An investigation into the effect of regular low-stakes recall testing on performance in knowledge and understanding examination questions in A-level Computer Science

Candidate number: 26725452
Word count: 3,984

Introduction

All A-level qualifications in the UK follow rules set out by Ofqual. These include a strict list of assessment objectives (AOs) and, in Computer Science, there are three. Between them, they detail the assessment of knowledge, understanding, application, design, analysis and evaluation (Ofqual, 2014). I am a Computer Science teacher in a large sixth form college in the south of England, hereafter referred to as “the college”. The Computer Science course at the college follows the AQA A-level specification. For this investigation, I was particularly interested in AO1. In my course, this accounts for 37.5% of marks available in examinations and covers assessing how students “demonstrate knowledge and understanding of the principles and concepts of Computer Science” (AQA, 2019). This often amounts to simple factual recall. However, perhaps surprisingly, analysis of the college’s 2018 examination results showed that students performed worse in questions targeted at AO1; on average, students scored 35% for AO1, 55% for AO2, and 55% for AO3 (AQA, 2018).

After this analysis, I was interested in investigating the discrepancy further and aimed to improve factual recall in my students. I had a project proposal signed off before my investigation began. This report outlines my action research: it starts with a review of some relevant literature, before looking at how that informed my methodology decisions. Finally, it analyses my results and puts forward some recommendations.

The aim of this project was to measure whether using regular low-stakes recall testing of technical language and ideas improves students’ performance in AO1-style examination questions. Although I only studied Computer Science classes, the topic researched could be applicable to other subjects. More traditional science subjects with similarly high levels of factual recall may benefit from my research.

Literature Review

Action Research

It was important to first explore my interpretation of action research. Perhaps one of the key contributors to the field was Lawrence Stenhouse (1975, p.142), who envisaged “each teacher [being] a member of the scientific community”. Action research is different from traditional, empirical research in that, “in action research, researchers do research on themselves” (McNiff, 2002). They reflect upon and modify patterns in their practice. In fact, many agree that “action research is simply a form of self-reflective enquiry” (Carr & Kemmis, 1986, p.162). Schön (1983) introduced reflection-in-action: adjusting in the moment using indescribable knowledge gained through action. However, much literature on action research focuses on systematic conscious reflection, using what Schön may have called “reflection on action” (Schön, 1983, p.55). Practitioners follow a sequence of reflective steps to act, then measure and evaluate the effects of their actions. Carr & Kemmis (1986, p.162) extended that process, saying that “a self-reflective spiral of cycles” is needed. Jean McNiff took it even further by adding another dimension, first introducing the “three-dimensional network of spirals” in 1984 (p.46). According to her, action research is not a linear process, nor a simple spiral one, but a set of spirals comprised of multiple reflective processes. My understanding of action research is most closely aligned with McNiff’s because it emphasises the wider picture of the practitioner’s role. Figure 1 is adapted from McNiff & Whitehead’s work (2009, p.41 & 2011, p.9) and shows my interpretation of action research. This small project formed part of one of the many interwoven spirals, shown magnified in the figure.

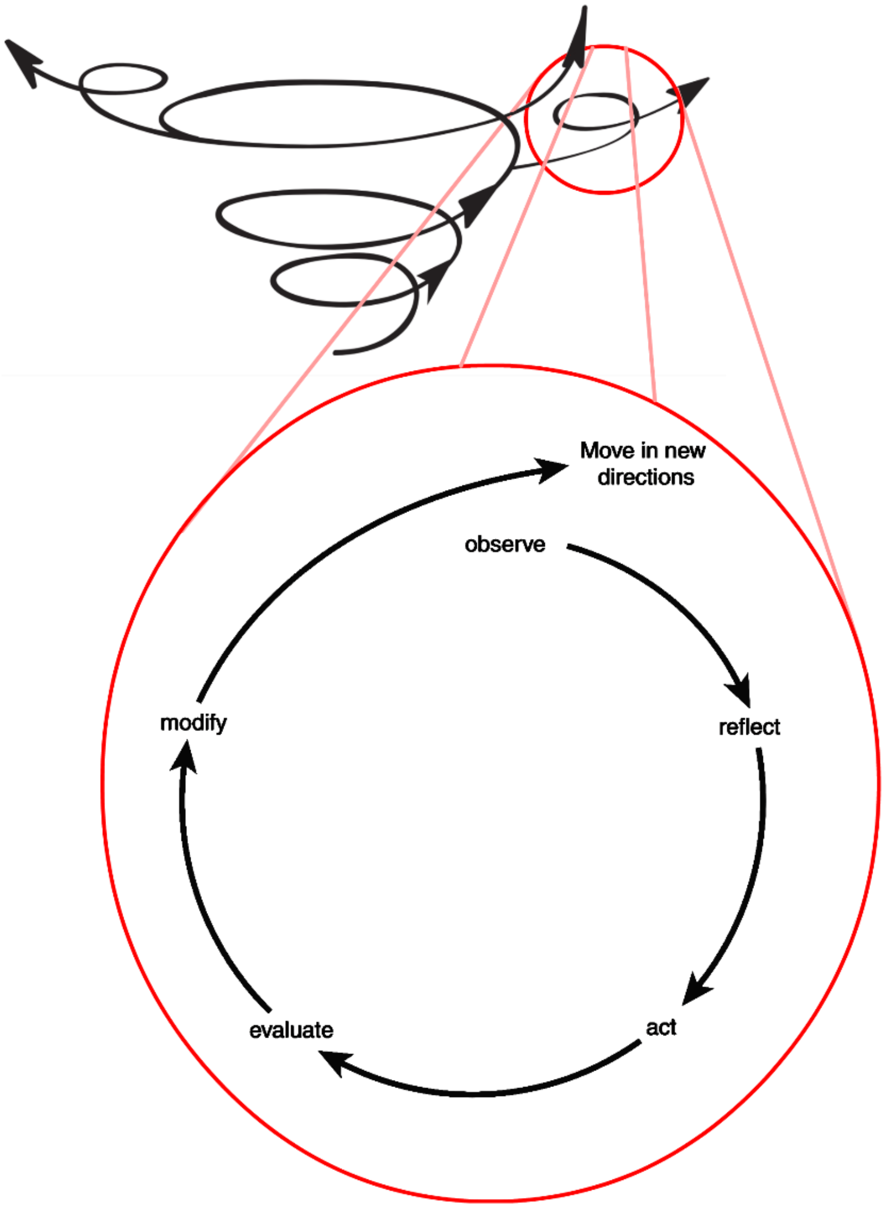

Figure 1

Another aspect of this model that I like is how it acknowledges allowing the teacher to join and leave the spirals at will, rather than being bound by a rigid structure (McNiff, 1984). From my little experience of teaching and of teachers so far, I see that the demands of the profession make it difficult to always think and act in a reflective manner. Although it seems to be a mindset very conducive to professional and personal development, sometimes practitioners are simply too tired to maintain the cognitive stamina needed for constant conscious reflection. It is crucial for development in the role to regularly take stock of one's practice, but it should not be a sign of a poor teacher to not be thinking often about improving.

Computer Science is a relatively new subject and is undergoing a period of drastic change (Royal Society, 2017). Therefore, there is a relatively small body of literature on its education, especially that which does not focus on Primary or Higher education. Furthermore, there are very few accessible and relevant reports of action research being undertaken by teachers of the subject. There does exist a community of practice called Computing At Schools, part of which encourages Computer Science teachers to share best-practice (Sentence & Humphreys, 2018). For this project, I had to rely on research not directly relevant to Computer Science. By no means did this render it invalid, but it was something I had to take into consideration when implementing some of the ideas in the literature.

Knowledge Recall

There were two aspects to my chosen action for this project, the first of which being to improve knowledge retention. My decisions here were rooted in the research of memory. This is a huge field, so I could only review a fraction of the relevant literature. Much of it is still based on the seminal work of Ebbinghaus (1885). He introduced the forgetting curve: the more times someone thinks about something, the longer they will be able to successfully recall it for. My chosen action was heavily influenced by Bjork's (1994, 2011) work on 'desirable difficulties' for improving learning, especially the generation effect: the simple "long-term benefit of generating an answer, solution, or procedure versus being presented that answer" (Bjork & Bjork, 2011, p.61). He concluded that being forced to recall information pushes it deeper into the long-term memory than just being told the information again, even though it intuitively feels like less learning happens. This was the basis for my regular tests. Sometimes also called the 'testing effect', a study by Roediger &

Karpicke (2006, p.249) found that, in the longer term, “prior testing produced substantially greater retention than studying”.

There were some other aspects of Bjork’s work that I could have drawn inspiration from, with spacing and interleaving being two examples. By spreading taught content across a longer period and interleaving other topics in the gaps, demonstrably better retention was measured (Bjork & Bjork, 2011, p.61). However, I decided that overhauling part of the college’s scheme of work would have been too ambitious, given the resources and time I had available. Another interesting option for action was experimenting with ‘flipped learning’ using teaching videos for pre-work. For Computer Science A-level, the teachers behind the relatively popular *craigndave* YouTube channel found that “watching the video, making notes in a book and then completing simple activities gives students three opportunities to understand the subject matter” (Sargent & Hillyard, 2017). This was clearly another application of Ebbinghaus’ (1885) forgetting curve to improve learning. However, for this project, I unfortunately did not have time to create enough teaching videos.

The second aspect studied in this project was creating ‘low-stakes’ tests. My decisions here were based on Ruth Butler’s (1988) work on feedback. Her study concluded that providing comments as feedback for work was the most effective for improving performance and engagement. By introducing marks or grades, those who achieved lower marks were hindered. Although this source is quite old, not based in the UK, and studied children with a “mean age 11.10” (Butler, 1988, p.4), I used it because it was taken seriously in John Hattie’s more popular work on feedback (Hattie & Timperley, 2007, p.92).

Methodology

For the project, I followed a sequence of steps like those pictured in Figure 1: I first found evidence for the issue; then decided what my options for action were; then found evidence to show what influence my action had; then aimed to explain the influence; and finally reflected on how to move forward in my practice (based on Jack Whitehead's action plan in (McNiff, 2002)). For the experiment, I had access to two first-year classes, each containing 22 students. One of the classes served as a *control* group in that my teaching continued as normal with them. The other class was my *active* group, which received four weekly low-stakes recall tests and encouragement to memorise definitions for them. The tests were low-stakes in the sense that there was no indication of marks available for each question, no total number of marks for each test, and no grade. Their purpose was to give students an opportunity to check their knowledge, giving them the desirable difficulty of the generation or testing effect (Roediger & Karpicke, 2006 and Bjork & Bjork, 2011). Students received written feedback on the tests from each other after peer-marking (based on Butler, 1988).

In order to gather sound evidence for the effect of these tests, and to quantify the magnitude of that effect, I created two assessments to give to both classes before and after the experiment. Both assessments contained examination-style questions aimed at AO1 (knowledge and understanding). The first of these assessments served as a pre-measure for each class; since the average examination performance differed between the two classes, it would have been invalid to simply compare both classes' average performances at the end of the experiment. This is because the 'better' class could have naturally performed better regardless of any difference in my teaching. Inter-group variation is one of the known disadvantages of the independent measures design (McLeod, 2017). However, my project timescale prevented me from assessing the same students at different times. Instead, I compared the average difference in assessment score before and after the experiment separately for each class. Then, it was these differences that were compared between classes. If the active class had improved more, this would have indicated that the action had a positive effect on their AO1 performance.

Assessment test scores were to be excluded for students who were absent for one or both assessments. This is because their individual improvement could not have been included in the class averages. For example, if there was a relatively strong student, I would expect them to perform above average for both assessments. However, if they missed the first assessment, the first average would not have been increased by them, yet the final average would. This would have resulted in an invalid apparent increase in their class's average improvement.

Since this project aimed to measure how the tests affect examination performance, it was important that my assessments accurately reflected real Computer Science examinations. It would have been infeasible in the required timescale to wait a year for both classes to sit their final A-level examinations, although further research in this area might involve that. It was not always possible to use questions from past papers due to the taught topic having not yet been examined much under the current specification. Despite this, I tried to structure the pre- and post-measure assessments as similar as possible to real AO1 questions. If given more time for this project, I would have carried it out when teaching a topic with many available past paper questions.

There were many sources of variation to control in order to improve the validity of my results. As well as trying to ensure that all other aspects of my teaching remained consistent across both classes, I made both initial and final assessments the same length and aimed to give them a similar difficulty. This was to mitigate the effects of question length and difficulty for different students. For example, it could be the case that more students in one class perform better in longer assessments. If so, the average score would be skewed higher for that class on the longer assessment and I would not be able to discern the true cause for the increase. My sample size within each class was not large enough to moderate this kind of variation. So, by keeping both assessments as similar as possible in structure, I eliminated this potential factor from affecting my results.

However, there were some aspects that were out of my control. Because of the structure of the scheme of work around the time of the project, the initial assessment needed to contain questions about a range of topics from the last 6 months. However, the final assessment could contain only questions on a single topic studied. This fact had to be taken into consideration when analysing the results. One class may have taken to the recent topic more than the others. Even if the pre-measure was based on one topic, it would still have been different content to the post-measure. Nonetheless, if I had more time, I would have carried out the research between two well-defined topics in the scheme of work.

The two AO1 assessments formed a quantitative method of my research. In addition, to gain a Kirkpatrick 'reaction' level (Kirkpatrick & Kirkpatrick, 2006) evaluation of my tests in the active class, I created a structured survey containing several questions about the impact of the tests. These used an ordinal scale to measure level of agreement, based on the original Likert (1932) scale. The difference was that there were only four choices, thus forcing the students to take either an 'agree' or 'disagree' stance. However, using simple ordinal scales for agreement cannot accurately capture feelings or opinions and the data from this must not be taken too seriously. With more time, I would have held a structured focus group to gauge student reaction from a representative sample.

It was important to also gather some qualitative perspectives. The survey contained some open-ended questions to allow students to give their general thoughts about the tests. In addition, I conducted a semi-structured interview with the subject leader for Computer Science at the college. This aimed to capture an experienced teacher's view on my identified issue and chosen action.

For this project, an ethical framework was followed (BERA, 2018). I had an ethics checklist signed off, which confirmed that the project would remain within typical teaching activities. One of the ways in which I adhered to a code of good ethics was ensuring that all student data was suitably anonymised. Obviously, I may use data about specific students to inform my own teaching practice, but it will not be possible to identify any individuals from results in this report.

Results

Before starting data collection, I informally interviewed an experienced Computer Science teacher at the college, who corroborated the existence of the issue that I inferred from the 2018 examination results. In our communication, she said that the issue was "definitely not unique to 2018" and that it had previously been "noticed in all college assessments", especially in the second year of the A-level. She was also supportive of my chosen actions, which helped to motivate it as an appropriate thing to try.

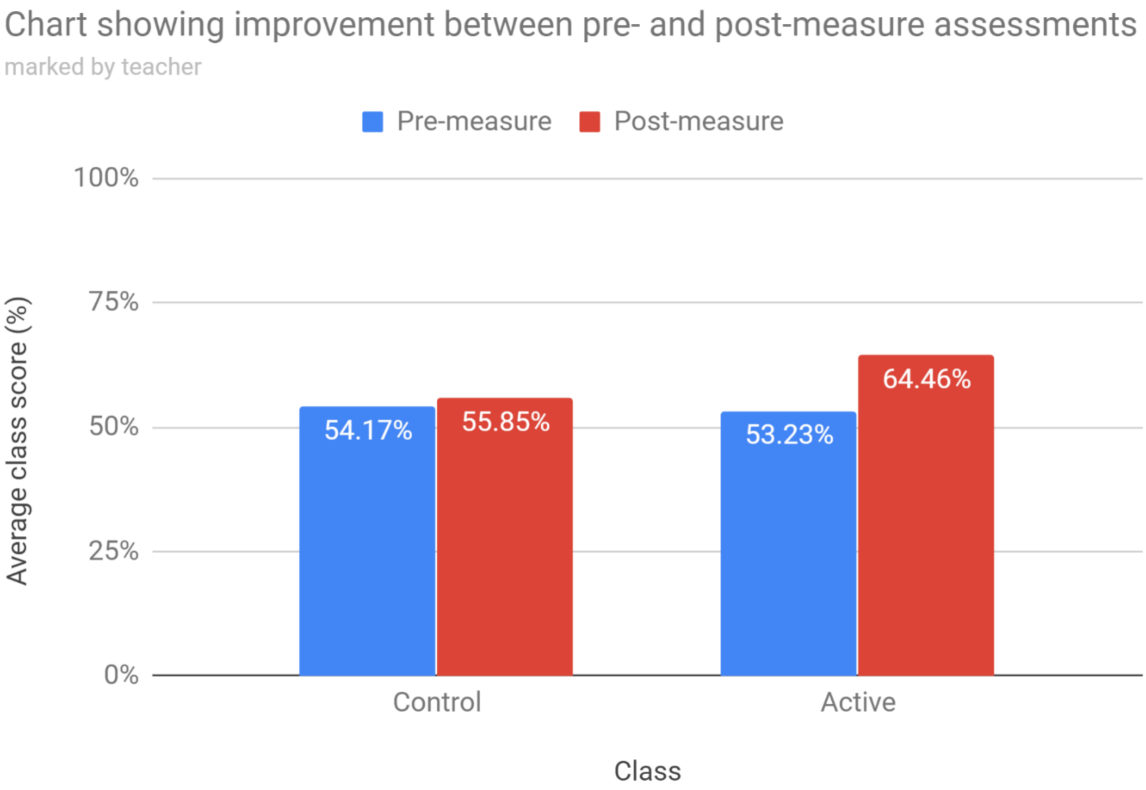

Figure 2

Figure 2 shows the average class improvement between the two AO1 assessments given at the start and end of the experiment. The first thing to draw from this chart is that both classes improved on average, implying that the final assessment was slightly easier. The control group improved by 1.68% and the active group by 11.23%. The difference shows that the active group performed *9.55% better* than would have been expected if they had improved by the same amount as the control group.

I did not solely collect data about the two assessments; I also recorded attendance for the weekly recall tests in the active class. It could be argued that students who missed one or more of the weekly tests did not receive the 'full dose' of my action. These students may not have performed as well in the final assessment as they would have done if they received all four tests. By excluding such students from the dataset, I can more truthfully see the potential effect of receiving weekly recall tests. Figure 3 shows the same data after excluding the four students who missed some recall tests. The result is an even more pronounced difference. Here, the active group performed *14.11% better* than would be expected.

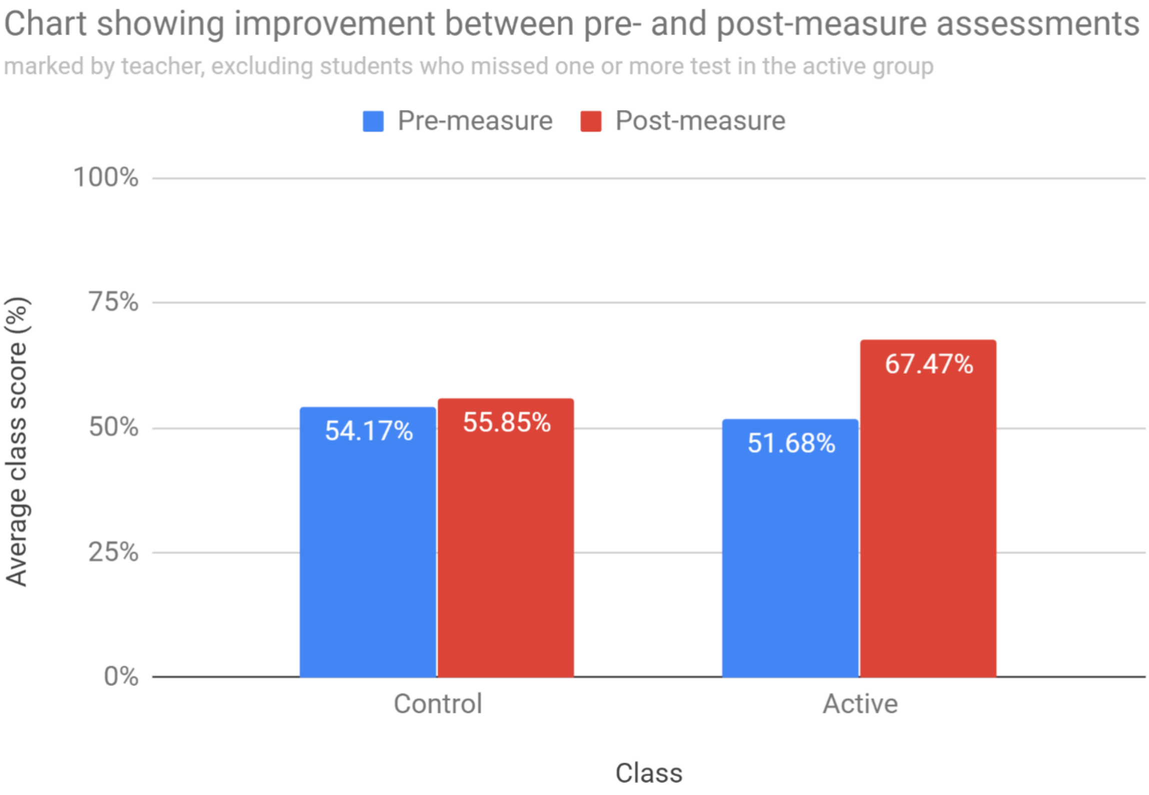

Figure 3

One thing to note is that, by only excluding four students from the dataset, there is a significant apparent improvement in the active group. This could partly be symptomatic of my small sample size: only 14 students remained in the active group dataset after all exclusions, so small modifications could cause large effects. However, it could also suggest that those students who did not have a high attendance performed significantly worse than those who did. This is by no means a surprising discovery, but it highlights how the effects of attendance are applicable to my A-level Computer Science classes and it is something that should be investigated further.

The aim of this project was to investigate the potential effect of my weekly recall tests on AO1 performance. My results could indicate that my action had a significant positive influence on performance in these types of examination questions. My results are consistent with the results of repeated testing found in the literature and there are different possible explanations for this improvement. One is that the desirable difficulty of being forced to recall information is conducive to better learning (Bjork, 1994). Another reason might be that the "testing permitted practice of the skill required on future tests and hence enhanced performance" (Roediger & Karpicke, 2006, p.254). However, it is important that, in this action research project, my conclusion remains tentative. I cannot make any strict claims of causation because there would be too many assumptions made. These include assuming that my assessments were fully representative of final examination questions, that my month-long timescale was representative of a full two-year A-level course, and that my two classes were representative of all Computer Science students in the country. Despite all these assumptions, my results are certainly promising and provide sound justification for deeper investigation into this topic at my college.

The results from my student survey were generally in favour of the weekly tests, with 19 out of 20 students agreeing that they helped with remembering content. The same proportion said that they wanted the tests to continue in the future. However, some questions received a more split opinion. Figure 4 shows that roughly half of students believed that having no marks available for the questions made the tests *more* stressful.

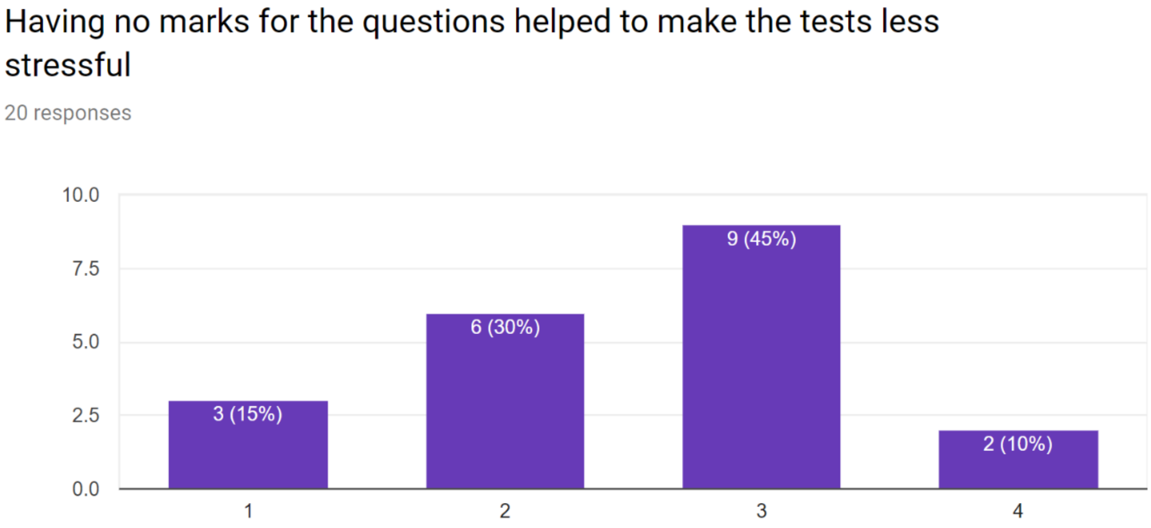

Figure 4: Strongly disagree (1) to Strongly agree (4)

The power of qualitative data was shown here, by introducing potential new ideas that I had not initially thought about. One student added that, for them, marks indicate the “amount of significant points to be made in my answers”. Three other students shared this view, with one adding that the marks “can help me see better how well I know certain topics”. This certainly did not line up with Butler’s (1988) findings about feedback. However, it could be that, even though the students seemed not to like it, it was useful for them to focus on their own progress. These findings suggest that my interpretation of ‘low-stakes’ could be improved. Given the age-range that I teach (16-18), it could be that my students are desensitised to marked tests from school. Perhaps the ‘stakes’ of each test can be lowered simply by having a larger number of tests.

The collection of data for this project was disturbed by a college trip happening on the same day as the first test was scheduled. I decided to postpone the test by a few days in order to maximise the number of students who received it. However, in doing so, I introduced some avoidable variation into the study. There is no way of knowing if this affected the results. If I had access to more students for the project, my sample size would have been high enough to make the lower attendance for the first test negligible.

Also during the data collection phase, in order to save time, I had the students peer-mark each other's assessments and recall tests. However, when collating the data, it became clear that most of the assessments had been marked incorrectly. After marking the 80 assessments myself, I saw that only 14 of them had been marked without error. This meant that roughly 80% of the assessments were marked incorrectly, more often too leniently. The average difference in marks awarded by a peer compared with the teacher was 9.46% across all 80 assessments. The potential implications of this error are worrying. For example, if a student's answer on a test was not detailed enough to gain full marks, but their peer-marker awarded them full marks, the student would have had no reason to believe they could not perfectly answer that question. This might misguide their revision away from the topics they need to work on. Perhaps more work needs to be done to simplify the mark schemes that are used by students, or the amount of peer-marking needs to be reduced.

This discovery may highlight an issue with my method. For the active group, I encouraged students to peer-mark the regular tests. This was in order to save time, and to contribute to them being 'low-stakes'. However, given the potential marking error, the tests may have been less effective than they would have been if peer-marking was not used. In my student survey, everyone in the active group said that they thought the tests led them to a better understanding of what areas they needed to work on. Yet, if their tests were marked incorrectly, they may have been led to an incorrect understanding. If I were to carry out this study again with more resources, I would mark the regular tests myself too. Not only would I be able to measure any marking error in them, but I would be able to compare performance in the tests to overall improvement.

Conclusions and Recommendations

The main result from this project was that introducing regular knowledge-recall tests to Computer Science A-level lessons may have caused an increase in performance of 14.11% compared with what would have been expected. My findings were consistent with the literature. Based on my research and subject specialist interview, the issue is certainly real, and these tests seem to be a promising strategy for improving the situation. Students appeared to be mostly happy with the tests, although they said they would have preferred a clear number of marks available for questions. My results seem to indicate that continuing with additional recall testing would be both beneficial and preferable to most students, and so I recommend that my department at the college do so. The main bulk of the work here would be in creating and compiling sets of tests. To lighten the burden, this work could be shared between teachers in the department and spread over a long period of time.

However, my main recommendation to the college is to broaden and continue my investigations. I would like to see more classes studied and more topics tested, over a longer timescale. Furthermore, there would be merit in exploring the accuracy of peer-marking in the department and the effect it has on students when done poorly, especially as the department makes extensive use of the peer-marking approach. The effects of low attendance on outcomes in Computer Science should also be investigated.

At its first conception, this project was just about measuring a cause-and-effect relationship between recall testing and examination performance. However, having now been through part of the action research process, I can see that this project has been exactly that: *a part of the process*. I now understand what McNiff & Whitehead (2009, p.41) meant by saying action research "is not about aiming for behavioural outcomes (a feature of traditional research) but about generating new, interesting questions that open up new possibilities”. From it I have not only learnt something new about how to improve my own practice, but I now have several further areas for investigation in the future.

References

Assessment and Qualifications Alliance. (2018). *AQA Enhanced Results Analysis. Marks analysis: skills and topics analysis*. Retrieved April 12, 2019, from https://results.aqa.org.uk /eAQA_ERA

Assessment and Qualifications Alliance. (2019). *AS and A-level Computer Science Specification: Specifications for first teaching in 2015*. Retrieved April 12, 2019, from https://filestore.aqa.org.uk/resources/computing/specifications/AQA-7516-7517-SP-2015.PDF

Bjork, R. A. (1994). 'Memory and metamemory considerations in the training of human beings'. In Metcalfe, J., & Shimamura, A. (Eds.), *Metacognition: Knowing about knowing* (185–205). Cambridge, MA: MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). 'Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning'. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(1), 59-68.

Butler, R. (1988). 'Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance'. *British journal of educational psychology*, 58(1), 1-14.

British Educational Research Association. (2018). *Ethical Guidelines for Educational Research*, fourth edition, London. Retrieved April 16, 2019, from https://www.bera.ac.uk/ researchers-resources/publications/ethicalguidelines-for-educational-research-2018

Carr, W., & Kemmis, S. (1986). *Becoming critical: education knowledge and action research*. Deakin University Press.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology*. Translated by Ruger, H. A., & Bussenius, C. E. (1913). New York: Teachers College, Columbia University. Retrieved April 18, 2019, from http://nwkpsych.rutgers.edu/~jose/courses/ 578_mem_learn/2012/readings/Ebbinghaus_1885.pdf

Hattie, J., & Timperley, H. (2007). 'The power of feedback'. *Review of educational research*, *77*(1), 81-112.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating Training Programs: The Four Levels Third Edition*. San Francisco: Berret-Koehler Publishers Inc.

Likert, R. (1932). 'A Technique for the Measurement of Attitudes'. *Archives of Psychology*, 22(140), 5-55.

McLeod, S. (2017). *Experimental design*. Retrieved April 14, 2019, from https://www.simplypsychology.org/experimental-designs.html

McNiff, J. (1984). ‘Action research: a generative model for in-service support’. *Journal of In-Service Education*, 10(3), 40-46.

McNiff, J. (2002). *Action research for professional development: Concise advice for new action researchers.* Retrieved April 12, 2019, from http://www.jeanmcniff.com/ar-booklet.asp

McNiff, J., & Whitehead, J. (2009). *You and your action research project, 3rd Edition*. Routledge.

McNiff, J., & Whitehead, J. (2011). *All you need to know about action research, 2nd edition.* SAGE Publications.

Office for Qualifications and Examinations Regulation. (2014). *GCE Subject Level Guidance for Computer Science*.

Roediger III, H. L., & Karpicke, J. D. (2006). ‘Test-enhanced learning: Taking memory tests improves long-term retention’. *Psychological science*, *17*(3), 249-255.

The Royal Society. (2017). *After the reboot: computing education in UK schools*.

Sargent, C., & Hillyard, D. (2017). *craigndave: Our pedagogy – Challenge*. Retrieved April 18, 2019, from http://craigndave.org/our-pedagogy/challenge/

Schön, D. (1983). *The Reflective Practitioner: How professionals think in action*. London: Temple Smith.

Sentence, S., & Humphreys, S. (2018). ‘Understanding professional learning for Computing teachers from the perspective of situated learning’, *Computer Science Education* [online], 28(4), 345-370. Retrieved April 17, 2019, from https://doi.org/10.1080/08993408.2018.1525233

Stenhouse, L. (1975). *An introduction to curriculum research and development*. Heinemann.