USING COMPUTER AIDED ASSESSMENT TO TEST HIGHER LEVEL LEARNING OUTCOMES

Terry King and
Emma Duke-Williams

Using Computer-Aided Assessment to Test Higher Level Learning Outcomes

Terry King and Emma Duke-Williams

Department of Information Systems
University of Portsmouth
Buckingham Building
Portsmouth PO1 3HE
Email: terry.king@port.ac.uk
Tel: +44 (0) 23 9284 6426

Abstract

This paper sets out an approach using a revised Bloom's taxonomy of learning objectives for the careful design of objective questions to assist in the assessment of higher learning outcomes (HLO's) and details the creation and evaluation of a variety of such questions. This has been done within the context of two constraints; the use of popular, commercially available computer-aided assessment (CAA) software, specifically Question Mark Perception and Half-Baked Hot Potatoes, and the assumption of limited learning technologist support. It examines the problems inherent in the design of objective questions for HLO's (specifically at the levels of Application, Analysis and Evaluation) and introduces a framework by which systematic design may be carried out. It examines the mode of design, construction and evaluation of 22 such objective questions devised for formative assessment for two post-graduate course units in Information Systems, involving two groups of students, 37 in total. The results from the trialing of these questions raises issues of crucial differences in CAA software for feedback, scoring and delivery. The paper examines key statistical indicators of question quality (facility and discrimination indices), as well as the results of interviews with students, to drawn conclusions about the best use of question types, student preparation for tests, and discusses key issues in question preparation. It concludes with an examination of the main advantages and disadvantages of using CAA for HLO's referring to the resource overheads needed and the problems inherent in the process.

Keywords

Computer-aided Assessment, Bloom's Taxonomy, Learning Objectives, Objective Testing, Computer-based Testing

Background

Students in higher education engage in a variety of learning activities with the aim of attaining certain defined learning outcomes. As students progress to Level 3 and

post-graduate work, it is accepted that they engage increasingly in activities designed to develop skills and abilities which are considered to be of a higher cognitive complexity (Zakrzewski and Steven, 2001). Holzl (2000) places emphasis on the development of certain graduate attributes in the cognitive domain such as critical thinking and making informed judgements; in-depth knowledge; information management and the capacity to analyse and organise; interdisciplinary perspectives; and, problem solving with the requirement to evaluate and create. Development of these attributes requires both appropriate learning activities and 'fit-for-purpose' assessment. Certain needs therefore arise within the higher education system. Firstly to promote formative self-assessment of such abilities and, related to that, to find a way to practice such high level activities within the context of the relevant knowledge domain. By engaging in such activities and receiving appropriate feedback, learning can occur and students progress. However such feedback historically has been given face-to-face between tutor and student or student group; a situation which current resourcing of higher education with increased student numbers makes increasingly rare. A further need is the measurement of the level of attainment of the students in acquiring such abilities in the form of summative assessment. Traditionally this has been met though paper-based written examinations with their attendant problems of resource intensive marking, subjectivity, bias etc. It is therefore not surprising that the spotlight should be placed on objective testing to see whether this mode of assessment can be employed in these areas associated with more complex cognitive abilities. And, in addition, by employing the use of computer-aided assessment (CAA) and more interactive question types, not only to assist the process of assessment, but to actually enhance it. This is an interesting area and one perpetually challenged on quality grounds by both academics and external examiners alike who judge CAA by simple issues like the number of questions in a test, the number of factual or comprehension questions, and the appropriateness of CAA at Level 3 or higher on any basis whatsoever. As McKenna and Bull (2000) point out CAA has the "need to satisfy its critics of its pedagogical fitness-for-purpose".

This paper sets out an approach using a revised Bloom's taxonomy of learning objectives for the careful design of objective questions to assist in the assessment of higher learning outcomes, and details the creation and evaluation of a variety of such questions. This has been done within the context of two constraints; the use of popular, commercially available CAA software and the limited availability of learning technician/technologist support. It assumes an environment in which the authors and many other lecturers find themselves, where they need to engage deeply in the process of question creation themselves and any software application will need to be easy to use and require no programming skills.

Classification of Learning Outcomes

Bloom's taxonomy of learning objectives has been chosen as the framework for approaching the problem of assessing for higher learning outcomes (Bloom et al, 1956). The reason for this follows Delgano (1998) who considered other schemes for classifying learning outcomes but came down in favour of Bloom's taxonomy because it had sufficiently detailed categories to allow outcomes to be mapped clearly onto learning activities, and was in widespread use so designers did not have

to learn an additional scheme. This latter point was felt to be particularly important at the University of Portsmouth where assessment strategies have been linked completely with Bloom's taxonomy and considerable efforts have been made to familiarise lecturers with Bloom's categories when writing learning objectives.

Bloom's taxonomy of learning objectives (Bloom et al, 1956) attempted to classify forms of learning into three categories: cognitive, affective and psychomotor domains. Within the cognitive domain, Bloom identified six levels of learning which represented increasing levels of cognitive complexity from the lowest level of Knowledge (or remembering) through Comprehension, Application, Analysis, Synthesis and Evaluation. Each level encompassed those below it, so, for example, analysis could only occur after the ability to apply understanding of factual or other knowledge had been accomplished. The three lowest levels have been described as 'foundation thinking' which are used as a basis for the higher learning levels **(**Ryan and Frangenheim, 2000). Associated with each level were certain learning outcomes expressed as 'verbs' such as recall, draw, calculate, categorise, design, or assess. The demonstration of higher learning outcomes would be a reflection of the attainment of learning at more cognitively complex levels. It is often assumed that objective testing with its need to provide a correct answer is only applicable to the lowest learning levels. While this has never been true, with advances in CAA, the applicability of objective testing to the three highest levels of Analysis, Synthesis and Evaluation can now be more appropriately considered.

Objective Question Design for Higher Learning Outcomes

Advice is available for the design of effective objective tests by providing general guidelines for question design, and employing grids and matrices to plot content against learning levels and outcomes (Heard et al,1997; Rolls and Watts, 1998). However much less direction is given on how to design the assessment questions themselves. Because of this, the design of objective questions to test higher learning outcomes (HLO's) often follows one of three approaches:

Derivation from the verbs associated with HLO's. This can misleading. For example, any question with the verb 'judge' is assumed to be an Evaluation level question, but in fact students are asked to judge on a variety of criteria, and, if the criteria is understanding of a theory, then the question will be at a Comprehension level. Often such criteria are not made explicit to students so that they are left coping with ambiguity

Extrapolation from existing subjective examination questions. This can result in objective questions which are unclear as to which learning level they apply. This is hardly surprising as examination questions are generally not subjected to the rigorous examination applied to objective questions and exactly what they assess can be open to debate.

Use of exemplars. For example, those described by Mackenzie (1999), Carneson et al. (no date), and Heard et al.(1997) are effective, but are limited because, while offering a template for generating future objective questions of those types, they do

not offer the lecturer a formal approach to creating and designing their own new question formats.

None of these are particularly successful. What is needed is a systematic framework for positioning questions for particular learning outcomes.

A framework which offers many possibilities is the revision to Bloom's taxonomy by Anderson and Krathwohl (2001) and co-workers which results in the basic table shown in Figure 1. The six levels remain but each has been replaced by its matching verb – Remember, Understand, Apply, Analyse, Evaluate and Create – in order to facilitate the writing of learning objectives. Create is now the last and highest level of learning as they consider that evaluation is a necessary step which precedes any generative process.

The three higher learning levels which were initially of main interest for constructing CAA questions were subdivided in detail:

Analyse - encompassing differentiating or distinguishing, organising or structuring, and deconstructing (which concerns determining the values underlying presented material).

Evaluate - which breaks down into the two processes of checking for internal consistency, and critiquing which involves judging against external criteria.

Create - which involves generative processes such as hypothesizing, planning, designing, and producing or constructing.

The Knowledge Dimension		**Remember**	**Understand**	**Apply**	**Analyse**	**Evaluate**	**Create**
Factual Knowledge	Terminology						
	Specific Details				A		
Conceptual Knowledge	Categories						
	Principles		A		A, A	A, A, A	
	Theories			B	B, B	A, B	
Procedural Knowledge	Skills/ algorithms			A, A, C			
	Techniques/ methods		B	A	A, B		
	Criteria				A	C, C	
Metacognitive Knowledge							

Figure 1: Modified basic table of Bloom's learning objectives (Anderson and Krathwohl, 2001) showing the distribution of objective questions with HLO's used in the paper.

The table in Figure 1 has been extended by the addition of a knowledge dimension to help 'educators distinguish more closely what they teach' and by implication what they are assessing. This dimension is detailed in the columns on the left of Figure 1, with the different forms of knowledge covering:

Factual Knowledge. The basic details of the content of a course which students must know to make sense of the discipline, divided into knowledge of terminology or specific details and elements. Itemised knowledge before interrelationships are considered here.

Conceptual Knowledge. Encompassing the knowledge of classifications and categories, principles and generalisations, theories, models and structures.

Procedural Knowledge. This considers the knowledge of subject-specific skills, algorithms, techniques and methods, and also the knowledge of the criteria used in determining when to use specific procedures. It is the knowledge of 'how to do something'.

Metacognitive Knowledge. That class of knowledge by which students know how they come to know and learn. It includes the conscious application of cognitive strategies by students and their own self-knowledge of learning strengths and styles.

In using this framework for the construction of objective questions to test HLO's, certain areas of the table in Figure 1 have been excluded, viz. metacognitive knowledge which is not subject-specific, and the two lowest learning levels of learning which are not applicable to this investigation.

Also it was the original intention to include the highest level of learning 'Create' within the study. Anderson and Krathwohl (2001) are quite specific in the activities which are included within this level of learning, viz. generating, planning and producing. Each of the many results of such activities ie. alternative hypotheses, research plans, designed procedures, an invention or a construction, are by their very nature both unique and *equally valid*. There must be many 'correct' answers, which takes this level outside of the remit of objective testing which needs one (and only one) totally valid response, and makes automated marking within CAA possible. Attempts were made to duplicate and explore generative activity by asking students to construct a diagram by dragging optional markers onto a template (Figure 2) or to improve a poorly constructed layout (Figure 3) from a incorrect example given in the question.

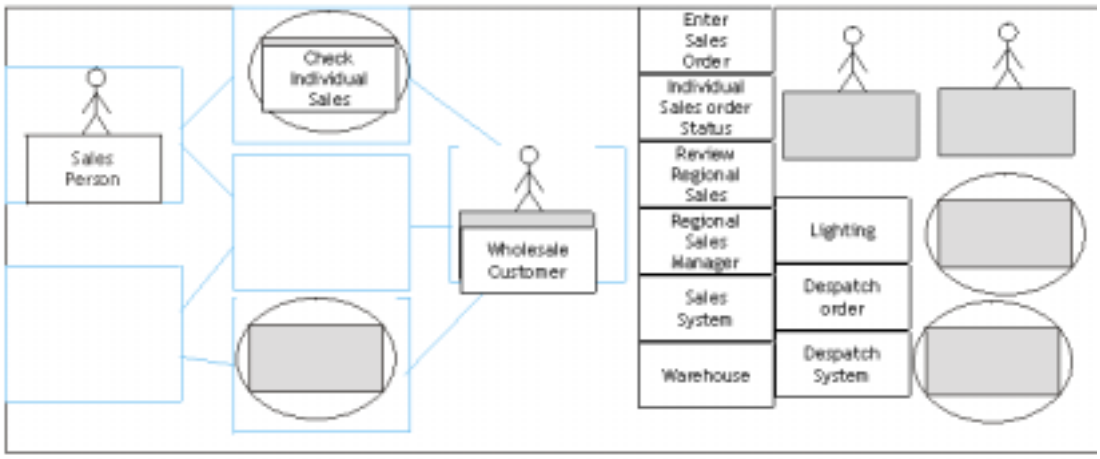

Figure 2: The student is required to drag symbols into specified locations to 'construct' a working diagram for a given problem.

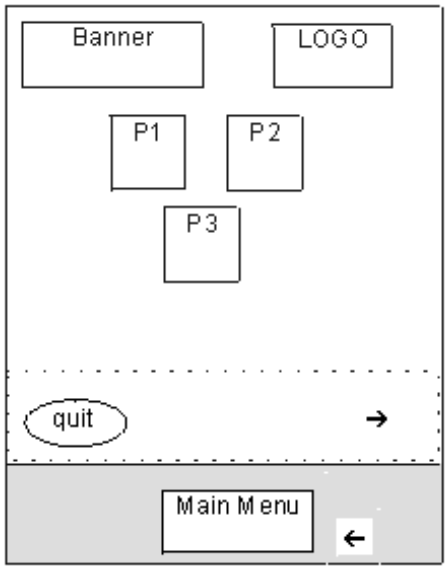

Figure 3: The student is asked to correct an incorrect layout according to specific criteria using drag-and-drop.

However further consideration of these questions against the framework lead to the conclusion that they were not at the highest learning level of *create* but essentially an *analysis* of material with a view to determining how elements are organised or (re)structured. In view of this it was decided to omit further consideration of the *create* learning outcome and to aim questions largely at the levels of *analyse* and *evaluate*, with some more exploration of *apply* level questions. The latter have the subcategories of either execution of a familiar task or implementing by applying a procedure to an unfamiliar task.

22 objective questions for assessing HLO's were constructed for implementation in two commercially available CAA software packages, *Question Mark Perception* and *Half-Baked Hot Potatoes.* The questions were devised for formative assessment for two course units on the MSc in Information Systems, one for basic multimedia

theory, and the other for educational theory underpinning the development of computer-aided learning. The questions fell into three categories:

A. 13 questions devised solely using the revised Bloom's framework.
B. 6 questions devised by using exemplars.
C. 3 questions adapted directly from past exam papers in multimedia theory.

Questions from *B* and *C* were later analysed according to the framework and the distribution of all the questions by learning level and knowledge type can be found in Figure 1.

Software Evaluation

The questions created were trialled in two software packages, *Question Mark Perception* and *Half Baked Hot Potatoes*, to evaluate the facilities offered by both, although only the Perception question set was available for student use. Both offer delivery of questions across the Web, but while *Perception* is aimed at large scale delivery and has many security features, Hot Potatoes has been developed primarily for language learning and comprises six programs one for each question type. With *Hot Potatoes* individual questions can be authored and included in web educational material. Significant differences between the packages include:

Feedback and Scoring. As *Perception* can be used for summative testing, the feedback is available after the student has indicated that they have “Finished” the question/ test. The feedback can relate to the answers given – this feature was used in the sample questions as they were primarily formative to assist with revision. *Perception* also allows great flexibility in the scoring. Thus particular scores can be assigned to best, reasonable and poor answers – which can include negative marking. With *Hot Potatoes* feedback varies between question types. Multiple choice questions give instant feedback – this can be altered for each choice. The other question types give a score and, in some cases, reset the incorrect answers, but leave the correct ones in place. The scoring in *Hot Potatoes* is set, and is calculated as a percentage. However, if a student has retried a question, the score will reflect the number of tried that a student made (and, if appropriate the number of hints given, though this feature was not used for these questions). The sophisticated features for giving feedback could be used to good effect for formative assessment. Scores however cannot be brought forward with *Hot Potatoes*.
Presentation. Both packages allow the addition of extra HTML material – as seen in several of the questions where external links to other sites were provided. While this could have been hand coded, it was easier to use an HTML editor (like *Macromedia Dreamweaver*), and then cut-and-paste the HTML code where necessary. *Hot Potatoes* had the option to add a “reading text” – which could contain further information.
Delivery. Both packages allow the creation of material for use on the Web. The output from *Perception* can be used with browsers from v3; however, there is no support for drag and drop (Hot Spot) questions in v3 browsers. *Hot Potatoes* allows the creation of several different versions (v3, v4, DHTML). If saved as v3, the drag and drop type questions used in JMatch, are saved with drop down lists. As both

packages create JavaScript based material this poses a potential significant difficulty to users of screen reading technology.
Overall, for formative testing both packages could be used. *Perception* clearly has greater flexibility, but there is sufficient scope with *Hot Potatoes* to allow creative users to create appropriate questions. However in order to use both effectively though a good knowledge of web authoring and graphic creation is very useful.

Evaluation of the Question Set

The questions were posted to the *Perception* web-server and made available over a 3 week period in two tests, to two groups of students: 19 educational technology students and 18 multimedia students. The latter used the tests as revision for their written exam. Students were asked to comment on questions individually and in summary. In addition, 6 multimedia students were invited to discuss the questions. The questions were also analysed using *Perception Reporter* to provide a facility and discrimination index result for each question.

Of the 37 students in total, results were obtained from 19, but 5 of these took the test more than once, giving 31 test answer sets in all. However only about one-third of test attempts were fully complete. Allowing for the split of students in two groups, the volume of data collected for each question was small, but some tentative conclusions can be drawn from the results obtained:

Facility (degree of difficulty). There were no trends by question category regarding the degree of difficulty of the questions, but students found questions increasingly hard from Factual Knowledge, through Procedural Knowledge to Conceptual Knowledge (the mean of the Facility Indexes were 0.55, 0.51, 0.37 respectively). Question type/ style did make a difference however. There were two questions where multiple selection required ranking of the responses. Where the options choices were textually long, then a drag-and-drop approach to creating the ranked list made the question easier than using drop-down lists. Students generally found the 'assertion-reason', 'observation-conclusion' style of question very difficult. On our limited sample these did not discriminate well between students and it is suggested that they are used cautiously and students are given practice in using them.

Discrimination (or Correlation). The questions did not discriminate well with a discrimination index mean = 0.34 and a mode =0.40. However this masked a very wide range and could be due to the nature of the more complex question styles especially for the education questions which used web pages and other resources as part of the question. It is possible that some students just guessed the answer rather than spend the time doing the question carefully. If such questions are to be used for formative assessment then students might need to be prepared in advance that such questions will take time, and perhaps they should be used either individually or in very small sets. Category A questions produced a less wide variation in discrimination values clustered around a mode figure of 0.40. This may suggest that designing questions systematically according to a framework may produce more consistency.

Student Opinions. 6 students were interviewed about the 14 revision multimedia questions. Of these they found 10 really useful, prompting reading or research; 2 were mixed, and only 2 not useful and so difficult that they simply gave up and guessed. The two poor questions although designed using the framework were based on ideas that had been used in examination questions, and both of them on inspection seemed to have retained rather mixed learning outcomes. The problem with using examination questions is that they can be generally worded (leaving space for students to justify their answers) or have space for more detailed description so that the criteria being used is more explicit. When converted for CAA neither of these are true so students are left with half the story which they find frustrating and generates ambiguities. There were many complaints about wording needing to be more precise and also a demand for better, more diagnostic feedback. More care certainly needs to be taken in the latter case if students are using CAA questions for formative assessment.

Question Preparation. When the project started it was the intention to write a question for each location in the table in Figure 1. However as the work progressed it became apparent that the nature of the knowledge being tested determined the level of learning outcomes which were relevant. As the material being used moved from Factual to Procedural areas, application and analysis questions became a natural mode of assessment, while for Conceptual knowledge analysis and evaluation seemed most appropriate. This 'drift' in the relationship between learning outcomes and the knowledge categories can be seen in Figure 4. The reverse was also true, in that it was found to be extremely difficult to write objective questions with HLO's for areas of Factual Knowledge. This may have implications for the range of assessment questions for HLO's which can be constructed for Level 1 undergraduate courses where question sets are open to criticism for containing too many low level learning outcome questions. Clearly the category of knowledge used in the course materials does have a bearing on what is possible in a CAA test, and the kind of questions which can be developed may be a measure of the course curriculum and mode of delivery.

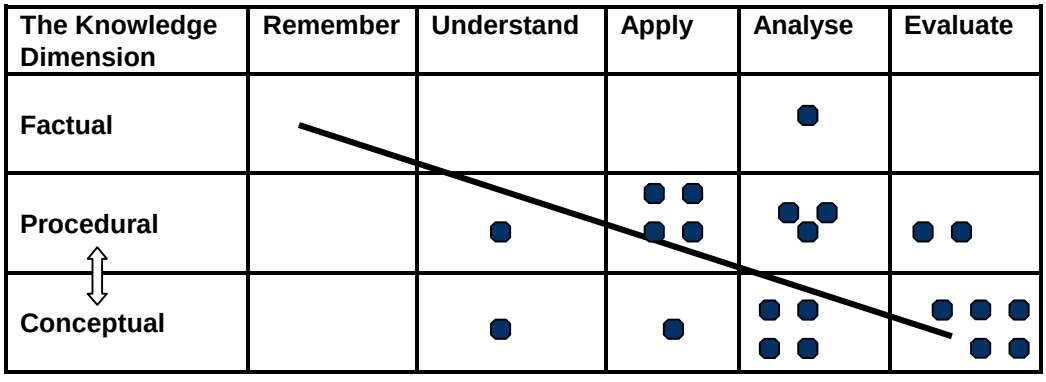

Figure 4: Apparent drift in the relationship between learning outcomes and knowledge categories when using the framework to devise objective questions to assess HLO's. (NB. The order of Procedural and Conceptual knowledge categories have been reversed).

As a final point, designing questions for HLO's is extremely time consuming, requiring 30-60 minutes for each question, depending on the complexity of the question type and whether other resources are to be used during the running of the question. Developing these questions can also be extremely difficult. Figure 1 reveals two questions at the *understand* level which were originally planned as *evaluate* questions but on re-analysis proved to be assessing a much lower level of learning. It became clear that training and experience is needed, which must be gained over time, to set such questions effectively. Just doing these questions intuitively is unlikely to be successful.

Conclusions

The results shown above reveal a number of issues about designing and constructing effective objective questions for HLO's to be delivered through CAA software. There is a considerable overhead in terms of training, time and expertise (both technical and pedagogical) if these questions are to be effectively and successfully developed by lecturers. It should also be noted that to use the features of even commercial packages to good effect some knowledge of scripting is extremely useful. However the student response was essentially very positive and clearly these questions have a place for formative assessment, albeit with the caveat that careful monitoring and evaluation of each question would be needed. With half of the questions in our sample comprising simple MCQ's, it is clear that CAA is not needed to deliver objective questions for higher learning outcomes. However CAA does offer enhanced features, for example, drag-and-drop, web links, systematic and selective feedback, and running additional software, as well as more sophisticated question types, such as multiple response and selection. The delivery of questions through CAA can also be cognitively advantageous. It was the original intention to use the experience gained on this study to alter the mode of the summative assessment for the post-graduate multimedia unit away from a paper-based examination to CAA. However the results, especially the relatively poor discrimination data, suggests that much trialling of questions would be needed before this could become a reality.

References

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

Carneson, J., Delpierre, G. and Masters, K. (No date) *Designing and Managing Multiple Choice Questions: Appendix C: MCQ's and Bloom's Taxonomy* <http://www.uct.ac.za/projects/cbe/mcqman/mcqappc.html>(9 Feb 2000).

Dalgarno, B. (1998) *Choosing learner activities for specific learning outcomes: A tool for constructivist computer assisted learning design.* In C. McBeath and R. Atkinson (eds), *Planning for Progress, Partnership and Profit*. Proceedings EdTech'98. Perth: Australian Society for Educational Technology. <http://cleo.murdoch.edu.au/gen/aset/confs/edtech98/pubs/articles/abcd/dalgarno.html> (25 April 2001).

Heard, S., Nicol, J. and Heath, S. (1997). *Setting effective objective tests*. MERTaL Publications, University of Aberdeen.

Holzl, A. (2000) *How do we Assess Graduate Attributes?* Paper presented at the TEDI Teaching and Learning Conference, *Effective Teaching and Learning at the University*, University of Queensland. November. <http://www.tedi.uq.edu.au/conferences/teach_conference00/papers/holzl-2.html>(24 April 2001).

Mackenzie, D. (1999) *Recent developments in the Tripartite Interactive Assessment Delivery System (TRIADS)*, in Danson, M. and Sherratt, R. (Eds), *Proceedings of the 3rd Annual CAA Conference*, Loughborough, 235-250.

McKenna, C. and Bull, J. (2000) *Quality assurance of computer-assisted assessment: practical and strategic issues.* Quality Assurance in Education, **8** (1) 24-31.

Rolls, D. and Watts, S. (1998) *Objective Question Design*. Publ.Computer-Assisted Assessment, Evaluation and Survey Service (CAAESS), Kingston University. < s.watts@kingston.ac.uk>

Ryan, T. and Frangenheim, E. (2000) *How we see Bloom's Taxonomy*, The E-Learning Lesson Planner< http://www.e-learning.com.au/portal/blooms2.asp>(19 January 2001).

Zakrzewski, S. and Steven, C. (2001) *A Model for Computer-based Assessment: the catherine wheel principle*, Assessment & Evaluation in Higher Education, **25**, 2, 201-215.

Question Mark Perception <http://www.qmark.com/>
Half-Baked Hot Potatoes <http://web.uvic.ca/hrd/halfbaked>