second, these views prevail without consideration of an overall academic program assessment framework.

Conclusion

This paper does not argue for replacement of assessment for learning coursework assessments i.e. coursework assessment by the unseen, timed invigilated examination - the use of a variety of assessment modes for assessment in a course is encouraged (Struyven, Dochy, & Janssens, 2008).

Development of evidence-based best practice guidelines for examinations will require certain calculated 'risk taking' behaviours by staff, with respect to setting of examination papers

within a framework of aligned teaching and overall course assessments and subsequent dissemination in peer reviewed literature of the approach and its evaluation. There are numerous websites and books which present examination strategies of individual academics. For example, in a collection of cases from US colleges one case describes students choosing and then taking examinations of their preferred style, either MCQ or essay style (Verosub, 1997). Evidence-based practice requires more than anecdotal descriptions.

This paper argues for development of evidence based, best practice guidelines for the use of unseen invigilated examination type of assessment, particularly for improved construction of this form of assessment. The argument is thus at odds with that which argues that we should design assessment, 'first, to support worthwhile learning and worry about reliability later. Standards will be raised by improving student learning rather than by better measurement of limited learning' (Gibbs & Simpson, 2004-05). Learning for assessment by examination need not be 'limited'.

References

Angelo, T.A. & Cross, K.P. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers*, 2nd edition. San Francisco: Jossey-Bass, (pp. 236-239).

Bers, T. H., & Smith, K. E. (1990). *Assessing assessment programs: The theory and practice of examining reliability and validity of a writing placement test.* Community College Review, 18, 17-28.

Biggs, J. (1999). *Teaching for quality learning at university*: Open University Press/Society for Research in Higher Education, Buckingham.

Black, H. T., & Duhon, D. L. (2003). Evaluating and improving student achievement in business programs: The effective use of standardized assessment tests. *Journal of Education for Business, 79*, 90-98.

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences*, (3rd edn.), National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*, 800-804.

Dickson, K. L., & Bauer, J. J. (2008). Do Students Learn Course Material During Crib Sheet Construction? *Teaching of Psychology, 35*(2), 117 – 120.

Dickson, K. L., & Miller, M. D. (2005). Authorized Crib Cards Do Not Improve Exam Performance. *Teaching of Psychology, 32*(4), 230 – 233.

Erbe, B. (2007). Reducing Test Anxiety While Increasing Learning: The Cheat Sheet. *College Teaching, 55*, 96 – 98.

Felder, R. M. (2002). Designing Tests to Maximise Learning. *Journal of Professional Issues in Engineering, Education & Practice, 128*, 1–3.

FIP. (2000). *FIP statement of policy on good pharmacy education practice*. (Retrieved 25/09/2008) http://www.fip.nl/www/?page=statements

Gibbs, G., & Simpson, C. (2004-05). Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education,* 1, 3–31. (Retrieved 13/06/2008) http://www.glos.ac.uk/departments/clt/lathe/issue1/index.cfm.

Hamed, K. M. (2008). Do You Prefer to Have the Text or a Sheet with Your Physics Exams? *The Physics Teacher, 46*, 290-293.

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal, 16*(2), 207 – 223.

Holsgrove, G., & Elzubeir, M. (1998). Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. *Medical Education.*32, 343-350.

James, R., McInnis, C., & Devlin, M. (2002). *Assessing learning in Australian Universities.* (Retrieved 25/09/2008) http://www.cshe.unimelb.edu.au/assessinglearning/docs/AssessingLearning.pdf

Kehoe, J. (1995). Basic Item Analysis for Multiple-Choice Tests. *Practical Assessment, Research & Evaluation,* 4(10). (Retrieved 30/07/2005) http://PAREonline.net/getvn.asp?v=4&n=10

Kirschenbaum, H. L., Brown, M. E., & Kalis, M. M. (2006). Programmatic Curricular Outcomes Assessment at Colleges and Schools of Pharmacy in the United States and Puerto Rico. *American Journal of Pharmaceutical Education, 70*, Article 08.

Lorusso, G. D. (2004). A Style Guide for Effective and Consistent Formatting of Multiple-Choice Questions. *Pathology Education, Volume 27*, 25-32.

Maclellan, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education, 29*(3), 311 – 321.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*, 709-712.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? *BMC Medical Education, 7*.

Phillips, G. (2006). Using open-book tests to strengthen the study skills of community-college biology students. *Journal of adolescent & adult literacy, 49*, 574-582.

Rakes, G. C. (2008). Open Book Testing in Online Learning Environments. *Journal of Interactive Online Learning, 7*, 1-9.

Ramsden, P. (1992). *Teaching and Learning in Higher Education*. Routledge:London.

Robles, M., & Braathen, S. (2002). Online assessment techniques. *Delta Pi Epsilon Journal, 44*, 39–49.

Schwartz, P. L., & Loten, E. G. (1999). Brief problem-solving questions in medical school examinations: is it necessary for students to explain their answers? *Medical Education, 33*(11), 823-827.

Struyven, K., Dochy, F., & Janssens, S. (2008). The effects of hands-on experience on students' preferences for assessment methods. *Journal of Teacher Education, 59*(1), 69-88.

Stupans, I. (2006). Multiple choice questions: Can they examine application of knowledge? *Pharmacy Education, 6*, 59 - 63.

Taylor, J. A. (2006). Assessment: a tool for development and engagement in the first year of university study. *Paper presented at the 9th Pacific Rim in Higher Education (FYHE) Conference: 'Engaging Students'*. Griffith University in conjunction with Queensland University of Technology, Gold Coast, 12-14 July 2006.

Toledo, C. A. (2006). Does your dog bite? Creating good questions for online discussions. *International Journal of Teaching and Learning in Higher Education, 18*, 150-154.

Verosub, K. L. (1997). Essay exam option with differential grading method for essay exam takers and multiple-choice test takers. In S. Tobias & J. Raphael (Eds.), *The Hidden Curriculum: Faculty-made Tests in Science* (pp. 143-145): Springer.

Wass, V., McGibbon, D., & Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Medical Education, 35*(4), 326-330.

Williams, J. B. (2006). The place of the closed book, invigilated final examination in a knowledge economy. *Educational Media International, 43*, 107–119.

Williams, P. (2008, 1–13, iFirst Article). Assessing context-based learning: not only rigorous but also relevant. *Assessment & Evaluation in Higher Education,* 2008, 2001–2013.