Making sense of Learning Management System's quiz analytics in understanding students' learning difficulties

Antonette Mendoza, Harald Søndergaard and Anne Venables
School of Computing and Information Systems, The University of Melbourne, Vic. 3010, Australia
Corresponding Author Email: mendozaa@unimelb.edu.au

CONTEXT

In engineering and other tertiary education programs, it is generally held that formative assessment can be a strong driver of active participation and learning. Two years ago we added a new assessment component in our algorithms subject, namely mandatory weekly quizzes. The aim was to encourage students to stay abreast with lectures and tutorials, and to identify common misconceptions in a subject that deals with numerous difficult algorithmic concepts. It was hoped that the quizzes would expose and challenge student misconceptions, and more generally support the mastery of important algorithmic tools and techniques. The quizzes are hosted within our institution's Learning Management System (LMS) which also provides learning analytics tools, including metrics for student engagement and performance.

PURPOSE

After rolling out the student quizzes we turned to evaluation. Apart from measuring the extent of student engagement, we were interested in the (initially vaguely defined) "learning value" of individual quiz questions. Through the LMS metrics, we wished to evaluate the quizzes' influence on student engagement with subject materials, and to discover which questions best expose common misconceptions harboured by students.

APPROACH

We have adopted an action research approach. LMS statistics were gathered and analysed in three successive semesters. For each weekly quiz, the data collected included students' participation statistics, records of multiple attempts made by every student, and LMS measures of question discrimination and difficulty. These data were analysed across the three iterations.

RESULTS

There were marked differences between the LMS statistics of the pilot and two following semesters. Unsurprisingly, student participation grew when quizzes were made mandatory. More significantly, the LMS inbuilt measures of question difficulty and discrimination were found to be extremely susceptible to the number of allowed quiz attempts allowed. We have found them relatively unreliable and unhelpful in identifying useful questions that challenge student misconceptions and we have had to find alternative metrics.

CONCLUSIONS

Feedback from our students strongly suggests that mandatory weekly quizzes do promote student engagement with learning materials. However, in our attempts to gauge question quality, we find that the LMS metrics for question analysis do not support a "learning value" assessment. They are so strongly influenced by the number of possible quiz attempts that they are of little practical use. This study therefore illustrates that using the LMS metrics may not be sufficient or straightforward to assess the effectiveness of quizzes on student learning outcomes. Further analytics and investigations will need to be conducted for a deeper understanding for best exposing common misconceptions harboured by students.

KEYWORDS

Quiz questions, LMS metrics, formative assessment

Introduction

Active learning is increasingly promoted within higher education institutions to support students in linking knowledge to meaning and the development of higher order thinking skills. Active learning involves: students engaged in more than listening; less emphasis being placed on transmitting information and more on developing students' skills; higher order thinking and engagement (Bonwell and Eison, 1991). However, encouraging active learning can be a challenge for both educators and students, particularly in large, lecture-based classes (Klein, 2003, Buckley et al. 2004). Much has been written about the use of assessment, particularly formative assessment, to drive active learning in higher education programs (Boud, 2010; Gibbs, 2010; Boud and Falchikov, 2007; Falchikov, 2005; Huba and Freed, 2000).

Formative assessment comes in many different shapes, encompassing a variety of practices, including self-assessments and peer-assessments; Black and Williams (1998) review no less than 250 articles on the topic. Often characterized by its informal techniques and mechanisms to encourage student participation, it is not a pre-condition that formative assessment be tied to summative assessments, although this is often the case (Dunn and Mulvenon, 2009). Gibbs and Simpson (2004) suggest that formative assessments in the form of frequent assignments or tests to distribute student effort across the duration of the semester, often on a weekly basis, promote students' participation and enhance learning. More specifically, several studies have shown that there is a high level of student engagement with regular quizzes; many report upon their positive role in encouraging, for example, the completion of prescribed reading in various programs (Scheyvens et al., 2008; Bonwell and Eison, 1991; Hanson and Moser 2003).

In the computing disciplines, many students struggle with complex algorithmic concepts. To promote students' engagement with teaching materials and engender understanding of important computational methods and theories, we designed a set of weekly quizzes. Each quiz comprised a variety of questions; some were crafted as revision materials, others were set to probe students' learning and challenge their perceptions and interpretations. Students could attempt each quiz multiple times and receive information about which questions they had answered correctly, including some hints for questions they had answered incorrectly. Their reflections on this formative feedback were expected to influence their follow up attempts.

Here we report on our efforts to evaluate the introduction of these quizzes and our attempts to gauge the usefulness of individual quiz questions in challenging common misconceptions and in helping students learn content and concepts.

Context and purpose

The introduction of quizzes, as a formative assessment component, took place within a graduate subject on Algorithms and Complexity. The aim of the subject is to develop student familiarity and competence in assessing and designing software for computational efficiency. Historically, many students struggle with the concepts and topics of this subject. Introduction of quizzes was a mechanism to offer more opportunity for students to be engaged with the subject content and, importantly, to challenge developing mental models of computational processes. Eleven quizzes were devised and set up; one for each week starting in the second week of a 12-week semester. On-line quizzes have been found to be an effective mechanism for incentivizing student completion of work and are relatively time efficient from the perspective of the educator (Wolt and Mason, 2003). As most virtual learning environments provide quiz frameworks for local customization, we hosted these weekly quizzes using the Learning Management System (LMS) of our institution. The LMS offers analytics tools to track student engagement with the quizzes along with some automated tools for question analyses.

Our plan was to use these LMS data to gauge the effectiveness of the quizzes in engaging students with teaching materials and, for individual questions, to assess "learning value", including a question's ability to pinpoint misconceptions. To this end, an action research approach was to be adopted. In the first instance, the quizzes would be piloted; then

adjustments to delivery would be made over the course of two subsequent iterations, whilst maintaining the same pool of questions over this investigation.

We note that the quizzes complement lectures, tutorials, and other continuous-assessment components in the subject. In particular, two written assignments challenge students to find good algorithmic solutions to relatively difficult problems. The main aim of the quiz component has been to encourage students to stay abreast during semester, and while individual questions are designed to identify misunderstandings, they are not intended to be difficult.

Approach

LMS data were collected and investigated for every quiz across the three semester iterations of their use. Two separate sets of LMS data were of interest. Firstly, LMS reports of students' participation activities were used as a measure engagement with the quizzes. Secondly, through the LMS metrics of discrimination and difficulty, we hoped to identify the most useful quiz questions for best exposing common misconceptions harboured by students.

Results and evaluation

Conveniently, the LMS allows teaching staff access to student participation statistics for the entire cohort. When a student logs in to access a weekly quiz in Algorithms and Complexity, an attempt is recorded, regardless of whether the attempt is complete or not. For each attempt, the LMS records whether each question is answered correctly or otherwise, and assigns a nominal score as decided by the teaching staff.

As a pilot, the weekly quizzes were trialled in semester 2, 2015 and students were permitted to make up to three separate attempts per quiz. Participation was voluntary, in that there was 'no' mark attached to quiz participation or to the number of answers correct in each quiz. Throughout the semester, teaching staff actively promoted the benefits of ongoing quiz involvement to their students. Of 151 students who completed the subject in the pilot semester, 121 students made a quiz attempt in week 2. Participation statistics are given in Table 1, where a decline is observed over the pilot semester culminating in 66 students attempting the final week 12 quiz, even though this quiz was timed nearest to the examination sitting.

The role of the pilot semester was to fine-tune the delivery of the quizzes before they would become mandatory. The quiz questions were unchanged throughout, but we needed to decide on the best parameters for delivery, including scrambling of questions, windows-of-access, and number of attempts allowed. As pointed out by Gibbs and Simpson (2004), the relationship between marks and effort is not straightforward for students, and as little as 5% of student time may be allocated un-assessed tasks. If too few marks (in this case, 0 marks) are allocated to preparatory work, many students may make the strategic decision to forego those marks and instead focus their time on other pieces of assessment. Thus, care is needed in the design of incentive mechanisms to ensure students balance extrinsic rewards or sanctions with intrinsic motivations to maximize their outcomes.

In the following two iterations, the quizzes were simply a hurdle requirement for the subject. More precisely, a student must successfully complete at least 8 of the 11 online quizzes to be eligible to sit the final examination. No mark was attached to quiz participation or to the number of answers correct in each quiz for these following two semester iterations.

Since the quizzes are intended as learning support rather than gate-keeping, we decided that "successful completion" would mean "getting each of the (3-5) quiz questions right in a single attempt", but also that an unbounded number of quiz attempts would be allowed, as long as the student met the weekly deadline. The questions are mixtures of multiple-choice, multiple-answer, matching, and numeric answer questions, so in general a large number of attempts are needed if a student decides to search exhaustively for the right combination of answers. In one instance, as student had 49 attempts at one quiz!

Table 1 shows how participation rates improved in these semesters, compared to the pilot; for the commencement quiz in week 2,186 students participated in semester 1, 2016 and 175 students in semester 2. Notably, over both semesters, there are only slight declines with week 11 quiz participation recorded as 155 and 126 (semester 1 and 2). This improvement in participation is attributed primarily to the introduction of the hurdle requirement for quizzes, where eligibility to attempt the final assessment was tied to completed quiz attempts.

Table 1: Summary of student participation statistics with weekly quizzes over three iterations

Week	Pilot Semester 2, 2015	Semester 1, 2016	Semester 2, 2016
2	121	186	175
3	109	188	175
4	103	187	170
5	90	179	169
6	72	171	168
7	77	166	164
8	69	163	158
9	63	182	157
10	72	176	148
11	69	155	126
12	66	not delivered	127

LMS Analytics

For each quiz, the LMS's "Item Analysis" tool reports overall attempt statistics for that quiz and its component questions, by relating each cohort's performance on each question compared to their overall performances on the quiz hosting that question. The tool uses two metrics to assess each question: discrimination and difficulty. The following descriptions of each are verbatim from the LMS User Guide: *Discrimination* indicates how well a question differentiates between students who know the subject matter and those who do not. A question is a good discriminator when students who answer the question correctly also do well on the test. Discrimination scores range between -1 and +1. Any question that gets a discrimination score above +0.3 is considered Good. Good and Fair questions may be used to help determine student knowledge levels. Discrimination cannot be calculated for questions where everyone receives the same score (everyone gets a question right or wrong). *Difficulty* shows the percentage of students who answered the question correctly. If >80% of students get a question right it is listed as easy; if <30% of students get a question right it is listed as 'hard'.

The discrimination measure categorizes questions as being 'good', 'fair' or 'poor' and the difficulty measure classifies questions as being 'easy', 'medium' or 'hard'. 'Difficulty' as defined by the LMS does not relate to any learning taxonomies, such as Bloom's, SOLO or Neo-Piaget that have been used to categorize questions as to the degree of learning difficulty as described in the computing education literature (Gluga, Kay and Lister et al., 2013; Jimoyiannis, 2011).

Our main interest was in finding quiz questions that best exposed student misconceptions. We expected that the difficulty statistic reported by the LMS would hint at questions that students found most problematic and could lead to misconceptions.

Pilot

During the pilot in semester 2, 2015, Item Analysis Reports were run for each quiz. Figure 1 shows an example, the Week 10 Quiz Report. The Test Summary shows that: (1) 72 student attempts across four questions; (2) All four questions are 'good' for their discrimination scores, each being above +0.3; (3) Two easy difficulty questions where >80% of the students scored

it correct; and (4) Two ‘medium’ difficulty questions where >= 30% and <=80% of the cohort scored the questions correct.

The second section of a Weekly Item Analysis Report lists all questions, each with their description, question type, number of graded attempts, together with discrimination, difficulty, average score, standard deviation and standard error statistics. In addition, the reports visually indicate possible issues with individual questions using a set of symbols beside the question description. The legend for these question classification symbols is shown in Figure 2. In Figure 1, the yellow triangle symbol indicating that the question may have changed since quiz deployment is found beside all questions, and the ‘Linear probing’ and ‘Coin-row instance’ questions are indicated with a red dot signifying LMS recommendations for these questions to be reviewed, most likely for their high difficulty score, that is, ‘easy’ classification.

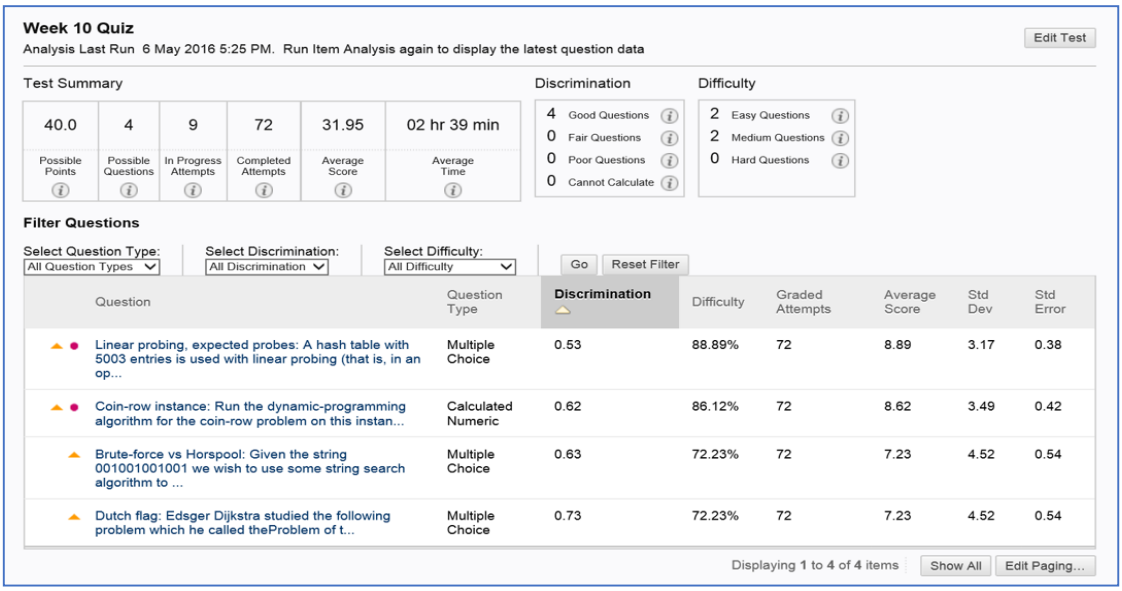

Figure 1: Summary statistics of week 10 Quiz in COMP90038 Algorithms and Complexity, semester 2, 2015 as reported in the pilot Item Analysis Reports

Figure 2: Legend used to classify questions in the Weekly Item Analysis Reports

For an overview of the quizzes and their questions, a collation of the Weekly Item Analysis Reports for the pilot semester is presented in Table 2. The Table lists the number of questions in each quiz, the discrimination classifications and difficulty classification of the quiz questions. In summary, it shows that: (1) All quiz questions have been classified as ‘good’ as their discrimination scores are above +0.3, indicating that students who answered the questions correctly also did well in their respective quiz; (2) All questions have been classified as ‘easy’ or ‘medium’ difficulty, where ‘easy’ questions saw over 80% of the students answered the questions correctly whereas for ‘medium’ difficulty questions, between 30% and 80% cohort answered correctly; and (3) Except for week 8, each LMS Item Analysis report tagged one or more questions as recommended for review; the tool advising that these questions should be more closely examined to assess their suitability in future iterations of the quizzes. Closer examination of how the LMS Analysis tool classifies difficulty shows that the questions ‘Recommended for Review’ are those classified as ‘easy’ in difficulty. These questions are shaded in Table 2.

Table 2: Summary statistics of the weekly Item analysis report for the pilot during semester 2, 2015. Shaded questions are those identified by the LMS as recommended for review.

Quiz	Questions					
	No.	Discrimination	Difficulty			All quiz questions with those 'Recommended for Review' shaded
			Easy	Medium	Hard	
2	5	5 Good	1	4	0	• Sorting time • Ranking functions by growth order • Big-Oh expressions • Sums and Theta • Big-Theta expressions
3	5	5 Good	3	2	0	• Assignment problem • Page number digits • Tower of Hanoi • Big theta for mixed iteration/recursion • Brute force string search
4	4	4 Good	1	3	0	• Find the non-dag • BFS_equals_DFS • Complexity (Theta) again • Topological sequence
5	5	5 Good	3	2	0	• Shellsort • Binary search • Insertion sort • Selection sort • Interpolation search
6	4	4 Good	4	0	0	• Hoare partitioning • Inorder traversal • Master theorem • Non-Master theorem
7	4	4 Good	2	2	0	• Bottom-up heap construction • Valid heaps • Nodes in complete tree • Pre/inorder sequences
8	4	4 Good	0	4	0	• Counting BSTs • AVL trees • BST-insertions • AVL tree traversals
9	4	4 Good	3	1	0	• 2-3 trees • Max-heap plus AVL • AVL shape • 2-3 shape
10	4	4 Good	2	2	0	• Linear probing • Coin-row instance • Brute-force vs Horspool • Dutch flag
11	4	4 Good	3	1	0	• Knapsack • Cost of minimum spanning tree • Edges in minimum spanning tree • Number of different spanning trees
12	3	3 Good	1	2	0	• Huffman AGCT • Dijkstra • Huffman codes

Following two iterations

For confirmation of quiz questions in need of attention, weekly Item Analysis reports were run for the same quizzes in the next two semesters. It was expected the question discrimination and difficulty scores for these reports would be like those of the corresponding weekly reports in the pilot semester. This was indeed the case for discrimination scores, in that students who answered the questions correctly did well in the quiz overall. There was a major difference in quiz question difficulty scores between the pilot and the following two semesters. In the pilot, some question difficulties were decided as 'easy' and others as 'medium', whereas in the following two semesters all questions have been classified as 'easy' difficulty, meaning over 80% of the cohort answered all questions correctly. Further, every quiz question was tagged by the software as 'recommended for review' due to its 'easy' classifications.

Why should the same question be classified with different 'difficulties' over various iterations? It was hypothesized that the disparity between iterations was an artefact of making quiz participation a hurdle requirement in the subject during the second and third iterations in 2016. Associated with the hurdle requirement came the opportunity for students to make an unlimited number of quiz attempts and, it seemed that students generally attempted the quiz as many times as they liked until they succeeded in getting all questions in a quiz correct.

This hypothesis was investigated by downloading all attempt statistics for all quizzes, for each iteration of the project. For each quiz question, students' attempt statistics were sorted and graphed. To make a 'like for like' comparison of question difficulties between the pilot and later iterations, the attempt statistics for those who answered correctly on their first, second or third attempts were taken. For illustration and discussion, the week 10 quiz questions Brute-force versus Horspool question of semester 1, 2016 is shown in Figure 3. This question is representative of questions that were rated in the pilot by the LMS software as having 'medium' difficulty score of 72.23% (Figure 1), but 'easy' in the following two iterations, 92.05% and 98.65% respectfully. In this example (Figure 3), 14 students were unable to answer the question correctly regardless of the number of attempts, while the remainder of the cohort took up to seven attempts to answer it correctly. 134 students of the cohort of 178 answered correctly on their first (50), second (60) or third (24) attempt, which yields an alternative percentage of 75.3 % that the LMS would associate with 'medium' difficulty.

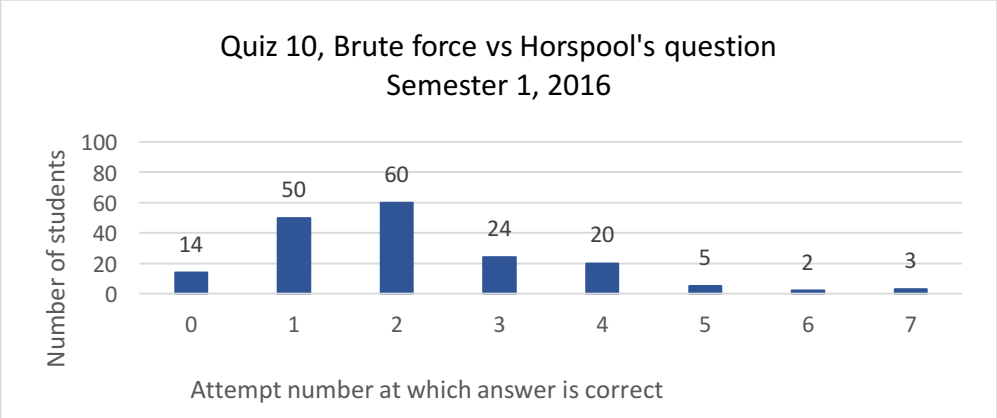

Figure 3: Number of attempt students to answer Brute force vs Horspool's in week 10 quiz

Further investigations into the discrepancy between difficulty classifications of 'medium' in the pilot semester and 'easy' in subsequent semesters revealed that the LMS difficulty classification is extremely susceptible to change of parameters in quiz delivery. The more attempts students can make, the more likely is the LMS to classify questions as 'easy'.

We had hoped that 'difficulty' classifications reported by the LMS would direct our search for the quiz questions that students found useful for challenging their misconceptions within the subject. Although it is disappointing not to be able to take LMS statistics at face value, our investigations of the raw attempt statistics for every question has pointed the way forward. Rather than dwell on the number attempts made for each quiz question, we are now looking more thoroughly at the incorrect alternatives chosen by the students in their attempts so that we may more correctly identify useful questions and common misconceptions.

Conclusion

Our experience with the use of learning analytics data from our institution's LMS to evaluate online quizzes has been mixed. The challenge we have found is not in the acquisition of data, but in making sense of the automated reports and discerning helpful information in identifying quiz questions useful in improving students' learning outcomes.

Mandatory quizzes in our algorithms subject have increased student engagement at motivated students to stay abreast with the subject material. However, we have been frustrated in our efforts using the inbuilt LMS question analysis tools of discrimination and difficulties to identify those quiz questions most helpful to students. In the end, the tool gave the same difficulty and discrimination classification to all questions. Analysis shows that the "difficulty" assessments are an artefact that is strongly influenced by the number of attempts that the software allows.

The issue may be broader than a particular learning analytics tool's rigidity. We would like metrics that are better aligned with the aims of formative assessment. While a focus on questions' "discrimination" values makes perfect sense in the context of summative assessment, its value in formative assessment is less clear. For our purpose, that is, for finding the "learning value" of a question, it makes better sense to study, as we ended up doing, students' response patterns. We do not have a metric to propose. However, loosely, a response pattern that, at least to us, suggests that a student has benefitted from a question is where many students fail to answer the question correctly in a first attempt, but then, perhaps based on some hint, get in right in a second attempt. When we ask students which questions they found useful, the questions they identify almost always follow that pattern.

Acknowledgements

This study was undertaken with Ethics Approval (ID 1648326) from the University of Melbourne and supported through the University's Learning and Teaching Initiatives grant scheme.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5(1),* 7-74.

Bonwell, C. C. & Eison, J. A. (1991). *Active learning: creating excitement in the classroom*. The George Washington University, School of Education and Human Development, Washington, D.C.

Boud, D. and Associates (2010). *Assessment 2020: Seven propositions for assessment reform in higher education.* Sydney: Australian Learning and Teaching Council.

Boud, D & Falchikov, N. (Eds). (2007). Rethinking Assessment in Higher Education: Learning for the Longer Term. Routledge.

Buckley, G. L., N. R. Bain, A. M. Luginbuhl, and M. L. Dyer. 2004. Adding an "Active Learning" Component to a Large Lecture Course. *Journal of Geography103*:231-237.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, *14*(7), 1-11.

Falchikov, N. (2005). *Improving Assessment through Student Involvement*. Routledge Falmer.

Klein, P. (2003). Active Learning Strategies and Assessment in World Geography Classes. *Journal of Geography 102*:146-157.

Gibbs, G. (2010). Using assessment to support student learning at University of East Anglia. Retrieved March 10, 2017, from https://portal.uea.ac.uk/documents/6207125/8588523/using-assessment-to-support-student-learning.pdf. Leeds Metropolitan University.

Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and teaching in higher education*, (1), 3-31.

Gluga, R., Kay, J., Lister, R., & Kleitman, S. (2013). Mastering cognitive development theory in computer science education. *Computer Science Education*, *23*(1), 24-57.

Hanson, S., & Moser, S. (2003). Reflections on a discipline-wide project: developing active learning modules on the human dimensions of global change. *Journal of Geography in Higher Education*, *27*(1), 17-38.

Huba, M. E.&. Freed, J. E. (2000). Learner-Centred Assessment on College Campuses: Shifting the Focus from Teaching to Learning. Allyn & Bacon.

Jimoyiannis, A. (2011). Using SOLO taxonomy to explore students' mental models of the programming variable and the assignment statement. *Themes in Science and Technology Education 4(2)*, 53-74.

Scheyvens, R., Griffin, A. L., Jocoy, C. L., Liu, Y., & Bradford, M. (2008). Experimenting with active learning in geography: Dispelling the myths that perpetuate resistance. *Journal of Geography in Higher Education*, *32*(1), 51-69.

Woit, D., & Mason, D. (2003). Effectiveness of online assessment. *ACM SIGCSE Bulletin, 35(1)*, 137-141.