I wish I could believe you: the frustrating unreliability of some assessment research

Practitioner Research in Higher Education
Special Assessment Issue

University of Cumbria
Vol 10(1) pages 13-21

Tim Hunt and Sally Jordan
The Open University
t.j.hunt@open.ac.uk

Abstract
Many practitioner researchers strive to understand which assessment practices have the best impact on learning, but in authentic educational settings, it can be difficult to determine whether one intervention, for example the introduction of an online quiz to a course studied by diverse students, is responsible for the observed effect. This paper uses examples to highlight some of the difficulties inherent in assessment research and suggests some ways to overcome them. Problems observed in the literature include: assuming that if two effects are correlated then one must have caused the other; confounding variables obscuring the true relationships; experimental approaches that are too far removed from reality; and the danger that self-reported behaviour and opinion is sometimes different from student's actual behaviour. Practical solutions include: the use of an experimental or pseudo-experimental approach; the use of mixed methods; and the use of meta-analysis.

Keywords
Research methodology

Introduction
Reviews of the literature (e.g. Black and Wiliam, 1998; Gibbs and Simpson, 2004-5) have identified conditions under which assessment seems to support learning, and a number of frameworks have been devised for use by practitioners in developing and auditing their assessment practice. The best known of these were proposed by Gibbs and Simpson (2004-5) and Nicol and Macfarlane-Dick (2006). However, our understanding of these matters is still incomplete, and so research continues into the effectiveness of assessment and feedback.

Although there are well-established educational research methods (e.g. Cohen et al., 2011, Punch, 2009), disentangling what is occurring in authentic assessment situations is not easy. It can be difficult to answer unconditionally even a seemingly simple question, such as "does the introduction of an online quiz to a course studied by diverse students actually improve attainment?" This paper uses examples of good and poor practice to highlight the inherent difficulties. It also makes suggestions for how they can be overcome, since practitioner research is something that should not be abandoned.

Feedback on errors, our own and others', is a powerful way to learn, and it is with that in mind that the observations here have been made. Neither author is an expert in research methods, making this paper rather presumptuous. Both are, however, practitioners with a numerate and scientific background[1] and shared some frustrations with parts of the assessment literature that seemed worthy of discussion.

1Tim Hunt is an Educational software developer with a PhD in mathematics, and Sally Jordan is a Professor of Physics Education.

Citation
Hunt, T., Jordan, S. (2016) 'I wish I could believe you: the frustrating unreliability of some assessment research, *Practitioner Research in Higher Education Journal*, 10(1), pp.13-21.

This is not a systematic review. The examples discussed were selected ad-hoc because they illustrated particular points in an interesting way. Even though some of the evidence used has only questionably supported the conclusions drawn, it should be noted that all the papers referred to have furthered collective understanding of assessment practice.

Problem one: Correlation does not imply causation

When stated as in the heading it becomes a cliché, but this assumption is one of the most common errors in the assessment literature. One cannot necessarily jump from the observation that two effects are correlated to the conclusion that one must be causing the other[2]. As Boyle (2007, p.98) says, "although several studies have claimed that use of eFA [e-formative assessment] materials is associated with learning gains, the bases on which they do so are generally not well founded". Unfortunately, some authors (e.g. Sly, 1999; Wilson et al., 2011) have fallen into the trap of assuming that because those students who chose to do an optional formative computer-marked quiz also did better in a later summative assessment, the formative quiz was the reason for the improved attainment. To analyse Sly's argument in more detail: the paper's title certainly is a claim that the present authors would like to believe: "Practice tests as formative assessment improve student performance on computer-managed learning assessments" (Sly, 1999). This research took place in the context of a large (614-student), first-year Economics course at Curtin University. There were two summative tests referred to as S01 and S02. Before the first of these, there was an optional practice test P01, which drew questions from the same test bank as S01. Thus the questions in S01 and P01 covered the same material, but were different. The results are summarised in Figure 1.

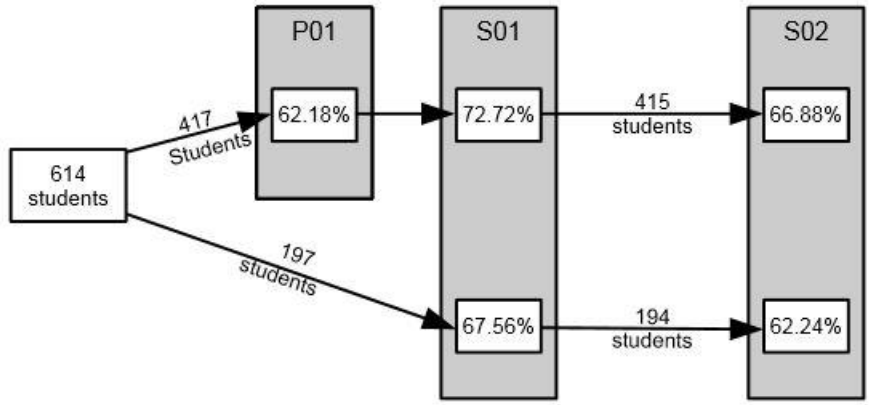

Figure 1. Average grades attained by Sly's students. For all tests and groups, standard deviation was between 15 and 17%. All differences significant to $p = 0.001$ or better.

The figure shows that the students who opted to do the practice test did better overall. What can be concluded from that? Perhaps it was just that the more able students chose to do the practice, and they would have got better marks anyway. Sly is aware of that possibility and argues as follows: Consider the very first test each student does, whether it be P01 or S01. Students who did not attempt P01 scored an average of 67.56% on the first test they encountered, compared to 62.18% for those who did P01. Therefore, it seems that students who chose not to do P01 were in general better than those who skipped it. Thus the impact of the practice test is doubly impressive. Not only did those students do better in the end, but initially they were weaker. In addition, improved performance also carried over to S02 which tested a different area of Economics.

2 An amusing way to explore this fallacy is at the 'Spurious correlation' web site:, which presents random combinations from a large set of unrelated time-series datasets which just happen to correlate. http://www.tylervigen.com/spurious-correlations

However, this argument is not entirely convincing. It assumes that students engage in the same way with summative and formative tests, although there is now much evidence to the contrary (Kibble, 2007; Jordan and Butcher, 2010; Jordan, 2011). An alternative interpretation would be: Since S02 covered different material, we can use that to control for different ability between the two groups. On test S02, student who chose to do P01 did about 5% better than those who did not. On S01 those students also did about 5% better. The practice test does not seem to have had much effect. However we can see that, for this type of test, students will concentrate more on a summative test, and so score about 10% more than for a similar practice test.

How can one tell which interpretation is correct? Unfortunately Sly's work did not produce enough information to determine.

Research technique one: use experiments

The scientific gold-standard for trying to tease out what, if anything, a statistically significant correlation might mean, is a randomised controlled experiment. The important features here are:

- **Experiment** – only change a single factor, while all other factors are held constant, so that any differences in outcome can be convincingly attributed to the one factor that was different.
- **Controlled** – experiments compare things, so the condition or intervention we are interested in must be compared to a control. This may be "do nothing", but it is often more meaningful to compare an innovation with current best practice.
- **Randomised** – while the ideal experiment would hold "all other factors constant" this is only really feasible in the physical sciences. When people are involved, as they are in educational research, there are many factors that are difficult or impossible to measure. There can be no guarantee that the different experimental groups are identical with regards to all other factors. The practical alternative is to randomly allocate participants to groups. Then it is likely that for most factors there will be no systematic bias between the groups.

There is much more that can be said, but that is best left to the research methods text books (e.g. Cohen et al., 2011, Punch, 2009). There are many good examples of well conducted experiments in the assessment literature (e.g. Angus and Watson, 2009; Mogey et al., 2012; Ćukušić et al., 2014). One area where many experiments all demonstrate similar results is with the "testing effect" (e.g. Roediger and Karpicke, 2006). Let us take Karpicke and Blunt (2011) as a typical example of a good experimental design. 80 participants were randomly assigned to four groups of 20, and each group subjected to a different 'condition'.

"In the study-once condition, students studied the text in a single study period. In the repeated study condition, students studied the text in four consecutive study periods. In the elaborative concept mapping condition, students studied the text in an initial study period and then created a concept map of the concepts in the text. … Finally, in the retrieval practice condition, students studied the text in an initial study period and then practiced retrieval by recalling as much of the information as they could on a free recall test. After recalling once, the students restudied the text and recalled again. The total amount of learning time was exactly matched in the concept mapping and retrieval practice conditions." (Karpicke and Blunt, 2011, p. 772-773).

This shows many of the hallmarks of a good experiment, including explicit control for the variable "time on task" which is known to have a big effect on how much is learned. To conclude the experiment, all students took a test containing a mixture of factual recall questions and "inference questions" to test understanding. In common with other similar experiments, those in the "recall

practice" group outperformed the others, providing yet another convincing replication of the testing effect.

Problem two: confounding variables

There have already been two examples in this paper of what are known as confounding variables. In Sly's paper, the innate ability of the students in each group could not be measured but affected the results. In the Karpicke and Blunt experiment, time-on-task was known to be an important variable and could be explicitly controlled. Above, it was argued that a correlation did not necessarily imply causation, but with confounding variables the situation is worse. The correlation seen may be the opposite of the real effect, a phenomenon known as Simpson's paradox.

Since this is so potentially dangerous and counter-intuitive, it is worth examining a simple example, even though we must step beyond assessment research to find one. In the fall of 1973, the admission figures for University of California, Berkeley were that 8442 men applied of whom 44% were admitted, and 4321 women applied of whom 35% were admitted. On this basis Berkeley was sued for bias against women.

When the data was analysed by department, however, the picture reversed, as shown in Table 1.

Table 1. Berkeley admission figures broken down by department.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	**825**	62%	108	**82%**
B	**560**	63%	25	**68%**
C	325	**37%**	**593**	34%
D	**417**	33%	375	**35%**
E	191	**28%**	**393**	24%
F	272	6%	**341**	**7%**

Most departments admitted about the same percentage of men and women, or had a bias in favour of women. However, different departments had very different numbers of applicants per place. Generally men applied to departments like Engineering or Chemistry which admitted more of their applicants, while women applied to more competitive departments such as English.
The initial analysis only considered two variables: gender, and percentage acceptance rate. A third (confounding) variable is the department applied to. When that is included, the direction of the bias (correlation) reverses. As stated above, with human subjects, it is not possible to measure, or even know, all the factors or variables that may confound the research. Therefore, the best protection is to allocate participants to conditions randomly.

Problem three: experiments may not represent reality

Striving to control for every possible confounding variable may, however, may lead to researching something rather far removed from the situation of interest. This could lead to experimental results that, while reliable, do not reveal anything valid about typical assessment practice.

As referred to above, the testing effect has been replicated in many experiments, and practitioners are now using it in their teaching. Wooldridge et al. (2014) decided to test the effect as it is

commonly implemented in classrooms, where the way questions are used for practice and assessment is somewhat different from the way they were used in the experimental procedures. They took a biology textbook (Phelan, 2009) which came provided with questions that could be used for retrieval practice. The experiment covered 19 sections of the text and the corresponding questions. Questions were of two sorts: factual and application. (An example of what they consider an application question is “Brown and white rabbits are born in snowy woods. This would produce what?”). The final test comprised 19 factual and 19 application multiple-choice questions, one for each section. For studying, students were split into five groups. The control group were just allowed to highlight the text while reading. The other four groups took an initial quiz of 19 short-answer questions, either factual or application, that were either essentially the same as the final multiple-choice questions, or only related to them.

The results only showed the testing effect in two cases. Students who had seen essentially the same factual questions in practice and the final test did much better on those, but did not do better on the application questions. Similarly, students who had practised the same application questions as in the final test did better on those but not on the factual questions. In all other cases, there was no significant difference in performance on the final test compared to the ‘highlight’ group. So, while the testing effect is well established, its effect is not as broad or transferable as had been hoped and assumed. Retrieval practice is a powerful tool for learning the specific set of facts that were covered when practising retrieval, but it seems to be of little help when trying to assimilate a broad area of knowledge.

Research technique two: ask the students

Given the limitations of experiments for understanding the complexities of real assessment practice, it is beneficial to employ a range of methodologies. Several research methods all generating evidence that supports the same conclusions is much more convincing than a single observation. Despite the claim by Walker et al. (2008) that student expectations and perceptions of e-assessment have been under-researched, much of what is written (e.g. Marriott, 2009; Holmes, 2015) about the benefits of computer-marked assessment relies on student opinion and self-reported behaviour, rather than on the way in which students actually engage with assessment tasks. Student opinion is important (Dermo, 2009) but care is needed when drawing certain types of conclusions because students' reports of their own behaviour and motivation do not always match their actions (Jordan, 2014).

Problem four: what students say does not always match what they do

Jordan (2011) reports on the results of a questionnaire that sought to ascertain student perception of interactive computer-marked assignments (iCMAs). Of the 151 student responses that were received, 128 (85%) agreed or strongly agreed with the statement “If I get the answer to an iCMA question wrong, the computer-generated feedback is useful” and only 8 (5%) disagreed.

However, when students were observed answering such questions in a usability laboratory, many were seen not to use the feedback provided. The questions that students were observed using were early developmental versions and as a result the answer-matching was not completely accurate. This led to a situation when two students were told that an answer was correct but it was actually completely wrong. One of the students commented “It’s so nice to get the first one right”. He then appeared to read the full answer, but completely missed the fact that it was the opposite of the answer he had given. The other student ignored the feedback and explained in the post-iCMA interview: “Yes, you think you get it right so you ignore this”.

Another case arose when a student typed “deceased” instead of “decreased” and was marked as incorrect. He failed to see his spelling mistake and was quite upset that the computer had marked

him as incorrect. He read the specimen answer, commented that he thought his answer was the same and ticked the box that allowed students to indicate that they thought they had been incorrectly marked. This student did not read any of the other final answers. These findings provide support for Kulhavy and Stock's (1989) concept of "response certitude", which argues that feedback is most helpful when a student is confident that their answer is correct but it turns out to be incorrect.

Jordan's paper gives further evidence from the learning analytics to support the notion that many students were not using the feedback. This disparity between the students' self-reported behaviour and their actual behaviour may have been as a result of the questionnaire's relatively low response rate (approximately 20%); perhaps the students who replied to the survey were those who actually did read and respond to the feedback? However it is almost certainly the case that more students report that they find feedback useful than actually make good use of it, in line with the bias in self-reported behaviour that is observed in medicine and business (Jordan, 2014, p. 69). This means that care must be taken when interpreting results such as the finding that 90% of students agreed with the statement "Positive comments have boosted my confidence" (Weaver, 2006, p. 386) and that 93% of students agreed with the statement "I find the immediate reporting of my test result valuable" (Marriott, 2009, p. 243). (Strangely, these statistics about assessment always seem to be around 90%.) It is very important to take note of student opinion and self-reported behaviour and attitudes; however it must always be remembered that actual behaviour might be rather different. There are many cognitive biases, particularly when it comes to self-knowledge. This is another reason to use a variety of research methods to try to understand the effects of assessment.

Research technique three: meta-analysis

Using mixed methods in one study to get a range of evidence about a particular intervention is good, but much more can be learned if we combine evidence from many studies. This is often done as a review article where one author qualitatively combines results from many papers in a field. During the last century the new technique 'Meta-analysis' was invented for quantitatively combining the results of many different studies. The process is briefly as follows:

1. Perform a systematic search of the literature. Define which databases will be searched, and which search terms will be used, and so get a long-list of papers to consider. This systematic approach is to guard against confirmation bias.
2. Read all the papers found to verify that they actually relate to the intervention of interest, and report the results of each experiment in sufficient detail to be used. The criteria for selection should again be decided in advance and applied as objectively as possible.
3. For each separate result found, compute the 'effect size'. We omit the statistical details here (e.g. Cohen et al., 2011, Chapter 17) but roughly the effect size is the difference in result divided by the standard deviation. So, in test S01 of Sly's paper, the difference in average grade between the groups was 5.16%, and the standard deviation was 15.67%, so the effect size is $d = 0.33$.
4. Combine all the different effect sizes for all the separate studies to get an overall effect size for the intervention of interest.

Meta-analysis was first widely adopted in medicine (see Goldacre (2008) for a popular summary) thanks in part to an organisation called the Cochrane Collaboration. In medicine, evidence-based practice is now the norm. However, in the presence of great commercial pressures, even such a supposedly objective process can suffer from systematic problems (Goldacre, 2012). In education, meta-analysis is taking longer to become common practice, though there are now many good examples (e.g. Kluger and DeNisi,1996; Nyquist, 2003; Hattie and Timperley, 2007; Hattie, 2013). An interesting finding from Hattie (2013) is that in education almost anything has a positive effect. It is

very difficult to make time-on-task actually harmful. Therefore, Hattie suggests, we should generally be seeking practices that give an effect size of more than $d = 0.4$.

Problem five: ethical and practical considerations when experimenting on students

All research on humans must be ethical. If it is hoped that an intervention will be beneficial to students, is it ethical for us to only make it available to some of them? That is what must be done in order to conduct an experiment. The counter argument is that until the research has been done, it is not known if the intervention actually is beneficial. Therefore, the long-term benefit of doing the experiment outweighs the short-term cost of some participants not getting any possible benefits now. This is the accepted justification in medical trials where the benefits and side-effects can be a matter of life and death.

In education it may be possible to make things reasonably fair even when only one group is receiving an intervention. Returning to Sly's example, the scenario could be extend as follows: Add an additional practice test P02 before test S02. Then, randomly allocate the class into two groups, G1 and G2. Students in group G1 must do test practice test P01 before S01, but do not get to do practice test P02. For group G2 it is the opposite, they cannot attempt P01, but must attempt P02. This is reasonably fair. In addition, between the time when students attempt S02 and the end of course exam, both P01 and P02 could be made available to all students for use as a revision aid. Similar designs are possible in many cases (e.g. Mogey et al., 2012). Alternatively, there is a range of less pure designs, collectively referred to as 'quasi-experiments' which may be easier to use in the situations where practitioners operate. Once again, however, there is not sufficient space to give details (see, for example, Cohen et al., 2011 or Punch, 2009).

Summary and suggestions for the future

This paper has looked at some common problems seen in the assessment literature: assuming causation when there is a correlation; the dangers of confounding variables; the risk that experiments become so abstracted from practice that the results are not useful; the fact that what students say and what they actually do can differ; and the difficulty of researching students ethically.

Also considered are some of the ways these issues can be mitigated, such as doing experiments (or pseudo-experiments) with as good a methodology as possible in the assessment context; using mixed methods to get a variety of perspectives on the assessment practice and checking that the different observations all support the same conclusions; and being aware how new research fits into the wider literature, which at the most formal end involves meta-analysis.

This paper started with the observation that assessment research in authentic settings is difficult. Many practitioners are only secondarily educational researchers, alongside their primary specialism. This is what makes their contributions to the field so interesting, but it also, as has been seen, brings risks. Given that it is unrealistic to expect everyone to take a full course on research methods, what practical steps might be taken to improve matters. Here are three small suggestions:

- Some assessment conferences have workshops before the main proceedings start. Such a workshops could be arranged to offer practitioner researchers a chance to discuss their proposed study with a research methods expert, and receive some advice about how best to proceed. Alternatively, institutions wishing to promote research among their staff could run their own workshop along these lines.
- The published assessment literature does not appear to contain any reviews of the available research methods, and of how well those methods are applied. Such a survey would be interesting, and would shed a more systematic light on the risks described here.
- Any increase in awareness of the issues around research methods would help improve the

field. All practitioners can contribute by occasionally thinking and talking about the topic. Hopefully this paper has provided one such opportunity for readers. That is the spirit in which it was written.

Acknowledgements

We would like to thank the audience of our talk on this topic at the Assessment in Higher Education Conference, June 2015. Their response gave us the confidence to proceed with writing this paper. We would also like to thank the anonymous reviewers for their comments, which helped us to clarify the focus of our argument.

References

Angus, S. D. and Watson, J. (2009) Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set, *British Journal of Educational Technology*, 40(2): 255–272.

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

Boyle, A. (2007) The formative use of e-assessment: Some early implementations, and suggestions for how we might move on. *Proceedings of the 11th International Computer Assisted (CAA) Conference.* Loughborough, 8th–9th July 2007. Available at http://caaconference.co.uk (Accessed 27th September 2015).

Cohen, L., Manon, L. and Morrison, K. (2011) *Research methods in education, 7th Edition*. London: Routledge.

Ćukušić, M., Garača, Ž., and Jadrić, M. (2014) Online self-assessment and students' success in higher education institutions, *Computers and Education*, 72: 100–109.

Dermo, J. (2009) e-Assessment and the student learning experience: A survey of student perceptions of e-assessment, *British Journal of Educational Technology*, 40(2): 203–214.

Gibbs, G. and Simpson, C. (2004-5) Conditions under which assessment supports students' learning, *Learning and Teaching in Higher Education*, 1: 3–31.

Goldacre, B. (2008) *Bad Science*. London: Fourth Estate.

Goldacre, B. (2012) *Bad Pharma*. London: Fourth Estate.

Hattie, J. (2013) *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J. and Timperley, H. (2007) The Power of Feedback. *Review of Educational Research*, 77(1): 81–112.

Holmes, N. (2015) Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module, *Assessment and Evaluation in Higher Education*, 40(1): 1–14.

Jordan, S. (2011) Using interactive computer–based assessment to support beginning distance learners of science, *Open Learning*, 26(2): 147–164.

Jordan, S. (2014) *E-assessment for learning? Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics*. Unpublished doctoral thesis. Milton Keynes: The Open University. Available at http://oro.open.ac.uk/41115/ (Accessed 27th September 2015).

Jordan, S. and Butcher, P. (2010) Using e-assessment to support distance learners of science. In Raine, D., Hurkett, C. and Rogers, L. (ed.) *Physics Community and Cooperation: Selected Contributions from the GIREP-EPEC and PHEC 2009 International Conference.* Leicester: Lula/The Centre for Interdisciplinary Science: 202–216.

Karpicke, J. and Blunt, J. (2011) Retrieval practice produces more learning than elaborative studying with concept mapping, *Science*, 331(6018): 772–775.

Kibble, J. (2007) Use of unsupervised online quizzes as formative assessment in a medical physiology course: Effects of incentives on student participation and performances, *Advances in*

Physiology Education, 31(3): 253–260.

Kluger, A. N. and DeNisi, A. (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory, *Psychological Bulletin*, 119(2): 254–284.

Kulhavy, R. W. and Stock, W. A. (1989) Feedback in written instruction: The place of response certitude, *Educational Psychology Review*, 1(4): 279–308.

Marriott, P. (2009) Students' evaluation of the use of online summative assessment on an undergraduate financial accounting module, *British Journal of Educational Technology*, 40(2): 237–254.

Mogey, N., Cowan, J., Paterson, J. and Purcell, M. (2012) Students' choices between typing and handwriting in examinations, *Active Learning in Higher Education*, 13(2): 117–128.

Nicol, D. and Macfarlane-Dick, D. (2006) Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, *Studies in Higher Education*, 31(2): 199–218.

Nyquist, J. B. (2003) *The benefits of reconstruing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished doctoral dissertation. Nashville, TN: Vanderbilt University.

Phelan, J. (2009) *What is life? A guide to biology*. New York: W H Freeman.

Punch, K. F. (2009) *Introduction to Research Methods in Education*, London: SAGE Publications.

Roediger, H. L. and Karpicke, J. D. (2006) The power of testing memory: Basic research and implications for educational practice, *Perspectives on Psychological Science*, 1(3): 181–210.

Sly, L. (1999) Practice tests as formative assessment improve student performance on computer-managed learning assessments, *Assessment and Evaluation in Higher Education*, 24(3): 339–343.

Walker, D. J., Topping, K., and Rodrigues, S. (2008) Student reflections on formative e-assessment: Expectations and perceptions. *Learning, Media and Technology*, 33(3): 221–234.

Weaver, M. R. (2006) Do students value feedback? Student perceptions of tutors' written responses, *Assessment and Evaluation in Higher Education*, 31(3): 379–394.

Wilson, K., Boyd, C., Chen, L., and Jamal, S. (2011) Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment, *Computers and Education*, 57(2): 1493–1500.

Wikipedia (2015) *Simpson's paradox*, available at https://en.wikipedia.org/wiki/Simpson%27s_paradox (Accessed 24th September 2015).

Wooldridge, C., Bugg, J., McDaniel, M. and Liu, Y. (2014) The testing effect with authentic educational materials: A cautionary note, *Journal of Applied Research in Memory and Cognition*, 3(3): 214–221.