E-assessment for learning? Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics.

Thesis

Sally Elizabeth Jordan, BSc (Hons), FHEA, MInstP

E-assessment for learning?

Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics.

Submitted for the degree of Doctor of Philosophy by Published Work

(Part 1: Covering paper only)

Department of Physical Sciences

The Open University

1st August 2014

Contents

Acknowledgements

My adventures in e-assessment began early in 2001, when I introduced myself to Phil Butcher. Our joint work in developing assessment fit for the Open University module *Maths for science* (S151) built heavily on question types and behaviours that Phil had developed for *Discovering science* (S103), working with Stuart Freake, S103's production chair. I was truly "standing on the shoulders of giants". Phil has been involved in each subsequent chapter of my work in e-assessment, sharing my enthusiasm, gently pointing out when my ideas have been unrealistic, and understanding my belief in the importance of pedagogic research and evaluation. I am extremely grateful to Phil, Stuart and other colleagues in the Science Faculty (in particular Shelagh Ross, Pat Murphy and Isla McTaggart), Learning and Teaching Solutions (LTS) (in particular Greg Black and Spencer Harben) and IT Development (in particular Tim Hunt and Sam Mitchell), for their work in writing questions, designing systems, and co-authoring papers. I am also grateful for the funding from the e-OU Strategy Group that enabled our early development work.

The assessment strategy that we devised for S151 and its theoretical underpinning soon came to the attention of Steve Swithenby and Graham Gibbs, and I became part of the Formative Assessment in Science Teaching (FAST) project, funded by the Fund for the Development of Teaching and Learning (FDTL). Later, I was fortunate enough to have teaching fellowships with two of the Open University-led Centres for Excellence in Teaching and Learning (CETLs). Most of my work on short-answer free-text questions was supported and funded by the Centre for Open Learning of Mathematics, Science, Computing and Technology (COLMSCT), whilst COLMSCT and the Physics Innovations Centre for Excellence in Teaching and Learning (pi-CETL) jointly sponsored much of my analysis of student engagement with e-assessment. I gratefully acknowledge the financial support that the CETLs gave me, providing some respite from a busy "day job", but what I remember most is the sense of academic community, with sparky conversations with other teaching fellows, support staff, and the directors, Steve Swithenby and Bob Lambourne. I am also

grateful to Tom Mitchell of Intelligent Assessment Technologies Ltd. and to Barbara Brockbank who had a related COLMSCT teaching fellowship.

I now work in association with eSTEeM, with first Steve Swithenby and then Nick Braithwaite as Science Director, and endless support from Diane Ford, who also transferred from COLMSCT and is now eSTEeM Manager. Steve meanwhile, kindly agreed to act as adviser for this PhD submission, in which role he has continued to challenge and encourage me in equal measure, helping me to use the publications to go beyond mere reportage to give a message that, hopefully, will have some impact as well as telling a story.

I am extremely grateful to the staff tutors and assistant staff tutors who have covered for my routine work, giving me time to produce publications and this covering paper. I am also grateful to the module teams whose assessment strategies have fallen under my scrutiny, and to many associate lecturers and students for their direct involvement in some of the projects reported here, and for much wider inspiration.

Alongside Phil Butcher and Steve Swithenby, the third person who has been there from beginning to end of this enterprise is my husband, Richard Jordan. Our children were still living at home when I started work on S151 and its assessment, and Richard, Michael and Helen, in recent years along with Michael and Helen's spouses Heather and Tom, have all had to put up with my lack of time for domestic tasks, and with holidays where my work came too. I'd like to thank them all for their support, encouragement and love. However my family's help with this submission has gone way beyond the cooking and cleaning. Richard and Helen (now Dr Helen Ogden and at the University of Warwick's Department of Statistics) were employed as consultants for some of the data analysis, and co-authored Publication 7 with me. More generally, they have acted as unpaid advisers, especially with anything related to IT or statistics. Richard has produced most of the figures in this submission and has checked every word that I have written. Thank you!

Sally Jordan

Abstract

This submission draws on research from twelve publications, all addressing some aspect of the broad research question: “Can interactive computer-marked assessment improve the effectiveness of assessment for learning?” The work starts from a consideration of the conditions under which assessment of any sort is predicted to best support learning, and reviews the broader literature of assessment and feedback before considering the potential of computer-based assessment, focusing on relatively sophisticated constructed-response questions, and on the impact of instantaneous, tailored and increasing feedback. A range of qualitative and quantitative research methodologies are used to investigate factors which influence the engagement of distance learners of science with computer-marked assessment and computer-generated feedback.

It is concluded that the strongest influence on engagement is the student’s understanding of what they are required to do, including their understanding of the wording of assessment tasks and feedback. Clarity of wording is thus important, as is an iterative design process that allows for improvements to be made. Factors such as cut-off dates can have considerable impact, pointing to the importance of good overall assessment design, and more generally to the power and responsibility that lie in the hands of remote developers of online assessment and teaching.

Four of the publications describe research into the marking accuracy and effectiveness of questions to which students give their answer as a short phrase or sentence. Relatively simple pattern-matching software has been shown to give marking accuracy at least as good as that of human markers and more sophisticated computer-marked systems, provided questions are developed on the basis of responses from students at a similar level. However, educators continue to use selected-response questions in preference to constructed-response questions, despite concerns over the validity and authenticity of selected-response questions. Factors contributing to the low take-up of more sophisticated computer-marked tasks are discussed.

E-assessment also has the potential to improve the learning experience indirectly by providing information to educators about student engagement and student errors, at either the cohort or individual student level. The effectiveness of these “assessment analytics” is also considered, concluding that they have the potential to provide deep general insight and an early warning of at-risk students.

PART 1 Covering paper

1. Introduction

The interaction between assessment and learning [can be] likened to a three-legged race, in which neither partner can make progress without the other's contribution. (Harding & Raikes, 2002, p. 1)

This submission for the degree of PhD by Published Work draws together findings from work conducted over a period of more than twelve years, with the aim of exploring the potential of innovative forms of e-assessment to enhance learning. The twelve publications are wide-ranging, but all of them address some aspect of the broad research question:

Can interactive computer-marked assessment improve the effectiveness of assessment for learning?

The impact of e-assessment on learning can be measured by considering, directly, the ways in which students engage with computer-marked assessment and computer-generated feedback, and the factors, whether affective, technical or pedagogic, that influence that engagement. However, e-assessment also has the potential to improve the learning experience indirectly by providing information to educators about student engagement and student errors, at either the cohort or individual student level, and the effectiveness of these "assessment analytics" (Ellis, 2013) is also considered. In addition, conclusions are drawn about the appropriate scope of computer-marked assessment and the barriers to wider uptake.

My thesis is framed in terms of four slightly more specific research questions:

1. How do students engage with computer-marked assessment and what factors affect this?
2. How do students engage with computer-generated feedback and what factors affect this?
3. What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement?

4. What is the scope of computer-marked assessment in supporting learning and what are the barriers to wider take-up of sophisticated computer-marked tasks?

Chapter 3 considers these questions in turn. I can only claim to have considered a subset of the possibilities raised by each question and my work has been done in one particular context, using the e-assessment systems available to me. However the implications of my findings are wide-ranging, sometimes going beyond the confines of computer-marked assessment and computer-generated feedback to inform more general practice in assessment and learning design.

I have made extensive use of e-assessment in my teaching and have broad and longstanding interests in the use of assessment to support learning, in particular for distance learners of science. My thesis reflects this; it comprises a data-driven analysis of the practice of computer-based assessment, embedded within the theory of assessment for learning.

1.1 Structure of this covering paper

The requirements of a PhD by published work are that the covering paper should:

a) provide a summary of each publication;
b) outline the interrelationship between the publications;
c) give a critical review of the current state of knowledge and research in the field, and indicate how the candidate's work has contributed to the field;
d) comment on the reception of the publications, as indicated by citations and reviews, and the standing of the journals in which they were published.

Chapter 2 of this covering paper is a critical review of the current state of knowledge and research in the field (requirement (c)) and refers to my publications as appropriate. Chapter 3, the heart of the covering paper, further explores my contribution to the field (requirement (c)) and the inter-relationships between the publications (requirement (b)) by addressing in turn the above research questions, whilst Chapter 4 draws conclusions and makes suggestions for e-assessment practice in the future. Chapter 5 gives, for each of the 12 publications in chronological order, a summary of the publication itself (requirement (a)) followed by a discussion of the context in which the

publication was written and the reception of the publication (requirement (d)). Chapter 5 concludes with an overview of my developing influence in the field of computer-marked assessment and of the impact of my work. The publications themselves are reproduced in Part 2 of this submission. For ease of reference, Part 2 starts with a summary of the key points from each of the publications that have been used as part of this submission.

1.2 Definitions

Wiliam and Black (1996) discuss the inter-relationship between summative and formative assessment. They point out that whilst Scriven (1967, p. 51) first used the phrase "formative evaluation", it was Bloom, Hastings and Madaus (1971, p. 117) who arrived at the definition of formative assessment as assessment whose purpose is to help students to improve, in contrast to summative assessment which is "for the purpose of grading, certification, evaluation of progress". Unfortunately the use of "formative" and "summative" has become confused in the literature. The terms as defined above are not opposites; Wiliam (2011) gives examples of summative assessment which is effective in helping students to improve, and throughout the Open University's 45-year history, summative tutor-marked assignments have included feedback designed to help students in their learning. In addition, many authors extend the definition of formative assessment to encompass assessment which influences learning in any way, with either positive or negative impact. Barnett (2007, p. 37) notes that:

> Summative assessment is itself "formative". It cannot help but be formative. This is not an issue. At issue is whether that formative potential of summative assessment is lethal or emancipatory. Does summative assessment exert its power to disrupt and control, a power so possibly lethal that the student may be wounded for life?

To avoid confusion, I am using "**summative** assessment" simply to mean assessment in which the mark that is awarded counts towards the overall outcome, in contrast to "**formative** assessment", in which any mark that is given does not count. For formative assessment's original purpose, as contributing towards learning in a positive sense, I prefer the phrase "**assessment for learning**

(AfL)", reputed to have been used for the first time by Harry Black (1986), then developed through the work of the Assessment Reform Group which commissioned an influential literature review from Paul Black and Dylan Wiliam (1998). Assessment for learning is discussed further in Sections 1.3.1 and 2.1.

It is also important to distinguish between summative and **thresholded** assessment. In the latter, a student is required to pass a certain threshold (either on an individual assignment, or on a number of tasks, perhaps with weighting applied between them) but once this threshold has been reached, the scores do not contribute to the overall score (see Section 1.3.2). I use the phrase "**diagnostic** assessment" to refer to diagnosis prior to a programme or module of study, again acknowledging that there is a wider meaning.

The meaning of the word "feedback" on assessed tasks is also the cause of much debate, with many authors now agreeing with Ramaprasad (1983) and Sadler (1989) that unless the information that is provided by an external agent (for example, a human marker or a computer) results in action from the student, it cannot really be described as feedback, with Kluger and DeNisi (1996) instead using the phrase "feedback intervention" for the information provided on assessment tasks. As a physicist, familiar with the concept of feedback as a cyclical process (for example regulating the temperature of my house), I have much sympathy with this view. However, simply for convenience and in line with everyday use, for most of this submission I describe the information provided to a student simply as **feedback**. The literature of what makes feedback effective is discussed further in Section 2.2.

"**E-assessment**"**,** according to its widest definition, includes any use of a computer as part of any assessment-related activity (JISC, 2006). Thus its scope includes the use of technology for the submission of work or to enable or delay the release of grades or feedback (Parkin, Hepplestone, Holden, Irwin & Thorpe, 2012), the use of audio feedback (e.g. Lunt & Curran, 2010), screencasts (Haxton & McGarvey, 2011) and e-portfolios (Shephard, 2009), as well as the assessment of student reflection in blogs (Sim & Hew, 2010) and of collaborative activity in wikis (Caple & Bogle, 2013) and forum discussions (Kirkwood & Price, 2008). These uses of technology all have potential

to improve learning, but they are not the focus of this thesis. My focus is on online computer-marked assessment, with instantaneous computer-generated feedback and an opportunity for the student to try the question again in the light of that feedback. At the Open University, these concepts are encapsulated by referring to this type of assessment as "**interactive computer-marked assessment**" with the individual quizzes being known as "**interactive computer-marked assignments (iCMAs)**". The use of iCMAs at the Open University is discussed further in Section 1.3.3 whilst Appendix A lists the acronyms and other abbreviations used in this submission.

1.3 Context

This section considers the context for the work described in this submission, focusing in turn on the significance of when the work took place (the early 21st century), the institutional setting (the UK Open University), and the assessment systems and e-assessment tools that were used. Section 1.3.1 discusses factors linked to the fact that the work was done at the beginning of the 21st century, including wider changes in higher education, the rapidly increasing availability of computers and access to the internet, and the increasing emphasis on assessment for learning. Section 1.3.2 outlines the implications for the work of the Open University's large student numbers, distance-learning environment and open entry policy. Finally, Section 1.3.3 discusses the use of assessment and e-assessment in the OU Science Faculty.

1.3.1 When the work took place: the early 21st century

Higher education at the turn of the 21st century was characterised by increasing student numbers and comparatively reduced resources (Kirkwood & Price, 2005). The use of computer-marked assessment was being considered quite widely for its potential to save resources, in particular the time taken for human marking. Furthermore, mass higher education was "squeezing out dialogue with the result that written feedback, which is essentially a monologue, [was] now having to carry much of the burden of teacher–student interaction" (Nicol, 2010, p.503). Thus, the interactive potential of online assessment with instantaneous computer-generated feedback was also becoming important.

The changes to higher education funding that occurred in England in 2012 added urgency to the already present drivers of competition and consumerisation, with recognition that students deserve and are likely to demand a high-quality experience. Universities place great importance on retaining students through their programmes of study, and there is widespread concern about poor scores for "feedback" in the National Student Survey and similar national surveys in other countries (Yorke, 2013).

The growth in the use of e-assessment has relied on the recent rapid increase in access to computers and the internet. In particular, although the driver for the introduction of interactive computer-marked assessment into the Open University module *Maths for science* was to deliver instantaneous and tailored feedback (Publications 1, 2 and 3), it was only possible to move to an online system in 2002 because by then it was reasonable to assume that distance-learning students would be able to access the internet from their remote locations in order to complete their assignments (Publication 11, p. 3).

Computers are now ubiquitous and this means that "as students become more familiar with the presence of computers in all aspects of their lives, it is likely that their reaction to e-learning and e-assessment will alter; they may cease to regard the use of technology in learning and assessment as anything special" (Publication 8, p. 819, drawing on Conole, de Laat, Darby & Dillon, 2006; Beevers et al., 2010). Earlier work, for example Brosnan's (1999) study of the impact of computer anxiety, is likely to be of limited relevance. However, it should be remembered that there remains a considerable "digital divide" both between and within countries and this can have substantial implications for education (Eynon, 2009). It should also be recognised that not all students react to the use of computers in the same way (Hewson, 2012) and "whilst some learners use technology to good effect to support their studies, others find it a barrier and a distraction" (Sharpe, 2009, p. 181).

The past few decades have also been characterised by a growing realisation of the importance of assessment for learning, further discussed in Section 2.1. It is noteworthy that two of the Centres for Excellence in Teaching and Learning (CETLs) (the Assessment Standards Knowledge exchange

(ASKe) at Oxford Brookes University and the Centre for Excellence in Teaching & Learning in Assessment for Learning (CETL AfL) at the University of Northumbria) had a specific focus on assessment for learning. At the Open University, e-assessment was one of the themes of both the Physics Innovations Centre for Excellence in Teaching and Learning (piCETL) (Jordan, 2010) and the Centre for Open Learning of Mathematics, Science, Computing and Technology (COLMSCT) (Butcher, Swithenby & Jordan, 2009).

1.3.2 The institutional context: the UK Open University

Although wider conclusions are drawn, the research described in this submission took place at the UK Open University, and the students considered are Open University students, frequently completely or relatively new Science Faculty students. This section discusses the significance of the institutional context.

Most Open University modules attract relatively large student numbers, with registrations on the modules considered in the publications ranging from a few hundred to around 4000 students per year. Most modules have an expected life-time of 8-10 years. This means that the effort spent in developing high-quality computer-marked questions is worthwhile and large numbers of student answers are available for use in developing the answer matching when necessary (Publications 4, 5 and 9). Furthermore, sufficient data are available to support robust analysis and to provide incontrovertible evidence of student errors (Publication 12) and factors affecting student engagement (Publication 10).

Students at the Open University are distance leaners, frequently though not exclusively studying part-time alongside employment and/or caring responsibilities. The Open University's model of supported distance learning means that students have a personal tutor or study adviser, with some (usually optional) face-to-face or online tuition available in support of the printed and onscreen study materials. However, much Open University study takes place in isolation, frequently in a student's own home, but also sometimes in remote locations (e.g. military bases, oil rigs) and anywhere in the world.

Most Open University modules include several tutor-marked assignments (TMAs), which both contribute to students' overall continuous assessment score and act as a medium through which tutors can provide detailed and personalised feedback on their students' work. In addition, Open University tutors are available to provide one-to-one support to their students. However computer-generated feedback may have a particular role to play in offering instantaneous support to distance learners. This is not to deny the usefulness of e-assessment in all settings, discussed in Sections 2.3 and 2.4, rather to point out that the challenges facing distance-educators are particularly severe, and that lessons might be learnt for the wider sector, especially since remote online learning is becoming more common, in both formal settings and informally, for example in Massive Open Online Courses (MOOCs).

Much of the work described in the publications addresses the needs of completely or relatively new Science Faculty students. The Open University's "open" mission means that these students have a wide range of ages and backgrounds, with a sizeable proportion having previous educational qualifications below those normally required for university entry; Section 5.10.2 describes the student populations on two contrasting modules. Some students therefore come to the Open University lacking in time, study skills and confidence; the careful use of computer-based assessment and computer-generated feedback has the potential to make a significant difference to their study experience. However, working with adult novice learners also carries a risk; emotional reactions to assessment amongst adult part-time students have been described as the cause of "trauma and euphoria" (Cramp, Lamond, Coleyshaw & Beck, 2012, p. 516).

The Open University's curriculum has a modular structure, with a considerable amount of student choice over the modules that can be included in many degrees. The changes to higher education funding that took place in 2012 resulted in some substantial changes to the curriculum. Appendix B describes the introductory Science Faculty curriculum, pre- and post-2012, and also explains the codes used for the modules and module presentations that are mentioned elsewhere in this submission.

Sally Jordan

1.3.3 Assessment and e-assessment in the Open University Science Faculty

The Open University Science Faculty is in the process of introducing what is being called "formative thresholded" continuous assessment (Jordan et al., in press), in place of previous practice in which TMAs and iCMAs were notionally summative, but excellent performance in these assignments counted for little in the face of poor performance in a module's examinable component. "Formative" here explicitly emphasises that the primary purpose of the assessment is to support learning, whilst "thresholded" indicates that students are required to meet a threshold of some sort in the continuous assessment; their final grade is then determined by the end-of-module assessment or examination alone. Two different forms of thresholding are in use; for example in *Scientific and investigative skills in science* (S141), students are required to achieve a weighted mean for the TMAs and iCMA of at least 40%, whilst for *The quantum world* (SM358), students are required to demonstrate their engagement by achieving at least 30% in seven out of the ten available TMAs and iCMAs. In the latter case, iCMA questions can be repeated as many times as a student wishes. The move to formative thresholded assessment post-dates most of the publications, but some of the data analysis reported in the publications has been repeated for formative thresholded iCMAs and a few significant results are reported in this covering paper.

Open University TMAs themselves make use of e-assessment, albeit different types of e-assessment from the one considered in detail in this submission; TMAs may assess course components that are delivered electronically (e.g. a virtual fieldtrip) as well as online activities involving information literacy and group working skills (Kirkwood & Price, 2008). TMAs were initially submitted on paper through the postal system, and returned from tutor to student via the Open University's Milton Keynes headquarters, again by post, but they are now usually submitted and returned electronically via an eTMA file-handling system. This enables TMAs to be submitted and returned with a minimum of delay, saves printing and postage costs, and facilitates the use of the plagiarism detection software CopyCatch and Turnitin®. Despite problems with inputting and commenting on diagrams, graphs and symbolic notation (Freake, 2008), eTMAs have been well received by students (Publication 6, p. 149).

The interactive computer-marked assessment whose impact is considered in this submission has been developed using the Open University's OpenMark system (Publication 6, p. 149-151, and further discussed in Section 2.6.2) and its immediate precursor (Publications 1, 2 and 3). OpenMark-type questions were first delivered to Open University students by CD-ROM, alongside multimedia activities on the module *Discovering science* (S103) in 1998 (Freake, 1999; Lawless & Freake, 2001). The move to an online operation for *Maths for science* (S151) in 2002 meant that student interactions could be logged, enabling questions to be used summatively and also enabling the data analysis that underpins most of the publications in this submission. In the early days, the *Maths for science* questions ran as Java applets, raising concerns that some students were unable to access the questions because of difficulties in downloading the Java Runtime Environment (Publication 1, p. 9). OpenMark, introduced in 2005, is an entirely web-based system and this download is no longer necessary.

Most, but not all, of OpenMark's functionality is now available in the Moodle assessment system (Butcher, 2008; Hunt, 2012). At the Open University, all iCMAs are run from within the Moodle virtual learning environment, whether questions are written in OpenMark or Moodle. Moodle's Gradebook enables students to monitor their own progress and tutors to monitor the progress of their group (Publication 6, p. 149), as discussed in Section 2.9.

The modules whose use of computer-marked assessment is described in the publications have used computer-marked assessment in a range of summative, formative and diagnostic settings (as defined in Section 1.2). Even in summative use, the focus has always been on assessment for learning, so students are allowed multiple (usually three) tries at each question, with increasing feedback[1].

[1] Note that in recent years, a distinction has been made between "try" (the opportunity to alter an answer in response to feedback) and "attempt" (the opportunity, only sometimes available, to repeat the whole question, perhaps in a different variant). Unfortunately most of the publications use the word "attempt" for what would now be described as a "try".

1.4 Research methods

As befits a submission with a broader coverage than a conventional PhD thesis, a range of research methodologies has been employed. Qualitative methodologies have included questionnaires, interviews and direct observation of students in the Open University Institute for Educational Technology's usability laboratory. However, the submission has a quantitative focus, with many of the results and conclusions being derived from analysis of many thousands of student interactions with computer-marked assessment on a range of Open University modules. Whilst recognising that student opinion is important, this focus has enabled the research to go beyond self-reported behaviour to observe actual student engagement with assessment.

The research instruments and data analyses are described in more detail in Appendix C. All questionnaires and interviews followed approved institutional protocols, with the support of the Open University's Student Research Project Panel. The usability laboratory work had the agreement of both the Student Research Project Panel and the Human Research Ethics Committee, and conventional usability laboratory protocols were followed (e.g. Stone, Jarrett, Woodroffe & Mincoha, 2005). Participants interacted with the questions without assistance, but their interactions were observed live and recorded for subsequent analysis. A verbal think-aloud protocol was used, whereby the participants were asked to talk about what they were thinking and doing, and after the observation, each participant was asked to comment retrospectively on the reasons for their actions. However, the emphasis of the evaluation was on what the students actually did rather than what they said they did (Publication 8, p. 820).

The data analysis reported in the publications relied on the fact that the OpenMark system records every response that is entered plus the student's identifier (only used to link data and then destroyed), the time, whether the response was a first, second or third try, and whether it was marked as correct or incorrect. The system also records "score" for each question, the browser(s) used and the time(s) at which a browser was opened. The only significant weakness in the analysis related to time spent in answering questions; the time at which a student first views a

question is not recorded, which means that the time spent considering the question before giving a first response cannot be calculated. The difficulty is exacerbated by the fact that students are increasingly in the habit of leaving browsers open whilst doing some completely different task, so the recorded time between interactions is also of limited use.

2. Literature review

Whilst this literature review focuses on research into the effectiveness of computer-marked assessment and computer-generated feedback, it starts by considering the underpinning assessment and feedback literature (Sections 2.1 and 2.2). Moving on to consider e-assessment specifically, the review summarises evidence for and against the use of computer-based assessment (Section 2.3) and considers the impact of computer-generated feedback (Section 2.4). The pros and cons of selected-response and constructed-response assessment items are considered (Section 2.5), as well as the use of more sophisticated question types (Section 2.6), specifically those requiring answers in the form of a sentence or an essay (Section 2.7). Finally, Section 2.8 reviews ways in which computer-marked assessment items of all types can be used within an assessment strategy to support learning, and Section 2.9 reviews the literature of the recently established field of learning analytics. Sections 2.3, 2.6, 2.7, 2.8 and 2.9 draw heavily on my 2013 historical review of e-assessment (Publication 11), whilst taking a more selective and critical approach.

2.1 Conditions under which assessment supports learning

"Whether we like it or not, assessment has a profound impact on learning" (Publication 6, p. 147). Assessment can become "the tail that wags the dog" (Dysthe, 2008, p. 17) or define a "hidden curriculum" (Snyder, 1971), which may be different for different students because of their differing previous experiences (Sambell & McDowell, 1998). Boud (1995) has pointed out that whilst students may be able to escape the effects of poor teaching they cannot escape the effects of poor assessment. Yorke (2003) reasons that both formative and summative assessment can discourage students from developing to their full potential, because students can become demotivated and also over-dependent on their tutors. Gibbs (2006, p. 15) reminds us that students study what is assessed or "more accurately, what they perceive the assessment system to require", whilst Morgan and O'Reilly (1999, p. 46) take a more optimistic stance:

Learners will make strategic decisions about whether to skip sections, or even by-pass the study materials altogether, based on their perceptions of assessment requirements. This is not necessarily a problem – indeed, from an open learning perspective, it is highly desirable to encourage independent and lifelong learning skills.

Reviews of the literature (e.g. Black & Wiliam, 1998; Gibbs & 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 well known of which were proposed by Gibbs and Simpson (2004-5) and Nicol and Macfarlane-Dick (2006). Other frameworks and lists of hints for effective assessment either explicitly adapt ideas from Gibbs and Simpson (2004-5) and Nicol and Macfarlane-Dick (2006) (e.g. Re-Engineering Assessment Practice in Scottish Higher Education, 2007) or have similar themes (e.g. Race, Brown & Smith, 2005; National Union of Students, 2010).

Gibbs and Simpson's ten conditions under which assessment supports students' learning can be grouped into those considering the quantity, distribution and quality of student effort; those considering the quantity, timing and quality of feedback; and those considering the student response to that feedback. The conditions are reproduced here for ease of reference later in this covering paper:

1. Sufficient assessed tasks are provided for students to capture sufficient study time;
2. These tasks are engaged with by students, orienting them to allocate appropriate amounts of time and effort to the most important aspects of the course;
3. Tackling the assessed task engages students in productive learning activity of an appropriate kind;
4. Sufficient feedback is provided, both often enough and in enough detail;
5. The feedback focuses on students' performance, on their learning and on actions under the students' control, rather than on the students themselves and on their characteristics;
6. The feedback is timely in that it is received by students while it still matters to them and in time for them to pay attention to further learning or receive further assistance;

7. Feedback is appropriate to the purpose of the assignment and to its criteria for success;
8. Feedback is appropriate, in relation to students' understanding of what they are supposed to be doing;
9. Feedback is received and attended to;
10. Feedback is acted upon by the student.

In line with Boud (2000)'s definition of sustainable assessment as that which both meets the immediate needs of a course and establishes a basis for students to undertake their own assessment activities in the future, Nicol and Macfarlane-Dick (2006) emphasise the importance of reflection, dialogue and self-regulation. They also include the condition that feedback should be provided to teachers, to improve future teaching, which is of direct relevance to my third research question: "What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement?"

In more detail, Nicol and Macfarlane Dick's conditions (p. 205) state that good feedback practice:

1. Helps clarify what good performance is;
2. Facilitates the development of self-assessment in learning;
3. Delivers high-quality information to students about their learning;
4. Encourages teacher and peer dialogue around learning;
5. Encourages positive motivational beliefs and self-esteem;
6. Provides opportunities to close the gap between current and desired performance;
7. Provides information to teachers that can be used to shape teaching.

Although the focus of this thesis is on computer-marked assessment and computer-generated feedback, whilst the Gibbs and Simpson and Nicol and Macfarlane-Dick conditions apply more generally, there are clear links between my research questions and many of the conditions. So, for example, if a factor is identified that leads to better engagement with computer-marked assessment (Research Question 1), then Gibbs and Simpson's Condition 2 is likely to have been

met. I will return to Gibbs and Simpson's and Nicol and Macfarlane-Dick's conditions throughout this covering paper, as I reflect on progress towards answers to my research questions.

2.2 Feedback: All that fuss but what's the impact?

Price, Handley, Millar and O'Donovan (2010, p. 27) state that "much staff effort goes into producing assessment feedback, but very little effort is made to examine its effectiveness". However, as Yang and Carless (2013, p. 285) point out, there is a "rapidly burgeoning literature on feedback in higher education" and in fact research into the effectiveness of feedback interventions has been going on for around 100 years, with Kluger and DeNisi, back in 1996, making the point that such research has not produced much useful progress. Price et al.'s (2010) contention should therefore perhaps be reframed as "much staff effort goes into producing assessment feedback and in examining its effectiveness, but the outcomes of that examination are contradictory and poorly understood". A selection of the literature on assessment feedback is reviewed in this section, with a specific focus on computer-generated feedback coming later, in Section 2.4.

There have been a number of reviews of the feedback literature in recent years and each of these has made important points. Evans (2013) is right to be critical of research methodology reported in many published papers, with issues including small sample sizes and assumptions about causality that are not substantiated. Struyven, Dochy and Janssens (2005) emphasise the importance of making a distinction between a student's construction of reality and their actions, whilst Jonsson (2013) points out that there is a difference between feedback that students claim to find useful and feedback that they actually use. Shute (2008) reports conflicting evidence about what works and what does not. Li and De Luca (2014) highlight students' and tutors' different perceptions of feedback, whilst Hattie and Timperley (2007, p. 13) make the true but frequently overlooked observation that "providing and receiving feedback requires much skill by students and teachers".

Where evidence as to the effectiveness of feedback has been attended to, it frequently indicates a problem: Feedback does not seem to have been as effective as educators might have hoped. The effect of this is that "many current practices waste both student learning potential and staff resources" (Merry, Price, Carless & Taras, 2013, p. xx (i.e. p. 20)). Sadler (1989, p. 119) comments on "the common but puzzling observation that even when teachers provide students with valid and reliable judgements about the quality of their work, improvement does not necessarily follow. Students often show little or no growth or development despite regular, accurate feedback". Beth Crisp (2007) found only limited evidence of students making changes to their work which were in line with what the feedback suggested, and in a meta-analysis covering 131 papers, Kluger and DeNisi (1996) report the shocking result that in 38% of cases the feedback appeared not to have been helpful, resulting in a decreased performance. This supports the hypothesis of Bangert-Drowns, Kulik, Kulik and Morgan (1991, p. 214) that "feedback might have no benefit for learning or actually be *mathemathantic*, a term coined by Snow (1972) to describe processes that 'kill' learning."

Feedback, even when conceived simply as information given to students, can have different purposes, be of different types, and focus on different things (e.g. on the student or on their work) and, unsurprisingly, these differences appear to influence the feedback's effectiveness. Section 2.2.1 considers the different ways in which feedback can be classified, then Section 2.2.2 considers the evidence relating to factors that increase or decrease the effectiveness of feedback. Section 2.2.3 addresses the importance of alignment between educators and students when considering the nature and purpose of feedback, and points towards a model in which responsibility for the use and usefulness of feedback is in the hands of students not tutors, in line with several of Nicol and Macfarlane-Dick's conditions.

2.2.1 Classifying feedback

Price et al. (2010) classify the roles associated with feedback into five broad categories: correction; reinforcement; forensic diagnosis [the identification of problems]; benchmarking; and

longitudinal development. Carless (2006) identifies that feedback might be: advice for improving the current assignment; advice for future assignments; a means of explaining or justifying a grade; or simply a ritual. Brown and Glover (2006) interviewed students and found that even when students did not act on feedback, they still liked receiving it. This supports Kluger and deNisi's (1996) conclusion that people find feedback psychologically reassuring and so like to think that feedback interventions are helpful, even if they are not.

In a meta-analysis of 185 research studies on the effectiveness of feedback in higher education, Nyquist (2003) developed a typology of formative assessment, ranging from "weaker feedback only" (limited to a student's knowledge of their score), which was found to be least effective, to "strong formative assessment", in which students are given information about the correct results, some explanation and specific activities to undertake in order to improve, which was found to be most effective. Shute (2008) classifies the types of feedback that can be given as: verification of response accuracy; explanation of the correct answer; hints; and worked examples.

In terms of the focus of the feedback, Bangert-Drowns et al. (1991) classify feedback as: focusing on affective and motivational dimensions; focusing on self-regulated learning by cuing self-monitoring; or, most commonly, being informational. Hattie and Timperley (2007, p. 90) identify four different possible foci of feedback: feedback about the task (FT); feedback aimed at the processes used to create a product or complete a task (FP); feedback about self-regulation (FR); or feedback about the self (FS) e.g. "You are a great student". They argue that feedback about the self is the least effective. This point is discussed further in Section 2.2.2. In addition, Draper (2009b) identifies a difficulty for students in knowing which "regulatory loop" they should engage for corrective action after receiving feedback; should they, for example, just work harder or do they need to take more radical action?

2.2.2 Effective and ineffective feedback

Mutch (2003) calls for more research on how students respond to feedback, rather than on the feedback itself. If feedback is to be useful, first of all it needs to be understood. Walker (2009)

found that students were unable to understand 27% of comments on tutor-marked assignments. Whilst the difficulty on this occasion was caused in part by students being unable to read their tutors' handwriting, students can also encounter difficulties in interpreting the language that their tutors use, for example, Chanock (2000) found that 40% of students did not understand what their tutors meant when they said that their work required more "analysis". Weaver (2006) describes the problems caused as a result of comments that are too general or vague e.g. "you've got the important stuff right"; students want to know *what* it is that they got right, and wrong.

As mentioned in Section 1.2, Ramaprasad (1983) and Sadler (1989) recognise that the effectiveness of feedback is enhanced when the student acts on the feedback received, enabling them to reduce the gap between their previous understanding and what is required; indeed Ramaprasad (op. cit., p. 5) says "The information on the gap between the actual level and the reference level is feedback only when it is used to alter the gap". In this way, feedback becomes a process not a transaction (Havnes & McDowell, 2008, p. 118).

To enable students to act on the feedback whilst the assessment task is still fresh in their minds (Gibbs & Simpson, Condition 6), most authors agree that it is important for students to receive feedback quickly, though Bayerlein (in press) found that "timely" feedback may be as well received as "extremely timely" feedback, with no improvement in timeliness rating for feedback received in 2.5 days as opposed to 5 days.

Shute (2008, p.9) acknowledges that specific feedback is generally more effective than less specific feedback, but points out that:

> A related dimension to consider in generating feedback is one of length or complexity of the information. For example, if feedback is too long or complicated, many learners will simply not pay attention to it, rendering it useless. Lengthy feedback can also diffuse or dilute the message.

Race et al. (2005, p. 105) emphasise that feedback needs to be manageable, for both tutors and students.

It is important to consider the impact of emotion on the way in which students respond to feedback (Värlander, 2008), especially since emotion mediates cognition (Dowden, Pittaway, Yost & McCarthy, 2013). Winter and Dye (2004) suggest that students should be taught how to use feedback and to view it less emotionally; the idea of teaching students how to use feedback is gaining in popularity, but teaching students to respond unemotionally is likely to be a challenging task. Winter and Dye also highlight the importance of tutors being aware of the emotional impact of what they write.

One issue on which authors are divided is the role of praise within feedback. Whilst Weaver (2006) and Walker (2013) emphasise the importance of motivational comments on assessment tasks, Duncan (2007, p. 278) found "a preponderance of positive and encouraging comments on feedback sheets at the expense of clear advice on how to improve the quality of subsequent work". Dweck (1999); Rust, O'Donovan and Price (2005); and Draper (2009b) all identify a deeper problem with praise on assessment tasks; if students believe that they have done well because of high personal qualities (e.g. high intelligence) then when they do less well in future tasks, they feel stupid and their self-efficacy may be damaged. Shute (2008) suggests that feedback is best when it is unbiased and objective, and uses praise sparingly. Praise can direct a student's attention towards themselves and away from the task, leading to a decrease in the effectiveness of the feedback (Kluger & DeNisi , 1996).

Another contested point is whether the receipt of a numerical score impedes the effectiveness of feedback. Millar (2005) reports that students want formative feedback to include marks, whilst Smith and Gorard (2005) found that students who were not given marks did less well. Other authors have criticised Smith and Gorard's paper, saying that the reported effect was caused by the students receiving insufficient feedback. However, given the frequency with which students do not understand the feedback given, there is an argument in favour of providing a mark to indicate how well a student has done, particularly where sub-scores can indicate the areas

requiring further attention (Xiaoling & Xuning, 2013). There is a strong counter-argument: As early as 1913, Thorndike (p. 286), cited by Stobbart (2006, p. 142) said that "grades can impede learning". The advocates of "comment-only" marking argue that grades do not tell students how to move forward and may have a negative (de-motivational) effect; learning is likely to stop when a summative grade is awarded (Stobbart, 2006).

Several authors have shown that students are more likely to take note of feedback if it is surprising. Bevan, Badge, Cann, Willmott and Scott (2008, p. 7) quote a student who said: "If I expect a mark, low or high, and it's that, I don't really read the comments. If I get a mark that's really different from what I expected then I'll really read the comments." Jones and Gorra (2013), in a paper about feedback on demand, found that students were more likely to request additional feedback if the mark was different from what was expected, especially if it was lower than expected. Kulhavy (1977) and Kulhavy and Stock (1989) found that students took note if their confidence was high in an answer that turned out to be incorrect. The "hypercorrection effect", in which people remember the correct answer when they have received feedback on an incorrect response that was stated with high confidence, is hypothesised to work because the feedback is surprising (Butterfield & Metcalf, 2001, 2006; Fazio & Marsh, 2009). By contrast, Orsmond and Merry (2011) give an interesting example of a student who was so preoccupied with what she expected the feedback to say that she did not notice what it actually said; she expected the feedback to be about spelling, and reported that this was the case, but in fact spelling was not mentioned.

2.2.3 Shared understanding and sustainable feedback

Many authors emphasise the importance of a shared understanding between educators and students as to the nature and purpose of feedback (e.g. Carless, 2006; Orsmond & Merry, 2011; Adcroft & Willis, 2013), but this is frequently not present. The Formative Assessment in Science Teaching (FAST) Project found that, at two universities and in very different contexts, students

and lecturers simply had a different understanding of the word “feedback” (Publication 2, p.484; Holden & Glover, 2013, p. 13). This is discussed further in Section 3.2.

It is also important to ensure that students know how to use the feedback that is provided (Burke, 2009). Sadler (1998, p. 78) sums this up well:

> Students should also be trained in how to interpret feedback, how to make connections between the feedback and the characteristics of the work they produce, and how to improve their work in the future. It cannot be assumed that when students are “given feedback” they will know what to do with it.

Whilst deeper learning is generally considered to take place when students are intrinsically rather than extrinsically motivated (Harlen, 2006; Broadfoot, 2008), Higgins, Hartley and Skelton (2002) point out that the interrelationship between intrinsic and extrinsic motivation is complex and poorly understood. They suggest that students are “conscientious consumers”, interested in the feedback provided as well as their grade and attempting to learn from the feedback; however, they do not always know how to do so.

Draper (2009b) comments that students may simply want to know if they are making satisfactory progress. If they are on course to pass the module, perhaps they do not want to do anything differently, and so perhaps the oft-reported finding that students do not look at feedback comments but only at their mark (e.g. Orsmond, Merry & Reiling, 2005), is not unreasonable. “To both educationalists and teachers, this seems dysfunctional and the mark of a bad student. But is it?” (Draper, 2009b, p. 312). In focus group discussions, Scott (2014, p. 54) found that students see feedback as the means by which they can gauge their progress at each stage of the course and that “some students may want to know how they are tracking simply so as to ensure they pass the course, not necessarily in order to rank at the top of the class”.

Back in 1989, Sadler (p. 130), recognised that there was likely to be an “optimal gap” which would depend on the learner’s current state and aspiration, and that:

> If the learner perceives the gap as too large, the goal may be regarded as unattainable. The same gap (in absolute terms) may, however, provide a powerful stimulus for another motivated and confident student, who would not be put off by a sequence of initial failures. Conversely, if the gap is perceived as too small, closing it might be considered not worthy of any additional effort.

Wiliam (2008, p.276) echoes this view, pointing out that if the feedback persuades a student that "he has reached some sort of threshold and that further progress is unlikely, then the overall effect will be negative".

Nicol and Macfarlane-Dick (2006) emphasise the importance of self-regulated learning, moving away from a transmission model of assessment and feedback to one in which the student takes responsibility. The "Agenda for change" in feedback practice, established following a meeting of experts brought together by the Assessment Standards Knowledge exchange (ASKe) acknowledges that "if students do not learn to evaluate their own work they will remain completely dependent upon others" (Price, Handley, O'Donovan, Rust and Millar, 2013, p. 41). Robinson and Udall (2006, p. 93) observe:

> Our experience is that if the feedback process is driven largely by the teacher, even when integrated into the learning experience, then students fail to engage fully with the meaning of that feedback. Where the feedback process is driven by the learner's own enquiry, through critical reflection, their focus becomes the progress they are making towards the intended learning outcomes of the unit of study.

Hounsell (2007) developed Boud's (2000) idea of sustainable assessment to define sustainable feedback, emphasising the role of the student in the process, whilst the teacher is reconfigured from "someone who mainly provides comments on a student progress, to a position of supporting students to make their own professional judgements" (Carless, 2013, p.120). McArthur and Huxham (2013) envisage feedback as a "spiral", occurring at many points (not just associated with

assessment) involving two-way dialogue, with teachers learning from students as well as the other way round. They comment on the fact that "such a conception blurs the distinction between 'feedback' and 'teaching'" and see this as "one of its attractions" (McArthur & Huxham, op. cit., p. 93).

Laurillard's (2002) "conversational framework" incorporates the notion of intrinsic and extrinsic feedback where intrinsic feedback is an embedded and unavoidable outcome of many actions in the physical and social world, and other authors recognise "that valuable and effective feedback can come from varied sources" (Price et al., 2013, p. 41), not just from the student's tutor (Kluger & DiNisi, 1996). Bangert-Drowns et al. (1991, p. 215) point out that intentional feedback can be delivered through direct interpersonal action (teacher to student or student to student) or delivered by an intervening agent, in which category they include computers.

2.2.4 Review of Section 2.2 and implications for computer-generated feedback

Rust et al. (2005, p. 234) state:

> Of the whole assessment process, the research literature is clear that feedback is the most important part in its potential to affect future learning and student achievement. But just as with assessment as a whole, there are many weaknesses and problems with feedback practice.

Section 2.2 of this literature review has indicated the size and complexity of the literature surrounding assessment feedback, pointing towards some factors that appear to make feedback more effective, whilst acknowledging that there is still a great deal of uncertainty. If what we are doing is creating "moments of contingency" (Black & Wiliam, 2009) or enabling "catalytic assessment", the use of simple questions to trigger deep learning (Draper, 2009a), it seems probable that the detail of what is written as feedback may be less important than the way in which it is delivered and the way in which students react to it.

In several aspects, the literature reviewed in this section has pointed towards a role for computer-generated feedback, with potential advantages including rapid delivery (Gibbs & Simpson,

Condition 6) and a lack of inter-personal emotion (Gibbs & Simpson, Condition 5). It is also the case that lessons can be learnt from the general feedback literature for improving the effectiveness of computer-generated feedback. These points are discussed further in Section 2.4, leading the way to a consideration, in Section 3.2, of Research Question 2: How do students engage with computer-generated feedback and what factors affect this? First of all, Section 2.3 reviews general research into the use of computer-marked assessment.

2.3 Computer-marked assessment: Drivers and distractors

In the past 10 years there have been several reviews of the e-assessment literature and, in addition to Publication 11, the following all pay significant attention to computer-marked assessment: Ridgway, McCusker and Pead (2004); Sim, Holifield and Brown (2004); Conole and Warburton (2005); Nicol (2008); Ripley, Harding, Redif, Ridgway & Tafler (2009); Stödberg (2012).

E-assessment is a natural partner to e-learning (Mackenzie, 2003) offering alignment of teaching and assessment methods (Gipps, 2005). Computer-marked assessment can be made available to students anywhere in the world, and can, with care, improve access for students with certain disabilities (Ball, 2009). It can add diversity (Sim et al., 2004) and enjoyment (Publication 6, p. 153) to the assessment package offered to students, and assessment tasks can be made more authentic by linking to simulations and multimedia activities (Ashton & Thomas, 2006; Redecker & Johannessen, 2013).

Quizzes can be administered weekly (McDaniel, Anderson, Derbish & Morrisette, 2007) or even daily (Leeming, 2002) and, in formative use, even the simplest of multiple-choice quizzes can enable students to check their understanding of a wide range of topics, whenever and wherever they choose to do so (Bull & McKenna, 2004). Students can study at their own pace, repeating questions whenever they want to (Sim et al., 2004), sometimes with different variants of the questions (Publication 7). This provides an excellent opportunity for ipsative assessment, defined by Hughes (2011, p. 353) as assessment which compares a student's existing performance in an assessed task with the same student's previous performance. Hughes points out that ipsative

assessment can have a motivational effect and invites an ongoing dialogue between student and tutor (Nicol & Macfarlane-Dick, Conditions 5 and 4). Feedback can be provided instantaneously (Gibbs & Simpson, Condition 6) and can be tailored to particular misunderstandings, with reference to relevant module materials (Nicol & Macfarlane-Dick, Condition 3). This provides, even for students who are studying at a distance, a virtual "tutor at the student's elbow" (Publication 3, p. 125).

Several authors highlight the potential of online assessment to engage and motivate students (e.g. Marriott, 2009; Holmes, in press), to build confidence (Cassady & Gridley, 2005) and to help them to pace their study, in line with Gibbs and Simpson's Conditions 1–3. Students can use the online assignments to check their understanding and so to target future study, but the mere act of taking tests has been shown to improve subsequent performance more than additional study of the material, even when tests are given without feedback. This is the so-called "testing effect" and research in this area is reviewed in Roediger & Karpicke (2006). Concern that unsuccessful retrieval attempts might impair learning has been shown to be unfounded (Kornell, Hays & Bjork, 2009).

The testing effect has been demonstrated in rigorous controlled tests, and also in more authentic settings (McDaniel, Roediger & McDermott, 2007; Lyle & Crawford, 2011). However, as with many educational interventions, it can be difficult to prove a causal benefit of the introduction of computer-based assessment; in particular, 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, Boyd, Chen & Jamal, 2011) fall into the trap of assuming that because the students who have chosen to do an optional formative computer-marked quiz also do better in a later summative assessment, the formative quiz has necessarily caused the later improvement. Sly also assumes that it is weaker students who do the formative quiz because the average mark obtained by them in this quiz is lower than that obtained by the students who did not engage in the formative quiz when they first encounter a computer-marked quiz, in the form of the

summative test. This disregards the point that many students engage more thoroughly with summative tests than formative ones (Kibble, 2007; Jordan & Butcher, 2010; Publication 6).

Despite the fact that Walker, Topping and Rodrigues (2008) claim that student expectations and perceptions of e-assessment have been under-researched, much of what is written (e.g. Marriott, 2009; Holmes, in press) 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 it is not the end of the story; students do not always engage with computer-marked assessment and computer-generated feedback in the way that they say they do (Publication 6).

There have however been a number of more rigorous studies that have demonstrated a positive influence of computer-based assessment. Angus and Watson (2009) used a retrospective regression methodology and found that regular low-mark online testing significantly improved learning, as measured by a final examination. Ćukušić, Garača and Jadrić (2014) looked at examination results for three consecutive cohorts of students, before and after the introduction of online self-assessment quizzes. Whilst improvements from one presentation to the next could possibly be attributed to a changing student population, the examination results were significantly better after the introduction of the formative tests, despite the fact that the students did less well in other similar modules, so appear to have been generally weaker. The authors suggest that a wider implementation of quizzes of this type could have a significant positive impact on retention. Van Gaal and De Ridder (2013) also found that examination scores improved following the introduction of regular online assessments, more so for weaker students than for stronger ones.

Computer-marked assessment thus has much to offer in terms of improving the student learning experience. However it is interesting to note that the phrase “objective questions”, used to describe multiple-choice questions in particular, reflects the fact that the early use of multiple-choice questions came from a desire to make assessment more objective. The earliest multiple-

choice tests were probably E.L. Thorndike's Alpha and Beta Tests, used to assess recruits for service in the US Army in World War I (Mathews, 2006). Multiple-choice testing as an educational tool gained in popularity during the 20th century as researchers became more aware of the limitations of essays (Bacon, 2003). Ashburn (1938) noted a worrying variation in the grading of essays by different markers, an oft-repeated finding (e.g. Black & Wiliam, 2006). Human markers are inherently inconsistent and can also be influenced by their expectations of individual students (Read, Francis & Robson, 2005; Orrell, 2008). Computerised marking brings objectivity and a consistency that can never be assured between markers or, over time, for the same human marker (Bull & McKenna, 2004, Publication 5).

Alongside increased reliability, computer-marked assessment can bring savings of time and thus of resource (Dermo, 2007), although writing high-quality questions should not be seen as a trivial task (McKenna & Bull, 2000). Computer-marked assessment is particularly useful for large class sizes (Whitelock & Brasher, 2006; Miller, 2008), because then the effort in writing high-quality questions is worthwhile (Publication 1, p. 15) and the marking time saved can help to "avoid meltdown" (Ridgway et al., 2004, p.17) and to enable practitioners to make more productive use of their time (JISC, 2010). Drever and Armstrong (2000, p.3) point out that:

> While it is reasonable to expect [human-markers to provide] detailed, individual feedback in subjects with a small number of students, meaningful individualised feedback and comments in subjects/units/courses with large numbers of students is time consuming for academics, repetitive and can lead to inconsistencies, particularly if a number of markers are involved.

There is some anxiety that inappropriate use of computer-marked assessment may have a negative effect on learning, in particular encouraging a surface approach (Gibbs, 2006). Some of these concerns appear to be founded on the belief that computer-marked assessment is limited to the indiscriminate use of multiple-choice questions. However, even with the increased availability of more sophisticated e-assessment items, it remains important that online assessment is not seen as a panacea, but rather only used when appropriate and in balance with

other types of assessment (Mackenzie, 2004). It is noteworthy that Beevers et al. (2010, p. 7) comment that "overall, e-assessment could be much more of a panacea than it is currently allowed to be; it could be a catalyst at the centre of performance data and a driver of improvement." This is in line with Draper's (2009a) conception of catalytic assessment and Black and William's (2009) idea that assessment can create "moments of contingency", as mentioned in Section 2.2 and revisited in sections 2.5 and 4.2.

A potential disadvantage of computer-marked assessment is that there is no tutor to interpret a student's actions. Draper (2009b) reminds us that, with assessed tasks of any kind, students do not always do what the question setter expected, and interactive computer-marked assessment is particularly vulnerable in this regard because it removes the human marker's ability to make retrospective adjustments. In reviewing interviews with various e-assessment practitioners, Gilbert, Gale, Warburton and Wills (2009, p. 31) comment: "Several people mentioned that with traditional assessment, if a question is badly worded the human marker and moderator can ensure that students are treated fairly. But with e-assessment you need to foresee all the problems up front."

Bull and Danson (2004, p. 5) point out that the "promotion of a new form of assessment usually invokes criticism rarely considered in the traditional assessment process". One example of this is a reluctance by some staff at the Open University to use interactive computer-marked assignments (iCMAs) summatively, because it is not currently possible to be absolutely certain that the person completing an iCMA remotely is who they say they are. However exactly the same is true of tutor-marked assignments, whilst responses to computer-marked short-answer questions can be checked by anti-plagiarism software (Publication 6, p. 160).

In considering the scope of computer-marked assessment it should be remembered that not all computer-marked assessment is the same. This point is considered in more detail in later sections, and Ashton et al. (2006b, p.117) point out that the emphasis on the use of objective multiple-choice questions to offer consistent and time-saving solutions for large groups of

students has “diverted attention away from many of the key benefits that online assessment offers to learning”. It is increasingly recognised that the authoring of high-quality questions requires both skill and time (Boyle & Hutchison, 2009) with the principal barrier to institution-wide take-up being one of academic staff time and training (Whitelock & Brasher, 2006); in describing the need for staff development in authoring STACK (System for Teaching and Assessment using a Computer algebra Kernel) questions, discussed in Section 2.6.2, Sangwin and Grove (2006) describe teachers as “neglected learners”.

The time taken to write high-quality questions means that the use of computer-marked assessment front loads the assessment cycle (Bull & Danson, 2004). However, the front loading is not complete; time needs to be allowed to monitor the behaviour of computer-marked assessment in use and resources need to be set aside to make changes when problems are identified (Publication 7, p. 10). Small things can make a difference, for example Richardson et al. (2002) report that computer-based problem solving tests for gifted and talented children were well received, but where the interface was not the same as that in standard internet provision, for example there was no “back” button, the children noticed the difference. It is also the case that the technology needs to be robust; as Gilbert et al. (2009, p. 19) report, students “hate it when the system crashes”.

Ridgway et al. (2004, p. 7) comment that “when we consider the introduction of e-assessment we should be aware that we are working with a very sharp sword”. The primary driver should not be financial gain or the use of new technologies for the sake of doing so, but rather a desire to improve student learning. Redecker and Johannessen (2013) call for development to be led by pedagogy not technology; sadly this has all too often not been the case, with Wood (1991, p. 247) giving an example of an “act first, research later” innovation. By contrast, Voelkel (2013) gives an excellent example of cyclical action research project, a model for development that has much to commend it (Benson & Brack, 2010)

This section has considered a broad range of research into the use of computer-marked assessment, with advantages including objectivity and savings of time and resource, especially for

large class sizes. Of more relevance when considering Research Question 1 (How do students engage with computer-marked assessment and what factors affect this?) is the use of computer-based assessment to motivate and engage students. In addressing Research Question 4 (What is the scope of computer-marked assessment in supporting learning and what are the barriers to wider take-up?), the major barrier that has been identified is the lack of academic time and skill in writing high-quality questions. Research indicates that computer-marked assessment is best used in balance with other types of assessment, and that different question types may have different advantages and disadvantages. Some of the pros and cons of different question types are discussed in Sections 2.5, 2.6 and 2.7., after Section 2.4's more detailed consideration of the effectiveness of computer-generated feedback.

2.4 A role for computer-generated feedback?

The literature reviewed in Section 2.2 indicates that much tutor-generated feedback is expensive to provide, of limited effectiveness and that sometimes students' emotional responses to the feedback prevent them from learning. It therefore seems sensible to explore the use of computer-generated feedback.

Computer-generated feedback can be provided "tirelessly" (Mason & Bruning, n.d.) and instantaneously and students can immediately be given an opportunity to repeat the task, or to perform a similar one, so as to learn from the feedback provided. This approach provides alignment with Gibbs and Simpson's Conditions 6, 9 and 10. In addition to the Open University's work on OpenMark and Moodle (Publication 1 and 6; Butcher, 2008), there have been a number of innovative approaches such as "try once, refine once", in which students are encouraged to engage with feedback before refining a previous answer (Fowler, 2014). It is also important to recognise that computer-based assessment has a role to play in providing information to students to assist in their self-regulation (Nicol & Milligan, 2006), with Miller (2008, p. 182) having a vision for a framework in which students "actively monitor, regulate, and control their cognition, motivation and behaviour".

However, relatively little is known about the affective impact of feedback from a computer. Gipps (2005, p. 197) asks "How does the student react to electronic feedback? What is the assessment relationship when the assessor is the computer?" One way forward is to consider the "Computers as Social Actors" (CASA) framework, which hypothesises that people respond socially and naturally to computers, as if they were human, even though they know that they are not (Reeves and Nass, 1996). Ferdig and Mishra (2004) found that even very simple human-computer interactions sometimes engendered feelings of unfairness, anger and spite.

Mishra (2006) found some support for the CASA hypothesis in students' reaction to affective feedback from a computer, though the evidence was mixed, with people tending to accept feedback from a computer at face value whilst being more interpretative of feedback received from a human. Lipnevich and Smith (2008) found no significant variation in overall performance as a result of whether students believed feedback to come from a computer or a human, though again there were differences when the detail was considered.

In focus group discussions, students reported that feedback they believed to come from a computer was helpful, unless they disagreed with the marking (typically because they thought they were right and the computer was wrong) at which point they ignored the feedback (Lipnevich & Smith, 2009). Praise from a computer was not particularly well received. However, some individuals felt that it was good that their tutors had not seen their work and some felt that the computer was fairer. Other researchers report that students prefer feedback from a computer because it enables them to make mistakes in private (Miller, 2008), and the feedback is perceived to be impersonal (Earley, 1988) and non-judgemental (Beevers et al., 2010).

Not only does feedback from a computer appear to be less subjective than feedback from a human marker, it *is* less subjective. Sim et al. (2004, p. 217) point out that "the emotional and subjectivity issues that are evident in human centred marking may be removed via automatic marking", whilst Mason & Bruning (n.d.) say that "unlike feedback from an instructor or tutor, this feedback can remain unbiased, accurate and non-judgmental, irrespective of student characteristics or the nature of the student response". The obvious counter-argument, which

appears to be missing from the literature, is that a tutor is able to alter his or her feedback in response to the individual student, responding to both the student's ability (so, for example, deliberately not commenting on all the errors in a weak script) and what they know of a student's personality and emotional state. However the effectiveness of feedback from a human-marker can be compromised if the student loses confidence in the tutor giving the feedback (Poulos & Mahony, 2008). Furthermore, Mason & Bruning (n.d.) point out that:

> The interactive ability of computer-based instruction has the potential for enhancing the quality and type of feedback that can be implemented, limited only by the ingenuity and energy of course designers. Thus computer-based feedback can, at least theoretically, be adapted to the learning styles of each individual student.

The way in which feedback on computer-marked tasks is provided can provide specific benefits over the worked solutions given in printed materials. Hattie & Timperley (2007, p.82) comment that "feedback has no effect in a vacuum; to be powerful in its effect, there must be a learning context to which feedback is addressed". Thus, in self-assessment, it is reasonable to assume that computer-marked assessment with feedback is more effective than just looking up the answer in the back of the book. Kulhavy (1977) points out that pre-search availability (the ability to look at feedback before attempting the question) must be controlled if feedback is to be effective.

Many of the issues surrounding student use of feedback in general (Section 2.2) are reflected in the literature that is specific to computer-generated feedback. Walker et al. (2008) found that students considered computer-generated feedback to be useful, and when it was not present they adopted a trial and error approach. However, students do not necessarily use feedback in the way that educators had hoped; for example, Bull and McKenna (2004, p. 62) report a case in which, "when faced with CAA self-assessments, students adopted what their tutors called a 'smash and grab' technique, 'punching any key' to 'strip off' the feedback and correct answers".

As with tutor-generated feedback, lengthy computer-generated feedback is not necessarily more effective than shorter feedback. Fowler (2008, p. 2), citing Van der Linden (1993), comments that "some practitioners have said that it is important not to overwhelm users with feedback…it is certainly true that learners are more likely to attend to feedback if it is concise and precise".

Various authors have attempted to classify the different types of feedback that can be provided on computer-based assessment. Bull & McKenna's (2004) feedback types include: "correct/incorrect"; "the answer is…"; directive feedback which tells the student where to find the correct answer; and non-directive feedback which prompts the student with relevant hints. "Correct/incorrect" feedback can be aligned with "knowledge of response" feedback (Clariana, 1993), which Hsieh and O'Neil (2002) found to be less effective than feedback that also explained where students had improved since their last access to the feedback. "The answer is…" feedback can be aligned with Clariana's "knowledge of correct response", or Ferreira and Atkinson 's (2009) "answer giving" feedback, which Ferreira and Atkinson found to be less effective than "answer prompting" feedback. This finding is supported by a comment from an Open University student, reported by Butcher et al. (2009, p.8): "Interactive assessment [is] a particularly good learning tool, if you are getting the answer wrong it doesn't give you the answer right away, but gently coaches you till you get it right for yourself".

Another approach is to allow students to repeat questions, perhaps as many times as they want to or until the answer is correct. Clariana (1993) found this "multiple-try" or "answer until correct" approach (with the student implicitly knowing that their previous answer was incorrect, but nothing more) to be more effective in supporting learning than no feedback, whilst commenting that some students are more likely to benefit than others from this approach, and noting uncertainty over the relative benefits of allowing just one additional try at a question or unlimited tries. Malmi, Karavirta, Korhonen and Nikander (2005) observed that the provision of an unlimited number of tries caused some students to waste time.

One of the advantages of computer-generated feedback is that different levels of feedback can be provided following subsequent tries. OpenMark (Publication 6) usually allows three tries, with

feedback after a first incorrect try being simply “Your answer is incorrect”. After a second incorrect try, feedback is intended to prompt the student, usually with a reference to the module materials, and wherever possible the feedback is tailored to the error that has been made. After a third incorrect try, or whenever a correct response is given, a full solution is provided. Sangwin (2013, p. 139) comments that “it is intriguing that both the OpenMark and CALM [Computer Aided Learning of Mathematics] teams independently concluded that three attempts is, in general, an optimum balance between opportunities to try again and helping students who are stuck to progress”. The use of CALM and OpenMark are further discussed in Section 2.6.

Section 2.4 has highlighted the complex and poorly understood nature of student engagement with computer-generated feedback (Research Question 2), indicating that this is a topic that is ripe for further research. My contribution to this work is summarised in Section 3.2.

2.5 Question types: Selected-response or constructed-response?

During the 20th century, large-scale multiple-choice tests were administered by the batch-processing of responses that had been entered onto machine-readable forms. These systems, which are still in use, enabled objectivity and resource saving, but the advantages of immediacy of feedback and student engagement, significant in many of Gibbs and Simpson’s conditions, were not yet present. Since around 1980 there has been a rapid growth in the number and sophistication of computer-marked assessment systems (Publication 11) with a move to web-based delivery in the 1990s and 2000s (Kleeman, 2013; Publication 1). As more and more learning took place online, the use of virtual learning environments (VLEs) became common, with most VLEs incorporating their own assessment tools. For example, the Moodle learning management system was first released in 2002 and its quiz system has been in constant development since its first release in 2003 (Hunt, 2012). Hunt identifies around thirty different question types available within Moodle, but yet an ad hoc survey of more than 50,000,000 questions from around 2,500 Moodle sites found that about 90% of the questions in use were selected-response questions i.e. question types like multiple-choice or drag-and-drop where options are presented for a student

to select from, in contrast to “constructed-response” where students construct their own response. This section reviews the evidence, which sometimes appears contradictory, concerning the pros and cons of selected-response and constructed-response questions.

Selected-response questions can be used to assess a large breadth of knowledge (Betts, Elder, Hartley & Trueman, 2009; Ferrão, 2010) whereas a test comprising constructed-response questions is likely to be more selective in its coverage. Use of selected-response questions also avoids issues of data-entry, particularly problematic in constructed-response questions when symbolic notation is required, for example in mathematics (Publication 3; Beevers & Paterson, 2003, Sangwin, 2013). In addition, selected-response questions may avoid the feeling of helplessness and frustration that can arise when students do not know what word to use in a free-text question (Walker et al., 2008).

Selected-response questions also avoid issues with incomplete or inaccurate answer matching. Constructed-response answers may sometimes be incorrectly marked (Publication 5), although the issue of the marking of incorrect spelling, identified by Sim et al. (2004) and which Walker et al. (2008) felt was resulting in students being assessed on their spelling ability rather than the intended learning outcome, can now be satisfactorily handled (Publication 9). However, in some systems, constructed-response questions require answers to be entered in a non-standard or incomplete form to facilitate answer matching. Gill and Greenhow (2008) report the worrying finding that students who had learnt to omit units from their answers because these could not be recognised by the assessment system, continued to omit units thereafter.

In some multiple-choice questions, the correct option can be selected by working back from the options, for example, a question that asks students to integrate a function can be answered by differentiating each of the options provided (Sangwin, 2013, p. 3). For all selected-response questions, especially those requiring a calculation or an algebraic manipulation, if a student obtains an answer that is not one of the options provided, they are given an early indication that there is likely to be something wrong with their answer (Bridgeman, 1992). Even when testing the well-established force-concept inventory (Hestenes, Wells & Swackhamer, 1992), Rebello and

Zollman (2004) found that in equivalent open-ended tests, students gave answers that were not provided in the selected-response test.

Curiously Simkin and Kuechler (2005) consider it an advantage from the student point of view that they can guess the answers to selected-response questions; other authors see guessing as a problem, with Geoffrey Crisp (2007, p. 113) pointing out that this means that the teacher has no way of telling what the student really understands. Downing (2003) is unconcerned about the impact of guessing on score, pointing out that it would be very difficult for a student to pass a whole assignment by guesswork alone. However Burton (2005) points out that a successful guess has the potential to make a significant difference to the outcome for a borderline student.

Funk and Dickson (2011) used exactly the same questions in multiple-choice and short-answer free-text response format, and found that students' performance in the multiple-choice items was significantly higher ($p < 0.001$) than performance on the same items in the short-answer test. However Ferrão (2010) found high correlation between scores on a multiple-choice and an open-ended test. Others have suggested that selected-response questions advantage particular groups of students, especially those who are more strategic or willing to take a risk (Hoffman, 1967). Different gender biases have been reported, for example by Gipps and Murphy (1994) who found that 15-year old girls disliked multiple-choice questions whereas 15-year old boys preferred them to free-response types of assessment. Kuechler and Simkin (2003) found that students for whom English was a second language sometimes had difficulty dissecting the wording nuances of multiple-choice questions.

Conole and Warburton (2005) discuss the difficulty of using selected-response questions to assess higher order learning outcomes, though some have tried (e.g. Gwinnett, Cassella & Allen, 2011). Williams (2006) discusses the potential of assertion-reason multiple-choice questions for this purpose, but again points out that for some students these may actually be a test of the finer points of the English language.

The points discussed above indicate that although selected-response questions may appear to have high reliability, they may not always be assessing the learning outcome that the question-setter thinks they are assessing. Furthermore, Nicol (2007) and Publication 4 identify a fundamentally different cognitive process in answering selected-response and constructed-response questions. This is in line with the so-called "generation effect" in which it has been demonstrated that responses that have been self-generated by people are retained better than those presented by the experimenter (Slamecka & Graf, 1978; Kang, McDermott & Roediger, 2007). This provides one explanation for the diminished "testing effect" observed when multiple-choice questions were used (Roediger & Marsh, 2005; Marsh, Roediger, Bjork & Bjork, 2007). This result is more commonly attributed to the fact that students are remembering the distractors rather than the correct answer.

Scouller (1998) argues that the use of selected-response questions can encourage students to take a surface approach to learning, although Kornell and Bjork (2007) found no support for the idea that students consider essay and short-answer tests to be more difficult than multiple-choice and so study harder for them. Struyven et al. (2005) point out that students' perceptions of assessment, however flawed, are important and that most students, though not all, prefer multiple-choice questions to essays. They take the view that if students perceive simple quizzes to be fun, they will "enable rather than pollute" (p. 332) and they will result in learning. Jordan (2009a, p. 16) reports a student who would prefer multiple-choice questions to short-answer free-text ones, saying that "in multiple choice, obviously you know the answer is there somewhere, it's just a matter of finding it, so there is an element of I'm not going to be completely out".

Perhaps the most damnatory indictments of selected-response questions are those that query their authenticity. In commenting on the widespread use of multiple-choice questions in medical schools, Mitchell, Aldridge, Williamson and Broomhead (2003, p.252) quote Veloski (1999): "Patients do not present with five choices." Bridgeman (1992, p.271) makes a similar point with

reference to engineers and chemists: They are seldom "confronted with five numerical answers of which one, and only one, will be the correct solution".

Back in 1995, Knight (1995, p. 13) pointed out that "what we choose to assess and how, shows quite starkly what we value". Wiliam (2008, p. 273) comments:

> In terms of traditional conceptualizations of validity, and especially when we take utility into account, the argument in favour of multiple-choice tests seems unanswerable. However, when we take value implications into account, adopting a mutliple-choice test…sends the message that only the aspects of the domain that can be assessed via multiple-choice tests are important.

Ridgway et al. (2004, pp. 19-20) point out that it is illusory that multiple-choice questions are cheap and that their over-use can be very expensive, "if it leads to a distortion of the curriculum in favour of atomised declarative knowledge". Murphy (2008) discusses the danger that the high stakes multiple-choice tests of writing will lead to actual writing beginning to disappear from the curriculum.

The apparent contradictory results of investigations into the effectiveness of selected-response questions may be because the questions not are homogeneous (Simkin & Kuechler, 2005). Different questions need different question types, with some questions (e.g. "Select the three equivalent expressions") lending themselves particularly to a selected-response format. Burton (2005, p.66) points out that "It is likely that particular tests, and with them their formats and scoring methods, have sometimes been judged as unreliable simply because of flawed items and procedures". Whatever question type is used, it is important that high-quality questions are written (McKenna & Bull, 2000). For multiple-choice questions this means, for example, that all distractors should be equally plausible.

This section has reviewed the advantages and disadvantages of selected-response and constructed-response question types, considering student engagement and student opinion,

which are linked to each other and to Research Question 1: How do students engage with computer-marked assessment and what factors affect this? Issues of reliability, validity and authenticity have also been considered; if these matters are not taken seriously, the scope of computer-marked assessment will be compromised (Research Question 4).

Even relatively simple multiple-choice questions can create "moments of contingency" (Black & Wiliam, 2009; Dermo & Carpenter, 2011) and Draper's (2009a) concept of catalytic assessment is based on the use of selected-response questions to trigger subsequent deep learning without direct teacher involvement. Some of the techniques that can be used to increase the effectiveness of selected-response questions are discussed in Section 2.8., but first of all Sections 2.6 and 2.7 consider research into the use of constructed-response question types.

2.6 More sophisticated computer-marked assessment

Educators and educational researchers now have available to them a wide range of different question types and test systems. It is not appropriate for this critical review simply to describe the systems that are in use; Publication 11 gives more detail whilst Geoffrey Crisp (2007, pp. 69-74) provides a comprehensive list. Instead, this section reviews the theoretical underpinning and pedagogic research that has accompanied two particular groups of systems: those emanating from the CALM (Computer Aided Learning of Mathematics) Project at Heriot-Watt University; and OpenMark and Moodle at the Open University, now incorporating STACK (System for Teaching and Assessment using a Computer algebra Kernel) which was originally developed by Chris Sangwin, then at the University of Birmingham. Research on the automatic marking of short-answer questions and essays is reviewed in Section 2.7.

2.6.1 The CALM Family of systems: Focus on breaking a question down into "Steps"

The CALM Project started at Heriot-Watt University in 1985, and various computer-marked assessment systems have derived at least in part from it, including CUE[2], Interactive Past Papers, PASS-IT (Project on ASsessment in Scotland – using Information Technology), i-assess, and

[2] CUE takes its name from the leading initials of CALM, UCLES (University of Cambridge Local Examinations Syndicate) and EQL International.

Numbas (Foster, Perfect & Youd, 2012). Some of the systems have been used in high-stakes summative testing, but the focus has always been on supporting student learning (Ashton et al., 2006b). From the early days, constructed-response questions have been favoured, with hints provided to help students (Beevers & Paterson, 2003). One of the signatures of the CALM family of assessment systems is the use of "Steps", allowing a question to be broken into manageable steps for the benefit of students who are not able to proceed without this additional scaffolding (Beevers & Paterson, 2003; Ashton, Beevers, Korabinski & Youngson, 2006a). This use of Steps provides students with "opportunities to close the gap between current and desired performance" (Nicol & Macfarlane-Dick, Condition 6).

McGuire, Youngson, Korabinski and McMillan (2002) compared the results for schoolchildren taking computer-marked tests in the CUE system with three different formats (no Steps, compulsory Steps or optional Steps) and with the partial credit they would have obtained by taking the corresponding examinations on paper. On this occasion no penalty was applied for the use of Steps. The overall marks for tests without Steps were lower than those in which Steps were available and they were also lower than the marks for the corresponding paper-based examinations. McGuire et al. (p. 228) concluded that "this means that without Steps the current marking schemes for paper-based examinations cannot, at present, be replicated by the current computer assessment packages". They found no evidence of a difference in marks between what would be obtained from a paper-based examination or from a corresponding computer examination with Steps, whether optional or compulsory. However they commented (p. 229) that even if the marks were similar "this does not mean that the candidates have shown the same skills. In particular, the use of Steps provides the candidate with the strategy to do a question".

2.6.2 OpenMark, Moodle and STACK: Focus on interactivity and computer algebra

As discussed in Section 1.3.3, the OpenMark system at the Open University was launched in 2005 following the success of interactive questions delivered to students by CD-ROM and the use of a precursor online system. In addition to OpenMark's emphasis on instantaneous and relatively

detailed feedback, wherever possible tailored to the response that was given, the provision of multiple tries seeks to improve interactivity and enables students to act immediately on the feedback that has been received (Gibbs & Simpson, Conditions 6, 9 and 10). In common with other systems such as TRIADs (TRipartite Interactive Assessment Delivery system) (Mackenzie, 1999) and CALM, a wide range of question types is available with the aim of "using the full capabilities of modern multimedia computers to create engaging assessments" (Butcher, 2008, p. 4). Most questions exist in several variants, to limit opportunities for plagiarism and to provide extra opportunities for practice (Publication 6, p. 150). In addition, OpenMark assignments are designed to enable part-time students to complete them in their own time and in a manner that fits in with their everyday life; hard cut-off dates are used to help students keep up to date, but the assignments are available for a period of least several weeks, with "no limit to the amount of time spent actually working on the iCMA within that period" (Publication 6, p. 149). They can be interrupted at any point and resumed later from the same location or from elsewhere on the internet (Butcher, 2008).

The Open University's Science Faculty was the first to embrace the use of OpenMark interactive computer-marked assignments, so the importance of answer matching and targeted feedback for units and precision in numerical answers was quickly realised (Publication 3, p. 128). This feedback is usually given after a student's first try, even where most feedback is reserved for the second or third tries. Research into student engagement with the questions and feedback is discussed later in this covering paper and in many of the submitted publications. An earlier analysis of individual student responses (Jordan, 2007) has led to improvements to the questions as well as giving insight into student misconceptions.

OpenMark's emphasis on the provision of a range of question types, and on multiple tries with feedback, influenced the development of Moodle's assessment system (Butcher, 2008), though Hunt (2012, p. 2) recognises that "the increasing support for assessment for learning can also be seen as an attempt to bring Moodle's CMA tools in line with its social constructivist pedagogy (Dougiamas et al., 2012)". Hunt identifies question type (e.g. "numerical" or "drag-and-drop")

and question behaviour (e.g. whether the question is to be run in "interactive mode" with instantaneous feedback and multiple tries or in "deferred mode" with just one try permitted and no feedback until the student's answers have been submitted) as separate concepts, and the combination of a question type and a question behaviour to generate an assessment item is a unique feature of the Moodle assessment system.

When free-text mathematical expressions are to be assessed, there are three ways in which a student's response can be checked. The original CALM assessment system evaluated the expression for particular numerical values. This is a reasonable approach, but is likely to lead to some incorrect responses being marked as correct (note, for example, that $2x$ and x^2 both have a value of 4 when $x = 2$). OpenMark uses string matching. This has worked effectively but relies on the question-setter thinking of all the equivalent answers that should be marked as correct and all the equivalent incorrect answers that should generate the same targeted feedback. Since around 1995, a number of computer-marked assessment systems have made use of a mainstream computer algebra system (CAS) to check student responses (Sangwin, 2013), for example AiM (Assessment in Mathematics) uses the computer algebra system Maple (Strickland, 2002).

STACK uses the open source computer algebra system Maxima (Sangwin, 2013); STACK was first released in 2004 as a stand-alone system but since 2012 it has been available as a Moodle question type (Butcher, Hunt & Sangwin, 2013). Aspects considered important by the Moodle system developers, in particular the provision of feedback and the analysis of student responses, are also important in STACK. Furthermore, it was felt insufficient to restrict question authors to the use of static responses written in advance, so STACK is able to use the computer algebra system to generate targeted feedback based on the student's response. Sangwin (2013, p. 104) expresses his surprise that "not all systems which make use of a CAS enable the teacher to encode feedback" and a focus group at Aalto University, Finland, considered the immediate feedback to be the best feature of STACK.

The CALM family of systems, OpenMark, the Moodle assessment system and STACK were all designed with an emphasis on assessment for learning, and each has, in some way, extended the scope of computer-based assessment to support learning (Research Question 4). Evaluation of their use by academics has provided insight into barriers to wider take-up (Research Question 4), whilst evaluation of their use by students has provided insight into student engagement with computer-based assessment and computer-generated feedback (Research Questions 1 and 2). Finally, each of these systems has the capability to provide information to educators (Nicol and Macfarlane-Dick, Condition 7), thus addressing Research Question 3: What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement? My research into each of these areas, using the OpenMark system, is discussed in Chapter 3. Much of my work has focused specifically on student engagement with short-answer free-text questions; Section 2.7 sets this in context.

2.7 Short-answer questions and essays

Alongside the use of computer algebra systems for more sophisticated mathematical questions, the introduction of software for marking short-answer questions and essays has extended the range of constructed response questions that can be used (Research Question 4). "Short-answer" is usually taken to mean questions requiring answers of a sentence or two in length, and since 2009, the allowed length of responses at the Open University has been restricted to no more than 20 words, partly to give an indication to students of what is required and partly to discourage responses including both correct and incorrect aspects (Publication 8, p. 825). Mitchell, Russell, Broomhead and Aldridge (2002, p. 245) first recognised the incorrect qualification of a correct answer as a potentially serious problem for the automatic marking of short-answer questions; this point is further discussed in Section 3.4.

Software for marking short-answer questions includes c-rater (Leacock & Chodorow, 2003; Sukkarieh & Blackmore, 2009) and systems developed by Intelligent Assessment Technologies Ltd. (IAT) (Mitchell *et al.* 2002; Publication 4) and by Sukkarieh, Pulman and Raikes (2003, 2004). These systems, reviewed by Siddiqi & Harrison (2008), are all based to some extent on computational

linguistics. For example, the IAT software draws on the natural language processing (NLP) techniques of information extraction, and compares templates based around the verb and subject of model answers with each student response. However IAT provide an authoring tool that can be used by a question author with no knowledge of NLP.

By contrast, OpenMark's PMatch (Publication 5; Publication 9) and the Moodle Pattern Match question type are simpler pattern-matching systems, based on the matching of keywords and their synonyms, sometimes in a particular order and/or separated by no more than a certain number of other words, and with consideration paid to the presence or absence of negation. So, for example, "There are no unbalanced forces" and "The forces are balanced" can be marked as correct whilst "The forces are unbalanced" is marked as incorrect. Similarly "Kinetic energy is converted into gravitational energy" can be marked as correct whilst "Gravitational energy is converted into kinetic energy" can be marked as incorrect. As for any OpenMark or Moodle question type, students can be provided with instantaneous targeted feedback, increasing in response to subsequent tries.

PMatch and the Moodle Pattern Match question type make use of a dictionary-based spell checker which notifies students if their response contains a word that is not recognised; however, the standard string-matching techniques of allowing missing or transposed letters remain useful, to cope with situations where a student accidentally uses a word that is slightly different from the intended one (e.g. "decease" instead of "decrease") (Publication 9, p. 3). Good marking accuracy has been obtained, always comparable or better than the marking of human markers and on a par with the more sophisticated IAT FreeText Author (Publication 5; Publication 9); this is discussed in more detail in Section 3.4. As with most of the systems used for the automatic marking of textual responses, the fact that real student responses are used in developing answer matching is regarded as being of crucial significance. Although most of the research interest in short-answer free-text questions continues to be concentrated on more sophisticated systems, Siddiqi's (2013)

findings are in line with those of Publication 5, noting only a slight reduction in marking accuracy when a linguistic features analyser was not used.

Automated systems for the marking of essays are characteristically different from those used to mark short-answer questions, because with essay-marking systems the focus is frequently on the writing style, and the required content can be less tightly constrained than is the case for shorter answers. This means that the marking of essays is in some senses easier than that of short-answer questions (Raikes & Harding, 2003), although Warburton and Conole (2005, p. 8) regard automated essay marking as the "holy grail" of computer-marked assessment.

Many systems exist for the automatic marking of essays, for example e-rater (Attali & Burstein, 2006) and Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2003), with reviews by Valenti, Neri and Cucchiarelli (2003), Dikli (2006), and Vojak, Kline, Cope, McCarthey and Kalantzis (2011). Further systems are under development and some, for example OpenEssayist (Field, Pulman, Van Labeke, Whitelock & Richardson, 2013), put the focus on the provision of feedback rather than grading, in line with Gibbs and Simpson's Condition 4, Nicol and Macfarlane-Dick's Condition 3, and the use of comment-only marking, as discussed in Section 2.2.2.

Deane (2013) points out that little is known about the impact of the automatic marking of essays on the test-taker. Scharber, Dexter and Riedel (2008) report on a study in which students initially liked the automated essay scorer, tried to "please" it, and were happy to engage with the system even though they did not have to do so. However, when they believed that the scorer was not accurately marking their essay improvements they became frustrated. Thus, in common with the findings reported in Section 2.4, there is evidence both for and against the Computers as Social Actors (CASA) hypothesis.

Shermis and Hammer (2012) report on a large-scale comparison of the marking accuracy of essay marking systems and human markers. The essay marking systems performed well, but Shermis and Hammer (p. 27) point out that "a predictive model may do a good job of matching human scoring behavior, but do this by means of features and methods which do not bear any plausible

relationship to the competencies and construct that the item aims to assess" which leads to the "manipulation by test-takers and coaches with an interest in maximizing scores". Systems that use simple proxies for writing style have been criticised, for example by Perelman (2008), who trained three students to obtain good marks for a computer-marked essay by such tricks as using long words and including a famous quotation (however irrelevant) in the essay's conclusion. Condon (2013) contends that until computers can make a meaningful assessment of writing style, they should not be used.

The literature reviewed in Section 2.7 is at the cutting edge of research into computer-marked assessment and computer-generated feedback. Some insights into student engagement (Research Questions 1 and 2) can be obtained, and my work in this area is summarised in Section 3.1 and Section 3.2. However, Section 2.7 is most directly relevant to Research Question 4: What is the scope of computer-marked assessment in supporting learning and what are the barriers to wider take-up of sophisticated computer-marked tasks? My work in response to Research Question 4, investigating the marking accuracy of short-answer free-text questions, is discussed in Section 3.4. However, when the assessed task is an essay, even when computerised marking is accurate, the fact that the software is marking proxies for writing style rather than the writing style itself appears to limit the acceptability of the process, with the fundamental objection being that essay writing is no longer being used as a means of communication between two people.

2.8 Using questions effectively

Hunt (2012) points out that a computer-marked assessment system comprises not just the bank of questions but also a question engine that presents each question to the student, grades their response and delivers appropriate feedback and a test system that combines individual items into a complete test (possibly with feedback at the test level). There are thus many different ways in which the assessment system can affect the student experience and this section reviews a wide range of literature, covering topics from confidence-based marking to adaptive assessment, the use of Peer Instruction and student-authored questions.

Various techniques have been used to compensate for the fact that students may guess the correct answers to multiple-choice questions. Simple negative marking (deducting marks or a percentage for incorrect answers) can be used, but there is then a danger that it is the students' answering strategies and risk-taking behaviours that are being assessed rather than the questions' intended learning outcomes (Burton, 2005; Betts et al., 2009).

More sophisticated techniques run similar risks, but at the same time they offer some potential to encourage students to engage more deeply with the assessment task, making it a productive learning activity (Gibbs & Simpson, Condition 3). Ventouras, Triantis, Tsiakas and Stergiopoulos (2010) constructed an examination using paired multiple-choice questions on the same topic (but not obviously so to students), with a scoring rule which gave a bonus if they got both questions right. This gave results that were statistically indistinguishable from the results of an examination with constructed-response questions. McAllister and Guidice (2012) describe another approach in which the options were combined for a number of questions, resulting in a much longer list (60 options for 50 questions in their case) and so a much lower probability of guessing the correct answer. However, in general, it may be difficult to find options that are equally plausible for a range of questions. Bush (2001) describes a "liberal multiple-choice test" in which students could select more than one answer to a question if they were uncertain of the correct one, and Walker and Thompson's (2001) "hedging" format took a similar approach. Whilst techniques of this sort are undoubtedly fairer, there is again concern over the emphasis on tactics rather than on knowledge and understanding of the correct answer (Bush, 2001).

It has long been recognised that the reliability of a test score can be increased by incorporating some sort of weighting for the appropriateness of a student's confidence (Ahlgren, 1969). Much work on "confidence-based" (or "certainty-based") marking has been done by Gardner-Medwin (2006) who points out that this approach does not favour the consistently confident or unconfident, but rather those who can correctly identify grounds for justification or reservation. Rosewell (2011) points to advantages in requiring students to indicate their confidence *before* the multiple-choice options are revealed whilst Archer and Bates (2009) included a confidence

indicator and also a free-text box into which students were required to give reasons for each answer. Nix and Wyllie (2011) incorporated both a confidence-indicator tool and a reflective log into a formative multiple-choice quiz, in an attempt to encourage students to regulate their own learning experience. They conclude (p. 111) that:

> The confidence-indicator tool and learning log were motivators for engagement in learning and assessment in different ways to different learners. Learners welcomed the opportunity to configure and control their own mode of learning, though they needed practice in the skills of self-regulation.

Fyfe et al. (2014) found a similarly varied response to an online reflective tool and they call for the scaffolded integration of reflection, especially for less experienced students.

The selection of questions from a question bank or the use of multiple variants of each question can provide additional opportunities for practice (Gibbs & Simpson, Condition 1) and discourage plagiarism (Publication 6, p. 150; Sangwin 2013, p. 38). However, especially in summative use, it is necessary to select questions that assess the same learning outcome and are of equivalent difficulty (Dermo, 2010; Publication 7). Dermo (2009) found a concern amongst students that the random selection of items was unfair.

Feedback can be given to students at the level of the quiz or test. Following the finding that students wanted more direct information about whether they were sufficiently prepared for a particular module (Publication 6, p. 161), feedback in the form of a “traffic light system” was introduced into a diagnostic quiz, indicating students’ preparedness in a number of different skill areas. Nicol and Macfarlane-Dick’s Condition 3 emphasises the importance of providing information to students about their learning.

Adaptive assessments (frequently described as “computer adaptive tests”) use a student’s responses to previous questions to make a judgement about his or her ability, and so to present subsequent questions that are deemed to be at an appropriate level (Geoffrey Crisp, 2007, p. 76).

Lilley, Barker and Britton (2004) found that students were not disadvantaged by a computer adaptive test and that they appreciated not having to answer questions that they considered too simple. Questions for computer adaptive tests are usually selected from a question bank and statistical tools are used to assign levels of difficulty (Gershon, 2005), thus most systems become complicated and rely on large calibrated question banks. Pyper and Lilley (2010) describe a simpler "flexilevel" system which applies fixed branching techniques to select the next item to present to a student at each stage.

Another use of adaptive testing is to create a "maze" in which questions are asked that depend on a student's answer to the previous question, without necessarily attributing "correctness" or otherwise. Wyllie and Waights (2010), cited by Publication 11 (pp. 11-12), developed a clinical decision-making maze to simulate the decisions that have to be taken, based on various sources of information, in deciding how to treat an elderly patient with a leg ulcer. This type of maze offers one way in which the authenticity of computer-marked assessments might be increased.

Nicol and Macfarlane-Dick's Condition 4 emphasises that good feedback practice "encourages teacher and peer dialogue around learning", and Nicol (2007) suggests that this dialogue can be achieved by initiating a class discussion of multiple-choice questions. This is one use to which "clickers" (also known as "electronic voting systems", "audience response systems" and "student response systems") can be put. Clickers have been in use in classrooms and lecture theatres since before 1970 (Publication 11, p. 6) and Littauer (1972) provided the questions before the lecture took place and noted students debating answers – an early indication of the type of approach later taken in Classtalk (Dufresne, Gerace, Leonard, Mestre & Wenk, 1996) and Peer Instruction (Mazur, 1991). Lantz (2010, p.557) noted that one of the benefits of clickers is that students respond in private, which means that students who are normally too shy to respond in class have an opportunity to do so i.e. all students are engaged. Many authors (including Wieman, 2010) attribute a profound positive effect on learning to the use of clickers, but Fies and Marshall (2006) call for more rigorous research in this area, whilst Beatty and Gerace (2009) argue that there are many different ways of using clickers and that these uses should not be lumped together. Peer

discussion is found to be particularly effective, making a lecture more interactive and students more active participants in their own learning processes (Dufresne *et al.*, 1996, Mazur, 1991; Crouch & Mazur, 2001; Lasry, Mazur & Watkins, 2008). Thus the observed learning gains are "not the result of the technology alone. Rather, these gains [are] the result of the application in class of teaching and learning principles centred on active engagement and dialogue", supported by technology (Nicol & Boyle, 2003, p. 472). Publication 11 (p. 6) points out that online classrooms such as Blackboard Collaborate™ now enable similar uses of voting in a virtual environment, whilst Conejo, Barros, Guzmán and Garcia-Viñas (2013) devised another approach for an online environment, by requiring each student to answer computer-marked questions twice, first by themselves and then following online collaboration with a partner.

Nicol (2007) also points out that dialogue around learning can be achieved by getting students to work in small groups to construct multiple-choice questions or to comment on some aspect of tests that others have written. PeerWise (Denny, Luxton-Reilly & Hamer, 2008b) is a system developed in the Computer Science Department at the University of Auckland but now in use worldwide, in which students author their own multiple-choice questions as well as using and evaluating questions written by their peers. Luxton-Reilly and Denny (2010) describe the pedagogy behind PeerWise, which rests on the premise that students shift from being consumers of knowledge to become participants in a community which is producing and sharing knowledge. Evaluation at Auckland showed that students consistently engaged with the PeerWise system more than they were required to do (Denny, Luxton-Reilly & Hamer, 2008c), that their questions were of remarkably high quality (Purchase, Hamer, Denny & Luxton-Reilly, 2010) and that there was a significant correlation between PeerWise activity and subsequent performance in multiple-choice and written examination questions (Denny, Hamer, Luxton-Reilly & Purchase, 2008a). These findings have been replicated for physics, chemistry and biology students at three UK universities, and there was some evidence that students of lower to intermediate ability may have gained particular benefit overall (Hardy et al., in press).

Most of the topics reviewed in Section 2.8 offer the potential to expand the scope of computer-marked assessment, and thus are of most relevance to Research Question 4. However, it is important to check that complex assignments are behaving in the expected way. Information about an assignment's reliability and validity can be provided to educators and is thus within the remit of Research Question 3 (What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement?), discussed further in Section 2.9.

2.9 Analysis and analytics

This section reviews literature on the use of data from assessment tasks to provide information that can be used to shape future teaching (Nicol & Macfarlane-Dick, Condition 7). Thus the review includes some recent literature on learning analytics and assessment analytics. At a more basic level, the information that passes from computer-based assessment to educator may simply relate to the way in which an assignment is working.

Gilbert et al. (2009) call for greater use of statistical measures of reliability and validity in summative computer-marked assessment. However there can be tensions between reliability and validity. Reliability refers to the notion that the outcome of a test should be the same, irrespective of when a student takes it, irrespective of who marks it and, if different questions are used to test the same learning outcome, irrespective of which set of questions the student receives (Black, 1998). Validity refers to the notion that a question or test is measuring "what the constructors say it is measuring and not something else instead or as well" (Wood, 1991, p. 3, quoting Gagné, 1970). Several dangers to validity have already been identified in this covering paper, for example when assignments assess risk-taking preferences or tactics (Hoffman, 1967; Bush, 2001). Tate (2005) points out that it may be easier to maintain trust in an assessment system that has high perceived reliability, but that this may be at the expense of lower validity, whilst Knight (2002, p. 278) points to the fact that "the quest for reliability tends to skew assessment towards the assessment of simple and unambiguous achievements".

The situation is further complicated by the possibility that questions may be unfair in their impact on different students, thus introducing bias. For example, a student from a cultural minority may fail to understand a question that is completely clear to others (Wood, 1991, pp. 166-183). Using variants of questions or selecting from a question bank, even when done for the best of educational reasons, may result in assignments whose difficulty varies from student to student; simply offering options to multiple-choice questions in a different order has been shown to produce different outcomes (Cizek, 1991). Gipps and Murphy (1994, p. 273) sum up the situation by saying "by now it should be clear that there is no such thing as a fair test, nor could there be: the situation is too complex and the notion simplistic". However they go on to say (p. 274) that "we can begin to work towards tests that are fair". Dermo (2010) describes an investigation into fairness and reliability when objectively marked questions were randomly selected from an item bank. He used Rasch analysis (Rasch, 1960) to calculate the difficulty rating for each item and this calculated individual students' test difficulty indices. He concluded that whilst actual differences in difficulty of the whole test might be small, the impact on some students in borderline situations could be significant. Dermo went on to recalculate student scores to account for the level of difficulty of the questions received. Publication 7 describes similar work, using tools to indicate when different variants of OpenMark questions were of significantly different difficulty, and considering the impact on students' overall scores. This work is discussed in more detail in Section 3.3.1.

In looking at the behaviour of questions at a statistical level, specific problems with individual questions have been detected and corrected (Publication 7). The monitoring of all types of computer-marked questions is important, both to find problems with the questions and to learn more about student behaviour and areas of student difficulty (Publication 10). As Conole and Warburton (2005, p.26) say "One of the benefits of CAA is the opportunity to record student interactions and analyse these to provide a richer understanding of learning". Ashton, Beevers, Schofield and Youngson (2004) describe the use of PASS-IT's summary reports in the cyclical question design process. If students consistently use Steps in a particular question this may

indicate that they are having difficulty starting this question. This may be due to a poorly designed question or it may reveal a widespread student difficulty. Similar work, discussed in more detail in Section 3.3.2, has investigated Open University students' mathematical errors (Jordan, 2007; Publication 12).

Computer-marked assessment can be used with the direct aim of learning about student misunderstandings. Thus, of the publications cited in Publication 12 that have analysed the mathematical misunderstandings of cohorts of children, some (e.g. Williams & Ryan, 2000) used responses to national tests whilst others (e.g. Ryan & Williams, 2007) used specially designed questions, sometimes supplemented by interviews (Hart et al., 1981). Similarly, in the higher education sector, whilst most of my work has investigated common student errors made on in-module assignment questions, Leopold and Edgar (2008) analysed cohort-wide responses to questions that were primarily for diagnostic use, and Tariq (2008) used specially designed questions to investigate the mathematical understanding of 326 first year biosciences students.

Clickers are commonly used in the classroom and lecture theatre to survey the understanding of a whole class with subsequent adjustment to teaching, whilst d'Inverno, Davis and White (2003, p. 167) report an unexpected outcome of the introduction of clickers to be "an increased understanding of the precise places where students were failing to understand". Walet and Birch (2012) used a model of "just in time" teaching in which a lecture set the scene then students did self-study prior to an online quiz. The results of the quiz informed the content of classes a few hours later.

As well as being used to investigate misunderstandings at the whole class level, computer-marked assessment can be used to diagnose misunderstandings of individual students. Sainsbury and Benton (2011) used diagnostic tests to assess young children's reading ability, developing a deeper descriptor of each child's ability than would be provided by a single number. They used latent class analysis to group children into four separate groups, each with characteristic abilities and difficulties. Many universities attempt to diagnose the mathematical difficulties of individual students on entry to their programmes (e.g. Appleby, Samuels & Treasure-Jones, 1997).

In addition to informing educators about the behaviour of cohorts of students and helping them to diagnose the abilities of individual students, information, for example from PASS-IT's summary reports (Ashton et al., 2004) or Moodle's Gradebook (Publication 6, p. 149), can inform the students themselves, and their tutors. For students, the ability to see how they are performing can be a driver towards independent learning (Ashton et al., 2004, p. 3).

The use of information gathered from computer-marked assessment to inform educators about the understanding of their students, at either the cohort or the individual level, falls within some definitions of the new field of learning analytics. Clow (2013, p. 683) considers learning analytics to be "the analysis and representation of data about learners in order to improve learning", whilst Ferguson (2012, p. 305) says that it is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs". Chatti, Dyckhoff, Schroeder and Thüs (2012) rightly point out that learning analytics is not a genuinely new research area, since it builds on other fields such educational data mining. Learning analytics' rise to prominence was linked to the growth of online learning and to the availability of statistical and computational tools to manage large data sets, but authors agree that the key point is that it uses these data to improve understanding of learning and thus to improve the learning itself.

For some reason, assessed tasks are sometimes excluded from the remit of learning analytics which seems bizarre, given that all students are expected to engage with assessment, which is not necessarily the case for other online tasks (Ellis, 2013). However, Publication 10 firmly aligns my analysis of student responses to iCMA questions with the field of learning analytics, and gives examples of data from assessment tasks that can provide a measure of the depth as well as the extent of student engagement. This point is further explored in Section 3.3.

Redecker, Punie and Ferrari (2012, p. 302) suggest that we should "transcend the testing paradigm"; data collected from student interaction in an online environment offers the possibility to assess students on their actual interactions rather than adding assessment separately.

Furthermore, Clow (2012, 2013) points out that learning analytics systems can provide assessment-like feedback even in informal settings and, for example, “any of the many tools under development to support marking of summative written assignments could be deployed in a formative way as cues for intervention” (Clow, 2013, p. 690). Thus online learning, assessment and feedback cease to be separate events.

I will return to this vision for the future in the conclusion of this thesis (Chapter 4). First of all, Chapter 3 considers the work described in the publications more specifically, building on the general assessment and feedback literature (Sections 2.1 and 2.2), the literature of computer-marked assessment (Sections 2.3 to 2.8) and the literature of learning analytics (Section 2.9). Given the data-driven focus of the work described in the publications, and its emphasis on understanding and optimising learning, the submission has a strong resonance with the field of learning analytics, as well as with the more established fields of assessment and feedback.

3. E-assessment for learning

This chapter addresses the four research questions in turn, with reference to the work that is described in more detail in the publications, which are summarised individually in Chapter 5 and reproduced in Part 2 of this submission. Sections 3.1 – 3.4 each end with a summary of generally-applicable findings, leading to my conclusions in Chapter 4.

3.1 How do students engage with computer-marked assessment and what factors affect this?

Many other authors have relied on students' self-reported behaviour of engagement with assessment tasks, but the work reported in the publications has gone beyond this by analysing students' actual actions, following direct observation in a usability laboratory (as reported chiefly in Publications 6 and 8) and by indirect observation, based on captured student responses to online questions (as reported chiefly in Publications 6, 8 and 10).

3.1.1 When students attempt questions

Throughout the publications, the impact of a fixed cut-off date on student engagement is clear. Activity on a summative assignment almost always increases in the run-up to a cut-off date. For a formative assignment, use frequently builds as the event for which students are practising, perhaps a summative assignment or an examination, approaches (Publication 6, Figure 3; Publication 10, Figure 2). This behaviour, both observed and reported, led to a concern, expressed in Publication 2, that for *Maths for science* (S151), with a single summative computer-marked end-of-module assessment (EMA) and a single "practice assessment" (PA), student effort was not well distributed throughout the module (Gibbs & Simpson, Condition 2). This is in line with the fact that most students reported using the PA mainly as a "mock" EMA rather than as a bank of additional questions (Publication 3, p. 129).

The publications report three ways in which better distribution of effort has been achieved. Firstly, Publication 10 reports a case in which, following the lengthening of the overall time available for study of a module, many students appeared to be waiting for the EMA to open; it appears that the opening of the assignment acted as a trigger for engagement as well as the approach of its closure.

Figure 1 illustrates two typical student behaviours (for "Student 1" and "Student 2") that are common for all iCMAs with a fixed cut-off date. The group of students exemplified by "Student 1" start the assignment (red dot) on its cut-off date and then attempt all the questions (blue dots) in one sitting; "Student 2" starts the assignment some time before the cut-off date and has several actions on each question, appearing to complete each question (frequently out of order) when they feel ready to do so. The behaviour labelled "Student 3" was only seen for one particular module (*Science starts here* (S154)); students attempted the questions in three distinct batches, corresponding to when it was suggested to them that they should complete the questions on each of three chapters. Scaffolding of this type is the second way in which an improvement in the distribution of effort has been achieved; this is particularly relevant for an introductory module like this one.

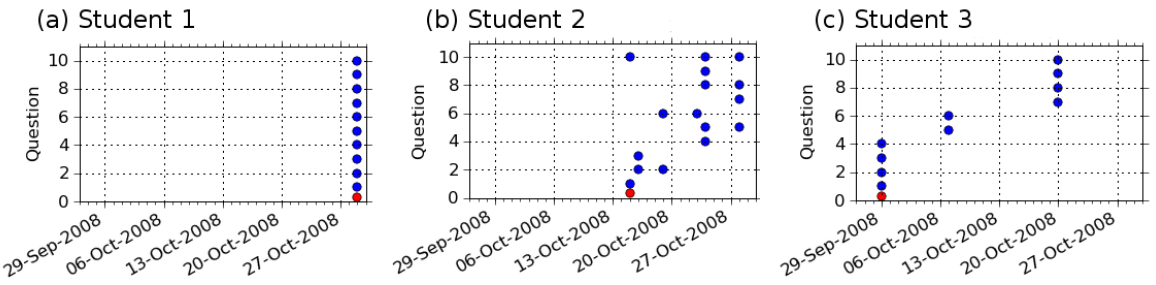

Figure 1. Three typical student behaviours exhibited in summative iCMAs on *Science starts here* (S154). Reproduced from Publication 6 Figure 4. Note: The cut-off date for this iCMA was 29 October 2008.

The third way in which distribution of effort has been achieved is by providing regular cut-off dates through a module's presentation. The effect of this is demonstrated in Publication 6 Figures 2 and 3, and more starkly in Figure 2, which shows iCMA engagement for a module which uses formative thresholded assessment, as described in Section 1.3.2. The iCMAs have fixed cut-off

dates and require students to meet certain thresholds of engagement, in this case by scoring more than 30% in 5 out of 7 iCMAs.

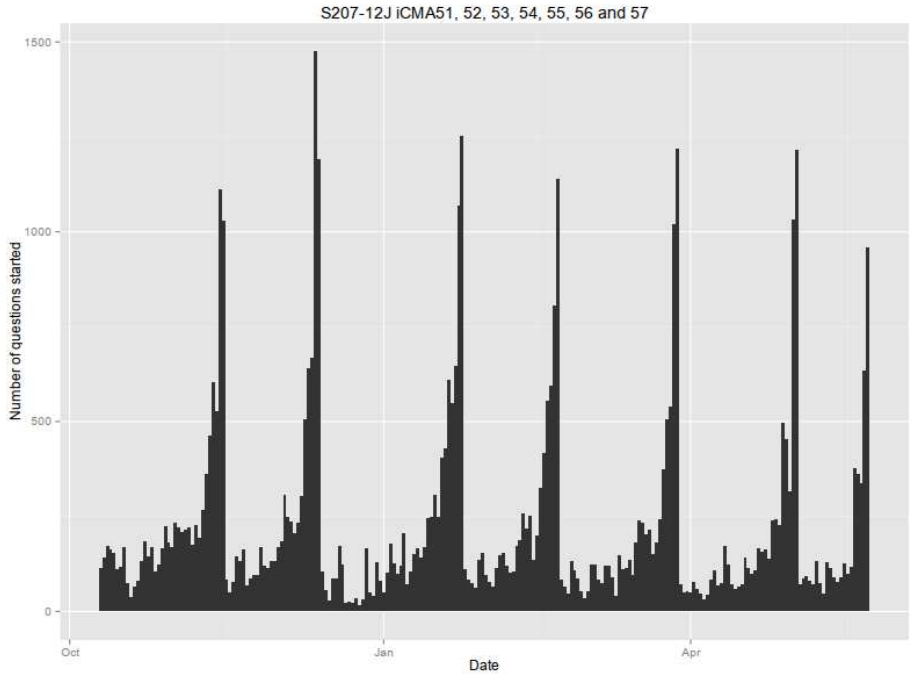

Figure 2. Overall actions on 7 iCMAs, with cut-off dates spread throughout the presentation (and each corresponding to a peak of activity) for a module with formative thresholded iCMAs.

It is possible to draw more general conclusions from these findings. Support has been provided for the hypothesis that students engage with assessed tasks in the way that they perceive their lecturers want them to, and sometimes just providing clear instructions can improve engagement. The "requirement of the teacher" is conveyed to students more powerfully by cut-off dates, but here it is relevant that the cut-off dates are imposed by a computer rather than a human marker. The computer's cut-off dates are non-negotiable, which leads to more equitable practice and removes the situation in which a tutor feels unable to refuse a student's request for an extension, but the student then falls further behind in their studies, to the point at which they feel unable to catch up and so drop out. Assessment design, mediated by a computer, has a role to play in improving student pacing through a module of study, and thus improving retention.

3.1.2 Time spent on questions

As discussed in Section 1.4, time spent on online computer-marked questions has become increasingly difficult to measure because easier and cheaper access to the internet has led to a growing tendency for students to log into an online assignment and leave it open whilst doing other things. In earlier work relating to *Maths for science* (S151), reported in Publications 1, 3 and 10, and making the assumption that if more than 30 minutes elapsed without activity then the student had "gone away", estimates were made for both active time and elapsed time. Active time spent on the 35-question EMA was found to vary from 48 minutes to more than 10 hours, with a median of about 3.5 hours, with a corresponding variation of elapsed time (from a student first accessing the EMA to pressing "submit") and the number of separate sessions spent on the EMA (Publication 1). Correlations between time spent on the EMA and score have been found to be slight (Publication 3, Figure 10.3), and if anything negative (Publication 10, Figures 8a and 8b) i.e. students who spend longer on the EMA score slightly less well. Students who submit early tend to do better (Publication 10, Figure 9), though the early submission is unlikely to be causing the higher score; this correlation is likely to be as a result of more able students submitting early. Similarly, it is not surprising that students who engage with the practice assessment do better on the summative EMA, and that there is a positive correlation between number of practice assessment questions attempted and score on the EMA (Publication 10, Figure 11 and 12). For some students, time spent on the practice assignment was clearly useful, if only in building confidence (Nicol & Macfarlane-Dick, Condition 5); one student was observed to have spent nearly 24 hours active time on the practice assessment, and Publication 1 (p. 12) comments that "it is reassuring that this student passed the course".

3.1.3 Number of questions attempted

Perhaps the most obvious difference in the signatures of use for summative and formative computer-marked assessment is the number of questions that students attempt. Whilst in summative use, most students attempt all the questions in the iCMA, in formative use, there is a typically a noticeable drop in use as the assignment progresses, as shown for different modules in

Publication 6 Figure 5 and Publication 10 Figure 4, with a reduction both in the number of students who have completed each question at least once and in the number of times questions are repeated. Furthermore, there appear to be some users who access the assignment but do not complete any questions at all, for example, 768 students were found to have accessed a formative iCMA in which just 640 students attempted Question 1 and 668 students attempted Question 2 (Publication 6, p. 156 & p.158). The reason for this behaviour is not known (no students admitted to it in interview) but it is speculated to occur when students decide for themselves that the questions are either too trivial or too difficult. An important general conclusion is that the accessing of an assignment (or any other online resource) should not be taken as a proxy for deep engagement.

For a single formative iCMA available for the duration of the module, some of the observed drop in use will be as a result of students withdrawing from the module or ceasing to study it (a behaviour known as "passive withdrawal"). It has also been suggested that students prefer short iCMAs (Ekins, 2008) and that the observed attrition may be partly a result of having a single long iCMA; however, Figure 3(a) shows a drop in use both throughout and between assignments for a module with separate short practice iCMAs.

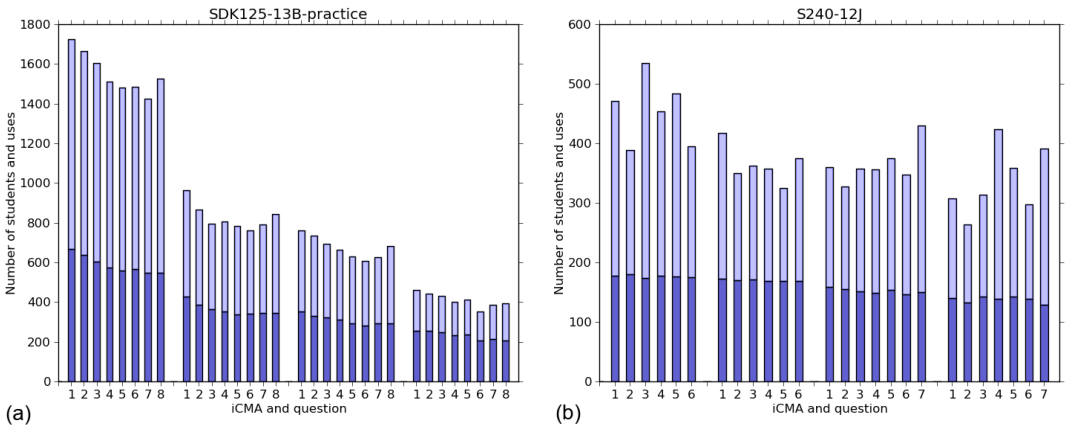

Figure 3. Contrasting patterns of engagement on iCMAs in (a) purely formative use; (b) formative thresholded use. The number of users of each question are shown in dark blue whilst the number of separate uses are shown in light blue. Compare with Publication 10 Figure 4.

Figure 3(b) again shows that formative thresholded assessment may be effective in encouraging deeper engagement, with considerably less reduction in use than shown in Figure 3(a). In the module whose iCMA usage is illustrated in Figure 3(b), students are required to score more than 30% in 3 out of 4 iCMAs and they are also explicitly reminded that the questions are effective revision tools for the module's examinable component.

When students are allowed to repeat questions as often as they wish, which is usually the case in formative and formative thresholded use, the light and dark lines in Figure 3 (also in Publication 6 Figure 5 and Publication 10 Figure 4) show that most questions have between 1 and 3 times as many uses as users. However care should be taken not to interpret this as meaning that each student attempts each question around two times. Figure 4 (below) and Publication 10 Figure 5 illustrate a finding that has been observed in every question whose use has been analysed: Most students attempt the question just once but a few students attempt the question many times and it is this that leads questions to have an average of between 1 and 3 times as many uses as users.

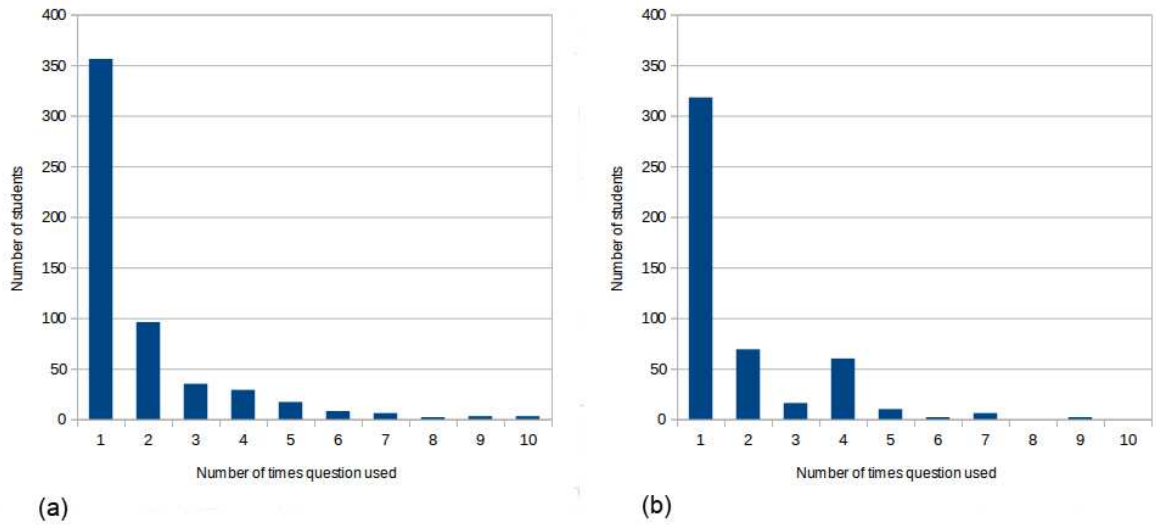

Figure 4. Distributions showing the extent of repeating, for two iCMA questions on *The Physical World* (S207), which requires students to demonstrate engagement by achieving at least 30% in 5 out of 7 iCMAs. Compare with Figure 3 and with Publication 10 Figure 5.

The behaviour illustrated in Figure 4(a) is widely observed; the behaviour in Figure 4(b) is less common and this figure is included here because it shows an interesting result which post-dates the publications but is derived from data-visualisation tools that were developed as part of my

earlier work. Figure 4(b)'s peak on 4 uses is thought to be as a result of the fact that the question had three variants and some students just "clicked through" until they received a variant that they had seen before, and presumably then just copied the answer. When the focus of an assignment is formative, the view can be taken that students are only "cheating themselves" if they engage in behaviour of this sort; fortunately it appears to be uncommon (Jordan et al., in press).

3.1.4 Student responses to short-answer questions

Publication 8 reports on an analysis of engagement with questions in which students were required to give their answer as a short phrase or sentence. A number of features of the answers given were used to indicate the depth of a student's engagement.

Whilst the short-answer free-text questions were being developed, they were used formatively; later they were employed in a range of diagnostic, formative and summative iCMAs. Answers were generally longer in summative use than in formative, but a greater difference in length was observed between responses to different questions. For some questions, the distribution of length was bimodal, with each peak corresponding to different typical responses. So for one question in formative use, whose length distribution is shown in Publication 8 Figure 3, typical responses were "balanced" and "equal and opposite", whilst in summative use (Publication 8, Figure 6) typical responses were "The forces are balanced." and "The gravitational force is balanced by air resistance."

For some questions, the distributions of length were quite different for correct and incorrect responses. For example, Publication 8 Figure 9 illustrates a length distribution for correct answers with peaks on 9 words and 20 words, whilst incorrect answers had a broad length distribution, peaking on 4–5 words. This can be explained by the fact that whilst correct answers specified the direction and extent of a change (e.g. "The force is decreased by a factor of four.") and sometimes included an explanation (e.g. "The force decreases by a factor of four since distance and electric force are related by an inverse square law."), incorrect answers tended to include just one

element of the answer (correct or incorrect) and to give no explanation (e.g. “The force is increased” or “The force is decreased”).

The powerful impact of question wording on student responses was seen when the following text was added to questions:

> You should give your answer as a short phrase or sentence. Answers of more than 20 words will not be accepted.

This wording accompanied the introduction of a filter to limit answers to no more than 20 words, so very long answers, which had been seen with earlier versions of the question, were no longer present. However, the average response length was found to increase for most questions. For the question shown in Publication 6 Figure 6, the initial median response length was 10 words with a range of 0–129 words; with the filter and additional wording in place the median response length was 13 words with a range of 0–20 words. This appeared to be as a result of students interpreting the “more than 20 words will not be accepted” as meaning that an answer of nearly 20 words was required. When the wording in the question was altered to “Please give your answer as a **short** phrase or sentence” the median word length was seen to reduce to 10 words with a range of 0–20 words (Publication 6, p. 160).

In usability laboratory observations, different students were seen to give their answers in different forms, with one student giving perfect sentences “in the same way that I would a tutor-marked assignment” (Publication 8, p. 821) whilst another consciously omitted wording that repeated the question, which he would have included in answer to a tutor-marked assignment. It was not clear whether this was because he knew that his answer would be marked by a computer or simply because the data-entry box was immediately beneath the question on the screen. However another student commented that “the computer is not concerned with my Ps and Qs” (Publication 8, p. 823).

Publication 8 Table 4 classifies student responses as empty (no response), a single word, a phrase, in note form, a sentence (containing a verb and a subject) or a paragraph (several sentences). The

major factor in determining response type was again found to be the question being asked, with considerably more answers in the form of sentences given in response to one question than to another question used on the same presentation of the same module; the second question had a larger proportion of responses given as phrases or in note form. Overall, there were a large proportion of responses given in the form of a sentence, and in work that was not published in the final version of Publication 8, a correspondingly large percentage (35%-62%) of responses were shown to start with a capital letter whilst 19%-46% of responses ended with a full stop. This could be interpreted as support for the Computers as Social Actors hypothesis (Section 2.4), with students behaving as if their work was going to be marked by a human. Alternatively it could be that students were uncertain what factors the computer would take into account when marking their answer and so had decided to take care, or simply that they had been told to give their answer as a sentence – and so were doing just that.

Some of the analysis reported in Publication 8 was done before the introduction of a dictionary to check for spelling, so it was possible to check answers for accuracy of spelling, whether or not any misspelling had been permitted (Publication 8, Table 6). As with length and form of answer, the main variation in frequency of spelling mistakes was between actual questions rather than mode of use, with the smallest number of spelling mistakes associated with questions that could be answered with “everyday” words whilst the largest number of spelling mistakes were associated with questions frequently answered with words that a lot of students seemed to find difficult to spell, in particular “align”, “crystal”, “separation” and “separated”.

3.1.5 Student opinion

Whilst the work described in the publications has concentrated on actual engagement with computer-marked assessment, student opinion is also important (Struyven et al., 2005) and this has also been investigated, especially in the work described in Publications 1 and 6. From the early days, students reported finding iCMA questions “fun” (even in *Maths for science*’s high-stakes summative EMA, as reported in Publication 1). Publication 6 Tables 1 and 2 show that this

was the majority view, with 64%–68% of students strongly or mostly agreeing with the statement "Answering iCMA questions is fun". More significantly, a large majority of students found iCMAs useful both directly (in helping them to learn skills or knowledge) and indirectly (in indicating when further study was needed) and a surprisingly large number were neutral (34%) or disagreed (20%) with the statement "I learn more by doing TMA questions than by doing iCMA questions" (Publication 6 Table 2). Points raised in free-text comments and interviews indicate that this may be to do with the fact that iCMA feedback is received instantaneously (Gibbs & Simpson, Condition 6), but also as a result of the fact that the questions are "not walkovers, not like an American kind of multiple-choice where you just go in and you have a vague idea but you know from the context which one is right" (Publication 6, p. 153). Overall, many students show their appreciation of the "assessment for learning" philosophy by saying things such as "It's more like having an online tutorial than doing a test" (Publication 6, p. 153), and the motivational effect is clear in comments such as "I had never touched differentiation before and when I got the questions on it … correct I felt as good as if I had won the lottery" (Publication 1, p. 140). Given the importance of the affective domain on learning (Kluger & deNisi, 1996, p. 261; Coutts, Gilleard & Baglin, 2011), the motivational impact of high-quality computer-marked assessment should not be underestimated (Nicol & Macfarlane-Dick, Condition 5).

However, it should not be forgotten that a small number of students do not consider iCMAs to be helpful or enjoyable. A decision was taken, in the work described in Publication 6, to interview as many as possible of the students whose questionnaire responses had indicated some disquiet with iCMAs. The emotive language used by two students in their questionnaire responses when describing their negative experience e.g. "really unfair", "unholy unfair" indicates the strength of their feelings. However, significantly, it transpired during the interviews that both of these students were completely happy with iCMAs in general, just not with the particular questions or aspects of their use (Publication 6, p. 154).

Factors that have been found to contribute to student dissatisfaction with iCMA questions include a flawed variant (resulting in all variants being zero-weighted); a student believing, rightly or

wrongly , that their response has been inaccurately marked (e.g. Publication 6, p. 160); and being penalised for what the student believes to be a minor error (Publication 3, p. 129). All of these things highlight the importance of high-quality questions and targeted feedback, and ensuring that students understand both the way the system works and the wording of the question and the feedback. These points will be re-visited several times in this covering paper.

Some of the technical points that students raised in the work described in Publication 1, for example, wanting to be able to answer questions in any order and wanting the availability of different background colours, had been addressed by the time of Publication 6. The concern, discussed in Section 1.3.2, that the need to install the Java Runtime Environment onto their computer was causing some students to give up without talking to anyone was also resolved by the move to web-based delivery. These points highlight the fact that student confidence in technical aspects is important (Gilbert et al., 2009), as is an iterative design process (Benson & Brack, 2010).

As discussed in Section 2.5, some students prefer selected-response questions. However they may be less reliable (Publication 7, p. 8) and they are not necessarily easier than constructed-response questions; Publication 9 Figure 4 (p. 8) shows that more students got short-answer questions right at first try than was the case for selected-response questions for all questions on one presentation of *Exploring science* (S104), whilst several of the *Maths for science* questions that Publication 12 Figure 3 indicates to have been particularly poorly answered were selected-response questions (Publication 12, p. 67). This is in line with Thiede's (1996) finding that students consider recall items to be more difficult than recognition items, even when the recall items were actually less difficult.

3.1.6 Summary of important findings from Section 3.1

How do students engage with computer-generated feedback and what factors affect this?

Care must be taken not to confuse the fact that a student has accessed an assignment (or any online resource) with deep engagement with it. Nevertheless, students have been observed to engage well with computer-marked assessment in a range of settings, and most students appear to like this method of assessment and to find it motivating. Students can be put off by flaws in questions or by believing that their answer was incorrectly marked, highlighting the importance of careful question writing and an iterative design process.

Students engage more when questions are in summative rather than purely formative use, but they may then be unreasonably anxious about the minutiae of grading. Evidence has been given to support the trialling of innovative assessment strategies, such as the Open University Science Faculty's current investigation into the use of formative thresholded assessment.

The strongest influence on student engagement with computer-marked assessment appears to be the impact of question wording and the assessment design (e.g. the use of cut-off dates), indicating both power and responsibility in the hands of academic developers of computer-marked assignments and assessment strategies.

3.2 How do students engage with computer-generated feedback and what factors affect this?

When asked whether they find computer-generated feedback useful, the percentage of students who respond in the affirmative is consistently around 80–90%. 89% of 270 students who responded to a survey strongly agreed or agreed with the statement "The interactive feedback associated with the ECA[3] questions helped me to learn" (Publication 2, p. 484) and when asked "What for you was the most positive aspect of [the module's] ECA?" many students commented on the instantaneous feedback. Similarly, 85% of 129 students who responded to another survey

[3] The *Maths for science* (S151) "end-of-module assessment (EMA)" was known as an "end-of-course assessment (ECA)" at the time this work was carried out and reported.

strongly or mostly agreed with the statement "If I initially get the answer to an iCMA question wrong, the hints enable me to get the correct answer at a later attempt" (Publication 6, Table 2) and the instantaneous receipt of feedback was the most commonly identified useful feature of iCMAs. One student contrasted iCMAs with computer-marked assignments in earlier modules, which were submitted and returned through the post, and so "the answers when they did return after being marked were of little interest as the course had moved on so far by that stage – the iCMA system I think is great – knowing instantly where you are going wrong" (Publication 6, p. 153).

However other findings have painted a less rosy picture of the use that students make of the feedback provided on iCMA questions. Whilst some students have been observed, directly in the usability laboratory and indirectly by changes to their responses following feedback, to make good use of feedback, others have appeared not to use it (Publications 6 and 8). Publication 2 reports low survey scores for "timeliness of feedback" and some students commented "as yet I have not yet had any feedback concerning this course" (Publication 2, p. 484). This section explores the reasons for the apparently contradictory findings and considers factors that influence student engagement with feedback.

3.2.1 Overall use and perception of feedback

The fact that around 90% of surveyed students claim to find feedback helpful is remarkably consistent with findings given in the literature for self-reported usefulness of feedback in other settings; for example Weaver (2006, p. 386) reports that 90% of students agreed with the statement "Positive comments have boosted my confidence", whilst Marriott (2009, p. 243) reports that 93% of students agreed with the statement "I find the immediate reporting of my test result valuable". 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 (e.g. van de Mortel, 2008) and business (e.g. Donaldson & Grant-Vallone, 2002).

Another consideration, mentioned in Section 2.2.3, is a misalignment of student and tutor understanding of the meaning of the word “feedback”. Some of the students who claimed to have received no feedback may have got all questions right at either the first or second try, and so have only received the typical first try feedback: “Your answer is incorrect” in response to incorrect answers, which they may not have perceived to be feedback at all, though they would also have received full worked answers when they got the question right. More significantly, these student comments may have had their origin in the fact that the students had not been told their overall grade (Publication 2, p. 484) and did not yet know whether they had passed the module. Students may have considered the hints and worked answers to be “teaching”, reserving “feedback” for knowledge of score and outcome. Glover, Macdonald, Mills and Swithenby (2005) report a different misalignment between students and lecturers in understanding of the word “feedback” in response to more conventional assignments, whilst Holden and Glover (2013) report on more recent work at Sheffield Hallam University to improve understanding of feedback processes.

The discrepancy reported in Publication 2 may be caused by more than semantics. If, in line with the literature discussed in Section 2.2.3, students just want to know whether they have passed the current assignment or, for formative assessment, whether their work is of a good enough standard to pass (Draper, 2009b), they may be more interested in assignment-level feedback rather than in the detail of what they have done wrong on each question. This emphasis on passing rather than learning is supported by the finding that more students appeared to use the *Maths for science* practice assessment as a trial run for the summative EMA than to practise their mathematical skills *per se* (Publication 3, p. 129). It appears that the high-stakes summative function of the assignment may be compromising learning quite widely, in line with the view of Stobbart (2006) and others, outlined in Section 2.2.2, that feedback is more useful if not accompanied by grades.

3.2.2 Changes to responses after receiving feedback

A consideration of students’ actual engagement with feedback adds a different perspective. For constructed-response questions, the extent to which the data-entry box is left blank and the

extent to which incorrect responses are left unchanged for a second or third try after feedback has been provided gives an indication of student engagement with the question and the feedback provided. (The data can be analysed in the same way for selected-response questions, but the results are less useful because students are more inclined to guess a new answer rather than to leave a previous response unchanged.) Figure 5 illustrates data reported in Publication 6 (p. 158) for the same question, which requires students to calculate a density, in (a) diagnostic use, (b) formative use and (c) summative use. In the figure, grey shading indicates blank responses; green shading indicates correct responses; red, orange or yellow shading indicates incorrect responses; and identical colour from first to second try or from second to third try indicates an unchanged response.

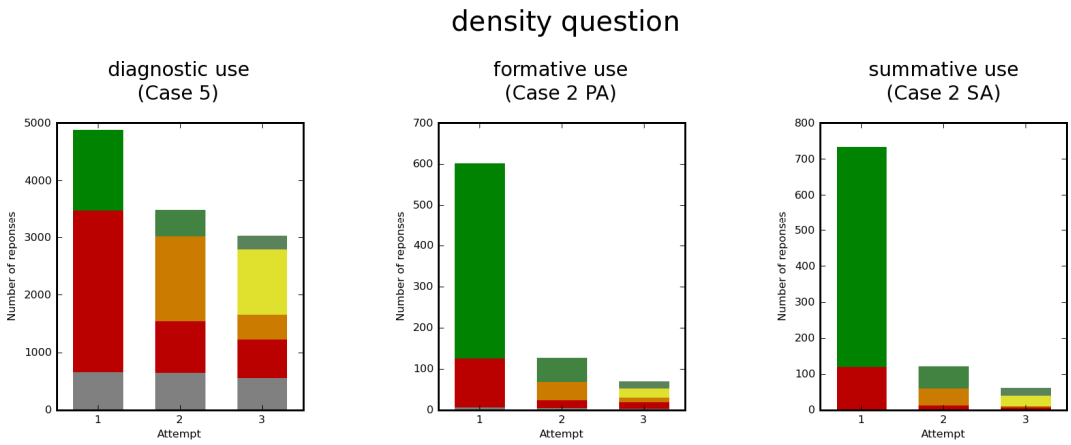

Figure 5. Blank and repeated responses for the same question in diagnostic, formative-only and summative use. Reproduced from Jordan & Butcher, 2010, Figure 14.

Thus, it appears that students are more likely to engage seriously with the feedback provided when the question is in summative than in formative use. In the case shown, in summative use, 21% of the third try responses were identical to those given previously, with 2% of them blank, whilst in formative use, 46% of the third try responses were identical to those given previously, with 7% of them blank (Publication 6, p. 158). However for this particular question, the greater proportion of repeated and blank responses in formative and particularly diagnostic use (where 55% of third try responses were identical to those given previously, with 19% of them blank) is

likely to also have been as a result of the fact that students needed to use a calculator in order to complete the question. In general, a high proportion of repeated and blank responses has been associated with:

- lack of seriousness of engagement;
- questions that are time-consuming to complete, perhaps with multiple boxes for completion, or which require students to access a course component such as a video;
- a lack of understanding of what the question wants or of the feedback provided, in which case leaving the data-entry box blank or repeating a previous response is an "I haven't got a clue" type of reaction from the student (Publication 10, p. 4).

Since the analysis whose outcomes are shown in Figure 5, a "skip to answer" button has been added for questions in diagnostic use, to remove the need for students to click through feedback hints when they have no idea how to answer a question (Publication 8, p. 828).

3.2.3 Impact of the wording of the feedback

Figure 6 illustrates the impact on question behaviour of improving the feedback that is provided to students, in particular after an unsuccessful first try.

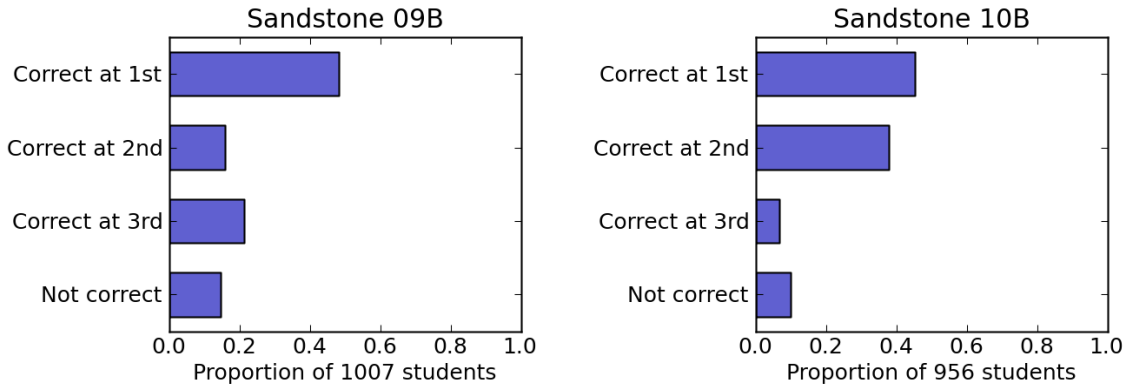

Figure 6. The proportion of students who got the same question right at first, second, third try, or not at all, for the same question in two presentations of a module (identified as 09B and 10B). Reproduced from Publication 8 Figure 11 and Publication 10 Figure 3.

In this case, students had been found to be dissatisfied with a short-answer question, despite the fact that its marking accuracy was known to be good. The question author had required a correct

answer to mention both dry conditions and the fact that the grains forming the rock had been wind-borne, so the answer that many students gave at first try: “It was formed in a desert” was incomplete. Unfortunately many students, when told that their answer was not completely correct, had assumed that the answer matching was faulty. Targeted feedback was added for responses that gave an incomplete answer of this type, for both first try and second try and, as shown in Figure 6, the question’s performance profile changed completely as a result, with many more students getting the question right at the second attempt. Students were also much more satisfied with the question.

Generally applicable insight into student engagement with feedback has also come from the way in which responses have been seen to be amended for a second or third try (Publication 8, Table 7). There are significant differences between the changes made from first to second try compared with from second to third try, between different questions, and as a result of changes to the feedback provided, such as described in the previous paragraph. All of the differences can be attributed to the differing feedback provided, and students’ understanding of the feedback. So, for example, the feedback “Your response appears to be incorrect or incomplete in some way. Have another go, remembering to express your answer as a simple sentence” may have encouraged students to rephrase, delete part of, or add to their answer even when this was not needed. More positively, following the change described in the previous paragraph, the percentage of responses which were added to between first and second try increased from 43.7% to 75.1%. In general, second- and third-try answers were observed to be longer than first-try answers, because they were being added to, and there were a higher proportion of spelling mistakes at second and third try than at first try (Publication 8, Table 7) because of typographical and spelling errors introduced as students amended their answers (Publication 8, p. 829).

Minimal first-try feedback can sometimes be useful, giving students an opportunity to work out their error for themselves; Publication 1 (p. 13) gives an example of a question that 78% of students got wrong at the first try, but nearly half of these students were observed to use the

simple “Your answer is incorrect” feedback to correctly recalculate the answer for the second try. Minimal first-try feedback was also seen to be useful in direct observation; Student C (Publication 8, p. 823) typed “Diffraction is greater” as her first try but commented “unless it’s the other way round”, and when she got the brief feedback that her answer was incorrect she immediately altered it to “Diffraction is less”. Elsewhere, the brief feedback appeared to confuse students, especially when their first answer was partially correct; for example, Student C typed the correct but insufficiently specific response “Differences between levels are higher with helium than hydrogen” at first try (Publication 8, p. 823), but the feedback caused her to alter it to a completely incorrect response “Differences between levels are lower with helium than hydrogen”. Thankfully this student was able to use the more detailed second-try feedback in order to give a correct answer “Differences between levels are 4 times the hydrogen”.

Students were seen to respond in different ways to second-try feedback, some reading it carefully and following references to module materials, whilst others read it more superficially. Targeted feedback appeared to be generally more useful than generic feedback. For example, Publication 2 (p. 483) reports on an analysis of responses to a question that asked for an answer to be given to a specified number of significant figures; targeted feedback resulted in all of the responses that contained an error in their significant figures being corrected by the next try.

3.2.4 Impact of response certitude

In the usability laboratory, one student was observed to read the final full-answer feedback only when his response had been incorrect, and another, who got most of the questions right, read the full-answer feedback on some occasions but not others, explaining that he read the full-answer feedback when he was less confident about his answers, “I’m just going to read the answer to make sure that my thinking was right instead of being lucky”. However, when told that an answer was correct, students appeared “blind” to the full answer. One particularly interesting example came about because the computer marked two students’ incorrect responses as correct. 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 interesting 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 at the time was present to allow 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. However, it was also noteworthy that the student appeared to lose confidence in the computer’s marking at this stage, in similar manner to that observed when the feedback was insufficiently targeted to point out a student’s error, so the student believed that he or she had been incorrectly marked (Publication 6, p. 160). It is possible that students have more innate confidence in human markers, even if computers are more consistent in their marking (Publication 5, p. 494). However, as discussed in Section 2.4, the effectiveness of feedback from a human-marker can also be compromised if the student loses confidence in the tutor giving the feedback (Poulos & Mahony, 2008). Assessment of any sort seems to be more effective when student confidence in the marker is high, a finding that reinforces the importance of accurate marking, whether by a human or a computer, and irrespective of whether the assessed task is formative, summative or thresholded.

3.2.5 Summary of important findings from Section 3.2

How do students engage with computer-generated feedback and what factors affect this?

More students report that they find assessment feedback useful than appears to actually be the case, and evidence has been provided of some misalignment between lecturers’ and students’ understanding of the nature and purpose of feedback. Students may be more interested in how

they are doing overall, and therefore in their grade, rather than in the detailed feedback on each question.

Nevertheless, students like instantly-provided computer-generated feedback, and have been observed to make sensible changes to their responses after receiving it. Feedback is more likely to be attended to when a student is well engaged with the assessed task (in particular if the task is summative), when the task is not too complex or time consuming, and when the question and feedback is well understood.

The detail of the wording of the feedback has been seen to have an extremely powerful impact both on student engagement with the feedback and on their satisfaction with the question. Targeted answer-prompting feedback is generally the most helpful. Answer-giving feedback (in this case, the full answer provided after three tries or when the answer is marked as correct) is more likely to be read if the student has been told that their answer is incorrect. It is important that students have confidence in the accuracy of the marking, a finding that reinforces the importance of ensuring that the marking is accurate and also of providing feedback that students are able to understand.

3.3 What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement?

In considering the potential of computer-marked assessment to provide useful information to educators (Nicol & Macfarlane-Dick, Condition 7), in line with the developing field of learning analytics (Section 2.9), this section presents examples of research into the performance of different variants of questions, common student errors, and different patterns of engagement by different student cohorts. These examples should be considered alongside those presented in Sections 3.1 and 3.2, which provide further illustration of the power of computer-marked assessment to give useful feedback to educators.

3.3.1 Information about the performance of different variants of questions

The analysis of the performance of different variants of computer-marked questions can be seen as a response to Gilbert et al.'s (2009) call for greater use of statistical measures. It arose following work to check whether the statistical tools developed in the 1980s for simple Open University computer-marked assignments (CMAs) were also valid for iCMAs, given their range of question types, provision of multiple tries and the inclusion of multiple variants. No reason was found to doubt the validity of any of the existing tools, though some revised approaches were suggested (Helen Jordan, 2009).

Prior to the work described in Publication 7, the various question-level statistics (facility index (mean), standard deviation, intended and effective weight, discrimination index and discrimination efficiency) had been made available separately for each variant, and in attempting to determine whether variants of a question were of equivalent difficulty, module teams did no more than glance at the different facility indices for different variants. This is a dangerous approach because it is not clear whether any difference in facility index is statistically significant. For this reason, two new tools were introduced: Plots showing the proportions of students with each score (usually 0, 1, 2 or 3); and a single figure (a probability *p*) expressing the likelihood that the observed variation between the variants might have arisen by chance. The use of a plot also brings a degree of alignment with the more recent field of learning analytics, with its sensible emphasis on data representation as well as data analysis (Clow, 2013; Section 2.9).

Figure 7 shows plots and corresponding probabilities for two questions in the same iCMA. The figure is similar to that given in Publication 7 Figure 2, but different questions have been chosen in order to demonstrate the benefits of the approach more powerfully. The two questions whose plots are given in Figure 7 each had 6 variants and there was a similar difference between the highest and lowest facility index for the variants (12 percentage points in each case). However, whilst there was a considerable difference in the behaviour of the variants for the question illustrated in Figure 7(a), statistically significant at the 5% level, there is no evidence to suggest

that the variants of the question shown in Figure 7(b) were behaving in a significantly different way.

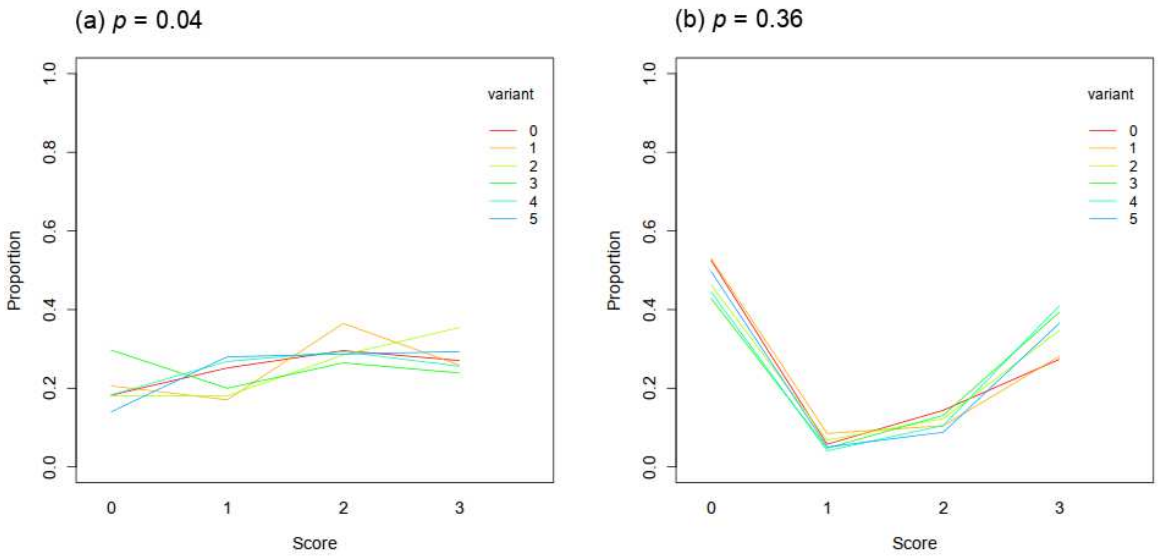

Figure 7. The proportion of students scoring 0, 1, 2 or 3 for each variant of two questions.

Publication 7 gives examples of the tools in use. For selected-response questions, where variants were generated by selecting correct and incorrect options from a longer list, the fact that some options were more obviously right or wrong than others led the variants to have different difficulty. For constructed-response questions, variants were sometimes easier than others because a common mistake led to an obviously incorrect answer, which students corrected for themselves before submitting their response. In other questions, variants were found to be of different difficulty because some of them included words that were more likely to be spelt incorrectly, or where some variants required numbers to be rounded up rather than rounded down. One question had variants of non-equivalent difficulty because of a problem with the font used in the module text; unsurprisingly, students confused Gln (where the middle letter is a lower case L) and GIn (where the middle letter is an upper case i)(Publication 7, p. 6). Thus analysis of the behaviour of variants of questions has led to insight into "the design of effective and equitable questions and assignments" (Publication 7, p. 10) and also into causes of student difficulty, including, but not limited to, problems with the module materials.

Inspired by the work of Dermo (2010), as discussed in Section 2.9, Publication 7 concludes with an analysis of the impact on students' overall score of having received different variants of questions. Factors contributing to the overall size of the effect included the number of individual questions with a significantly different variant difficulty (which, for one module, was caused by a large number of rather poorly written selected-response questions), the number of questions in each iCMA and the overall weighting of the iCMA component in the module score. In line with Dermo's findings, whilst the impact on most students would only be minimal, it is possible that a borderline student could be disadvantaged by the combination of variants that they had received. However, in a difference of opinion with Dermo, it was deemed unrealistic to attempt to correct for the effect, because "not all students are equally disadvantaged by 'difficult' questions nor equally advantaged by 'easy' questions" (Publication 7, p. 10); a better approach would be to alert examination boards when a borderline student has received a batch of questions of above average difficulty, and where question variants have been shown to cause problems they should be removed or improved for future presentations. This again emphasises the importance of an iterative design process; the authoring of high-quality computer-marked questions front-loads the assessment cycle (Bull & Dyson, 2004) but "the front loading is not complete; time needs to be allowed to monitor the behaviour of e-assessment questions once they are in use and resource needs to be set aside to make changes when problems are identified" (Publication 7, p.10).

3.3.2 Information about student misunderstandings

Several of the publications give examples of the way in which an analysis of student responses to computer-marked questions can be used to learn more about the errors that students make, and Publication 12 is a detailed analysis of the behaviour of the questions in the *Maths for science* assignments. The reliability of this approach was found to be greatest for constructed-response questions in summative use (so the students were less likely to be guessing) and when equivalent errors were seen in different variants of questions (Publication 10, p.3).

The initial analysis of *Maths for science* questions, as reported in Publication 1, revealed some unexpected errors, which "would not have been identified as anything other than slips were it not for the large amount of data available" (Publication 1, p. 13). The example given in Publication 1 is of a question in which 9% of all students (50% of those who got the question wrong) made a precedence error and calculated $3^6/3$ (giving an answer of 243) instead of $3^{6/3}$ (giving an answer of 9), or equivalent for different variants. The discovery of this error led to the introduction of targeted feedback for errors of this type (Jordan, 2007). Elsewhere, the discovery of specific errors has led to improvements to teaching materials. It was known that a common error made when students are adding two fractions is to add numerators and denominators separately (Hart , 1981, p. 75) and feedback was already provided for this. However an additional error was discovered in 1.9% of responses in which students correctly found a common denominator but then added the numerators and denominators separately. Inspection of the teaching text revealed that the teaching on this point was unclear, so this was amended for the new edition and in the first presentation of the revised module, just two responses out of 300 (0.7%) included an error of this type (Publication 10, p.3).

In addition to leading to improvements to the module materials and assessment, the analysis of responses has provided broader insight into the errors that students make, as described chiefly in Publication 12. Whilst it must be remembered that insight into errors does not categorically equate to insight into students' deeper misunderstandings (Hadjidemetriou & Williams, 2002, p. 69), the strength of the methodology is the unequivocal identification of errors made by large numbers of students who, at the Open University, are of all ages and come from widely varied backgrounds. This means that blame for weak mathematics cannot be attributed to a particular educational system (Publication 12, p. 64).

Many of the errors observed were similar to those noted by well-known authors on children's mathematical misconceptions (e.g. Sawyer, 1964; Hart et al., 1981), for example students were seen to confuse multiplication with raising a number to a power (e.g. confusing $3x$ and x^3) and to fail to square or cube the conversion factors when converting units of area and volume. Errors

seen in the questions on graphs, gradient and differentiation illustrated the difficulty that students have in discriminating between a value of a point on a curve and the gradient at that point (Kerslake, 1981; McDermott, Rosenquist & VanZee, 1987).

Students made fewer mistakes in rearranging equations than in simplifying them and the most persistent errors, seen in questions designed to assess a range of different skills, were in rounding numerical values to an appropriate number of decimal places or significant figures and in working out the units of an answer. Errors were often at a lower level than expected, so not only did students have difficulty expressing a number to an appropriate number of significant figures, they also had difficulty in simply stating the number of decimal places given. Howarth and Smith (1980) identify the widespread nature of mathematical problems at a more fundamental level than students are prepared to admit, whilst many tutors assume that students have a greater fluency in basic mathematics than is actually the case (Leopold & Edgar, 2008).

An attempt was made to classify the errors into separate categories that could be attributed (with some uncertainty) to careless mistakes, a lack of understanding of a method and deeper conceptual misunderstandings. The most common errors observed in *Maths for science* questions were frequently not related to the question's primary purpose, so errors in rearranging equations were seen in the trigonometry questions, faulty addition of fractions was seen in a probability question and precedence errors were seen in a statistics question. This finding raises some doubt concerning the practice of teaching of mathematics in a scientific context; students may need to hone their mathematical skills before transferring them to a different context (Britton, New, Sharma & Yardley, 2005; Hoban, Finlayson & Nolan, 2013).

It is of vital importance that educators have an accurate and detailed picture of their students' understanding, and the analysis of responses to computer-marked questions provides a good starting point. Whilst there may be doubt over the interpretation of the data, the data themselves do not lie! The tutor time freed by the use of computer-marked assessment can be used for

discussion with students, both to discover more about the reasons for their errors and to correct misconceptions.

3.3.3 Information about student engagement

Sections 3.1 and 3.2 summarised findings from the publications concerning the analysis of student engagement with computer-based assessment and computer-generated feedback in different ways. Having established proxies for student engagement, it became possible to compare the behaviour of two interactive computer-marked assignments, known to be of equivalent difficulty, in use on two modules with contrasting student populations . This opportunity came about because the rewritten *Maths for science* book (Jordan, Ross & Murphy, 2013) was used in two different modules, with one of the previously analysed iCMA-based end of module assignments (iEMAs) used, with only slight modification, as an iCMA on Module Y[4] whilst the other was used as the iEMA for Module Z. Whilst there were differences between the two modules and the way in which the assessment was embedded within them, Publication 10 hypothesises that the observed difference in behaviour was primarily a result of the different student populations and the intensity at which they were studying.

The different student populations of Module Y and Module Z are described in more detail in Section 5.10.2. In summary, students who started Module Z in October 2012 tended to be older and to have higher previous educational qualifications than those who started Module Y at the same time. The students on Module Y were also more likely to be new to the Open University and to be studying at higher study intensity.

Unsurprisingly, student engagement with the computer-marked assessment in Module Y and Module Z was different, but the extent of the difference was surprising. Module Y students started the assignment much closer to the cut-off date than Module Z students did, and Figure 8 shows the relationship between the date the computer-marked assignment was completed and

[4] Two modules were identified as A and B in Publication 7 and a different two modules were identified as A and B in Publication 10. To avoid confusion, all have been renamed. Modules A and B in Publication 10 are here referred to as Y and Z.

the student's score on it. Although the plots are superficially similar and have similar correlation coefficients, the behaviour was actually quite different. As with Figure 7, careful inspection of the visual representation of the data provides rich rewards.

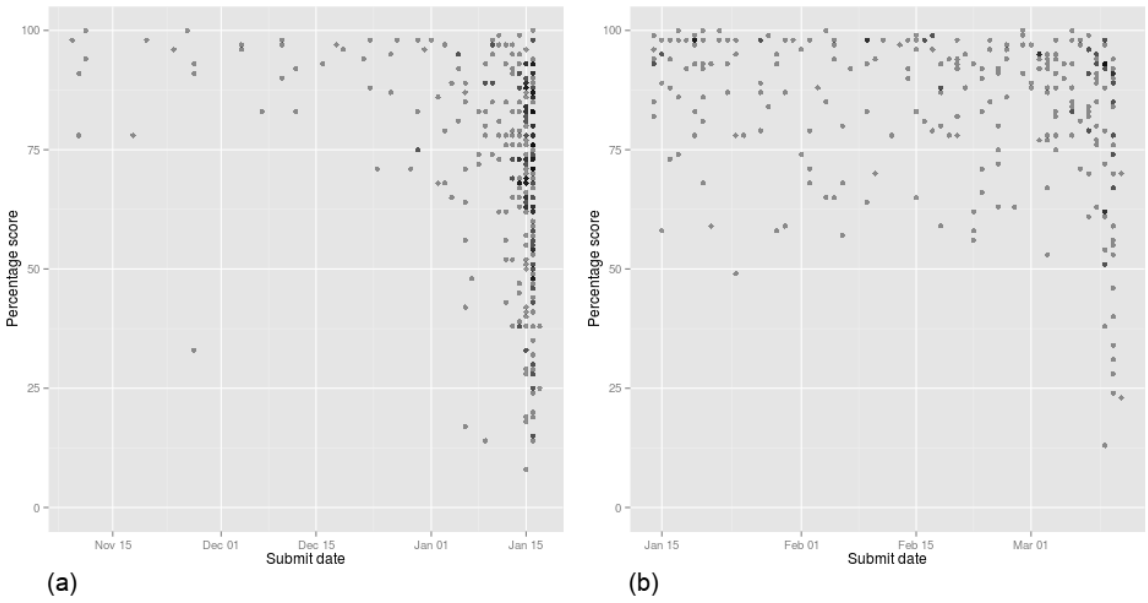

Figure 8. Scatter plot of score against date submitted for (a) Module Y iCMA (*n* = 316; Spearman rank correlation coefficient = −0.41; (b) Module Z iEMA (*n* = 272; Spearman rank correlation coefficient = −0.32). Reproduced from Publication 10 Figure 9.

For Module Y (Figure 8(a)), the small number of students who completed the iCMA more than a few days before the cut-off date all did well, but there were a large number of students who appeared to be rushing to complete the iCMA by the due date and who did not do well. By contrast, for Module Z, Figure 8(b) shows a good spread of marks, all above 50%, for the duration of the assignment. The nine students who submitted the iEMA right at the end were probably encouraged to do so by an email reminding students of the approaching due date, and they are unlikely to have been surprised to learn that they had failed the module.

As discussed in Section 3.2, the extent to which responses are left blank or repeated after receiving feedback is another good indicator of student engagement, and it was useful that the assignments for the two modules included one identical question. For this question, whilst only 2 students (both on Module Y) left the data-entry box empty, 22% of third try responses for Module

Y repeated a previously given response, compared with 14% for Module Z. These data are shown visually in Publication 10 Figure 2.

Finally, student engagement with the purely-formative *Maths for science* practice assignment (PA), shared by the two modules, was very different. For Module Y, just 36% of the students who submitted the iCMA had attempted the PA, in contrast with 80% of students for Module Z. This was despite the fact that the PA had been well advertised to both groups of students, if anything slightly more to Module Y students than to those studying Module Z. Inevitably, for both modules, those who had attempted the PA did better overall, and there was a positive correlation between the number of questions attempted in the PA and final score. This is unlikely to be causal, but it is unfortunate and noteworthy that so few Module Y students took up the opportunity for extra practice. As stated in Publication 10 (p. 11) "it is well known that busy students do not engage with aspects of a module that are not assessed, especially if they fail to appreciate the benefit to themselves of doing so".

All of these factors indicated good student engagement on Module Z, but that Module Y had many students who were lacking in the time and motivation to engage fully and whose success was compromised as a result. This was entirely consistent with the less well prepared and overcommitted student population on Module Y. Indeed, student behaviour on the Module Y assignments provided an early warning of later study difficulties on the module; this point is discussed further in Section 5.10.2.

More generally, the work described in this section has illustrated the power of assessment analytics to indicate student engagement with a module and to provide an early warning of at risk students for whom proactive intervention would be appropriate.

3.3.4 Summary of important findings from Section 3.3

What is the potential of computer-marked assessment to give feedback to educators about student misunderstandings and engagement?

Section 3.3 has added examples to those from Sections 3.1 and 3.2 of feedback provided to educators about student misunderstandings and engagement. Large data sets give powerful evidence, whilst the analysis and representation of the data for the purposes of understanding and optimising future learning aligns the work with the field of learning analytics.

Statistical analysis and visual representation of the behaviour of different variants enables equitable assessment, but it also provides feedback to educators that can lead to improved assessment design and to insights into the mistakes that students make. Further information about student misunderstandings can be obtained from the detailed analysis of responses to questions, and this is most powerful when the questions are constructed-response, in summative use, and when similar errors are seen in different variants of questions.

Knowledge of student errors can lead to improvements in both teaching materials and the assessment itself, provided time and resource is available to enable an iterative design process.

Assessment analytics have been shown to provide evidence of the depth as well as the extent of student engagement with a module, and can thus give an early warning when students are experiencing difficulties.

3.4 What is the scope of computer-marked assessment in supporting learning and what are the barriers to wider take-up of sophisticated computer-marked tasks?

Most of the research outlined in this section concerns one innovative type of computer-marked assessment, namely the automatic marking of questions in which students give their answer as a free-text phrase or sentence. The work is summarised in the context of the relevant literature in

Section 2.7 and described in more detail in publications 4, 5 and 9. Section 3.4 concludes by discussing the barriers to the take-up of short-answer free-text questions, many of which have implications for the wider adoption of other sophisticated computer-marked tasks.

Short-answer free-text questions at the Open University were initially developed using Intelligent Assessment Technologies Ltd. (IAT)'s FreeText Author software (Publication 4, pp. 373-375) operating within OpenMark for the provision of feedback (Publication 4, pp. 376-378). Later, OpenMark's own pattern-matching software (Publication 5, p. 499; Publication 9, p. 3), now called PMatch, was used. Initially answers of any length were allowed though this was later restricted to no more than 20 words. Publication 8 (summarised in Section 3.1) discusses student engagement with short-answer free-text questions, including the impact on the responses received of the wording of the questions and feedback.

3.4.1 Human and computer marking accuracy for short-answer free-text questions

Haley, Thomas, Petre and De Roeck (2009, p. 83) state that they "believe that a CAA system has good enough accuracy if its results agree with humans as well as humans agree with each other" and it is clear from research described in Publication 4 (pp. 378-380) and Publication 5 (pp. 492-493) that the IAT marking was generally more reliable than that of six course tutors. On some occasions, one human marker consistently marked in a way that was different from the others, while on other occasions an individual course tutor marked inconsistently, marking a response as correct, when an identical one had previously been marked as incorrect, or vice versa. Divergence of human marking could sometimes be attributed to insufficient detail in the marking guidelines that had been provided, or to uncertainty over whether to give credit for a partially correct solution. However, some errors were caused by slips and by poor subject knowledge or understanding.

Reasons for inaccuracies in the IAT system's marking ranged from the system's failure to identify a sentence structure or a misspelt word, to a completely missing "mark scheme template" i.e. when the question author had failed to think of a particular way in which a correct or incorrect answer

might be expressed, recognised as more likely to occur when insufficient responses from students had been used in question development (Publication 5, p. 493). One of the other sources of error was an artefact of the way in which the IAT and OpenMark systems were working in tandem: The IAT system set a flag when a response did not match a mark scheme template but was considered to be "close" to it, and a decision had been taken to give such responses the benefit of the doubt, i.e. to mark them as correct in OpenMark. Unfortunately, responses with the correct keywords but incorrect word order, or where the response included the word "not" in addition to a correct answer, were flagged as "close" by the IAT system and so marked as correct by the combined IAT/OpenMark system, resulting in a number of false positives for some questions.

The final reason for inaccuracies in IAT's marking, again resulting in false positives, was the difficulty of marking responses that included both correct and incorrect aspects, recognised as "potentially serious problem for free text marking" (Mitchell et al., 2002, p. 245). Publication 4 (p. 380) explains that "whilst any individual incorrect response of this nature can be dealt with by the addition of a 'do not accept' mark scheme in FreeText Author, it is not realistic to make provision for all flawed answers of this type". Publication 5 p. 494 gives an example of a response of this type: "That there is a balanced force acting on the hailstone with more downward force (gravity)". The first part of the answer is excellent, but the addition of "with more downward force (gravity)" indicates flawed understanding. It is in marking responses of this type that human markers are at an advantage to most computerised systems. However it should be emphasised that, "contrary to e-assessment folklore, responses of this type do not originate from students trying to 'beat the system'… but rather by genuine misunderstanding" (Publication 4, p. 380).

Having ascertained that the IAT system working within OpenMark was generally as accurate as human markers, the accuracy of this system, which employed some natural language processing techniques, was compared with that of two simpler algorithmically-based systems, OpenMark's own pattern-matching software (PMatch) and Regular Expressions. An undergraduate student (not of computer science) was employed and given a set of marked training responses to use in

developing the answer matching. The marking of IAT within OpenMark, PMatch and Regular Expressions was then compared by testing the PMatch and Regular Expressions answer matching against the same sets of responses that had been used for the human-computer marking comparison. This showed that each of the three systems had the best agreement for at least one question, with IAT/OpenMark and PMatch generally performing better than Regular Expressions, possibly because the student who had developed the answer matching had not had sufficient time to learn how to use Regular Expressions properly.

Optimisation of all the systems' answer matching and further testing confirmed that PMatch marking was generally as good as the IAT/OpenMark system (Publication 5, Table 5); furthermore, the PMatch answer matching had been developed more quickly than that for IAT/OpenMark and using freely available software. A decision was therefore taken to move the answer matching for the questions that were in use on Open University modules across to PMatch. In 2012 a further check of PMatch marking accuracy was performed (Publication 9, pp. 5-6). For each of eleven questions, between 1591 and 2218 responses were marked by a single human marker with good knowledge of the question author's intentions. The percentage agreement with the human "expert" ranged from 96.8% to 99.3% for the eleven questions and the kappa inter-rater statistic was greater than 0.9 for all but one question (Publication 9, Table 2)[5]. The kappa statistic for this question (κ = 0.84) was lower than would be expected on the basis of percentage agreement (98.2%) because the question was very well answered by students, so the computer and human were more likely to agree by chance. This was reflected in the kappa statistic.

3.4.2 Reflection on computer marking accuracy

Data presented in the publications has thus shown not only that computer systems can mark short-answer questions as accurately as human markers, but also that the computerised marking system does not need to be sophisticated in order to achieve good results. However, several points should be emphasised:

[5] A kappa statistic of greater than 0.8 is conventionally taken to indicate excellent agreement.

Firstly, whether IAT or PMatch software was being used, the answer matching was developed iteratively in the light of responses from students on the modules for which the questions were designed. This is of critical importance in achieving high marking accuracy, as explained in the final report of my investigation into the potential of short-answer free-text questions:

> Responses to the questions from friends, family members and colleagues have been extremely useful in helping us to develop the questions to a point at which we are able to offer them to students in the first place. However, for subsequent development, these responses are not an adequate alternative to answers from real students. However hard we ask them not to, academics try to make the questions "fall over" in a way that students do not seem to do, and they do not have the same misconceptions as students. (Jordan, 2009b, p. 8).

Indeed, on the one occasion when a few short-answer free-text questions were developed for formative use on the basis of the answers expected from students, rather than on the basis of actual student responses, for one question just 57% of the first batch of 848 student responses were accurately marked by the IAT software (unpublished result).

Secondly, although PMatch answer matching is simple and relies primarily on keywords, it is not a simple "bag of words" system; it is significant that it can also take word order and negation into account.

Finally, PMatch uses two complementary systems for checking and correcting spelling mistakes, as described in Section 2.7.

3.4.3 Limitations to the use of short-answer free-text questions

In principle, answer matching can be written for any question that has distinct correct and incorrect answers that can be expressed in the form of short phrases or simple sentences. Questions are likely to be unsuitable if there are less clear "right" and "wrong" answers, or if the expected responses are complex in nature. In addition, where there are many different ways of

expressing a correct or incorrect response, developing the answer matching rules can become tedious. The “Snowflake” question (Publication 9, p. 7) has 23 independent rules in PMatch, and very many hours have been spent in developing the answer matching for this question.

The iterative development of answer matching adds to the time required and makes this difficult to quantify. For the questions described in Publications 4, 5 and 9, the initial answer matching took between minutes and hours to write, whilst the refinement in the light of further responses took from hours to days. There is a tension between accepting that a question’s answer matching is “good enough” and matches the vast majority of student responses (which Publication 9 speculates might have been the case with 10 independent rules for the “Snowflake” question) and continuing to strive for perfection. In practice, the questions reviewed in Publication 9 were deemed to be in good enough shape to be used without further intervention for the lifetime of the module.

In the earlier iterative design process described in Publication 4, of the 78 questions originally authored, four were deemed unworkable. In a further 13 cases, changes were made to the wording of the questions themselves, again emphasising the importance of question wording (Section 3.1.6). In most cases the change in wording was minimal, but sometimes it was more substantial, acting to more tightly constrain the student responses. For example, the question “You are handed a rock specimen that consists of interlocking crystals. How would you decide, from its appearance, whether it is an igneous or metamorphic rock?” became “You are handed a rock specimen that consists of interlocking crystals. How could you be sure, from its appearance, that this was a metamorphic rock?” (Publication 4, p. 382).

Reliable answer matching was obtained for a question where a correct answer needed to contain three separate components: that the rock was formed from magma [*or* molten rock]; that the magma cooled [crystallised/solidified]; and that the cooling took place slowly [over a long period of time/deep underground] (Publication 4 Figure 4). However, if students are required to write about two or more separate concepts in one answer, matching can be difficult and the only way forward may be to split the question into two separate parts. Sometimes this can be achieved

without severe detriment to the assessed task but "in other cases, splitting the task would substantially alter its function and so is unlikely to be a desirable way forward" (Publication 4, p. 382). However, just because questions require a short answer, it does not follow that they can only be used to assess low-level learning outcomes (Publication 9, p. 8).

3.4.4 Overcoming the barriers to take-up of short-answer free-text questions and other types of sophisticated computer-marked assessment

Despite the fact that relatively simple software has been shown to provide more consistent and generally more accurate marking accuracy than human markers, that a pattern-matching question type is now available in Moodle (Publication 9, p. 3), and that relatively simple short-answer questions can assess a range of learning outcomes, most question authors continue to use selected-response questions (Hunt, 2012). What are the real barriers to wider take-up? This section proposes four possible reasons and suggests ways in which the barriers might be overcome.

Whilst good marking accuracy has been demonstrated in the publications, it will never be possible to obtain 100% marking accuracy for short-answer free-text questions. On one level this should not matter, since the marking accuracy is higher than that of human markers. However, as discussed in Section 3.2.4, there is some evidence that students are less sympathetic towards inaccurate marking from a computer than from a human marker, and more likely to assume that a computer has made a mistake, further emphasising the importance of making the marking as accurate as is possible, within reasonable resource constraints, and the importance of helpful feedback. A larger training set will increase the accuracy of the computer's marking (Publication 5, p. 497) but there is a law of diminishing returns and even for answer matching based on the inspection of thousands of responses, occasional answers have been observed to be incorrectly matched simply because they were expressed in a way that had not been seen before (Publication 9, p. 9). One way forward would be for the computer to only mark short-answer free-text when the response matches a specific rule, and to pass the others for human marking (Publication 5,

p. 498). There are other situations where a similar hybrid approach may be most effective (Publication 11, p. 14), for example, SaiL-M (Semi-automatic analysis of individual Learning processes in Mathematics) automatically monitors student interactions and then, if necessary, passes these to a tutor who supplies detailed feedback (Herding & Schroeder, 2012). The use of hybrid human-computer systems are further discussed in Section 4.2, as are situations where it may remain more appropriate to use human markers.

Another fundamental issue for the development of short-answer free-text questions is the need for a large number of marked responses for use in developing the answer matching, gathered from real students who are making a serious attempt at the questions. Publication 9 (p. 9) reports that the number of responses required to develop sufficiently robust answer matching varies considerably from question to question, but it is usually measured in hundreds of responses. For modules with large student numbers, long lifetimes, and some way of capturing student responses for use in question development, answer matching can be iteratively designed for as long as the question author has the time and motivation to do so. Modules with small student numbers face a more serious problem. Possible solutions include gathering responses from other forms of assessment (e.g. examinations), the use of peer marking, and sharing questions, though there is then a need to check that responses from different student cohorts are sufficiently similar (Jordan & Butcher, 2013, p. 4).

Linked to the need for a large number of marked responses is the fact that developing the answer matching takes a non-trivial amount of time. Again, this is likely to be more of an issue for modules with small student numbers and where questions cannot be re-used, since it is then less easy to justify time spent in developing answer matching in terms of future marking time saved. In some situations, the time spent developing answer matching may be prohibitive, but with the increased use of e-learning and in particular with the development of MOOCs (Massive Open Online Courses), it becomes both easy to collect responses and easier to justify the time spent in developing answer matching. Pulman and Sukkarieh (2005) have attempted to use machine learning to generate response-matching patterns, with limited success; I am aware of some more

recent work in this area, but evaluated and published work is urgently required. Using machine learning in this way would be a positive response to Phil Butcher's plea (quoted in Publication 11, p. 13-14) that we should start to "use the computer as a computer" in our e-assessment systems.

Perhaps the largest barrier to the wider take up of short-answer free-text questions, and other sophisticated question types, remains the fact that academic authors are not aware of their existence, and are lacking in the necessary time and opportunity for professional development to learn how to use to the authoring tools. The IAT software was selected for use at the Open University because it could be used by a question author with no knowledge of natural language processing (Publication 4, p. 374-375). However, learning how to use the authoring tool was undoubtedly time-consuming (Publication 5, p. 497) and "the writing of good e-assessment questions and embedding them within an appropriate assessment strategy that truly supports student learning are skills that should not be underestimated" (Jordan, 2009, p. 18).

3.4.5 Summary of important findings from Section 3.4

What is the scope of computer-marked assessment in supporting learning and what are the barriers to wider take-up of sophisticated computer-marked tasks?

Computerised systems can mark short-answer questions as accurately as human markers, and a computerised marking system does not need to be sophisticated in order to achieve good results. However, consideration should be paid to word order, negation and incorrectly spelt words. Most importantly, the answer matching should be developed iteratively, in the light of marked responses from students. This adds to the time and resource required and means that this type of assessment is more useful for large class sizes and where questions can be re-used.

In order to be appropriate for use as a short-answer free-text question, a question should have clearly defined right and wrong answers, and even then 100% marking accuracy is unlikely to be attainable. Question wording is again very important and if a question needs to be split into two it may alter the nature of the assessment task.

Suggestions for overcoming barriers to the use of sophisticated computer-marked assessment of this type include using hybrid systems in which responses that the computer cannot mark are passed to a human marker, the sharing of questions and the use of machine learning to assist with the development of answer matching. However the largest barrier to wider take-up remains an inherent conservatism, lack of confidence and lack of time which leads academics to continue to prefer selected-response questions over constructed-response ones. The resource and staff development implications of the use of high-quality computer-marked assessment are further discussed in the conclusion (Chapter 4), as are several of the exciting possibilities for the future of computer-marked assessment that have been mentioned previously.

4. Conclusions and suggestions for the future

4.1 Review of the research questions

My literature review started by introducing Gibbs and Simpson's (2004-5) conditions under which assessment supports learning and Nicol and Macfarlane-Dick's (2006) conditions for good feedback progress. My data-driven analysis of engagement with computer-marked assessment and computer-generated feedback at the Open University has led to findings of relevance to my individual research questions, summarised in Sections 3.1.6, 3.2.5, 3.3.4 and 3.4.5. I am now in a position to draw overarching conclusions and thus to review the contribution that my work has made to the field.

The strongest influence on student engagement with computer-based assessment and computer-generated feedback appears to be the student's understanding of what they were required to do and their understanding of the wording of the question and the feedback. Repeated evidence has been presented of students as "conscientious consumers" (Higgins et al., 2002), doing exactly what they thought a question or piece of feedback intended them to do, which was not necessarily what the question author had intended. This highlights the importance of checking that assessment items are behaving in the expected way, and of using an iterative design process (Benson & Brack, 2010). It also reinforces the importance of the careful wording of assessment items, targeted feedback, and a shared understanding with students as to the nature and operation of assessment practices.

In line with the findings of Kibble (2007), I have found that students engage more thoroughly with the tasks when they carry some summative weighting, but there is then a danger that the summative function of the assessment may over-ride the formative, as observed by Miller (2008, p.190). The use of formative thresholded computer-marked assessment, with fixed cut-off dates (Jordan et al., in press) shows potential as a way of encouraging engagement and helping students to pace their studies, in line with Gibbs and Simpson's Condition 2, whilst enabling them to focus

on the feedback provided and to repeat the questions as often as required. Computer-mediated cut-off dates are an example of the powerful influence of a remote "teacher" on student behaviour.

There has been some evidence of students losing confidence in the computer-marking software when they were sure that their answer was correct but it had (rightly or wrongly) been marked as incorrect. At this point, students appeared to be acutely aware that they were being marked by a computer not a human marker. Walker et al. (2008) and Lipnevich and Smith (2009) have reported similar results. Similarly, when students reworded answers after being told that their first try was incorrect, this was sometimes because they believed that the computer had not "understood" them, an assumption they may be less likely to make of a human marker. However, the credibility of tutor-marked assignments and the perceived usefulness of feedback provided have been found to relate to the students' perceptions of the lecturers themselves (Poulos & Mahony, 2008). Assessment of any sort appears to be more effective when the student confidence in the marker is high, a finding that reinforces the importance of high-quality questions, accurate marking, and the need for feedback to be understood by students.

The analysis of student responses has been shown to be a powerful tool in providing information to educators, indicating faulty assessment items and misleading course text as well as more general student misunderstandings of the subject matter. Engagement with computer-marked assessment has also been used as a proxy for more general engagement on a module, aligning the work with the newly defined field of assessment analytics (Ellis, 2013), and with a potential to act as an early-warning device for individual students in difficulty. Ellis points out that assessment is ubiquitous; the work described in this thesis also demonstrates that the analysis of students' use of computer-based assessment and computer-generated feedback can provide considerably deeper insight into their engagement with a module than is provided by considering simply whether a student has clicked on an online activity. I therefore contend that assessment analytics provide a particularly powerful type of learning analytics.

An initially surprising but persistent finding from my research has been that relatively simple computer-based systems can mark short-answer free-text questions more consistently and at least as accurately as human markers. Students have been observed to engage well with questions of this type. However, the wider take-up of short-answer free-text question has been poor, with many academics preferring to use selected-response questions, even when concern has been expressed over their validity and authenticity. The reasons for this poor uptake include the need for a large number of marked responses for use in developing the answer matching, and the fact that 100% accuracy of answer matching cannot be assured. These points are further explored in Section 4.2. More generally, we should not lose sight of the need for staff resource and professional development to encourage educators to use more sophisticated question types and to develop the skills to produce high-quality assessment items.

4.2 Suggestions for the future

Whatever the future of MOOCs (Massive Open Online Courses), a beneficial side-effect is that they are forcing the assessment community to consider appropriate methodologies for assessing the informal and online learning of huge numbers of students. Two common approaches are peer assessment and computer-marked assessment, but there is a danger that the rush to create courses quickly and at low cost will lead to poor quality assessment. To avoid this, MOOC systems should support interactive computer-marked assessment with instantaneous and meaningful feedback, different variants of questions as required, and a wide range of question types. One of the previous limitations to the wider uptake of short-answer free-text questions has been a shortage of responses from students to use in developing the answer matching; MOOCs solve that problem immediately! There is still a need to mark the responses and to develop the answer matching, but a creative use of peer marking to agree on acceptable and non-acceptable answers, followed by the use of machine learning to guide the development of matching rules, offer a promising way forward. Short-answer free-text questions also offer the potential to establish inventories of threshold concepts, such as the force-concept inventory, with greater accuracy

than can be achieved by the use of multiple-choice questions alone (Rebello & Zollman, 2004; Section 2.5).

What is the limit of appropriate use of computer-marked assessment? It is my view that computers have huge potential to assist learning, in particular to assess tasks where they can do so more accurately, consistently and sometimes more authentically than human markers. For example, when the task is to write a computer program, why should we not assess that task by running the code that the student has written, as Lobb (2013) has done? Neither human markers nor computer systems can mark free-text responses with 100% accuracy (Basu, Jacobs & Vanderwende, 2013), but a combined system of computer marking (for most of the responses) and human marking (for responses that are identified as "borderline" or "difficult" by some measure in the computer-marking system) is likely to attain more accurate outcomes than either computer or human alone. Finally, there is the contentious territory of assessed tasks (e.g. essays, experimental reports, proofs) where, for the foreseeable future, it may be best to recognise that the most appropriate and authentic markers are human. My conclusion on this point remains as it was in 2009 (Publication 4, p. 383):

> If course tutors can be relieved of the drudgery associated with marking relatively short and simple responses, time is freed for them to spend more productively, perhaps in supporting students in the light of misunderstandings highlighted by the e-assessment questions or in marking questions where the sophistication of human judgement is more appropriate.

The title of this thesis, "E-assessment for Learning" is unashamedly borrowed from my blog "e-assessment (f)or learning" (http://www.open.ac.uk/blogs/SallyJordan/), in which the bracketed 'f' hints at my long-held belief that the relationship between assessment and learning is complex, to the extent that it is meaningless to attempt to separate the two. Data logging in an online environment offers the possibility to assess students on their actual interactions rather than adding assessment as a separate event (Redecker et al., 2012) and learning analytics can also be used to generate assessment-like feedback (Clow, 2012). In this conception, high-quality

computer-marked questions have an important part to play in scaffolding learning, creating "moments of contingency" (Black & Wiliam, 2009) and providing accurate and timely information to tutors and, most importantly, to students themselves. Students retain responsibility for their own learning (Boud, 2000), "ushered" by tutors (McArthur & Huxham, 2013), with both students and tutors supported by high-quality computerised systems and with "feedback" regaining its rightful place as a process involving all elements in a continuous cycle of learning.

5. Summaries of the publications and their context and reception

5.1 Publication 1

Jordan, S. E., Butcher, P. G., & Ross, S. M. (2003). *Mathematics assessment at a distance.* Maths CAA Series. Retrieved 1st May 2014 from http://www.heacademy.ac.uk/assets/documents/subjects/msor/mathscaa_jul2003.pdf

5.1.1 Summary of Publication 1

This paper describes the original project that introduced web-based assessment into the Open University 10-credit module *Maths for science*, presented for the first time in September 2002. Students received immediate, targeted feedback on their answers, with an opportunity to have another go whilst the feedback was fresh in their mind. The paper also describes early evaluation, which included consideration of technical robustness, some analysis of student responses to the questions, and reflection on student comments received both anecdotally and as part of a short end-of-module survey at the end of the first presentation. The work described in the paper is of most relevance to Research Questions 1 and 2, though its feedback on student engagement and student misunderstandings contributed some early answers to Research Question 3, and its reflection on the success and limitations of the project is of relevance to Research Question 4.

There was a wish to move beyond the constraints imposed by the use of multiple-choice questions alone, so a range of different question types were used.

The summative end-of-module assessment (EMA)[6] was complemented by a purely formative practice assessment (PA), which students could access as many times as they wanted throughout the duration of the module. Students were encouraged to attempt the PA early in their studies,

[6] In Publications 1, 2 and 3, the "end-of-module assessment (EMA)" is described by its previous name of "end-of-course assessment (ECA)"

largely to check that the technology was working as it should be, and they were also encouraged to complete the EMA well before the deadline. Alternative provision was in place in case of intractable technical problems, but the assessments proved to be satisfactorily robust.

Data-analysis of responses showed that more than 90% of students who attempted the EMA had tried the PA first, but the way in which they used it varied. Some students spent just a few minutes on the PA prior to starting the EMA. Others made extensive use of it, working through each question many times. There was no obvious correlation between the amount of time spent on the PA and a student's success.

Analysis of responses to individual questions revealed situations where students were making use of the simple "Your answer is incorrect" feedback to correct their answer at the second try. An example is given of a particular precedence error which was seen in 50% of incorrect responses.

Survey responses included many complimentary comments e.g. "Perfect for distance learning"; "The ECA worked well, the immediate feedback was ace". Several students reported the EMA to have been "fun". Students reported using the PA both as a practice for the EMA and as a bank of additional questions, with most regarding the former purpose as more important.

The paper concludes that the outcomes of the introduction of interactive computer-based assessment into *Maths for science* were pleasing, but that the effort in getting the assignments up and running had been substantial. The time and expense could only be justified because of the large student numbers expected on the module and the predicted long-term savings in staff time.

5.1.2 Context and reception of Publication 1

This paper describes the beginning of my involvement with computer-marked assessment with feedback and it is included in this submission for that reason. I was module team chair and lead author of *Maths for science* (S151) and wanted to be able to give students meaningful feedback and to do so in timely fashion; S151 thus became the first OU module to use online assessment of this type. I was lead academic for the development, working closely with Phil Butcher, with a grant from the e-OU Strategy Group. We built on the work of others who had used similar

questions but delivered them on CD-ROM rather than online, but we developed the pedagogy further.

We chose to publish our early findings in the monthly online Maths CAA series, which was widely read by the pioneer developers and researchers at that time. The editor of the series, Professor Cliff Beevers of Heriot Watt University, accepted our paper for publication very quickly. The field was developing rapidly and we wanted to share our findings without the delays that would have been inevitable if we had waited for publication in a fully peer-reviewed journal.

My contribution to Publication 1, as agreed with my co-authors, was 70%.

Excluding self-citation, I am aware of eight publications that cite this paper. Most of these refer to Publication 1 for its description of the pedagogy underlying the development of the OpenMark system, recognising the ground-breaking nature of our work. Mills (2004, p. 34) comments:

> The impact of ICT on assessment is significant in a number of ways… Secondly, learning and assessment can be integrated in new ways. In a very interesting paper, Jordan, Butcher and Ross (2003) describe the development of a Web-based assessment system by which remote students take a credit-bearing test online at the end of a Maths for Science course. During the test, students receive immediate, targeted feedback on their answers and are awarded a mark which reflects the amount of help they have been given by the computer system in arriving at their answers. Clearly, as such systems are developed, the costs of assessment will be reduced and at the same time, students will get more feedback on their performance.

5.2 Publication 2

Jordan, S. E. & Swithenby, S. J. (2005). Online summative assessment with feedback as an aid to effective learning at a distance. In C. Rust (Ed.), *Improving student learning: Diversity and inclusivity. Proceedings of The 2004 12th Improving Student Learning Symposium* (pp. 480-485). Oxford: Oxford Centre for Staff & Learning Development.

5.2.1 Summary of Publication 2

This publication, of most relevance to Research Questions 1 and 2, discusses *Maths for science*'s online assessment in the context of Gibbs and Simpson's (2004-5) conditions for effective formative assessment, which were under development at the time of writing. It also considers the economic drivers for the introduction of summative computer-marked end-of-module assessment, enabling teacher marking to be eliminated, templates and question types to be re-used and, in principle, the random generation of questions with minimal effort.

Publication 2 compares the system's recognition of partial achievement, with a mark awarded that reflected the amount of help that had been given, with the "partial credit" described by Beevers, Wild, McGuire, Fiddes and Youngson (1999). It also aligns the decision to develop a practice assessment, in order to give students the opportunity to familiarise themselves with the technology and the question types, and thus to reduce their anxiety, with the work of Sly (1999).

Most significantly, Publication 2 describes an evaluation of student use and perception of the practice assessment (PA) and end-of-module assessment (EMA) by means of the Assessment Experience Questionnaire (Brown, Gibbs & Glover, 2003). This identified that student effort was not well distributed throughout the module, in contravention of two of Gibbs and Simpson's other conditions:

- Assessed tasks capture sufficient student time and effort;
- The tasks distribute student effort evenly across topics and weeks.

Measures had been introduced to improve distribution of effort, in particular greater flexibility in the order in which PA questions could be attempted, making the EMA available to students for longer and proactive contact with students, encouraging them to make greater use of both assignments.

Student views of the usefulness of feedback were superficially contradictory. 89% of the students agreed or strongly agreed with the statement "the interactive feedback associated with the ECA questions helped me to learn" and 79% of the students agreed or strongly agreed that "there

were occasions when the detailed feedback enabled me to amend my answer and so to provide the correct answer at the second or third attempt". Moreover, when asked "what for you, was the most positive aspect of the ECA?" many students commented on the instantaneous feedback. However scores for "timeliness of feedback" were low, and the students' comments included several items such as "As yet I have not had any feedback on this course". The paper suggests that this may have been due to a misalignment of student and lecturer understanding of the word "feedback". For students, "feedback" was associated with getting marks and what lecturers thought of as feedback was identified by students as "teaching". This finding was in line with that of Glover et al. (2005).

5.2.2 Context and reception of Publication 2

I had been asked if *Maths for science* and its assessment could be evaluated as part of the Formative Assessment in Science Teaching (FAST) Project, a three-year project funded by the Higher Education Funding Council under Phase 4 of the Fund for the Development of Teaching and Learning (FDTL4). The FAST project was led by The Open University and Sheffield Hallam University and ran from 2002-2005. One of the FAST Project's outputs was the much-cited Gibbs and Simpson (2004-5) review, and my two case studies are available from http://www.open.ac.uk/fast/.

Publication 2 was first presented at the Improving Student Learning Symposium in Birmingham in September 2004, as part of a symposium "Studying the impact of assessment on student learning in physical science and bioscience courses" with two other papers from the FAST Project, concerning the effectiveness of written feedback at the OU and student perceptions of the value of different modes of feedback at Sheffield Hallam University. The importance of timely feedback (in line with Gibbs & Simpson, 2004-5) was also identified in the other papers and the mismatch between staff and student perceptions of feedback was more widely reported.

According to their website (http://www.brookes.ac.uk/services/ocsld/isl/), "the major aim of the Improving Student Learning Symposia was to provide a forum which brings together those who

are primarily researchers into learning in higher education and those who are primarily practitioners concerned more pragmatically with improving their practice, but from whichever starting point, papers are only accepted if they take a sufficiently scholarly, research-based approach".

My contribution to Publication 2, as agreed with my co-author, was 60%.

Unfortunately, after publication in the symposium proceedings, Publication 2 only appears to have been cited by one publication other than those which include me as an author.

5.3 Publication 3

Ross, S. M., Jordan, S. E. & Butcher, P. G. (2006). Online instantaneous and targeted feedback for remote learners. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 123-131). London: Routledge.

5.3.1 Summary of Publication 3

This book chapter again considers the *Maths for science* assessment in the light of Gibbs and Simpson's (2004-5) conditions under which assessment supports student learning, in this case emphasising the importance of the conditions in a distance education context. Publication 3 also draws conclusions from 270 respondents to a postal questionnaire which focussed specifically on the *Maths for science* assessment and from further data analysis. It is thus of relevance to Research Questions 1, 2 and 3.

Timely and tailored feedback was considered particularly important as a strategy for confidence-building for the module's mature distance-learning students, many of whom were near the beginning of their Open University studies and had previous negative experiences of studying mathematics. The feedback aimed to simulate a "tutor at the student's elbow". Other factors that were considered important in developing the assessments were: interactivity; minimal use of multiple-choice formats; easy to use and robust software, with secure transmission to OU servers;

simple input, recognising that finding an easy way for students to input mathematical expressions is a widely-acknowledged problem (e.g. Beevers & Paterson, 2002).

It had been hoped that the practice assessment (PA) would help students to learn and to distribute their effort throughout the module (Gibbs & Simpson, Condition 2). However roughly three-quarters of students accessed the PA for the first time just before starting the end-of-module assessment (EMA), thus most of them were using the PA primarily as a "mock" EMA.

Questionnaires about *Maths for science*'s online assessment were sent to around 500 students on an early presentation, and 270 completed questionnaires were received. These revealed that students were generally happy with the format of the EMA and found the feedback useful. However a few had experienced technical difficulties and some felt they would have done better if they had received credit for working. Some students seemed to think that "feedback" related to their overall module score rather than to the teaching comments provided on each question.

5.3.2 Context and reception of Publication 3

I was approached by Graham Gibbs in 2004 with a view to writing a chapter for this book. The book was edited by Cordula Bryan and Karen Clegg, who describe our chapter thus:

> Shelagh Ross, Sally Jordan and Philip Butcher address the problem of providing rapid but detailed teaching feedback to large distance-education groups. Their case study researches online assessment of a "maths for science" module in which meaningful feedback was given to student answers on formative and summative assessment exercises. (Bryan & Clegg, 2006, p. 5)

Shelagh Ross kindly acted as lead author of Publication 3 because I was off work sick for much of 2004, but I was back at work in time to advise on final content. The work described was led by Phil Butcher and me. My contribution to Publication 3, as agreed with my co-authors, was 33%.

Excluding self-citation, I am aware of 13 publications that cite this chapter. As with Publication 1, most of these cite Publication 3 for its description of the pedagogy underlying the development of

the OpenMark system, noting the benefits of constructed-response question types and instantaneous feedback. For example, Kear (2010, p. 148) writes:

> Many students appreciate the challenge and immediate feedback that formative e-assessments can offer, and are therefore willing to take the time to complete them (Ross et al., 2006; Nicol & Milligan, 2006)

5.4 Publication 4

Jordan, S. E. & Mitchell, T. (2009). E-assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, *40*(2), 371-385.

5.4.1 Summary of Publication 4

This paper describes the introduction of short-answer free-text questions with tailored feedback at the Open University, using Intelligent Assessment Technologies (IAT)'s FreeText Author software linked to the OU's OpenMark software. It includes initial findings from a comparison of the accuracy of the computer's marking with that of six course tutors and a summary of observations of student volunteers attempting the questions in a usability laboratory. It reflects on the features of the questions that had been found to be successful and on the changes made in the light of early use. The research described in Publication 4 is thus of most relevance to Research Question 4.

The IAT software makes use of natural language processing (NLP) techniques of information extraction, performing a match between the student responses and a series of "model answers", each represented as a syntactic-semantic template. The templates are prepared offline, using an authoring tool designed to protect question authors from the complexity of NLP. The range of possible correct and incorrect answers to each short-answer question means that it is important to develop answer matching in the light of responses gathered from students at a similar level to those for whom the questions are intended, and at the OU responses were gathered from online developmental versions of the questions. The IAT software was used within OpenMark to enable

short-answer questions to be used alongside other question types and also to provide students with multiple tries at each question with the amount of feedback increasing at each try.

Between 92 and 246 student responses to each of seven free-text questions were marked independently by the computer system, by six course tutors and by the question author. Responses in which there was any divergence between markers (human or computer) were inspected in more detail to investigate the reasons for the disagreement. Chi-squared tests showed that, for four of the questions, the marking of all the markers was indistinguishable at the 1% level. For the other three questions, the markers were marking in a way that was significantly different, but in all cases the mean awarded by the computer was within the range awarded by the human markers. For six of the seven questions the computer system was in agreement with the question author for 94.7%-99.5% of the responses, whilst for the least well developed question there was 89.4% agreement.

Six student volunteers were observed in a usability laboratory as they attempted short-answer questions. Five of the six students entered answers as phrases rather than sentences whilst the sixth gave long and complex sentences. Use made of feedback was similarly variable.

Of the 78 questions originally authored, four were deemed unworkable and removed during developmental testing. In a further 13 cases, some changes were made to the question wording, sometimes to more tightly constrain the student responses. In order to be suitable for computer marking, it was recognised that questions should be easily answerable in a short sentence and that there should be a clear-cut distinction between correct and incorrect responses.

After the initial training phase, the author was able to write questions and appropriate answer matching in a time that varied between a few minutes and several hours, depending on the complexity of the question. Amending the question and the answer matching in the light of student responses was even more dependent on the complexity of a question, sometimes taking

more than a day. However, the accuracy of the answer matching was undoubtedly improved by its development in the light of real student answers.

The paper concludes that while acknowledging that computerised marking of free-text questions will never be perfect, the inconsistency of human markers should not be underestimated; computerised marking is inherently consistent. Furthermore, if course tutors can be relieved of the drudgery associated with marking relatively short and simple responses, time is freed for them to spend more productively.

5.4.2 Context and reception of Publication 4

I was asked to contribute a presentation on my investigation into the use and evaluation of short-answer free-text questions to an invited e-assessment symposium at the EARLI/Northumbria Assessment Conference in Potsdam in August 2008. After the conference, this paper, with co-author Tom Mitchell of Intelligent Assessment Technologies Ltd., was accepted for publication in a special issue of the *British Journal of Educational Technology* (BJET). BJET (2012 impact factor 1.313) is actually an international journal and it is one of the leading journals for research into educational technology. My contribution to Publication 4, as agreed with my co-author, was 50%.

Excluding self-citation, I am aware of 48 publications that cite this paper. These split into four roughly equal groups: (1) those that discuss the automatic marking of short-answer questions; (2) those that compare the marking accuracy with that of human markers; (3) those that discuss the benefits of constructed response questions over selected response; and (4) those that talk about the advantages of our rapid, detailed and graduated feedback. For example:

1. Handke and Schäfer (2012, p. 162 & p. 195), write (with translation from German following the original text):

> Es sollte jedoch nicht der Fehler gemacht werden, im E-Assessment immer nur “auf Nummer Sicher” zu gehen und ausschließlich die unproblematischen (aber auch restriktiven) Aufgabentypen einzusetzen. Ein Fortschritt wird nur stattfinden, wenn neue Typen kontinuierlich (weiter-) entwickelt und ausprobiert warden … Ein an der Open

University verfolgter Ansatz (Jordan, 2009; Jordan/Mitchell, 2009) ermöglicht frei formulierte textuelle Eingaben von ein bis zwei Sätzen Länge, die mit vordefinierten korrekten und inkorrekten Musterantworten abgeglichen warden, um Rückmeldungen des Systems zu generieren. Die Musterantworten sind als Templates definiert, die Schlüsselwörter (z.B. Nomen, Präpositionen, Verben etc.) und ihre Synonyme sowie die syntaktisch-semantischen Beziehungen zwischen ihnen beschreiben.

However it would be a mistake to always simply play it safe where machine-evaluation is concerned and to use only the unproblematic (but restrictive) task types. Progress will only take place if new types are continuously (further) developed and tested…The Open University approach (Jordan, 2009; Jordan/Mitchell, 2009) allows free-form text input of one or two sentences. These are checked against predefined correct and incorrect sample responses to generate system feedback. The sample responses are defined as templates that describe keywords (e.g, nouns, prepositions, verbs , etc.) and their synonyms as well as the syntactic and semantic relationships between them.

2. Gilbert, Whitelock and Gale (2011, p. 23) write:

Some academics query whether technology can be used to address higher learning skills effectively. This barrier to take up is challenged by authors working with more complex questions in a technology-enhanced environment. For example, Jordan and Mitchell (2009) and Butcher and Jordan (2010) discuss the merits of different tools for free-text technology-enhanced assessment. They conducted a study using a tool called FreeText Author (developed by Intelligent Assessment Technologies and the Open University) for short-answer free-text questions with tailored feedback. This study (evidence category 1a) found that the “answer matching has been demonstrated to be of similar or greater accuracy than specialist human markers”.

3. Bellotti, Kapralos, Lee, Moreno-Ger and Berta (2013, p. 6) comment:

> Short answers are responses to questions in the test takers' own words and therefore better reflect how well they understand the material since they have to provide their own response instead of choosing the most plausible of the alternatives, as with multiple choice questions [Jordan & Mitchell, 2009].

4. Redecker (2013, p. 19) states:

> Additionally, programs are being developed which not only author and reliably mark short-answer free-text assessment tasks, but also give tailored and relatively detailed feedback on incorrect and incomplete responses, inviting examinees to repeat the task immediately so as to learn from the feedback provided (Jordan & Mitchell, 2009).

5.5 Publication 5

Butcher, P. G. & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, *55*(2), 489-499.

5.5.1 Summary of Publication 5

This paper gives more detail about the comparison of the marking accuracy of IAT's FreeText Author software, as used at the Open University, with that of six course tutors. It also compares the marking accuracy of the IAT software with OpenMark's own pattern-matching software[7] and with Regular Expressions as implemented in Java, both of which are based on the manipulation of keywords. This research is of most relevance to Research Question 4.

To ensure that the human marking comparison did not assume that either the computer or human markers were correct, both the computer's and the tutors' marking of each response were compared against the median of all the tutors' marks for that response (used here as a way of indicating the majority view), the mark awarded by the IAT software and the mark awarded by question author, done "blind".

[7] Now called PMatch

For four of the questions, the marking of all the markers was indistinguishable at the 1% level, and where there was significant difference this appeared to be due to differences between the human markers. The kappa inter-rater statistic for agreement with the question author varied between 0.35 to 1 for individual human markers (question means of 0.67 to 0.99) whilst for the IAT system it was between 0.79 and 0.98. Markers were rated for each question by inter-rater statistic, and the mean ranks for each marker were compared. Analysis of variance showed that, at the 95% confidence level, only one marker marked consistently as well as the computer.

Although the extent of errors in human marking was alarming, it was not surprising to find that the computer's marking was more consistent than that of the human markers. Even in purely formative use, accuracy of marking matters because of the importance of giving correct feedback to students and of retaining students' confidence. Students appear to have more confidence in human marking than in computer marking, even if this is not justified.

For the comparison of the marking accuracy of the three computer systems, an undergraduate student (not of computer science) was appointed for several weeks and given the task of obtaining adequate answer matching in OpenMark and Regular Expressions. He was provided with a set of training responses, obtained from students on the module.

After the first development of answer matching, kappa inter-rater agreement with the question author varied between 0.76 and 0.98 for OpenMark and between 0.77 and 0.98 for Regular Expressions (with IAT between 0.79 and 0.98). Each system showed the best marking accuracy for at least one of the questions. The answer matching was improved for all three systems in the light of further student responses. This resulted in an improved marking accuracy for each system, with OpenMark, the simplest system, performing at least as well as the others. It is noted that OpenMark's response-matching algorithm, although intuitive to use, is not a simple "bag of words" system; it can cope with inaccuracies in spelling, with word order and negation.

The paper emphasises the fact that responses from real students were used in the development of answer matching for all three systems. A paradox is identified in that student responses to summative questions are likely to be the most useful in developing answer matching, but yet questions need to be well developed before being used summatively.

For six of the seven questions, between 100 and 250 responses were adequate for the purpose of developing the answer matching, but for the seventh question a greater number of responses was required. The complexity of the answer matching was variable (4–15 lines of OpenMark code for the questions under consideration). The final levels of marking accuracy were considered acceptable.

5.5.2 Context and reception of Publication 5

Following the surprising finding that OpenMark's own pattern-matching software provided answer matching as accurate as that provided by IAT's FreeTextAuthor, we decided to publish the detail of the human-computer marking comparison (introduced in Publication 4) alongside the detail of the computer-computer marking comparison. We selected the international journal *Computers & Education*, whose impact in the field of educational technology has been growing steadily, with a 2012 impact factor of 2.775.

My contribution to Publication 5, as agreed with my co-author, was 50%.

Excluding self-citation, I am aware of 19 publications that cite this paper. Most of these comment on the human-marking comparison with, for example, Redecker and Johannessen (2013, p.84) stating "For short-answer free-text responses of around one sentence, automatic scoring has also been shown to be at least as good as human markers (Butcher & Jordan, 2010)".

I consider the computer-computer marking comparison to have implications that are as important as those of the human-computer marking comparison; this point was not lost on one of the paper's (unnamed) reviewers, who said:

> I think this is a fine paper and offer my congratulations on this excellent work, which ranks in my view with the best sources you have cited…You have also raised some interesting questions that as far as I know have not surfaced before in the literature about free-text scoring...Your comment "Knowing that computational linguistics is behind the IAT response matching algorithms has provided an element of respectability to the marking process" is spot on: the debate is about gaining acceptance for these techniques. The late Prof Roger Needham once remarked (c 1998, in a seminar I arranged in Cambridge on these issues) that "counting counts, but syntax sucks", but he did not supply any published references (or none that I noted at the time). So it was known even then that techniques like those used by OpenMark are "unreasonably effective", and that attempts to employ artificial intelligence or even just syntactical analysis do not produce much, if any, improvement… Your results are a major step forward in justifying the use of such techniques.

It is also most pleasing that Publication 5, along with Publication 4, was highly ranked in a project, commissioned by the Higher Education Academy and described in most detail in Gilbert et al. (2011), which aimed to synthesise the main points from 124 references which had been recommended as evidence-based papers on assessment and feedback with technology enhancement. Of the 124 papers, only 15 (12.1%) were classified as category 1a (“Peer reviewed generalizable study providing effect size estimates and which includes (i) some form of control group or treatment and/or (ii) blind or preferably double-blind protocol”) and publications 4 and 5 were both amongst this 15.

5.6 Publication 6

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

5.6.1 Summary of Publication 6

This paper describes the use and evaluation of interactive computer-marked assignments (iCMAs) on a range of Open University science modules and for summative, formative and diagnostic purposes. The paper concludes with two case studies, illustrating an iterative process of assessment design, evaluation and improvement. The research is of relevance to Research Questions 1, 2 and 3.

A number of Open University science modules and their assessment strategies are described, generally with iCMAs being used alongside tutor-marked assignments (TMAs). In addition, a diagnostic quiz *Are you ready for level 1 science?*, intended to direct students to the most appropriate starting module, had been produced as three interlinked OpenMark quizzes.

A range of evaluation methodologies was employed including questionnaires and semi-structured interviews, the observation of students completing assignments in a usability laboratory, and an extensive analysis of data collected as students engaged with the iCMAs.

iCMA questions were generally very well received by students, with a substantial number (although a minority) reporting that they learned more from doing iCMA questions than from doing TMA questions, and a substantial majority feeling that their performance in iCMA questions reflected their ability and that answering questions was “fun” (see Publication 6 Table 2). The instantaneous receipt of feedback was the most commonly identified useful feature of iCMAs and even when individual students were unhappy with specific questions or aspects of their use, they were happy with the use of iCMAs in general. Most students felt that their iCMA scores should count towards their overall grade. Some said that they would engage with formative and summative iCMAs in exactly the same way whilst others said that if the iCMAs were purely formative they would not bother with them at all.

Moving beyond student opinion, overall activity on summative iCMAs was seen to increase as the cut-off date approached. Unsurprisingly, most students attempted all the questions on summative iCMAs, but in formative-only use, usage dropped off as the iCMA progressed, or if the

questions were presented in several smaller iCMAs, usage dropped off both during and between the iCMAs. Furthermore, there were always some students who accessed the iCMAs but did not complete any questions.

When students were asked whether they found the feedback useful, the number who replied in the affirmative was consistently around 90%. Observations in the usability laboratory presented a rather different picture. Whilst some students clearly made good use of the feedback provided, there were several instances when students did not pay sufficient attention to the feedback provided even though they appeared to read it. Being told that their answer was right appeared to distract students from the detailed comments provided.

The paper ends with two case studies. In the first, a short-answer free-text question was reworded several times in order to encourage students to work out the answer rather than searching for it on the internet, and to discourage over-long answers. Targeted feedback was added for a specific incomplete first-try response to make it clear to students that this response had been "understood". The second case study describes the simplification of the structure of the diagnostic quiz, following the evaluation finding that users were "getting lost" in the previous structure, and the addition of quiz level feedback, as requested by users.

The paper concludes that student perceptions of e-assessment and rigorous quantitative analysis of actual engagement are both important. As Kibble (2007) found, a light summative weighting can increase student engagement with e-assessment, but the summative function may overwhelm the formative. A very small change in wording can lead to a considerable improvement in the performance of a question. Interactive computer-marked assessment with feedback has huge potential to help distance learners.

5.6.2 Context and reception of Publication 6

By 2008, iCMAs were in use in a wide range of Open University Science Faculty Modules and I started a project, jointly funded by COLMSCT and piCETL, which used a range of qualitative and

quantitative methodologies to investigate student engagement with these iCMAs. Publication 6 was the result of an invitation to submit a paper to a special issue of *Open Learning*, focusing on distance and e-learning in science and related subjects. *Open Learning* describes itself as "a leading international journal in the field of open, flexible and distance learning" with articles and case studies "peer reviewed by an international panel of experts in the field". Following publication of the paper, I was invited to give a presentation about the work to an international audience at the University of London's Research in Distance Learning Conference.

Excluding self-citation, I am aware of 9 publications that cite this publication, mostly for its description of the pedagogy underlying OpenMark and the way in which iCMAs had been incorporated into Science Faculty modules, sometimes (e.g. Sancho-Vinuesa, Escudero-Viladoms & Masià, 2013) comparing our strategies with those in use at other universities.

More significantly, Sangwin (2013) and Sorensen (2013) refer to some of Publication 6's conclusions. Sangwin (2013, p. 140) writes:

> Students' attitudes to iCMAs are examined in more detail in Jordan (2011). She concludes "Interactive computer-marked assessment has huge potential to help distance learners to find appropriate starting points, to provide them with timely feedback and to help them pace their study…The use of e-assessment does not imply a tutor-less future, rather one in which tutors are freed from the drudgery of marking simple items to give increased support for their students, with information about student performance on e-assessment tasks used to encourage dialogue." (Jordan, 2011)

Sorensen (2013, pp. 173-174) writes:

> Jordan (2011) reported on a survey of distance learners and found that students engaged more with the online questions when they carried some weighting and most students felt that their marks should count towards their overall course score. Jordan also argued that student perception is very important and that it is important to monitor the use of e-assessment and to make changes (to individual questions, the structure of an assignment

or to the underlying e-assessment system) as and when appropriate as, frequently, a very minor change in wording can lead to a considerable improvement in the performance of a question.

5.7 Publication 7

Jordan, S. E., Jordan, H. E., & Jordan, R. S. (2011). Same but different, but is it fair? An analysis of the use of variants of interactive computer-marked questions. In *Proceedings of the 14th International Computer Assisted Assessment (CAA) Conference, Southampton, 5th-6th July 2011.* Retrieved 1st May 2014 from http://caaconference.co.uk/pastConferences/2011/

5.7.1 Summary of Publication 7

This paper builds on the work of Dermo (2010) in developing and trialling tools and techniques to verify that students are not disadvantaged because of the particular variants of computer-marked questions that they have received. It is of most relevance to Research Question 3.

Usual practice at the Open University is to produce at least five, hopefully equivalent, variants of each question. Different students therefore receive different versions of each summative interactive computer-marked assignment (iCMA) as a whole, reducing opportunities for plagiarism. In formative use, the different variants of the questions provide extra opportunities for practice.

A selection of tools was produced to check the equivalence of the different variants of each question. In particular, for each question, plots were produced showing the proportion of students with each available score for each variant, as illustrated in Figure 7 (Section 3.3.1), along with a single number (a probability *p*) that indicated the likelihood that any observed differences between variants had arisen by chance. If *p* was less than some set value, conventionally 0.05[8], then there was a reasonable certainty that the different variants of the question were not of equivalent difficulty. The different tools were used in conjunction with each other. In particular, if

[8] The repeated use made of the statistical tools meant that the Bonferroni correction was applied, requiring $p < 0.05/n$, where n was the number of questions in the iCMA, typically 10.

the value of *p* indicated that it was likely that the variants were not of equivalent difficulty, then it was helpful to inspect first the plots and then the actual student responses in order to determine the reason for the difference.

Factors identified as leading to variants with non-equivalent difficulty include the fact that it is easier to round a number down than up and that spelling words like "sulphur" is likely to cause more difficulty than other standard forms of chemical names. In addition, some variants of some questions were found to be testing additional skills over the other variants of the same questions, for example to calculate the value of $1 + 2 \times (3 + 4)^2$ requires understanding of the rules of precedence, whilst to calculate the value of $(2 + 3)^2 \times 4 + 1$, students simply need to work from left to right. Where variants of multiple-choice, multiple-response or true/false questions had been produced from longer lists of correct and incorrect options, variants were observed to be of different difficulty if some of the options were more clearly right or wrong.

The project also considered the likely impact on overall score of the variants that students received for all questions. For one module considered (Module W[9], with 1508 students on the presentation under consideration), the calculated variation in overall score was less than $\pm$0.5%. By contrast, for Module X (362 students) the calculated variation in overall score was about $\pm$5%. This was attributed to Module X having a larger proportion of individual questions whose variants were of significantly different difficulty (including a high proportion of multiple-response and true/false questions), fewer questions in each iCMA, and a higher weighting of the iCMA component.

All students are not equally disadvantaged by "difficult" questions or equally advantaged by "easy" questions, so it is not appropriate to make automatic adjustment to scores. The paper concludes that the most appropriate action when discrepancies are detected is to inform examination boards of particular students who may have been disadvantaged and to amend

[9] Two modules were identified as A and B in Publication 7 and a different two modules were identified as A and B in Publication 10. To avoid confusion, all have been renamed. Modules A and B in Publication 7 are here referred to as W and X.

individual variants of affected questions for the future, in the process learning more about the design of effective and equitable questions and assignments.

5.7.2 Context and reception of Publication 7

Publication 7 is the first of three publications in this submission that were accepted, following peer review, for presentation at the International Computer Assisted Assessment (CAA) Conference (http://caaconference.co.uk/). I have always struggled to know whether it is best to present my work at this specialist conference or whether to aim for a wider audience. From 2010 to 2013, I decided that it was important to align my work with other developments in e-assessment, so I attended the CAA Conference in Southampton each July. In 2010, my poster "Student engagement with e-assessment", reporting on early quantitative evaluation of student use (described in more detail in publications 6, 8 and 10), had won the prize for the best poster at the conference.

Publication 7 was inspired by a paper which John Dermo had given at the 2010 Conference (Dermo, 2010), and built on work that had been done by Helen Jordan (2009) in developing statistical tools for iCMA analysis, to investigate more thoroughly the fairness of Open University practice in offering students different variants of iCMA questions. My contribution to Publication 7, as agreed with my co-authors, was 50%.

Publication 7 was well received at the conference and was one of nine papers from the conference that were selected for publication in a special issue of the *International Journal of e-Assessment* (IJEA) in 2012. However, from the point of view of increasing citations to my papers, my decision to submit full papers to the CAA Conference rather than peer-reviewed journals may have been the wrong one! Excluding self-citation, I am aware of just 2 publications that cite Publication 7. More happily, one of the citations picks up on our reasons for using multiple variants of questions whilst the other identifies one of issues that this raises: Sangwin (2013, p. 139) comments "several variants are usually created for each question, carefully designed to be of similar difficulty; see Jordan et al. (2011). This is both an anti-plagiarism device in summative use

and for extra practice in formative use", whilst Voelkel (2013, p. 3), who cites the IJEA version, writes "indeed, Jordan, Jordan, and Jordan (2012) found that different variants of computer-marked questions can behave differently and that it is necessary to monitor performance of supposedly equivalent questions".

5.8 Publication 8

Jordan, S. E. (2012). Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions. *Computers & Education, 58*(2), 818-834.

5.8.1 Summary of Publication 8

This paper describes research that addressed the following questions:

- How do students actually engage with short-answer free-text questions?
- How do students use the feedback provided?
- The influence of factors such as question wording and whether the assessment is summative or purely formative.

The methodology included direct observation of six student volunteers in a usability laboratory (a more detailed report of work mentioned in publications 4 and 6) and a large scale analysis of thousands of student responses to seven short-answer questions and the changes made to these responses after feedback. The work was thus of most relevance to Research Questions 1, 2 and 3. However, in that the focus of the research was on engagement with an innovative question type, it was also of some relevance to Research Question 4.

In the usability laboratory, most students gave their answers as phrases rather than sentences but one composed his answers carefully and gave them in the form of sentences. Another student commented that, since his responses were being marked by a computer, he could take a more relaxed attitude to grammar and punctuation.

There were several instances where the students were seen to make use of the brief feedback after a first incorrect response to correct their answer. However, in other instances this feedback

confused students, in one case leading a student to modify what had been a correct but incomplete response and to give a completely incorrect response at the second try. Some students clearly read and acted on the more detailed feedback after an incorrect response at their second try, following references to the module materials and correcting their answer. Other students missed the "clues" that were given. Use of the full "model answer" given at the end of each question also varied from student to student, with the most consistent student reading the answer carefully when his answer had been marked as incorrect but not bothering when his answer had been marked as correct. Interesting insight came about as a result of the fact that the computerised answer matching was not well developed at the time of the observation: After being told that a correct (though misspelt) response was incorrect, a student was annoyed with the computer; he read the final answer and believed his answer to be the same as this (he did not notice the spelling mistake). By contrast, after being told that an incorrect response was correct, two students appeared to read the final answer, which contradicted the one they had given, but they did not notice the contradiction.

The analysis of responses showed that response length varied with question, permitted word-length, detail of question wording and mode of use (e.g. summative, formative, diagnostic), with the question being asked being the most significant factor. The importance of detailed question wording was illustrated when a word-limit of 20 words was observed to lead to an *increase* in the average word-length of responses, apparently because some students interpreted the guidance "Answers of more than 20 words will not be accepted" as meaning that their answer should be close to 20 words in length.

The types of response (blank i.e. no response, a single word, a phrase, note form, a sentence or a paragraph) given were found to vary with question, detailed wording, allowed word length and mode of use, with the question being asked again being the most significant factor. Some questions seemed to lend themselves more to answers of a particular form (e.g. as a single sentence) than was the case for other questions.

The way in which responses were altered in response to feedback was inspected, with second and third try responses being classified as unchanged, rephrased, [some text] deleted, [some text] added, changed or new (Publication 8 Table 8). There was considerable evidence of students altering their responses in a way that was in line with their understanding or lack of understanding of the feedback provided, e.g. adding to their response when they had been told that their previous answer was incomplete, rephrasing their answer when they believed (sometimes incorrectly) that the computer had not "understood" their previous response and making no change when there was insufficient feedback or they did not understand it.

5.8.2 Context and reception of Publication 8

Publication 8 is the second paper in this submission that I chose to submit to the journal *Computers & Education* (2012 impact factor 2.775). I see this paper as being very much at the "heart" of my thesis, because it links my work on student engagement with my work on short-answer free-text questions.

Excluding self-citation, I am aware of 5 publications that cite this publication. Abio and Barandela (2012, p. 3) pick up on my discussion of the different behaviour observed when a student has been told that their answer was correct or incorrect (with translation from Spanish following the original text):

> Eso corrobora lo que Jordan (2012) encontró en observaciones de usabilidad realizadas con seis estudiantes que recibían feedback automático inmediato en la tarea de responder preguntas cortas, pues en el caso en que la respuesta proporcionada es correcta existe la tendencia a no atender a la explicación posterior, pero de forma contraria, la atención es mayor para el feedback en el caso en que la respuesta proporcionada por el sistema fuese negativa cuando el estudiante creía que su respuesta estaba correcta.
>
> That corroborates what Jordan (2012) found in usability observations made with six students receiving immediate automatic feedback when answering short-answer

questions: when the answer given is correct there is a tendency not to pay attention to the explanation/feedback provided. Conversely, students pay more attention to the feedback when the system reports that their answer is incorrect but they believed it to be correct.

5.9 Publication 9

Jordan, S. E. (2012). Short-answer e-assessment questions: Five years on. In *Proceedings of the 2012 International Computer Assisted Assessment (CAA) Conference, Southampton, 10th-11th July 2012.* Retrieved 1st May 2014 from http://caaconference.co.uk/proceedings/

5.9.1 Summary of Publication 9

This paper updates previous work (Publication 4 and Publication 5), explores some of the misconceptions surrounding short-answer free-text questions and concludes with a discussion of the reasons why questions of this sort are not more commonly used. It is of most relevance to Research Question 4.

At the time of writing, 24 short-answer questions were in use in two Open University modules, with around 5000 individual student users per year. OpenMark PMatch was used for all the answer matching, with between 2 and 23 PMatch "rules" per question used for marking, plus up to 6 rules with the sole purpose of generating feedback. Two complementary methods were used for dealing with incorrect spelling: within words of greater than three letters, single incorrect, transposed, missing or extra letters could be allowed (if this was the question author's wish); in addition, a spell checker informed students when the word they had used was not recognised by the spell checker's dictionary (to which scientific words had been added when necessary) and offered suggestions for correct spelling. Responses were restricted to no more than 20 words.

An update of the earlier analysis of marking accuracy, for 3 questions included in the original analysis (Publication 5) and a further 8 questions in regular use, showed that PMatch marking accuracy was good for all 11 questions, with agreement with a human "expert" marker varying

between 97.9% and 99.3% and kappa inter-rater statistic between 0.84 and 0.98. “Borderline” responses caused the most difficulty for human and computer markers alike. Some responses that had not been predicted by the question author and so had previously been inaccurately marked (e.g. answers to a question about a child on a slide that instead talked about a child on a swing) were now accurately marked, following earlier analysis of actual student responses and subsequent alterations to the PMatch marking. Additional targeted feedback had also been provided.

In discussing the reasons why short-answer free-text questions are not more widely used, it was acknowledged that some academics will find it difficult to write sufficiently rigorous answer matching, especially given a reluctance to move beyond multiple-choice questions (Hunt, 2012). The perception that short-answer questions are more difficult for students was demonstrated to be unjustified for the questions in use in the Open University Science Faculty, and some of these questions assessed more than recall. Student confidence in the marking software was found to be important, with accurate targeted feedback helping them to appreciate that certain responses really were incorrect, rather than inaccurately marked by the computer. In providing sufficiently accurate answer matching, the system’s ability to deal with negation and word order were recognised as important features.

The real barriers to wider take up were identified as the time taken to develop the answer-matching rules iteratively and the need for a relatively large number of marked responses, where the actual time taken and number of responses required had been found to vary quite considerably from question to question. The developmental questions had each taken hours to write but longer to refine, and had reached a “good enough” state on the basis of several hundred responses per question. In principle, answer matching can be written for any question that has distinct correct and incorrect answers, but when there are many ways of expressing a correct (or incorrect) response, developing the answer-matching rules becomes tedious.

Machine learning offers the potential for removing the drudgery from the development of answer matching. For modules with small student numbers, collecting sufficient responses for answer

development represents a more serious problem, though collecting responses from other tasks e.g. examinations may provide a way forward. The paper concludes that, provided large numbers of responses are available, "following the data" (Halevy, Norvig & Pereira, 2009, p. 12) can provide robust and effective matching for short-answer free-text questions, using simple algorithmically-based answer matching.

5.9.2 Context and reception of Publication 9

Publication 9 is the peer-reviewed paper that accompanied my presentation at the 2012 International Computer Assisted Assessment (CAA) Conference (http://caaconference.co.uk/). Following the computer-computer marking comparison reported in Publication 5, we were now using OpenMark PMatch software for marking 24 short-answer free-text questions in use in OU Science Faculty modules, and a further analysis of the computer-marking accuracy showed that they were all behaving well. I chose to present this paper at the CAA conference because I wanted to encourage wider dialogue around the reasons for the disappointing take-up of short-answer free-text question, and to suggest some potential solutions.

Excluding self-citation, I am only aware of one publication that cites this paper. Hunt (2012) writes:

> The increasing support for assessment for learning [by Moodle] can also be seen as an attempt to bring Moodle's CMA tools more in line with its social constructionist pedagogy (Dougiamas et al 2012b)…While attempting an assessment, the student must actively engage with the subject matter. If the system allows students to answer in their own words (Jordan 2012), rather than just picking one of multiple choices, then they are constructing something, albeit for the computer, rather than another person, to see.

5.10 Publication 10

Jordan, S. E. (2013). Using e-assessment to learn about learning. In *Proceedings of the 2013 International Computer Assisted Assessment (CAA) Conference, Southampton, 9th-10th July 2013*. Retrieved 1st May 2014 from http://caaconference.co.uk/proceedings/

5.10.1 Summary of Publication 10

This paper highlights, by means of illustrative examples, some of the lessons that can be learnt by educators from the analysis of student responses to computer-marked assignments. Examples are given of detailed analysis of responses to particular questions, of assignment-level analysis and of correlations within and between assignments. Different patterns of engagement are illustrated in the context of two specific modules with different student populations. The research described has some relevance for Research Questions 1 and 2, but the focus on analytics means that it is most closely linked with Research Question 3.

Analysis of responses to individual questions can provide information about common and sometimes unexpected student errors, with an example given of a change made to teaching material following the discovery of a student misunderstanding in an interactive computer-marked end-of-module assessment.

Information about the extent to which responses are repeated or left blank can be associated with:

- lack of seriousness of engagement;
- lack of understanding of what the question wants, or of the meaning of the feedback;
- questions that are time consuming to complete, perhaps with multiple boxes for completion, or which require students to access a course component such as a video.

Analysis of the responses to all the questions in a purely-formative assignment showed the usage to drop off as the assignment progressed and, although most questions were seen to have between one and two times as many complete “usages” as they had distinct “users”, this effect was as a result of a small number of students repeating questions many times, whilst most

students attempted them only once. Cut-off dates and examinations were effective in encouraging activity on interactive computer-marked assignments (iCMAs), with Module Z[10] (with many students who appeared well motivated - see explanation in Section 5.10.2) also showing a peak in activity on the day the iCMA opened.

There was a slight negative correlation between the active time spent on an iCMA and score in the same assignment. There was also a negative correlation between submission date and score with, for Module Y in particular, students who completed the iCMA early doing much better than those who did not.

Some students were seen to use a purely formative practice assignment in preparation for a similar summative assignment, and students who had used the practice assignment did significantly better in the summative assignment. There was also a positive correlation between the number of questions answered in the practice assignment and score in the summative assignment.

For very similar assignments used by two different student populations, contrasting patterns of use were observed. Students on Module Y were less likely to alter their responses after receiving feedback, and were more likely to submit the iCMA just before the due date than were students on Module Z. Students on Module Y were also considerably less likely to attempt the practice assignment, a factor associated with a lower score on the corresponding summative assignment, as discussed above. All of these factors point towards good student engagement on Module Z, but to many students on Module Y who were lacking in the time and motivation to engage fully, and whose success was compromised as a result.

5.10.2 Comparison of the student populations of Module Y and Module Z

Table 1 gives information, not published in Publication 10, in support of the assertion (Publication 10, p. 2) that students who started Module Z in October 2012 tended to be older and to have

[10] Two modules were identified as A and B in Publication 7 and a different two modules were identified as A and B in Publication 10. To avoid confusion, all have been renamed. Modules A and B in Publication 10 are here referred to as Y and Z.

higher previous educational qualifications than those who started Module Y at the same time. The students on Module Y were also more likely to be new to the Open University and to be studying at a study intensity of 90 or 120 credits per year. Some Open University students are able to cope with 120 credits per year (equivalent to full-time study) but for those who are attempting to study alongside employment and/or caring responsibilities, a study intensity of more than 60 credits per year is a cause of concern, and efforts are made (including telephone calls and diagnostic quizzes, as in Publication 6, p. 152) to discourage students from over-commitment. Nevertheless, at least 54% of the students who commenced Module Y in October 2012 were both new to the Open University and studying this 30-credit module alongside at least one other, most commonly a 60-credit module.

Table 1. A comparison of the student populations of Module Y and Module Z at the start of each module in October 2012.

	Module Y	Module Z
Percentage of students with lower than standard conventional university entry qualifications (< A-level).	33%	22%
Percentage of students completely new to Open University study.	68%	7%
Age ≤ 21 years	19%	6%
Age 22-29 years	45%	21%
Age 30-39 years	20%	30%
Age 40-49 years	9%	25%
Age 50-59 years	5%	12%
Age ≥ 60 years	2%	6%

Student behaviour on the Module Y assignments provided an early warning of later study difficulties on the module. Data not included in Publication 10 indicates a 35 percentage point difference in completion rate between the students who were new to the Open University and studying at a high study intensity and those who were not new and had previously studied a relevant precursor module for Module Y. Thankfully, for the presentation of Module Y that started in October 2013, a considerably smaller percentage of the students were new to the University, and more were appropriately prepared.

5.10.3 Context and reception of Publication 10

Publication 10 is the peer-reviewed paper that accompanied my presentation at the 2013 International Computer Assisted Assessment (CAA) Conference (http://caaconference.co.uk/), which had as its theme the various "assessment analytics" that are possible using the data that systems collect following student use of computer-marked tasks.

I am not aware of any publications that cite Publication 10, but following further peer review, an expanded version of the paper was selected for publication in a special issue of the *International Journal of e-Assessment (IJEA)*, which was published in July 2014, just days before I submitted this covering paper.

5.11 Publication 11

Jordan, S. E. (2013). E-assessment: Past, present and future. *New Directions*, *9*(1), 87-106.

5.11.1 Summary of Publication 11

This review of e-assessment takes a broad definition, including any use of a computer in assessment, whilst focusing on computer-marked assessment. The review considers the scope of e-assessment so it is of most relevance to Research Question 4, though the use of assessment data within learning analytics is also discussed, which is firmly within the scope of Research Question 3.

Drivers for e-assessment include increased variety of assessed tasks and the provision of instantaneous feedback, as well as increased objectivity and resource saving. From the early use of multiple-choice questions and machine-readable forms, computer-marked assessment has developed to encompass sophisticated online systems.

The pros and cons of selected-response and constructed-response question types are discussed, with perhaps the most damning indictments of selected-response questions being those that query their authenticity. Some of the disadvantages of selected-response question types can be alleviated by techniques such as confidence-based or certainty-based marking (Gardner-Medwin, 2006). The use of electronic response systems ("clickers") in classrooms can be effective, especially when coupled with peer discussion (e.g. Mazur, 1991), conceptualised in terms of the principles of good feedback practice by Nicol (2007). Student authoring of questions e.g. using Peerwise (Denny et al., 2008b) can also encourage dialogue around learning.

More sophisticated computer-marked assessment systems have enabled mathematical questions to be broken down into steps , in particular in the CALM (Computer-Aided Learning of Mathematics) family of systems (Ashton et al., 2006a) and have provided targeted and increasing feedback, for example in OpenMark and Moodle (Publications 1 and 6; Butcher, 2008). Systems that use computer algebra, such as STACK (Sangwin, 2013) and those that provide answer matching for short-answer questions (e.g. Publications 4, 5 and 9) and essays (e.g. Attali & Burstein, 2006) are discussed.

In addition to considering question types, it is necessary to consider the way in which questions are combined. Computer-adaptive tests use a student's response to previous questions to alter the subsequent form of the test.

More generally, e-assessment includes the use of technology-enabled peer-assessment, audio assessment and assessed e-portfolios, blogs, wikis and forums.

The paper's conclusions and predictions for the future are:

1. That a beneficial side effect of massive open online courses (MOOCs) is that they are forcing the assessment community to consider appropriate methodologies for assessing huge numbers of informal learners. However it is important not to let standards slip in the rush to deliver assessment at scale, speed and at low cost.
2. That "assessment analytics" (Ellis, 2013) provide a means of finding out more about the misunderstandings of individual students and cohorts of students. Learning analytics provide data from student interactions in an online environment, but assessment is sometimes excluded. Redecker et al. (2012) emphasise the importance of assessing students on their actual interactions rather than adding assessment as a separate event.
3. That the boundaries between teaching, assessment and learning are becoming blurred.
4. That computers should be used to their full capacity in assessment when that is appropriate, but sometimes human marking is more appropriate. The well-placed use of computers can relieve human markers of some of the drudgery of marking and free up time for them to assess what they and only they can assess with authenticity.

5.11.2 Context and reception of Publication 11

Professor Derek Raine, the editor of the Higher Education Academy's *New Directions* journal, invited me to submit a review of the e-assessment literature. *New Directions* describes itself as a "showcase to disseminate innovation and research in the teaching of physical sciences in higher education" and I was pleased to accede to Derek's request because of the opportunity it provided to review a literature that I already knew quite well and to share the result with the physical science higher education community. Publication 11 is the result.

The review was published in September 2013, so it is slightly too soon to expect citations, though I am aware of one. However, it has been extremely well received: By October 2013 it was the "most read" *New Directions* paper, a position which it has held ever since and in November 2013, I received the following email from Vivien Ward, Head of Journals at the HEA:

> Your article has been topping the rankings of articles accessed on our journals platform – that's the cumulative score from when we launched last April. You're way out in the lead, well done!

I have given permission for a link to the review to be added to the Moodle Research Library and, following approval from HEA Journals, translation into Chinese is underway by the Open University of China, with publication due later in 2014.

5.12 Publication 12

Jordan, S. E. (2014). Adult science learners' mathematical mistakes: An analysis of student responses to computer-marked questions. *European Journal of Science and Mathematics Education*, 2(2), 63-87.

5.12.1 Summary of Publication 12

The research described in Publication 12 is of most relevance to Research Question 3. Inspection of thousands of responses to computer-marked assessment questions has brought insight into mathematical errors made by adult students on a *Maths for science* module. The students are from all over the world and with widely varying ages and mathematical backgrounds. This means that blame for weak mathematics cannot be attributed to a particular educational system, whilst there is every incentive to find out as much as possible about where the misunderstandings lie.

For each question in each assignment, two analyses were conducted:

- An analysis of the number of students who got the question right at the first try, second try, third try, or not at all. This was used to give a quick indication of question difficulty and the effectiveness of the feedback provided;
- An inspection of the actual correct and incorrect responses given. This was used to investigate in more detail the errors that students made.

Most of the questions analysed were in summative use and required students to construct their own response. Both of these things increased confidence in the reliability of the findings, as did the fact that similar errors were seen in different variants and in different questions.

Questions on logarithms, graphs and gradient, differentiation and standard deviation were poorly answered. However, the most persistent errors, seen in questions designed to assess a range of different skills, were in rounding numerical values to an appropriate number of decimal places or significant figures and in working out the units of an answer. Other errors included incorrect precedence in calculations, giving symbols in an incorrect case, and problems with reciprocation, adding fractions, and conversions between units of area and volume.

Many of the errors were similar to those reported by others, both at university and school level, and it appears that there are some very basic stumbling blocks that affect a wide range of students. Students appeared to make fewer mistakes in rearranging equations than they did in simplifying them, a finding in line with others' observations of students over-simplifying and cancelling inappropriately (e.g. Bradis, Minkovskii & Kharcheva, 1963; Schechter, 2009). Errors seen in *Maths for science* questions on graphs and differentiation demonstrated several of the difficulties reported by Kerslake (1981): "Graphical prototypes" (e.g. assuming that all graphs go through the origin); "slope-height confusion"; and problems with the equation of a straight line.

It is possible to attempt to classify the errors observed into separate categories, for example, careless mistakes, a lack of understanding of a method taught in the module materials (or a misremembered method from years ago) and deeper conceptual misunderstandings. However there is considerable blurring of the boundaries between the categories and there are many examples of questions where it is difficult to be certain whether the common errors observed were a result of conceptual misconceptions, faulty memory or careless errors.

The most common errors observed in *Maths for science* questions were frequently neither related to the question's primary purpose nor the errors that the question author had expected in that

question. Errors in rearranging equations were seen in trigonometry questions, faulty addition of fractions was seen in a probability question, precedence errors were seen in a statistics question etc. It is also the case that errors were often at a lower level i.e. with "easier" mathematics than expected, for example, not only did students have difficulty expressing a number to an appropriate number of significant figures, they also had difficulty in simply stating the number of decimal places given.

It is suggested that increased use of discussion with students might bring further insight into the reasons for errors as well as increasing student understanding directly.

5.12.2 Context and reception of Publication 12

I had spoken regularly on the lessons that can be learnt from the analysis of student responses to iCMA questions (e.g. Publication 10), but I had not published specific findings relating to students' mathematical errors since 2007 (Jordan, 2007). Responses to all *Maths for science* (S151) questions were analysed in 2010. I wanted to reach the appropriate science and mathematics education community, and I also hoped for rapid publication, so I chose to submit this paper to the newly founded *European Journal of Science and Mathematics Education*, on the basis of the reputation of several members of the Editorial Advisory Board and of the following statements from the journal's website (http://www.scimath.net/)

> The Journal aims to stimulate discussions on contemporary topics in science and mathematics education and to foster the application of the results in primary, secondary, and higher education. Research papers are welcome for rapid publication.
>
> All articles to be published in the European Journal of Science and Mathematics Education will undergo a rigorous and double-blinded peer review process by assessment of each article by a minimum of two referees.

This paper was accepted for publication without amendment, and published in April 2014.

5.13 More general reflection on the reception and impact of my work

My work with short-answer free-text questions became quite widely known, and featured as a case study in JISC's *Review of advanced e-assessment technologies (RaeAT)* (Ripley et al., 2009). It was also featured in JISC's guide to *Effective assessment in a digital age* (JISC, 2010) and was the case study selected to feature at the launch of this guide, at the Association for Learning Technology's annual conference (ALT-C) 2010 (Jordan, Butcher, Knight & Smith, 2010). JISC ran a series of workshops featuring some of the case studies from the guide to *Effective assessment in a digital age*, and I acted as facilitator of a session on "Designing interactive assessments to promote independent learning" in February 2011. I have been invited to give many presentations and webinars on my work with short-answer questions.

As indicated by my decision to align my work with the wider assessment and feedback literature, I have a longstanding interest in the use of assessment to support learning. Over the time-period covered by the publications in this submission, I have become increasingly respected in more general e-assessment and assessment circles, for example as one of the contributors to the JISC *Report on Summative E-assessment Quality* (Gilbert et al., 2009) and as a reviewer for *Computers & Education* and *Practitioner Research in Higher Education*.

I am a member of the organising committee for the biennial international Assessment in Higher Education Conference, leading the 'Assessment Technologies" strand. At the 2013 Conference, I was invited to run a master class on some aspect of e-assessment and chose to run this, with Tim Hunt, on "Producing high-quality computer-marked assessment", a topic on which I have spoken very many times to universities and other bodies across the UK. I have also given webinars with world-wide reach.

The data-driven research described in this thesis has led to some surprising but irrefutable conclusions, in particular the strength of the impact of question wording and assessment design on student behaviour and, more generally, the power of "assessment analytics" to provide evidence of the depth as well as the extent of student engagement with a module. These findings

have implications that extend considerably beyond computer-marked assessment, with relevance for the design of assessment of all types and a strong indication that the analysis of student responses to all assessment tasks has the potential to improve the effectiveness of learning analytics.

I have proved beyond reasonable doubt that relatively simple pattern-matching software can mark as accurately as human markers and more sophisticated software, but yet uptake of questions of this type and other constructed-response question types remains disappointingly low, despite concerns regarding the validity and authenticity of selected-response questions. Alongside further research into assessment analytics, there is a critical need for further work in this area, for example, into the use of peer review as a means of establishing the correctness of responses, into the use of machine learning to develop answer matching rules, and into the potential of short-answer free-text questions to assess and engage students on MOOCS and as a more accurate way than multiple-choice questions of establishing inventories of threshold concepts.

6. References

Abio, G. & Barandela, A.M. (2012). Algunas reflexiones sobre aprendizaje, evaluación formative y mediación technológica. *MarcoELE, 15*, 1-11.

Adcroft, A. & Willis, R. (2013). Do those who benefit the most need it the least? A four-year experiment in enquiry-based feedback. *Assessment & Evaluation in Higher Education*, *38*(7), 803-815.

Ahlgren, A. (1969, February). *Reliability, predictive validity, and personality bias of confidence-weighted scores*. Paper presented at the American Educational Research Association Convention, Los Angeles, California, 5th-8th February 1969. Retrieved 3rd March 2014 from http://files.eric.ed.gov/fulltext/ED033384.pdf

Angus, S. D. & 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.

Appleby, J., Samuels, P., & Treasure-Jones, T. (1997). DIAGNOSYS: A knowledge-based diagnostic test of basic mathematical skills. *Computers & Education*, *28*(2), 113-131.

Archer, R. & Bates, S. (2009). Asking the right questions: Developing diagnostic tests in undergraduate physics. *New Directions*, *5*, 22-25.

Ashburn, R. (1938). An experiment in the essay-type question. *The Journal of Experimental Education*, *7*(1), 1-3.

Ashton, H. S., Beevers, C. E., Korabinski, A. A., & Youngson, M. A. (2006a). Incorporating partial credit in computer-aided assessment of mathematics in secondary education. *British Journal of Educational Technology*, *37*(1), 93–119.

Ashton, H. S., Beevers, C. E., Milligan, C. D., Schofield, D. K., Thomas, R. C., & Youngson, M. A. (2006b). Moving beyond objective testing in online assessment. In S. C. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies from higher education, K-12 and corporate* (pp. 116-127). Hershey, PA: Information Science Publishing.

Ashton, H. S., Beevers, C. E. Schofield, D. K., & Youngson, M. A. (2004). Informative reports: Experience from the PASS-IT project. In *Proceedings of the 8th International Computer Assisted Assessment Conference, Loughborough*, *6th-7th July 2004.* Retrieved 1st March 2014 from http://caaconference.co.uk/pastConferences/2004/proceedings

Ashton, H. S. & Thomas, R. C. (2006). Bridging the gap between assessment, learning and teaching. In *Proceedings of the 10th International Computer Assisted Assessment Conference, Loughborough*, 4th-5th July 2006. Retrieved 1st March 2014 from http://caaconference.co.uk/pastConferences/2006/proceedings

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning & Assessment*, *4*(3).

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36.

Ball, S. (2009). Accessibility in e-assessment. *Assessment & Evaluation in Higher Education*, *34*(3), 293-303.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Barnett, R. (2007). Assessment in higher education: An impossible mission? In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education* (pp. 29-40). London: Routledge.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391-402.

Bayerlein, L. (in press). Students' feedback preferences: How do students react to timely and automatically generated assessment feedback? *Assessment & Evaluation in Higher Education*. DOI: 10.1080/02602938.2013.870531

Beatty, I. D. & Gerace, W. J. (2009). Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology. *Journal of Science Education and Technology*, *18*(2), 146-162.

Beevers, C. E. & Paterson, J. S. (2002). Assessment in mathematics. In P. Khan & J. Kyle (Eds.), *Effective Teaching and Learning in Mathematics and its Applications* (pp. 49-61). London: Kogan Page.

Beevers, C. E. & Paterson, J. S. (2003). Automatic assessment of problem-solving skills in mathematics. *Active Learning in Higher Education, 4*(2), 127-144.

Beevers, C. E., Wild, D. G., McGuire, G. R., Fiddes, D. J., & Youngson, M. A. (1999). Issues of partial credit in mathematical assessment by computer. *ALT-J*, *Research in Learning Technology*, *7*, 26-32.

Beevers, C. E. et al. (2010). What can e-assessment do for learning and teaching? Part 1 of a draft of current and emerging practice: Review by the E-Assessment Association expert panel. In *Proceedings of the 13th International Computer Assisted Assessment (CAA) Conference, Southampton, 20th-21st July 2010.* Retrieved 25th January 2014 from http://caaconference.co.uk/pastConferences/2010/

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction*, *2013*, Article ID 136864, 1-11.

Benson, R. & Brack, C. (2010). *Online learning and assessment in higher education: A planning guide.* Oxford: Chandos Publishing.

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, *34*(1), 1-15.

Bevan, R., Badge, J., Cann, A., Willmott, C., & Scott, J. (2008). Seeing eye-to-eye? Staff and student views on feedback. *Bioscience Education*, *12*. DOI: 10.3108/beej.12.1

Black, H. (1986). Assessment for Learning. In D. Nuttall (Ed.), *Assessing educational achievement* (pp. 7-18). London: Falmer Press.

Black, P. (1998). *Testing: Friend or foe? The theory and practice of assessment and testing*. London: RoutledgeFarmer.

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

Black, P. & Wiliam, D. (2006). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (pp. 119-131). London: Sage.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5-31.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds.). (1971*). Handbook of the formative and summative evaluation of student learning*. New York: McGraw Hill.

Boud, D. (1995). Assessment and learning: Contradictory or complementary? In P. T. Knight (Ed.), *Assessment for learning in higher education* (pp. 35-48). London: Kogan Page.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151-167.

Boyle, A. (2007). The formative use of e-assessment: Some early implementations, and suggestions for how we might move on. In *Proceedings of the 11th International Computer*

Assisted (CAA) Conference, Loughborough, 8th-9th July 2007. Retrieved 11th February 2014 from http://caaconference.co.uk/pastConferences/2007/proceedings

Boyle, A. & Hutchison, D. (2009). Sophisticated tasks in e-assessment: What are they and what are their benefits? *Assessment & Evaluation in Higher Education*, *34*(3), 305-319.

Bradis, V., Minkovskii, V., & Kharcheva, A. (1963). *Lapses in mathematical reasoning* (J.J. Schorr-Kon, Trans.). Oxford: Pergamon Press. (Translation of 2nd Russian edition, published 1959.)

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*(3), 253-271.

Britton, S., New, P., Sharma, M., & Yardley, D. (2005). A case study of the transfer of mathematics skills by university students. *International Journal of Mathematical Education in Science and Technology*, *36*(1), 1-13.

Broadfoot, P. (2008). Assessment for learners: Assessment literacy and the development of learning power. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 213-224). New York: Routledge.

Brosnan, M. (1999). Computer anxiety in students: Should computer-based assessment be used at all. In S. Brown, P. Race, & J. Bull (Eds.), *Computer-assisted assessment in higher education*. London: Kogan Page.

Brown, E., Gibbs, G., & Glover, C. (2003). Evaluation tools for investigating the impact of assessment regimes on student learning. *Bioscience Education*, *2*. DOI: 10.3108/beej.2003.02000006

Brown, E. & Glover, C. (2006). Evaluating written feedback. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 81-91). London: Routledge.

Bull, J. & Dyson, M. (2004). *Computer-aided assessment (CAA)*. York: LTSN Generic Centre.

Bull, J. & McKenna, C. (2004). *Blueprint for computer-aided assessment*. London: RoutledgeFalmer.

Burke, D. (2009). Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education*, *34*(1), 41-50.

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education*, *30*(1), 65-72.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, *25*(2), 157-163.

Butcher, P. G. (2008). Online assessment at the Open University using open source software: Moodle, OpenMark and more. In *Proceedings of the 12th International Computer Assisted (CAA) Conference, Loughborough, 8th-9th July 2008.* Retrieved 25th January 2014 from http://caaconference.co.uk/pastConferences/2008/proceedings

Butcher, P. G., Hunt, T. J., & Sangwin, C. J. (2013*).* Embedding and enhancing eAssessment in the leading open source VLE. In *Proceedings of the HEA-STEM Annual Conference, Birmingham, 17th-18th April 2013* (pp. 58-62)*.* DOI: 10.11120/stem.hea.2013.0020

Butcher, P. G., Swithenby, S. J., & Jordan, S. E. (2009, June). *eAssessment and the independent learner.* Paper presented at the 23rd ICDE World Conference on Open Learning and Distance Education, Maastricht, The Netherlands, 7th-10th June 2009. Retrieved 28th February 2014 from http://www.openhogeschoolnetwerk.com/Docs/Campagnes/ICDE2009/Papers/Final_Paper_278Butcher.pdf

Butterfield, B. & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491-1494.

Butterfield, B. & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69-84.

Caple, H. & Bogle, M. (2013). Making Group assessment transparent: What wikis can contribute to collaborative projects. *Assessment & Evaluation in Higher Education*, *38*(2), 198-210.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, *31*(2), 219-233.

Carless, D. (2013). Sustainable feedback and the development of student self-evaluative capacities. In S. Merry, M. Price, D. Carless & M. Taras (Eds.), *Reconceptualising Feedback in Higher Education: Developing dialogue with students* (pp. 113-122). London: Routledge.

Cassady, J. C. & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *Journal of Technology, Learning, and Assessment*, *4*(1).

Chanock, K. (2000). Comments on essays: Do students understand what tutors write? *Teaching in Higher Education, 5*(1), 95-105.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5), 318-331.

Cizek, G. J. (1991, April). *The effect of altering the position of options in a multiple-choice examination*. Paper presented at the Annual Meeting of the National Council on Measurement in Education in Chicago, 4th-6th April 1991. Educational Resources Document Reproduction Service (ERIC) #ED333024.

Clariana, R. (1993). A review of multiple-try feedback in traditional and computer-based instruction. *Journal of Computer-Based Instruction*, *20*(3), 67-74.

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In *LAK 2012: Proceedings of the 2nd International Conference on Learning Analytics & Knowledge, Vancouver, 29th April – 2nd May 2012* (pp. 134-138). Retrieved 17th March 2014 from http://dl.acm.org/citation.cfm?doid=2330601.2330636

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683-695.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100-108.

Conejo, R., Barros, B., Guzmán, E., & Garcia-Viñas, J. I. (2013). A web based collaborative testing environment. *Computers & Education*, *68*, 440-457.

Conole, G., de Laat, M., Darby, J., & Dillon, T. (2006). *JISC LXP: Student experience of technologies: Final report*. Bristol: JISC.

Conole, G. & Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J, Research in Learning Technology*, *13*(1), 17-31.

Coutts, R., Gilleard, W., & Baglin, R. (2011). Evidence for the impact of assessment on mood and motivation in first-year students. *Studies in Higher Education*, *36*(3), 291-300.

Cramp, A., Lamond, C., Coleyshaw, L., & Beck, S. (2012). Empowering or disabling? Emotional reactions to assessment amongst part-time adult students. *Teaching in Higher Education*, *17*(5), 509-521.

Crisp, B. R. (2007). Is it worth the effort? How feedback influences students' subsequent submission of assessable work. *Assessment & Evaluation in Higher Education*, *32*(5), 571-581.

Crisp, G. (2007). *The e-Assessment Handbook*. London, Continuum.

Crouch, C. H. & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970-977.

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

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7-24.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008a). PeerWise: Students sharing their multiple choice questions. In *Proceedings of the Fourth international Workshop on Computing*

Education Research, Sydney, 6th-7th September 2008 (pp. 51-58). Retrieved 3rd March 2014 from http://dl.acm.org/citation.cfm?doid=1404520.1404526

Denny, P., Luxton-Reilly, A., & Hamer, J. (2008b). The PeerWise system of student contributed assessment questions. In *Proceedings of the 10th Conference on Australasian Computing Education* (pp. 69-74). Retrieved 25th January 2014 from http://dl.acm.org/citation.cfm?id=1379255

Denny, P., Luxton-Reilly, A., & Hamer, J. (2008c). Student use of the PeerWise system. *ACM SIGCSE Bulletin*, *40*(3), 73-77.

Dermo, J. (2007). Benefits and obstacles: Factors affecting the uptake of CAA in undergraduate courses. In *Proceedings of the 11th International Computer Assisted Assessment Conference, Loughborough, 10th-11th July 2007*. Retrieved 1st March 2014 from http://caaconference.co.uk/pastConferences/2007/proceedings

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.

Dermo, J. (2010). In search of Osiris: Random item selection, fairness and defensibility in high-stakes e-assessment. In *Proceedings of the 13th International Computer Assisted Assessment (CAA) Conference, Southampton, 20th-21st July 2010.* Retrieved 25th January 2014 from http://caaconference.co.uk/pastConferences/2010/

Dermo, J. & Carpenter, L. (2011). e-Assessment for learning: Can online selected response questions really provide useful formative feedback? In *Proceedings of the 2011 International Computer Assisted Assessment (CAA) Conference, Southampton, 5th-6th July 2011*. Retrieved 23rd January 2014 from http://caaconference.co.uk/pastConferences/2011/

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

d'Inverno, R., Davis, H., & White, S. (2003). Using a personal response system for promoting student interaction. *Teaching Mathematics and its Applications*, *22*(4), 163-169.

Donaldson, S. I. & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245-260.

Dowden, T., Pittaway, S., Yost, H., & McCarthy, R. (2013). Students' perceptions of written feedback in teacher education: Ideally feedback is a continuing two-way communication that encourages progress. *Assessment & Evaluation in Higher Education*, *38*(3), 349-362.

Downing, S. M. (2003). Guessing on selected-response examinations. *Medical Education*, *37*(8), 670-671.

Draper, S. (2009a). Catalytic assessment: Understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology*, *40*(2), 285-293.

Draper, S. (2009b). What are learners actually regulating when given feedback? *British Journal of Educational Technology*, *40*(2), 306-315.

Drever, M. F. & Armstrong, B. (2000). An electronic feedback tool for distance education students called "MarkIt". In *Proceedings of Learning to Choose – Choosing to Learn: 17th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE), Coffs Harbour, NSW, 11th-13th December 2000* (pp. 467-475). Lismore, NSW: Southern Cross University Press.

Dufresne, R. J., Gerace, W. J., Leonard, W. J., Mestre, J. P., & Wenk, L. (1996). Classtalk: A classroom communication system for active learning. *Journal of Computing in Higher Education*, *7*(2), 3-47.

Duncan, N. (2007). "Feed-forward": Improving students' use of tutors' comments. *Assessment & Evaluation in Higher Education*, *32*(3), 271-283.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Dysthe, O. (2008). The challenges of assessment in a new learning culture. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 15-28). New York: Routledge.

Earley, P. C. (1988). Computer-generated performance feedback in the magazine-subscription industry. *Organizational Behavior and Human Decision Processes*, *41*(1), 50-64.

Ekins, J. (2008). Interactive mathematics e-quizzes using OpenMark. *MSOR Connections*, *8*(3), 21-24.

Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology*, *44*(4), 662-664.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70-120.

Eynon, R. (2009). Mapping the digital divide in Britain: Implications for learning and education. *Learning, Media & Technology*, *34*(4), 277-290.

Fazio, L. K. & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88-92.

Ferdig, R. E. & Mishra, P. (2004). Emotional responses to computers: Experiences in unfairness, anger, and spite. *Journal of Educational Multimedia and Hypermedia*, *13*(2), 143-161.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5), 304-317.

Ferrão, M. (2010). E-assessment within the Bologna paradigm: Evidence from Portugal. *Assessment & Evaluation in Higher Education*, *35(*7), 819-830.

Ferreira, A. & Atkinson, J. (2009). Designing a feedback component of an intelligent tutoring system for foreign language. *Knowledge Based Systems*, *22*, 496-501.

Field, D., Pulman, S., Van Labeke, N., Whitelock, D., & Richardson, J. (2013). Did I really mean that? Applying automatic summarisation techniques to formative feedback. In *Proceedings of the International Conference on Recent Advantages on Natural Language Programming, Hissar, Bulgaria, 9-11th September 2013* (pp. 277-284). Retrieved 2nd March 2014 from http://lml.bas.bg/ranlp2013/proceedings.php

Fies, C. & Marshall, J. (2006). Classroom response systems: A review of the literature. *Journal of Science Education and Technology*, *15*(1), 101-109.

Foster, B., Perfect, C., & Youd, A. (2012). A completely client-side approach to e-assessment and e-learning of mathematics and statistics. *International Journal of e-Assessment*, *2*(2).

Fowler, A. (2008). Providing effective feedback on whole-phrase input in computer-assisted language learning. In *Proceedings of the 12th International Computer Assisted (CAA) Conference, Loughborough, 8th-9th July 2008.* Retrieved 24th February 2014 from http://caaconference.co.uk/pastConferences/2008/proceedings

Fowler, A. (2014) Pattern: Try once, refine once. In Y. Mor, H. Mellar, S. Warburton & N. Winters (Eds.), *Practical design patterns for teaching and learning with technology* (pp. 323-327). Rotterdam: Sense Publishers.

Freake, S. (1999). Discovering science: A distance-learning course with integrated interactive multimedia. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (EdMedia), Seattle, 19th-14th June 1999 (pp. 1489-1490). Chesapeake, VA: Association for the Advancement of Computers in Education.

Freake, S. (2008). Electronic marking of physics assignments using a tablet PC. *New Directions*, *4*, 12-16.

Funk, S. C. & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, *38*(4), 273-277.

Fyfe, G., Fyfe, S., Meyer, J., Ziman, M., Sanders, K., & Hill, J. (2014). Students reflecting on test performance and feedback: An on-line approach. *Assessment & Evaluation in Higher Education*, *39*(2), 179-194.

Gardner-Medwin, A. R. (2006). Confidence-based marking: Towards deeper learning and better exams. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 141-149). London: Routledge.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109-127.

Gibbs, G. (2006). Why assessment is changing. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 11-22). London: Routledge.

Gibbs, G. & Simpson, C. (2004-5). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, 3-31.[11]

Gilbert, L., Gale., V., Warburton, B., & Wills, G. (2009) *Report on summative e-assessment quality (REAQ).* Southampton: University of Southampton. Retrieved 3rd March 2014 from http://www.jisc.ac.uk/media/documents/projects/reaqfinalreport.pdf

Gilbert, L., Whitelock, D., & Gale, V. (2011). *Synthesis report on assessment and feedback with technology enhancement: Report commissioned by the Higher Education Academy.* Southampton: University of Southampton and the Open University.

Gill, M. & Greenhow, M. (2008). How effective is feedback in computer-aided assessments? *Learning, Media and Technology*, *33*(3), 207-220.

[11] Previously published with the title "Does your assessment support your students' learning?"

Gipps, C. V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, *30*(2), 171-180.

Gipps, C. V. & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.

Glover, C., Macdonald, R., Mills, J., & Swithenby, S. (2005). Perceptions of the value of different modes of tutor feedback. In C. Rust (Ed.), *Improving student learning: Diversity and inclusivity. Proceedings of The 2004 12th Improving Student learning Symposium* (pp. 486-494). Oxford: Oxford Centre for Staff & Learning Development.

Gwinnett, C., Cassella, J., & Allen, M. (2011). The trials and tribulations of designing and utilising MCQs in HE and for assessing forensic practitioner competency. *New Directions*, *7*, 72-78.

Hadjidemetriou, C. & Williams, J. (2002). Children's graphical conceptions. *Research in Mathematics Education*, *4*(1), 69-87.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.

Haley, D., Thomas, P., Petre, M., & De Roeck, A. (2009). Human fallibility: How well do human markers agree? In *Proceedings of the 11th Australasian Conference on Computing Education, Wellington, New Zealand, 20th-23rd January 2009* (pp. 83-92). Retrieved 3rd March 2014 from http://dl.acm.org/citation.cfm?id=1862727

Handke, J. & Schäfer, A. M. (2012). *E-Learning, E-Teaching und E-Assessment in der Hochschullehre: Eine Anleitung*. München: Oldenbourg Verlag.

Harding, R. & Raikes, N. (2002). *ICT in assessment and learning: The evolving role of an external examinations board*. Maths CAA Series. Retrieved 30th January 2014 from http://www.heacademy.ac.uk/assets/documents/subjects/msor/mathscaa_feb2002.pdf

Hardy, J., Bates, S. P., Casey, M. M., Galloway, K. W., Galloway, R. K., Kay, A. E., Kirsop, P., & McQueen, H. A. (in press). Student-Generated Content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*. DOI: 10.1080/09500693.2014.916831

Harlen, W. (2006). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 61-80). London: Sage.

Hart, K. M. (1981). Fractions. In K. M. Hart (Ed.), *Children's understanding of mathematics 11-16* (pp. 66-81). London: John Murray.

Hart, K. M., Brown, M. L., Kuchemann, D. E., Kerslake, D., Ruddock, G., & McCartney, M. (1981). *Children's understanding of mathematics: 11-16*. London: John Murray.

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

Havnes, A. & McDowell, L. (2008). The dilemmas of assessment practice in educational institutions. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 115-119). New York: Routledge.

Haxton, K. J. & McGarvey, D. J. (2011). Screencasts as a means of providing timely, general feedback on assessment. *New Directions*, *7*, 18-21.

Herding, D. & Schroeder, U. (2012). Using capture and replay for semi-automatic assessment. *International Journal of e-Assessment*, *2*(1).

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141-158.

Hewson, C. (2012). Can online course-based assessment methods be fair and equitable? Relationships between students' preferences and performance within online and offline assessments. *Journal of Computer Assisted Learning*, *28*(5), 488-498.

Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, *27*(1), 53-64.

Hoban, R., Finlayson, O., & Nolan, B. (2013). Transfer in chemistry: A study of students' abilities in transferring mathematical knowledge to chemistry. *International Journal of Mathematical Education in Science and Technology*, *44*(1), 14-35.

Hoffman, B. (1967). Multiple-choice tests. *Physics Education*, *2*, 247-51.

Holden, G. & Glover, C. (2013). Fostering institutional change in feedback practice through partnership. In S. Merry, M. Price, D. Carless & M. Taras (Eds.), *Reconceptualising Feedback in Higher Education: Developing dialogue with students* (pp. 160-171). London: Routledge.

Holmes, N. (in press). Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module*. Assessment & Evaluation in Higher Education*. DOI: 10.1080/02602938.2014.881978

Hounsell, D. (2007). Towards more sustainable feedback to students. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education* (pp. 101-113). London: Routledge.

Howarth, M. J. & Smith, B. J. (1980). Attempts to identify and remedy the mathematical deficiencies of engineering undergraduate entrants at Plymouth Polytechnic. *International Journal of Mathematical Educational in Science & Technology*, *11*(3), 377-383.

Hsieh, I. L. G. & O'Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior*, *18*(6), 699-715.

Hughes, G. (2011). Towards a personal best: A case for introducing ipsative assessment in higher education. *Studies in Higher Education*, *36*(3), 353-367.

Hunt, T. J. (2012). Computer-marked assessment in Moodle: Past, present and future. In *Proceedings of the 2012 International Computer Assisted Assessment (CAA) Conference,*

Southampton, 10th-11th July 2012. Retrieved 23rd January 2014 from http://caaconference.co.uk/proceedings/

JISC. (2006). *e-Assessment Glossary (Extended)*. Retrieved 1st April 2014 from http://www.jisc.ac.uk/uploaded_documents/eAssess-Glossary-Extended-v1-01.pdf

JISC. (2010). *Effective assessment in a digital age: A guide to technology-enhanced assessment and feedback*. Retrieved 25th January 2014 from http://www.jisc.ac.uk/publications/programmerelated/2010/digiassess.aspx

Jones, O. & Gorra, A. (2013). Assessment feedback only on demand: Supporting the few not supplying the many. *Active Learning in Higher Education*, *14*(2), 149-161.

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63-76.

Jordan, H. E. (2009*). iCMA statistics report*. Retrieved 24th January 2014 from http://www.open.ac.uk/blogs/SallyJordan/wp-content/uploads/2011/04/report1.pdf

Jordan, S. E. (2007). The mathematical misconceptions of adult distance-learning science students. In D. Green (Ed.), *Proceedings of the CETL-MSOR Conference, Loughborough University, 11th-12th September 2006* (pp. 87-92). Birmingham: Maths, Stats & OR Network.

Jordan, S. E. (2009a). Assessment for learning: Pushing the boundaries of computer based assessment. *Practitioner Research in Higher Education*, *3*(1), 11-19.

Jordan, S. E. (2009b). *Investigating the use of short free text questions in online assessment. COLMSCT final report*. Retrieved 21st April 2014 from http://www.open.ac.uk/opencetl/resources/colmsct-resources/jordan-s-2009-colmsct-final-report-investigating-the-use-short-free-text-questions-online-assessment

Jordan, S. E. (Ed.). (2010). *e-Assessment: Compilation of final reports on Open University Physics Innovations CETL Projects*. Milton Keynes: The Physics Innovations Centre for Excellence in Teaching and Learning.

Jordan, S. E., Bolton, J. P. R., Cook, L. J., Datta, S. B., Golding, J. P., Haresnape, J. M., Jordan, R. S., Murphy, K. P. S. J., New, K. J., & Williams, R. T. (in press). *Thresholded assessment: Does it work? Report on an eSTEeM Project.* Will be available in August 2014 from http://www.open.ac.uk/about/teaching-and-learning/esteem/projects/themes/innovative-assessment/thresholded-assessment-does-it-work

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

Jordan, S.E. & Butcher, P.G. (2013*).* Does the Sun orbit the Earth? Challenges in using short free-text computer-marked questions. In *Proceedings of the HEA-STEM Annual Conference, Birmingham, 17th-18th April 2013 (pp. 53-57).* DOI: 10.11120/stem.hea.2013.0012

Jordan, S. E., Butcher, P. G., Knight, S., & Smith, R. (2010). "Your answer is not quite correct, try again": Making online assessment and feedback work for learners. In *"Into something rich and strange" – making sense of the sea-change. The 17th Association for Learning Technology Conference (ALT-C 2010), University of Nottingham, 7th-9th September 2010* (p. 121). Abstract retrieved 25th January 2014 from http://repository.alt.ac.uk/798/2/Abstracts_Handbook_web.pdf

Jordan, S. E., Ross, S. M., & Murphy, P. J. (2013). *Maths for science*. Oxford: Oxford University Press in association with the Open University.

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Sally Jordan

Kear, K. (2010). *Online and social networking communities: A best practice guide for educators*. London: Routledge.

Kerslake, D. (1981). Graphs. In K. M. Hart (Ed.), *Children's understanding of mathematics 11-16* (pp. 120-136). London: John Murray.

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.

Kirkwood, A. & Price, L. (2005). Learners and learning in the Twenty-First Century: What do we know about students' attitudes towards and experiences of information and communication technologies that will help us design courses? *Studies in Higher Education*, *30*(3), 257-274.

Kirkwood, A. & Price, L. (2008). Assessment and student learning: A fundamental relationship and the role of information and communication technologies. *Open Learning*, *23*(1), 5-16.

Kleeman, J. (2013). *Assessment prior art.* Retrieved 22nd February 2014 from http://assessmentpriorart.org/

Kluger, A. N. & 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.

Knight, P. T. (Ed.). (1995). *Assessment for learning in higher education*. London: Kogan Page.

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education*, *27*(3), 275-286.

Kornell, N. & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219-224.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989-998.

Kuechler, W. L. & Simkin, M. G. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education*, *14*(4), 389-400.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2), 211-232.

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

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295-308.

Lantz, M. E. (2010). The use of "Clickers" in the classroom: Teaching innovation or merely an amusing novelty? *Computers in Human Behavior*, *26*(4), 556-561.

Lasry, N., Mazur, E., & Watkins, J. (2008). Peer instruction: From Harvard to the two-year college. *American Journal of Physics*, *76*(11), 1066-1069.

Laurillard, D. (2002). *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. 2nd edition. London: RoutledgeFalmer.

Lawless, C. & Freake, S. (2001). Students' use of multimedia activities in an Open University introductory science course. *Journal of Educational Media*, *26*(2), 117-141.

Leacock, C. & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers & Humanities*, 37(4), 389-405.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*(3), 210-212.

Leopold, D. & Edgar, B. (2008). Degree of mathematics fluency and success in second-semester introductory chemistry. *Journal of Chemical Education*, *85*(5), 724-731.

Li, J. & De Luca, R. (2014). Review of assessment feedback. *Studies in Higher Education*, *39*(2), 378-393.

Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education*, *43*(1), 109-123.

Lipnevich, A. A. & Smith, J. K. (2008). *The effects of differential feedback on students' performance.* Princeton, NJ: Educational Testing Service.

Lipnevich , A. A. & Smith, J. K. (2009). "I really need feedback to learn": Students' perspectives on the effectiveness of the differential feedback messages*. Educational Assessment Evaluation & Accountability*, *21*, 347-367.

Littauer, R. (1972). Instructional implications of a low-cost electronic student response system. *Educational Technology: Teacher and Technology Supplement*, *12*(10), 69-71.

Lobb, R. (2013). Coderunner [computer software]. Retrieved 22nd April 2014 from https://github.com/trampgeek/CodeRunner

Lunt, T. & Curran, J. (2010). "Are you listening please?" The advantages of electronic audio feedback compared to written feedback. *Assessment & Evaluation in Higher Education*, *35*(7), 759-769.

Luxton-Reilly, A. & Denny, P. (2010). Constructive evaluation: A pedagogy of student-contributed assessment. *Computer Science Education*, *20*(2), 145-167.

Lyle, K. B. & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*(2), 94-97.

McAllister, D. & Guidice, R. M. (2012). This is only a test: A machine-graded improvement to the multiple-choice and true-false examination. *Teaching in Higher Education*, *17*(2), 193-207.

McArthur, J. & Huxham, M. (2013). Feedback unbound: From master to usher. In S. Merry, M. Price, D. Carless & M. Taras (Eds.), *Reconceptualising Feedback in Higher Education: Developing dialogue with students* (pp. 92-102). London: Routledge.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200-206.

McDermott, L., Rosenquist, M., & VanZee, E. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, *55*(6), 503-513.

McGuire, G. R., Youngson, M. A., Korabinski, A. A., & McMillan, D. (2002). Partial credit in mathematics exams: A comparison of traditional and CAA exams. In *Proceedings of the 6th International Computer-Assisted Assessment Conference, Loughborough, 9th-10th July 2002.* Retrieved 1st March 2014 from http://caaconference.co.uk/pastConferences/2002/proceedings

McKenna, C. & Bull, J. (2000). Quality assurance of computer-assisted assessment: Practical and strategic issues. *Quality Assurance in Education*, *8*(1), 24-32.

Mackenzie, D. (1999). Recent developments in the Tripartite Interactive Assessment Delivery system (TRIADs). In *Proceedings of the 3rd International Computer-Assisted Assessment Conference, Loughborough*, *June 1999.* Retrieved 1st March 2014 from http://caaconference.co.uk/pastConferences/1999/proceedings

Mackenzie, D. (2003). Assessment for e-learning : What are the features of an ideal e-assessment system? In *Proceedings of the 7th International Computer-Assisted Assessment (CAA) Conference, Loughborough, 8th-9th July 2003*. Retrieved 24th January 2014 from http://caaconference.co.uk/pastConferences/2003/procedings

Mackenzie, D. (2004). *Online assessment: Quality production and delivery for higher education*. Retrieved 7th April 2014 from www.enhancementthemes.ac.uk/documents/events/20040416/Mackenziepaper-revised.pdf

Malmi, L., Karavirta, V., Korhonen, A., & Nikander, J. (2005). Experiences on automatically assessed algorithm simulation exercises with different resubmission policies. *Journal on Educational Resources in Computing*, *5*(3), Article 7.

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.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194-199.

Mason, B. J. & Bruning, R. (n.d.). *Providing feedback in computer-based instruction: What the research tells us.* Retrieved 1st March 2014 from http://dwb.unl.edu/Edit/MB/MasonBruning.html

Mathews, J. (2006, November 14). Just whose idea was all this testing? *The Washington Post*, A06.

Mazur, E. (1991). *Peer instruction: A user's manual*. New Jersey: Prentice-Hall.

Merry, S., Price, M., Carless, D., & Taras, M. (2013). *Reconceptualising feedback in higher education: Developing dialogue with students*. London: Routledge.

Millar, J. (2005). *Engaging students with assessment feedback: What works? An FDTL5 Project literature review.* Oxford: Oxford Brookes University.

Miller, T. (2008). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, 34(2), 181-192.

Mills, R. (2004). Looking back, looking forward: What have we learned? In J.E. Brindley, C. Walti & O. Zawacki-Richter (Eds.), *Learner Support in Open, Distance and Online Learning Environments* (pp. 29-37). Oldenburg: BIS-Verlag der Carl von Ossietzky Universität Oldenburg.

Mishra, P. (2006). Affective feedback from computers and its effect on perceived ability and affect: A test of the computers as social actor hypothesis. *Journal of Educational Multimedia and Hypermedia*, *15*(1), 107-131.

Mitchell, T., Aldridge, N., Williamson, W., & Broomhead, P. (2003). Computer based testing of medical knowledge. In *Proceedings of the 7th International Computer-Assisted Assessment (CAA) Conference, Loughborough, 8th-9th July 2003*. Retrieved 24th January 2014 from http://caaconference.co.uk/pastConferences/2003/proceedings

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer-Assisted Assessment Conference, Loughborough, 9th-10th July 2002.* Retrieved 24th January 2014 from http://caaconference.co.uk/pastConferences/2002/proceedings

Morgan, C. & O'Reilly, M. (1999). *Assessing open and distance learners*. London: Kogan Page.

Murphy, S. (2008). Some consequences of writing assessment. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 33-49). New York: Routledge.

Mutch, A. (2003). Exploring the practice of feedback to students. *Active Learning in Higher*

National Union of Students. (2010). Charter on feedback and assessment. Retrieved 3rd April 2014 from http://www.nusconnect.org.uk/asset/news/6010/FeedbackCharter-toview.pdf

Nicol, D. (2007). E-assessment by design: Using multiple choice tests to good effect. *Journal of Further & Higher Education*, *31*(1), 53-64.

Nicol, D. (2008). *Technology-supported assessment: A review of research*. Unpublished manuscript. Retrieved 6th April 2014 from http://www.reap.ac.uk/Portals/101/Documents/REAP/Technology_supported_assessment.pdf

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501-517.

Nicol, D. & Boyle, J. (2003). Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom. *Studies in Higher Education*, *28*(4), 457-473.

Nicol, D. & 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.

Nicol, D. & Milligan, C. (2006). Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 64-77). London: Routledge.

Nix, I. & Wyllie, A. (2011). Exploring design features to enhance computer-based assessment: Learners' views on using a confidence-indicator tool and computer-based feedback. *British Journal of Educational Technology*, *42*(1), 101-112.

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

Orrell, J. (2008). Assessment beyond belief: The cognitive process of grading. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 251-263). New York: Routledge.

Orsmond, P. & Merry, S. (2011). Feedback alignment: Effective and ineffective links between tutors' and students' understanding of coursework feedback. *Assessment & Evaluation in Higher Education*, *36*(2), 125-136.

Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education*, *30*(4), 369-386.

Parkin, H. J., Hepplestone, S., Holden, G., Irwin, B., & Thorpe, L. (2012). A role for technology in enhancing students' engagement with feedback. *Assessment & Evaluation in Higher Education*, *37*(8), 963-973.

Perelman, L. (2008). Information illiteracy and mass market writing assessments. *College Composition and Communication*, *60*(1), 128-141.

Poulos, A. & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education*, *33*(2), 143-154.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277-289.

Price, M., Handley, K., O'Donovan, B., Rust, C., & Millar, J. (2013). Assessment feedback: An agenda for change. In S. Merry, M. Price, D. Carless and M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 41-53). London: Routledge.

Pulman, S. G. & Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP, Ann Arbor, June 2005* (pp. 9-16). New Brunswick, NJ: Association for Computational Linguistics.

Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010). The quality of a PeerWise MCQ repository. In *Proceedings of the 12th Australasian Conference on Computing Education, Brisbane, 18th-22nd January 2010* (pp. 137-146). Retrieved 3rd March 2014 from http://dl.acm.org/citation.cfm?id=1862238&CFID=298176117&CFTOKEN=11936262

Pyper, A. & Lilley, M. (2010). A comparison between the flexilevel and conventional approaches to objective testing. In *Proceedings of the 13th International Computer Assisted (CAA) Conference, Southampton, 20th-21st July 2010.* Retrieved 3rd March 2014 from http://caaconference.co.uk/pastConferences/2010/

Race, P., Brown, S., & Smith, B. (2005). *500 tips on assessment*. 2nd edition. London : Routledge.

Raikes, N. & Harding, R. (2003). The horseless carriage stage: Replacing conventional measures. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 267-277.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Read, B., Francis, B., & Robson, J. (2005). Gender, bias, assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, *30*(3), 241-260.

Rebello, N. S. & Zollman, D. A. (2003). The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*, *72*(1), 116-125.

Redecker, C. (2013). *The Use of ICT for the Assessment of Key Competences*. European Commission Joint Research Centre, report JRC76971. Retrieved 24th January 2014 from http://boletines.prisadigital.com/JRC76971.pdf

Redecker, C. & Johannessen, Ø. (2013). Changing assessment: Towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79-96.

Redecker, C., Punie, Y., & Ferrari, A. (2012). eAssessment for 21st Century Learning and Skills. In A. Ravenscroft, S. Lindstaedt, C.D. Kloos & D. Hernandez-Leo (Eds.), *21st Century Learning for 21st Century Skills* (pp. 292-305). Berlin: Springer.

Re-Engineering Assessment Practice in Scottish Higher Education. (2007). *Assessment principles: Some possible candidates*. Retrieved 2nd February 2014 from http://www.reap.ac.uk/reap/resourcesPrinciples.html

Reeves, B. & Nass, C. (1996). *The media equation*. Stanford, CA: Center for the Study of Language and Information.

Richardson, M., Baird, J. A., Ridgway, J., Ripley, M., Shorrocks-Taylor, D., & Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behavior*, *18*(6), 633-649.

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment*. Bristol: Futurelab.

Ripley, M., Harding, R., Redif, H., Ridgway, J., & Tafler, J. (2009). *Review of advanced e-assessment technologies (RAeAT): Final report*. Retrieved 6th April 2014 from http://www.jisc.ac.uk/media/documents/projects/raeat_finalreport.pdf

Robinson, A. & Udall, M. (2006). Using formative assessment to improve student learning through critical reflection. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 92-99). London: Routledge.

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

Roediger, H. L. & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155 -1159.

Rosewell, J. P. (2011). Opening up multiple-choice: Assessing with confidence. In *Proceedings of the 2011 International Computer Assisted Assessment (CAA) Conference, Southampton, 5th-6th July 2011.* Abstract retrieved 3rd March 2014 from http://caaconference.co.uk/pastConferences/2011/

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education*, *30*(3), 231-240.

Ryan, J. & Williams, J. (2007). *Children's mathematics 4–15: Learning from errors and misconceptions*. Maidenhead: Open University Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education*, *5*(1), 77-84.

Sainsbury, M. & Benton, T. (2011). Designing a formative e-assessment: Latent class analysis of early reading skills. *British Journal of Educational Technology*, *42*(3), 500-514.

Sambell, K. & McDowell, L. (1998). The construction of the hidden curriculum: Messages and meanings in the assessment of student learning. *Assessment and Evaluation in Higher Education*, *23*(4), 391-402.

Sancho-Vinuesa, T., Escudero-Viladoms, N., & Masià, R. (2013). Continuous activity with immediate feedback: A good strategy to guarantee student engagement with the course. *Open Learning*, *28*(1), 51-66.

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford: Oxford University Press.

Sangwin, C. J. & Grove, M. J. (2006). STACK: Addressing the needs of the "neglected learners". In *Proceedings of the First WebALT Conference and Exhibition, Technical University of Eindhoven, Netherlands*, 5th-6th January 2006 (pp. 81-95). Retrieved 8th April 2014 from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=3F4E0172FCD4B94BC259110A8A080963?doi=10.1.1.115.3418&rep=rep1&type=pdf

Sawyer, W. (1964). *Vision in elementary mathematics*. Harmondsworth: Penguin Books.

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' experiences with an automated essay scorer. *The Journal of Technology, Learning and Assessment*, *7*(1).

Schechter, E. (2009). *The most common errors in undergraduate mathematics*. Retrieved 25th January 2014 from http://www.math.vanderbilt.edu/~schectex/commerrs/

Scott, S. V. (2014). Practising what we preach: Towards a student-centred definition of feedback. *Teaching in Higher Education*, *19*(1), 49-57.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453-472.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler (Ed.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago: Rand McNally.

Sharpe, R. (2009). The impact of learner experience research on transforming institutional practices. In T. Mayes, D. Morrison, H. Mellar, P. Bullen, & M. Oliver (Eds.), *Transforming higher education through technology-enhanced learning* (pp. 178-190). York: Higher Education Academy.

Shephard, K. (2009). E is for exploration: Assessing hard-to-measure learning outcomes. *British Journal of Educational Technology*, *40*(2), 386-398.

Shermis, M. D. & Hammer, D. (2012). *Contrasting state-of-the-art automated scoring of essays: Analysis.* Retrieved 2nd March 2014 from http://www.scribd.com/doc/91191010/Mark-d-Shermis-2012-contrasting-State-Of-The-Art-Automated-Scoring-of-Essays-Analysis

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153-189.

Siddiqi, R. (2013, December). *Impact of automated short-answer marking on students' learning: IndusMarker, a case study.* Paper presented at the 5th International Conference on Information & Communication Technologies (ICICT), Karachi, Pakistan, 14th-15th December 2013. Retrieved 15th February 2014 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6732782&tag=1

Siddiqi, R. & Harrison, C. J. (2008, September). *On the automated assessment of short free-text responses*. Paper presented at the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge, 7th-12th September 2008. Retrieved 1st March 2014 from http://www.iaea.info/papers.aspx?id=71

Sim, G., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: Lessons from the literature. *ALT-J, Research in Learning Technology*, *12*(3), 215-229.

Sim, J. W. S. & Hew, K. F. (2010). The use of weblogs in higher education settings: A review of empirical research. *Educational Research Review*, *5*(2), 151-163.

Simkin, M. G. & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What Is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73-98.

Slamecka, N. J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592-604.

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

Smith, E. & Gorard, S. (2005). "They don't give us our marks": The role of formative feedback in student progress. *Assessment in Education: Principles, Policy & Practice*, *12*(1), 21-38.

Snyder, B.R. (1971). *The hidden curriculum*. New York: Alfred A. Knopf.

Sorensen, E. (2013). Implementation and student perceptions of e-assessment in a chemical engineering module. *European Journal of Engineering Education*, *38*(2), 172-185.

Stobbart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). London: Sage.

Stödberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education*, *37*(5), 591-604.

Stone, D., Jarrett, C., Woodroffe, M., & Mincoha, S. (2005). *User interface design and evaluation*. San Francisco: Morgan Kaufman.

Strickland, N. (2002). Alice Interactive Mathematics, *MSOR Connections*, *2*(1), 27–30.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, *30*(4), 325-341.

Sukkarieh, J. Z. & Blackmore, J. (2009). C-rater: Automatic content scoring for short constructed responses. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society (FLAIRS) Conference, Sanibel Island, Florida, 19th-21st May 2009* (pp. 290-295). Retrieved 1st March 2014 from http://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/view/122/302

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003, October). *Auto-marking: Using computational linguistics to score short, free-text responses*. Paper presented at the 29th International Association for Educational Assessment (IAEA) Annual Conference, Manchester, October 2003.

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2004, June). *Auto-marking 2: Using computational linguistics to score short, free-text responses*. Paper presented at the 30th International Association for Educational Assessment (IAEA) Annual Conference, Philadelphia, June 2004.

Tariq, V. (2008). Defining the problem: Mathematical errors and misconceptions exhibited by first-year bioscience undergraduates. *International Journal of Mathematical Education in Science and Technology*, *39*(7), 889-904.

Tate, N. (2005, October). *Maintaining trust in public assessment systems: An international perspective.* Paper presented at the Cambridge Assessment Conference, 17th October 2005. Retrieved 3rd March 2014 from http://prd.cambridgeassessment.org.uk/Images/126036-dr-nicholas-tate.pdf

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(4), 901-918

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319-330.

van de Mortel, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, *25*(4), 40-48.

Van Gaal, F. & De Ridder, A. (2013). The impact of assessment tasks on subsequent examination performance. *Active Learning in Higher Education*, *14*(3), 213-225.

Värlander, S. (2008). The role of students' emotions in formal feedback situations. *Teaching in Higher Education*, *13*(2), 145-156.

Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. *Computers & Education*, *54*(2), 455-461.

Voelkel, S. (2013). Combining the formative with the summative: The development of a two stage online test to encourage engagement and provide personal feedback in large classes. *Research in Learning Technology*, *21*.

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, *28*(2), 97-111.

Walet, N. & Birch, M. (2012). Using online assessment to provide instant feedback. In *Proceedings of the HEA-STEM Annual Conference, London, 12th-13th April 2012*. DOI: 10.11120/stem.hea.2012.095

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

Walker, D. M. & Thompson, J. S. (2001). A note on multiple choice exams, with respect to students' risk preference and confidence. *Assessment and Evaluation in Higher Education*, *26*(3), 261-267.

Walker, M. (2009). An investigation into written comments on assignments: Do students find them usable? *Assessment & Evaluation in Higher Education*, *34*(1), 67-78.

Walker, M. (2013). Feedback and feedforward: Student responses and their implications. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising Feedback in Higher Education: Developing dialogue with students* (pp. 103-112). London: Routledge.

Warburton, B. & Conole, G. (2005). Whither e-assessment? In *Proceedings of the 9th International Computer-Assisted Assessment Conference, Loughborough, 5th-6th July 20035.* Retrieved 1st March 2005 from http://caaconference.co.uk/pastConferences/2005/proceedings

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

Whitelock, D. & Brasher, A. (2006). *Roadmap for e-assessment: Report for JISC*. Retrieved 24th February 2014 from http://www.jisc.ac.uk/whatwedo/programmes/elearningpedagogy/assessment.aspx

Wieman, C. (2007). Why not try a scientific approach to science education? *Change: The Magazine of Higher Learning*, *39*(5), 9-15.

Wiliam, D. (2008). Balancing dilemmas: Traditional theories and new applications. In A. Havnes & L. McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 267-281). New York: Routledge.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3-14.

Wiliam, D. & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, *22*(5), 537-548.

Williams, J. (2006). Assertion-reason multiple-choice testing as a tool for deep learning: A qualitative analysis. *Assessment & Evaluation in Higher Education*, *31*(3), 287-301.

Williams, J. & Ryan, J. (2000). National testing and the improvement of classroom teaching: Can they coexist? *British Educational Research Journal*, *26*(1), 49-73.

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

Winter, C. & Dye, V. (2004). An investigation into the reasons why students do not collect marked assignments and the accompanying feedback. Wolverhampton: University of Wolverhampton. Retrieved 5th April 2014 from http://wlv.openrepository.com/wlv/bitstream/2436/3780/1/An%20investigation%20pgs%20133-141.pdf

Wood, R. (1991). *Assessment and Testing: A survey of research commissioned by the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Xiaoling, Z. & Xuning, Z. (2013). Effect of different score reports of web-based formative test on students' self-regulated learning. *Computers & Education*, *66*. 54-63.

Yang, M. & Carless, D. (2013). The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education*, 18(3), 285-297.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, *45*(4), 477-501.

Yorke, M. (2013). Surveys of "the student experience" and the politics of feedback. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising Feedback in Higher Education: Developing dialogue with students* (pp. 6-18). London: Routledge.

Appendix A Abbreviations

AEQ	Assessment Evaluation Questionnaire (developed by the FAST Project)
AfL	Assessment for Learning
AiM	Assessment in Mathematics, a web-based assessment system based on the Maple computer algebra system.
ALT-C	The Association for Learning Technology's annual conference
ASKe	Assessment Standards Knowledge exchange (Oxford Brookes University)
CAA	Computer-aided assessment *or* Computer-assisted assessment
CALM	Computer-Aided Learning of Mathematics Project (Heriot-Watt University)
CAS	Computer algebra system
CASA	Computers as social actors
CETL	Centre for Excellence in Teaching and Learning
CMA	Computer-marked assignment (not necessarily interactive)
COLMSCT	Centre for Open Learning of Mathematics, Science, Computing and Technology
CUE	An e-assessment system, developed collaboratively by the CALM project, UCLES (University of Cambridge Local Examinations Syndicate) and EQL (a company)
ECA	End-of-course assessment (now called an end-of-module assessment (EMA))
EMA	End-of-module assessment
eSTEeM	The OU's centre for scholarship, innovation and enterprise in science, technology, engineering and mathematics subjects
eTMA	Tutor-marked assignment submitted electronically
FAST	Formative Assessment in Science Teaching Project (OU and Sheffield Hallam University)
FDTL	Fund for the Development of Teaching and Learning
IAT	Intelligent Assessment Technologies Ltd.
iCMA	Interactive computer-marked assignment

iEMA	End-of-module assignment delivered as an iCMA
JISC	Historically stood for “Joint Information Systems Committee”, now just JISC
MOOC	Massive open online course
NLP	Natural language processing
NUMBAS	An e-assessment and e-learning system, developed at Newcastle University
OU	The UK Open University
PA	Practice assessment or Practice assignment
PASS-IT	Project on ASsessment in Scotland – using Information Technology
piCETL	Physics Innovations Centre for Excellence in Teaching and Learning
PMatch	OpenMark’s pattern-matching question type
REAP	Re-engineering Assessment Practices in Higher Education Project (University of Strathclyde)
SCHOLAR	An online learning programme, originating at Heriot-Watt University
STACK	System for Teaching and Assessment using a Computer algebra Kernel (developed at the University of Birmingham but now a Moodle question type)
TMA	Tutor-marked assignment
TRIADs	TRipartite Interactive Assessment Delivery system (University of Derby)
VLE	Virtual learning environment

Appendix B OU Science Faculty modules and codes

From 2008-2012, students who wanted to study for a Natural Sciences degree had a choice of first module, starting either with the 60-credit module *Exploring science* (S104), which introduces students to physics, astronomy, chemistry, biology, Earth and environmental science, and to the mathematical and study skills needed for subsequent study, or, if they did not have the basic scientific and mathematical skills required by S104, with the 10-credit "gateway" module *Science starts here* (S154). A diagnostic quiz, *Are you ready for S104?* (initially part of *Are you ready for level 1 study?*), was used to direct students to the appropriate starting point (Publication 6, p. 151-152 & p. 160-161). S104 was complemented by the 10-credit residential school module *Practising science* (SXR103), which included a week at residential school in which students gained experience in the laboratory and the field. Students who wanted additional practice in mathematical skills but who did not want to take 30 credits of mathematics before higher level study could choose to study the 10-credit *Maths for science* (S151) and similarly, the 10-credit *Scientific investigations* (S155) was available to students who wanted more experience of investigative science or, latterly, as an alternative to SXR103. In addition, a series of 10-credit subject based "Science short modules" were available to students, on subjects such as *Understanding the weather* (S189) and *Galaxies, stars and planets* (S177).

The changes to higher education funding in England in 2012 resulted in the discontinuation of most 10-credit modules, including S154, S189, ST174. The starting point for most students is now *Exploring science* (S104), though a new 30-credit *Science, technology and maths access module* (Y033) has been available since October 2013. *Are you ready for S104?* is still in use, now directing students who are insufficiently prepared for S104 either to some online preparatory material or to Y033. After studying S104 (or if they have sufficient time, concurrently), all natural sciences students study *Investigative and mathematical skills in science* (S141), which includes the content of *Maths for science* (S151), *Understanding the weather* (S189) and *Scientific investigations* (S155), repackaged into a single 30-credit module.

The codes for the modules mentioned in the publications or covering paper are as follows:

S103	*Discovering science* (1998−2007) (replaced by S104)	60-credit level 1 module
S104	*Exploring science* (2008−)	60-credit level 1 module
S141	*Investigative and mathematical skills in science* (2012−)	30-credit level 1 module
S151	*Maths for science* (2002−2012; new edition 2012−)	10-credit level 1 module
S154	*Science starts here* (2007−2012)	10-credit level 1 module
S155	*Scientific investigations* (2010−)	10-credit level 1 module
S177	*Galaxies, stars and planets* (2012−)	10-credit level 1 module
S189	*Understanding the weather* (2008−2013)	10-credit level 1 module
S207	*The physical world* (2000−)	30-credit level 2 module
S240	*Analytical science* (2012−)	30-credit level 2 module
S279	*Our dynamic planet: Earth and life* (2007−2013)	30-credit level 2 module
SDK125	*Introducing health sciences* (2008−)	30-credit level 1 module
SXR103	*Practising science* (Residential school; 2001−2012)	10-credit level 1 module
SM358	*The quantum world* (2007−)	30-credit level 3 module
Y033	*Science, technology and maths access module* (2013−)	30-credit access module

The start date of each presentation is indicated by the year and a letter to indicate the month, so 2008B indicates a presentation which started in February 2008, since B is the second letter of the alphabet and February is the second month, whilst 2012J indicates a presentation which started in October 2012J, since J is the tenth letter of the alphabet and October is the tenth month.

Appendix C Research methods

Quantitative measures:

- Analysis of computer-marked assessment interactions by students on the first (2002I; 245 registered students), second (2002K; 497 students) and third (2003B; 379 students) presentations of S151 (Publications 1 and 2).
- Analysis of computer-marked assessment interactions on a range of presentations and modules, including S104 2008B (1498 registered students), S154 2008J (977 students), SDK125 08J (530 students), S279 07B (496 students), SM358 09B (380 students) (Publications 7 and 10); S207 12J (657 students), S240 12J (208 students), SDK125 13B (971 students) (Section 3.1.1).
- Analysis of computer-marked assessment interactions by 316 students on "Module Y" and 272 students on "Module Z" in 2013 (Publication 10).
- Analysis of interactions by 89,565 users of *Are you ready for level 1 science?* (2007-2009).
- Statistical analysis of the behaviour of tests, questions and different variants of questions on a range of modules in 2009, in particular for 1508 students on "Module W" and 362 students on "Module X" (Publication 7).
- Analysis of responses submitted by 274 students to the January 2003 S151 end-of-module assessment (EMA) (Publication 1). Similar analyses of responses submitted to the February 2010 EMA (262 students), May 2010 EMA (478 students), September 2010 EMA (285 students), November 2010 EMA (415 students) and the January 2013 EMA (322 students) (Publication 10 and Publication 12).
- Analysis of a total of 44,220 responses to 7 short-answer free-text questions in formative, summative and diagnostic questions, with the detail as described in Publication 8 Table 1.
- Comparison of the human and computer marking of between 92 and 248 responses to seven short-answer free-text questions (Publications 4 and 5).

- Comparison of the accuracy of marking of three computer systems for between 361 and 849 responses to seven short-answer free-text questions (Publication 5).
- Comparison of the PMatch marking of between 1591 and 2218 responses to eleven questions with that of a human "expert" (Publication 9).

Qualitative measures:

- Postal questionnaire sent to 201 students on the first (2002I) presentation of S151, and returned by 92 students (46%) (reported in Publication 1).
- More detailed postal questionnaire sent to 500 students on the second (2002K) presentation of S151, and returned by 270 students (54%) (Publications 1, 2, 3 and 6).
- Assessment Evaluation Questionnaire (AEQ) (Brown et al., 2003) sent to 73 S151 students in October 2003 and 150 S151 students in October 2004, returned by 33 (45%) and 69 (46%) students respectively, with follow-up emails/telephone calls to 50 students (Publication 2).
- Online questionnaire sent to 400 students on each of the first (2008B) and second (2008J) presentations of S104, returned by 87 (21%) and 61 (16%) students respectively (Publication 6).
- Online questionnaire sent to 400 SXR103 students in 2009, returned by 129 students (32%). SXR103 was a residential school module and was selected because students could be assumed to have studied a range of other modules, with different assessment strategies. Eight students who made interesting points in their responses were interviewed by telephone in late 2009 (Publication 6).
- Observation of six students in the Institute for Educational Technology (IET) usability laboratory in June 2007 (Publications 4, 6 and 8).