Chapter 6

Computer-marked assessment and concept inventories

Abstract

After discussing the use of computers in assessment more generally, this chapter concentrates on online computer-marked assessment, in which responses are automatically marked and feedback is instantaneously provided to students. Concept inventories, designed to assess students' conceptual understanding, can make use of the same technologies. I consider the advantages and disadvantages that computer-marked assessment can offer and ways in which its quality can be improved by thorough assessment design, choice of appropriate question types, and careful question writing. I conclude that computer-marked assessment is not a panacea, but nevertheless has much to offer to teachers and learners in large STEM classes.

6.1 Introduction

When considering the needs of large classes and their teachers, the use of a computer to mark and deliver feedback to both students and their teachers is an immediately attractive proposition. However, this use of technology requires particularly careful implementation. Chapter 5 illustrated the potential that assessment has to bring learning benefits, but also pointed out some of the risks. There is a danger that even well-meaning assessment and feedback interventions may create a barrier to learning rather than enabling it (Bangert-Drowns *et al* 1991). These risks are greater when the assessment is delivered remotely or marked by electronic means, especially when there is no human intermediary. As Ridgway *et al* (2004, p. 7) comment, "when we consider the introduction of e-assessment we should be aware that we are dealing with a very sharp sword".

In this chapter I will explore the advantages and disadvantages of computer-marked assessment and concept inventories. I will discuss ways in which the advantages can be reinforced, and the disadvantages ameliorated, by careful assessment design, choice of appropriate question types and careful question writing. Throughout, I will summarise some of the underpinning theory, while also offering practical hints that derive from my own experience.

6.2 Definitions and history

To avoid any confusion later in the chapter, in this section I will define what I mean by "computer-marked assessment", "concept inventory", and some related terms. I will also give a brief history of work in this area.

Since the early years of the 21st Century, the term "e-assessment" (electronic assessment) has been used to include any use of a computer as part of any assessment-related activity (JISC 2006). Terms such as "digital assessment", "computer-based assessment" and "technology-enhanced assessment" are similarly broad, though each brings a slightly different emphasis. These terms encompass, among many other things, the use of e-portfolios, the delivery of audio or video feedback, and the online delivery of written assessments to a teacher for marking, and their later return to students. In recent times, much attention has been given to remote online examinations and the detection of plagiarism and contract cheating (Dawson *et al* 2020).

More specifically, computer-marked assessment refers to situations in which students' responses are automatically marked and, in some cases, feedback is automatically generated. The earliest

computer-marked multiple-choice questions were probably E.L. Thorndike's Alpha and Beta tests used to assess recruits for service in the US Army during the First World War; during the 20th Century multiple-choice questions also gained in popularity as an educational tool. By the 21st Century the focus became online computer-marked assessment, which enables instantaneous interaction between a student and the system on which the online assessment sits. At the same time, the number of providers of computer-marking assessment systems has grown and the range of question types has extended way beyond multiple-choice questions. For a more detailed historical approach, see Jordan (2013).

A particular use of computers in assessment, common in STEM, is in concept inventories. A concept inventory is an instrument designed to assess students' conceptual understanding, usually with the aim of measuring the learning gain that has occurred across a class as a result of a particular piece of teaching (Sands *et al* 2018). Following the practice established in the Force Concept Inventory (FCI) (Hestenes *et al* 1992), believed to have been the first instrument of this type, most current concept inventories consist of a series of multiple-choice questions, each with one correct answer and a number of incorrect answers, known as distractors, based on common student misconceptions. In order to measure learning gain, they are presented to students as a "pre-test" before the pedagogical intervention and then repeated as a "post-test". Although concept inventories are still sometimes run as paper-based exercises, they are increasingly also run online, which makes them quick and easy to administer, even when considering large class sizes. Students are not usually given direct feedback on their answers to concept inventories, but instead feedback is provided to the teacher, thus fulfilling one of the functions that has been identified if assessment is to support learning (Nicol and Macfarlane-Dick 2006). It is therefore logical to consider concept inventories in this chapter alongside other uses of computer-marked assessment in supporting effective teaching in large class sizes.

6.3 Advantages and disadvantages of computer-marked assessment

6.3.1 Why use computer-marked assessment?

Computer-marked assessment brings many affordances. It has the potential to mark and deliver feedback on students' work, while saving academic time and therefore money. Indeed, Boitshwarelo *et al* (2017) see the recent growth in the use of online quizzes as an inevitable corollary of the dual drivers of reduced resources for teaching and growth in student numbers. Online quizzes can be re-used from year to year at minimal cost. Furthermore, provided a student has access to an appropriate device and the internet, computer-marked assessment can be completed from any location, something which became highly relevant during the Covid-19 Pandemic.

The phrase "objective questions", used historically to describe multiple-choice questions, reflects the fact that the early use of computer-marked assessment came from a desire to make assessment more objective. Ashburn (1938) noted a variation in the grading of essays by different markers, and issues with the reliability of human grading of essays (Brown 2010) and short-answer questions (Butcher and Jordan 2010) remain a persistent concern. Human markers are inherently inconsistent (Bloxham *et al* 2016) and they can be influenced by their expectations of individual students. Even for more open-ended questions, computerised marking brings objectivity and a consistency that can never be assured between human markers (inter-rater reliability) or, over time, for the same human marker (intra-rater reliability).

Computer-marked assessment also has the potential to enhance students' learning. Computer-generated feedback can be provided tirelessly and instantaneously, without the delay inevitably caused by a human marker, and students can be given the opportunity to immediately repeat the task or to perform a similar one and so to learn from the feedback provided. Thus, the feedback is received by students "while it still matters to them and in time for them to pay attention to further learning or to receive further assistance", fulfilling one of Gibbs and Simpson's conditions under which assessment supports learning (Gibbs and Simpson 2005, p. 18).

Tests can be offered to students regularly and, in formative use, even the simplest of quizzes can enable students to repeatedly check their own understanding, encouraging self-regulated learning (Nicol and Macfarlane-Dick 2006). Regular computer-marked assessment can also help students to pace their study, and many authors speak of its role in engaging students and motivating learning (e.g. Holmes 2015) and building self-efficacy and confidence (Cassady and Grindley 2015). Riegel and Evans (2021) found that students experienced a range of positive emotions such as hope and pride more strongly when responding to an online mathematics quiz rather than a conventional test, and a range of negative emotions such as anxiety, anger and hopelessness were all rated less strongly. One student commented that "online quizzes make me relax and it is enjoyable to me" (Riegel and Evans 2021, p.82). In survey I conducted (Jordan 2011, p.153), 64-68% of students agreed with the statement that answering quiz questions was "fun", and comments such as "It's more like having an online tutorial than doing a test" and "give[s] you confidence that you're heading on the right lines" were received.

Computer-generated feedback is inevitably impersonal and non-judgemental, and many students appreciate being able to make mistakes in private (Miller 2009). Although the feedback may be relatively detailed and targeted to specific errors in the student's answer, the fact that it is not responding to a specific person inevitably means that the focus is on the student's performance rather than on the student themselves, which is seen as a key feature in enabling the student to learn (Gibbs and Simpson 2005).

Modern conceptions of effective assessment feedback see this as a *process* with student sense-making and action to improve at the centre rather than as a *product* delivered to students (Sadler *et al* 2022; Winstone and Carless 2020). The feedback provided on computer-marked assessment is most effective when regarded as information that is provided to students, for them to make sense of and act on themselves. Although the assessment is automatically marked and feedback is automatically generated, the student remains in control of their own learning.

Many of the claims for the positive impact of computer-based assessment on learning rely on student opinion; this is important, but in addition to improving student satisfaction, we want to improve learning itself. Furthermore, even when there is a demonstrable correlation between some claims of improved course outcomes and engagement with computer-marked assessment, it can be difficult to be sure that the link is causal. For example, in the case of a purely formative online quiz, it is likely that the more diligent students will both engage with the quiz and do better in the course overall; that does not mean that engagement with the quiz led to the better overall result. However, there have been a number of more rigorous studies that have demonstrated a positive influence of computer-based assessment (e.g Van Gaal and de Ridder 2013). In addition, research into the so-called "testing effect" shows that the mere act of taking tests leads to an improvement in subsequent performance that is greater than additional study of the material, even when the tests are given without feedback (Roediger and Karpicke 2006; Prisacari 2015).

Most computer-marked assessment systems are also able to provide information to teachers about the performance of individual students and of the class as a whole, both on individual questions and on the assessment overall. At the whole-class level, this enables improvements to be made to the assessments and the feedback provided. It can also inform the teacher's teaching. For example, if a quiz or concept inventory is used with a class before a lecture or other intervention, it can also inform the teacher where they need to concentrate their efforts. If the use of the concept inventory is repeated after the intervention, its effectiveness can be gauged.

In recent years, the growth of data about students and their online behaviour, and the ability to analyse this, has led to the new field of learning analytics, defined by the Society for Learning Analytics Research as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs". As the prevalence of online teaching and learning increases, there is an increased potential to make use of an individual student's digital footprint, including their use of e-assessment of all types, to target advice and teaching interventions to their particular needs. This approach raises some ethical concerns (Kitto and Knight 2019) but has been seen to markedly increase retention and completion rates (de Oliveira *et al* 2021).

6.3.2 Why not?

For all its advantages, concerns have been expressed about increasing reliance on computer-marked assessment (Or and Chapman 2022). In this section I will discuss the potential disadvantages. Perhaps the most significant of these relate to lack of authenticity. In a scathing comment about the over-use of multiple-choice questions in medical education, Mitchell et al (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". Even when more sophisticated question types are used, there is some anxiety that use of computer-marked assessment will encourage a surface approach to learning.

Other criticisms of multiple-choice questions include the fact that students can guess the answer, or work backwards from the distractors. For example, if a question asks students to integrate an expression and offers five possible answers, they do not necessarily have to know how to integrate to work out the correct answer; they could instead differentiate each of the options until they arrive at the original expression (Sangwin 2013). In this case, the question is not actually addressing what it set out to assess.

Even so-called constructed-response question types (questions that require an answer to be entered into the system, in contrast to selected-response questions in which the students pick an option or options from a choice that is made available) usually require the entry of a single answer. Thus, it is the answer that is being marked, not the working, something which is at variance with much assessment in STEM subjects, where students are encouraged to show their working and explain their logic. Various attempts have been made over the years to replicate the way in which humans mark but I am not aware of any that have fully succeeded. The most common approach has been to break a question down into constituent steps, which brings the advantages of scaffolding for less confident students (Dawkins *et al* 2017) but it does not truly replicate the more open-ended task it was seeking to replace. This leads to various potential problems. If the assessment is summative, it is not possible to give credit for partially correct answers. If the focus is formative, it may not be possible to identify the source of a student's difficulty and so to give appropriate feedback; were they floundering at the start of the problem or did they make a careless slip near the end? Finally,

this lack of evidence for the working behind a final answer can, in principle, make it easier for a student to cheat. It is easier for a student to copy a single answer from a fellow student or a "homework" website. Plagiarism is discussed further in Section 6.3.3.

In Section 6.3.1, I discussed the preference of some students to receive feedback from a computer rather than a person. However, the lack of a human intermediary between the assessment and the marking engine also brings disadvantages. While it is possible for a tutor to be available to explain an ambiguous question to a student, in much the same way as they would for any type of assessment, there is not usually a human available to explain an unexpected response from a student to the computer! This can lead to inaccurate marking, even for multiple choice questions. More commonly, the fact that feedback is usually pre-prepared in the expectation of the wrong answers that will be given, rather than being given in response to the actual answer received, means that the feedback may fail to respond in a helpful way to the answer given by a particular student.

6.3.3 Disadvantage or advantage?

In considering the disadvantages of computer-marked assessment, which undoubtedly do exist, it is important not to also blame the use of technology for difficulties that have other causes (Bull and Dyson 2004). Many concerns over plagiarism fall into this category. During the Covid-19 Pandemic, it became necessary for students to complete assessed tasks from their own home rather than in invigilated examination halls, and this led to an increased incidence of plagiarism (Montenegro-Rueda *et al* 2021). However, the fact that many institutions increased their use of computer-marked assessment in response to the need to assess remotely does not mean that the computer-marked assessment was the cause of the rise in plagiarism, but rather the necessarily remote location from which it was being completed. Indeed, different variants of computer-marked questions can often be produced for minimum effort, as discussed later, meaning that it is relatively easy for different students to receive subtly different assessments, limiting opportunities for plagiarism.

Similarly, the use of computer-based assessment, in common with any educational intervention which relies on access to a computer and the internet, raises some concern over equitable accessibility for all students. If students are completing the assessment on their own device and from their own home, then there are indeed legitimate concerns relating to digital poverty and the digital divide. However, the use of a technology which means that students can complete an assessment without travelling can save transport costs and make the assessment more accessible to those with certain disabilities as well as to those who are unable or reluctant to travel for other reasons, for example caring responsibilities. In addition, most computer-marked assessment systems include features designed to increase accessibility for those with eyesight problems or dyslexia etc. Accessibility is discussed further in Section 6.4.4.

The feedback from students that has given me the most pleasure, relating to the use of computer-marked assessment in the Open University's STEM Faculty, is that which describes positive enjoyment in completing an assessment (even when with a high-stakes summative function) and that which talks about the assessment feeling as if there was a human tutor there to guide their learning. However, the feedback we receive is not uniformly positive. Careful analysis of the negative feedback received, in the light of the questions and feedback it relates to, has led me to conclude that student complaints usually originate as a result of a particular question, most commonly because the wording is ambiguous or the student has been given misleading feedback (Jordan 2011). My own experience of online quizzes in everyday life strengthens my view that whether computer-marked assessment is a "good thing" or a "bad thing", and the extent of its effectiveness, depend on

the details of its operation, and context, that can all too easily by overlooked. Sections 6.4-6.6 explore these points more fully.

6.4 Assessment design and integration with teaching

In seeking to develop high-quality computer-marked assessment, it is important to start by thinking about why you are taking this approach in the first place, and how it links to your teaching and other assessment. It is also important to think about any limitations imposed and affordances offered by the particular system that you are using, and to remember that the computer-marked assessment needs to be accessible to all students. In this section I will consider these points in turn, in each case building on my own experience.

6.4.1 Motivation and purpose

I used computer-marked assessment for the first time in the Open University (OU) course *Maths for Science* that ran from 2002 to 2018. The course was the first in the University to use online computer-based assessment of the type described in this chapter, though the colleagues on whose shoulders we built had previously sent questions out to our distance-learning students by CD-ROM and DVD, and for many years we had been using multiple-choice questions which students answered on a prepared computer-readable form. In making the decision to move to online computer-based assessment, we had to assume that students had access to a computer and the internet, at home, work or via an "internet café" or library. We also relied on the University's systems capability to record student attempts at the computer-marked assessment and to transfer information about performance to the student's record. Our decision to go ahead thus assumed a certain level of technological readiness, but the primary reason for the move was pedagogical not technological; I had a very strong wish to provide instantaneous and targeted feedback to our students on their mathematical ability, with an opportunity to repeat the question, and the move to online computer-marked assessment enabled this. Similarly, although students were required to achieve a certain overall mark in order to pass the module, my motivation for introducing computer-marked assessment was for its formative not its summative function.

More generally, before deciding to use any particular type of assessment, it is important to reflect on what your purpose is: summative ("assessment of learning") or formative ("assessment for learning")? Or is your aim to find out what your students already know, or to ascertain their conceptual understanding so as to investigate the effectiveness a teaching intervention? It is also important, though surprisingly uncommon, to think about the learning outcomes that you are seeking to assess. As a general rule, learning outcomes at the lower end of Bloom's taxonomy (Bloom et al, 1956) such as those related to recall are perhaps more easily assessed by simple computer-marked question than higher-order learning outcomes.

6.4.2 The context in which the assessment is used

Individual teachers may or may not be able to influence the overall assessment strategy of a course or a qualification. However, apparently small differences in assessment strategy can make a very large difference to the effectiveness of individual components, including computer-marked assessment. An example of this was the introduction of a formative thresholded assessment strategy across the Open University's Science faculty from early 2010s. This strategy required students to reach a modest threshold on the course's continuous assessment, but their continuous assessment score did not contribute to their overall course outcome. The detail of and rationale for this strategy is not relevant to this chapter, but the key point is that two separate models were used, one of

which simply required students to reach a threshold of just 30% in, say, 5 out of 7 of the interactive computer-marked assignments (iCMAs). In addition to being allowed three tries with increasing feedback, students could also repeat questions or the whole iCMA as many times as they wanted to, with multiple variants available. This model was found to be very successful, with many students repeating the questions in different variants and, apparently in consequence, also doing better on the overall course outcomes (Jordan and Bolton 2023).

In addition to its direct use in formative and summative assessment, computer-marked assessment has a role to play in many of the effective teaching techniques mentioned throughout this book, for example in Peer Instruction (introduced in Chapter 2) and Spaced Testing (introduced in Chapter 3). The function, use and effectiveness are different depending on context. This serves as a useful reminder of the fact that technology is merely a tool that can be useful in supporting teaching, learning and assessment when appropriate. Similarly, any particular type of assessment is simply a tool in the hands of the teacher.

Thus it is that concept inventories, while potentially using the same technology as other types of computer-marked assessment, have a different function, namely to provide feedback to the teacher about their students' understanding at a specific point in the course and so (usually) to measure learning gain. For this reason, and in contrast to most uses of computer-marked assessment, common practice is for no direct feedback on their performance on a concept inventory to be given to students, partly because of concerns about widespread circulation of questions and their answers.

Another way in which computer-marked assessment can be used in effective STEM teaching is by requiring students to author the questions, for example using the Peerwise system, which enables students to create questions for their fellow-students to answer. Students may also be encouraged to provide feedback to question authors, and to engage in discussion about the questions with their peers. A significant positive correlation has been found between engagement with PeerWise and overall attainment on a range of STEM courses, even after controlling for ability prior to the course (Kay *et al* 2020).

6.4.3 How does the assessment run?

When writing computer-marked assessment, it can be tempting to think about question type but not to give much thought to the way in which the questions run within the overall quiz. The growth of research into the impact of gamification in education (Dichev and Dicheva 2017) reminds us of the need to think more broadly.

There is some variation between the functionality of different quiz engines and virtual learning environments, but most now allow a particular instance of a question to be attempted several times, with retry allowed after receipt of a feedback hint. In some systems the feedback can be varied depending both on how many attempts the student has had and on the answer they give. I most commonly use questions that operate in the manner illustrated in Figure 6.1. After their first incorrect try (shown in the top left image), students are simply told that their answer is incorrect, to give them an attempt to find and rectify their error for themselves. After a second incorrect try, more detailed feedback is given, where possible targeted to the student misunderstanding. After the third try, whether the answer given is correct or incorrect, a full answer is given. If a numerical score is required, either for summative use or for the purpose of generating feedback on the quiz overall, a decreasing score can be awarded depending on whether the answer is correct, partially correct or incorrect at each of the three tries.

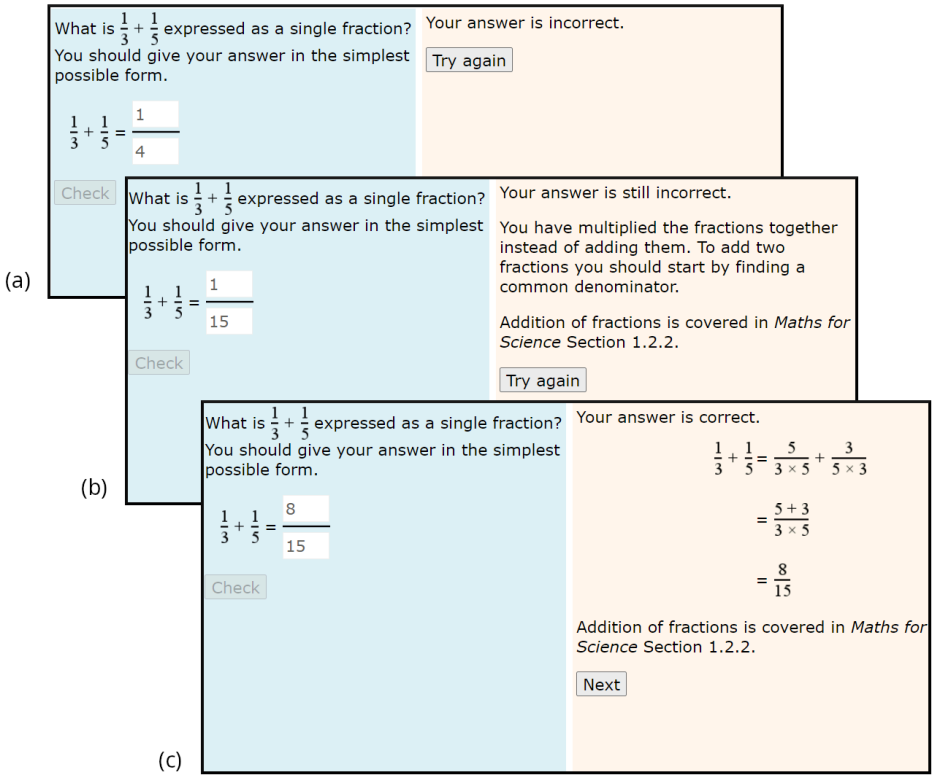

Figure 6.1 A simple question, showing three tries at a question with increasing feedback. This question was written in the OU's OpenMark system, whose functionality informed the development of the Moodle Quiz Engine.

If allowed by the assessment system and the overall assessment strategy, it may also be possible to repeat the whole question, hopefully in a different variant. So, in the simple example shown in Figure 6.1, if a student chose to repeat the whole question, they might be asked to find $1/4 + 1/3$. When the focus is formative, some authors will write variants of questions that differ by a greater amount than simply changing the numbers, to provide greater variety. Alternatively, it may be possible to select different questions from a question bank. In summative use, different variants of questions can be used as an anti-plagiarism device, moving towards a situation in which each student receives a different assessment. However, in this case, it is important to ensure that the different variants used assess the same learning outcome and are of similar difficulty. Up to a point, this can be achieved by careful review of the variants, for example recognising that calculations involving very small numbers (with negative powers of 10 when expressed in scientific notation) tend to be more difficult than those involving very large numbers. Sometimes, variants that are more difficult than others (perhaps because there is some additional skill being assessed) are not spotted until the performance analysis that should be done after the assessment has run. This is discussed further in Section 6.7.3.

I have never imposed tight time constraints on quizzes that I have authored, partly because of the context in which I operate: OU students are frequently studying part-time alongside other responsibilities and they are usually studying from home, so interruptions can be difficult to avoid. However, my reluctance to impose a strict time limit goes beyond my own context; the pressure caused by an awareness that time is running out can impair a student's ability to complete the quiz to the best of their ability and can act as a barrier to learning. I have however, imposed hard cut-off dates on quizzes (days, weeks or months from when the quiz was made available), to encourage student pacing through the course. The consideration of time limits and deadlines is yet another

matter that depends on the context in which you operate, the purpose of the assessment, and the learning outcomes being assessed. You may, for example, be explicitly aiming to assess a student's ability to work under pressure.

As well as generating feedback on a student's answer to each question, it is usually possible to generate appropriate feedback based on a group of questions assessing the same learning outcome, or on the whole quiz.

6.4.4 Accessibility

As mentioned in Section 6.3.3, the use of computer-marked assessment brings some advantages with regards to accessibility, but there are also some issues that require consideration. Early computer-marked assessment systems relied on internal systems to enable the magnification of text, to alter the text or background colour (which some students with dyslexia find helpful) or to produce a plain text version suitable for feeding into a screen reader. Modern web-based systems more commonly make use of the accessibility systems in the browser that the student is using, which brings alignment with other online tools, but also requires question developers to be more aware of the wider provision available.

Some question types are more difficult to use than others for students with limited dexterity. For example, "drag and drop" questions, which require students to drag an option into place, require relatively fine motor skills, including the use of a mouse or touchpad. However, from a functional if not an aesthetic perspective, a drop-down list of options can be provided as a substitute for the draggable options, and this list can be navigated by keyboard functions or read by a screen-reader.

As for any sort of teaching or assessment resource, figure and graph descriptions should be provided for the use of those who are blind or partially sighted, or who benefit from spoken versions of the questions for any reason. However, it is worth noting that the use of a figure or graph description may result in a change to what the question is assessing. For those who do not need to use a screen-reader but whose eyesight is sub-optimal, generous tolerances should be placed on the range of acceptable answers to any questions that require students to read values from a graph etc.

At the OU, there remain a small number of students for whom access to online resources is problematic. This group includes a few students with particular disabilities, e.g. epilepsy, students studying in some prisons and other secure institutions, where access to the internet is not allowed, and - rarely – students who are unable to access the internet because of an unexpected technical problem. We make alternative versions of the questions available in these circumstances, while recognising that the assessment alters as a result, and the students do not benefit from the instantaneous feedback. Designing assessments to be useable from mobile phones and tablets enables wider accessibility.

Taking a broad definition of "accessibility", I would also emphasise the importance of minimising the use of extraneous contextual information in questions. There is a popular belief that this adds interest, which may be the case for a small number of students. However, it more often confuses students, for example if the context is a sport which is unfamiliar to those from different cultural backgrounds.

6.5 Question types

Most computer-marked assessment systems offer a range of question types. In this section I will consider the most common and introduce some less common question types which have particular

potential for assessing large STEM classes. Some of the question types introduced here will be discussed in more detail in the case study chapters later in this book.

6.5.1 Selected-response questions

Selected-response questions, defined as those in which a student selects from pre-defined options, are most commonly multiple choice or multiple-response (in which students are required to select more than one option) but this category also includes true/false questions, questions which require students to match one statement to another, and drag and drop questions. These question types, especially multiple-choice, are generally considered easier and faster to write than constructed-response questions. Care must still be taken in question writing, and the criticisms of lack of authenticity, being able to work backwards, and being able to arrive at the answer by guesswork generally apply to all selected-response questions. However, it becomes more difficult to arrive at the correct answer by guesswork if students are required to select several options.

Other ways of discouraging guesswork include requiring students to explain their answer in a free-text box. This answer is not necessarily marked, but it is available to the teacher should they wish to check that a student understands why an answer is correct. Similarly, students can be asked to upload a file containing their working. Another way of discouraging guesswork is so-called confidence-based (or certainty-based) marking, in which students are required to rate their confidence as well as to give an answer. A correct but unconfident answer receives a lower score than a correct confident answer, whereas an incorrect confident answer is more heavily penalised than an incorrect unconfident one (Gardner-Medwin 2019).

For all the criticisms of them, even simple selected-response questions can lead to “moments of contingency”, formative interactions that can improve cognition (Black and Wiliam 2009). This enables “catalytic assessment”, the use of simple questions to trigger deep learning (Draper 2009). In addition, there are some situations in which a selected-response question is the most suitable type to use. A carefully worded selected-response question, such as the one shown in Figure 6.2, can require a certain amount of logical reasoning and thus assess learning outcomes of higher order than simple recall questions.

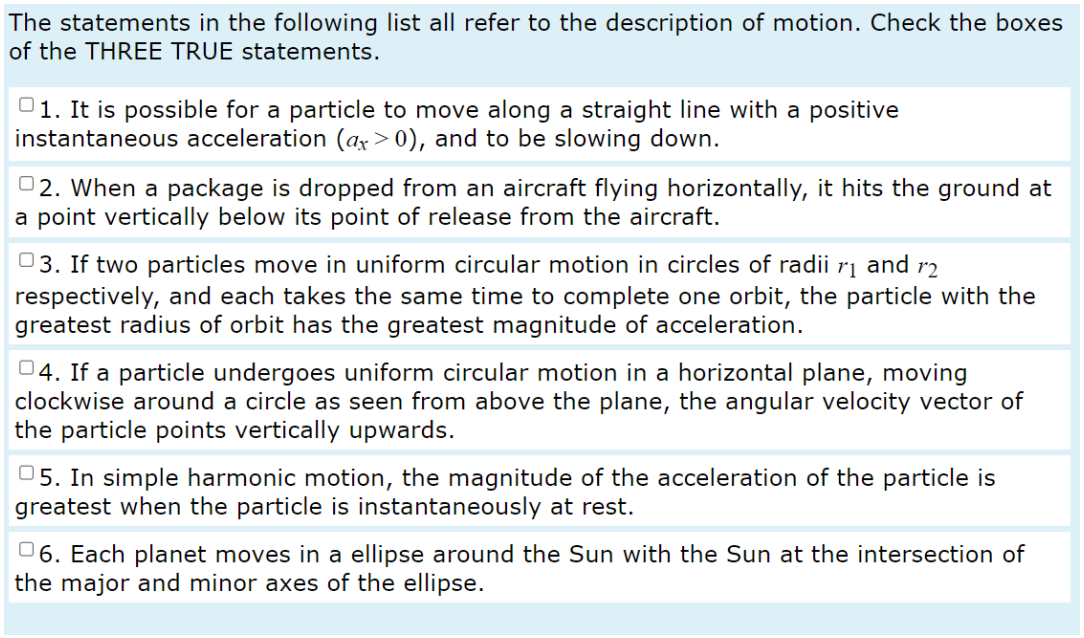

Figure 6.2. A multiple-response question

6.5.2 Simple constructed-response questions

The answer-matching required for a question in which the student enters a simple numerical answer should be straightforward, rendering it unnecessary to ask this sort of question in multiple-choice form. It becomes more complicated if you require the answer to be in scientific notation, or to a particular precision. However, many modern quiz engines and virtual learning environments now include specific functionality to enable such questions to be automatically marked.

Similarly, questions that can be answered in the form of one or a small number of letters and other symbols e.g. “Give the standard abbreviation for the SI unit of mass” can be automatically marked by relatively straightforward means such as string matching. However, immediate thought must be given to whether the case of the letters and order in which they are written is significant: in the example given, kg is a correct answer but KG and gk are not, though I would give targeted feedback for KG. If the correct answer is an algebraic expression e.g. “vt”, then “tv” is also correct. Some systems have specific functionality to check for the correct units alongside numerical values.

6.5.3 Questions based on computer-algebra

The introduction of assessment underpinned by computer-algebra systems (CAS) has revolutionised the computer-marked assessment of mathematics and subjects like physics and engineering. The work described in Chapter 9 is based on the use of one of the leading and well-established systems, STACK, which is underpinned by Maxima. The system enables students to enter an answer as a mathematical expression, which STACK then asks the student to verify as being the answer they want to submit, before it is marked. The underlying CAS removes any anxiety that alternative correct answers will be missed, which is always a concern when relying on string matching, while good CAS-based systems still leave decisions about what to mark as correct with the question author, for instance deciding whether an unfactorized answer would be considered correct. STACK goes further in also enabling checks for units, precision and the steps used in a calculation (Sangwin and Harjula 2017).

6.5.4 Assessing words, phrases and essays

Many computer-marked assessment systems now include a question type in which students can type their answer as a single word, though the quality of these systems is somewhat variable. The best allow misspelling, if and only if the question author wants this, which allows the assessment of accurate spelling where this is important, while also not unfairly penalising students who misspell common English words, which may be as a result of dyslexia, English not being the student’s first language, or a slip which is not relevant to the learning outcome being assessed. Where the correct answer is not a specific technical term, it is also important that the system allows synonyms.

The question shown in Figure 6.3 is one that I wrote as part of a project that looked at the automatic marking of free-text answers of phrases and sentences, usually up to 20 words in length. We investigated the use of two contrasting technologies, one making use of artificial intelligence (AI) and one using a “bag of words” approach i.e. looking for words (strings of characters) while also considering negation and word order. This meant that, in response to another question, answers such as “The forces are balanced” and “There are no unbalanced forces” could be marked as correct, while “The forces are unbalanced” could be marked as incorrect. Somewhat to our surprise, my colleagues and I found that the relatively simple answer matching was at least as accurate as both human markers and the more sophisticated system, provided that the answer matching was based on human-marked responses from actual students on a similar course (Butcher and Jordan 2010).

Despite this finding, subsequent developments in AI and machine learning have suggested interesting avenues for future development (Süzen *et al* 2020).

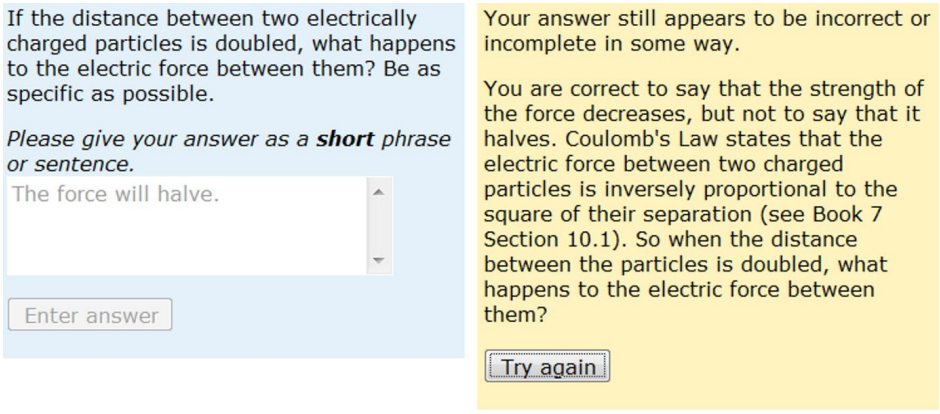

Figure 6.3 An automatically marked question requiring a free-text answer of a few words. This question was written in the PMatch question type, the precursor to Moodle's Pattern Match.

My colleagues and I have used the same technology that underpins the question shown in Figure 6.3, which is available as the Pattern Match question type within Moodle, to develop a version of the FCI in which some questions are replaced by automatically marked short-answer free-text questions (Parker *et al* 2023). Development work is ongoing, but the tool is approaching sufficient reliability and we hope to use it to gain deeper understanding into conceptual understanding and to investigate some of the known demographic differences in outcome as measured by the conventional FCI.

It is generally considered to be technically easier to obtain accurate automatic marking for essays than it is for short-answer questions. If content is marked at all (which not all essay-marking systems do), simply looking for keywords is often sufficient. Details such as word-order and negation are generally found to be less important than is the case for short-answer questions. When essays are marked for style, it is usual to make use of proxies such as sentence and paragraph structure (Shermis and Burstein 2013).

6.5.5 More advanced question types

As technologies develop, so too does the potential for increasingly sophisticated computer-marked assessment questions, for example those assessing mathematical proof. In addition, the growing understanding of the importance of authenticity in assessment has been rewarded by systems such of Coderunner (Lobb and Harlow 2016), which rather than assessing students' understanding of programming by asking questions about it, asks them to write a simple program, which is evaluated according to whether it works as required.

6.6 Writing questions and feedback

After the assessment has been designed and appropriate question types selected, it still remains to actually write the questions. In this section I offer some tips for question authors, emphasising the importance of checking your questions and evaluating their performance.

6.6.1 Writing questions

Many of the problems that students experience with computer-marked questions stem from question wording that is in some sense unclear, ambiguous or requires good understanding of a

particular culture or language. In addition to avoiding unnecessary contextual information, where possible, I recommend avoiding the use of double negatives. Questions which the author may consider to be “clever” can all too easily end up assessing an ability to understand the question rather than knowledge or understanding of the course.

Figure 7.4 shows two fictional questions, both deliberately written in “nonsense language” but illustrating points that I have seen in all too many real questions. Despite the fact that the question has no meaning, it is clear that the correct answer to Question 1 (Figure 6.4a) is Option B, because of the length of the explanation provided relative to the lack of explanation in the other options. To find the correct answer to Question 2 (Figure 6.4b) requires an understanding of the English language that should be familiar to those who are native speakers, but maybe not to others; the correct answer is Option C, as this completes the sentence “The bfeld links to the mnoge by means of a tanag” in a grammatically correct way. All the other options require the final word of the question stem to be “an” not “a”; ““The bfeld links to the mnoge by means of a elland” is not grammatically correct.

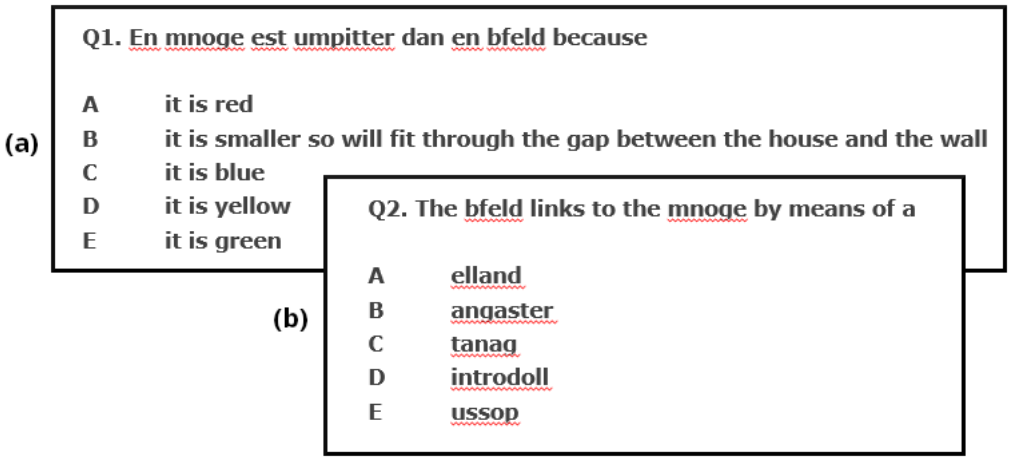

Figure 6.4 Two fictional multiple-choice questions

The examples given in Figure 6.4 may seem trivial, but a colleague tells the story of being able to achieve 65% in a multiple-choice assessment despite knowing effectively nothing about the subject.

Checking your own questions should reveal many issues, like these, but it can be particularly difficult for any of us to spot our own mistakes, so I would always advocate checking and rechecking, but also asking a colleague to check your questions.

6.6.2 Distractors, correct and incorrect answers and feedback

When writing multiple-choice questions, the distractors should be plausible answers, preferably based on common misconceptions and mistakes. For constructed-response questions, there is also a need for consideration of both correct and incorrect responses that students are likely to give (or, even better, that students have been observed to give). This enables the question author to ensure that all correct answers are marked as such, and that appropriate feedback can be given.

In much the same way as for the wording of the question itself, it is important that feedback is clear and understood by the student. If they are to learn from it, it is also important that, whenever possible, the feedback makes sense to the student in the context of the answer they have given.

Perhaps the largest single source of student frustration is when they are told that an answer is incorrect, but the feedback is too general and does not relate to the student's error. This is particularly irritating to students when their error is minor, or perceived to be. This is exemplified by the following student feedback, received in a survey I conducted:

"I had a go at practice quiz one and when I got to question 2 I got the answer wrong. I spent over an hour going over it and trying to work out where I was going wrong to no avail. Eventually I had to give up, only to discover that my answer was the same as the quiz had except I had expressed my answers to one significant figure more. I was convinced I had lacked understanding of the concept, I was very frustrated and demoralized by this."

6.6.3 Evaluation and iterative design

In addition to asking a colleague to check your questions before use, it is important to monitor them in use, to detect any serious issues and (in summative use) to check that variants of questions are of equivalent difficulty. Various statistical techniques are available to help with this and most computer-marked assessment systems provide basic management information on student performance on different questions and variants. If you are fortunate enough to be able to re-use a question from year to year, you will be able to improve your questions in the light of observed student performance. Figures 6.5 and 6.6 illustrate the effectiveness of one such modification to a question I wrote. The question shown in Figure 6.5 did not originally give targeted feedback for the common partially correct answer "It was formed in a desert". The answer matching was always acceptable, but simply being told that their answer was incorrect caused much student frustration and, as shown on the left in Figure 6.6, very few students were able to correct their answer between and 1st and 2nd try. The simple addition of the targeted feedback shown in Figure 6.5, resulted in a marked improvement in question performance and also to considerably less student frustration.

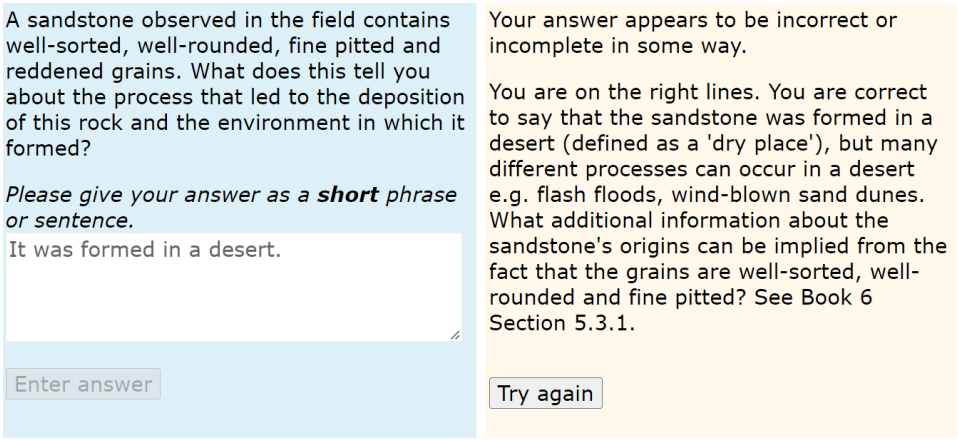

Figure 6.5. A question which has been amended to give targeted feedback on the common partially correct answer shown.

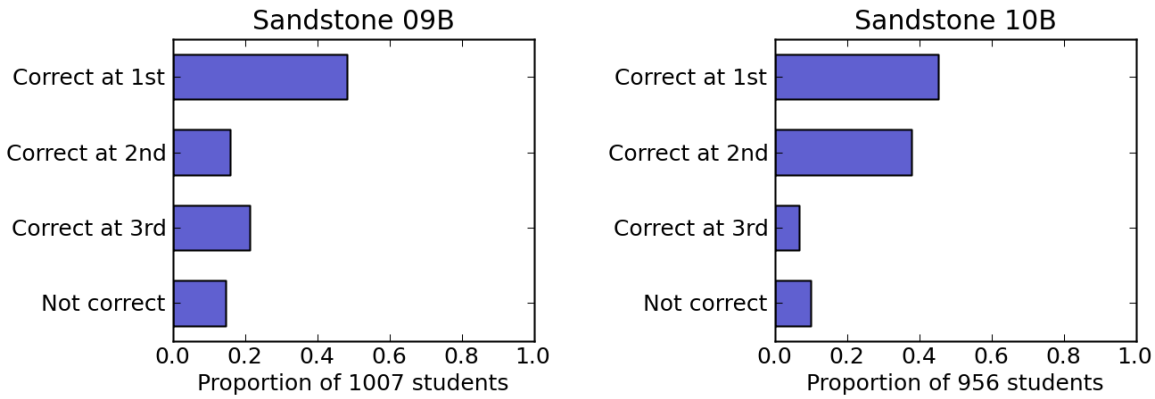

Figure 6.6 The change in question performance as a result of the addition of targeted feedback, between one year (labelled 09B) and the next (10B).

Analysis of student responses to computer-marked assessment questions can be a rich source of information about student understanding of the topics being taught. Furthermore, analysis of engagement with the system more generally can provide useful information about student engagement with the course as a whole.

6.7 How far should we go?

In this chapter I have outlined the development of computer-marked assessment, pointing towards an increasingly sophisticated future in which we can effectively assess, for example, essays and proof. I have also observed a growing interest in the use of AI to assess students' actual engagement with online teaching activities, rather than considering assessment as a separate event. At one level, this development excites me. However, just because we can do something, it does not mean that we should.

Perelman (2008) famously tricked an essay-marking system by using the proxies that the system was looking for, while demonstrating no understanding of the subject. Particularly in STEM, the subject content is important. Furthermore, essays are intended as a means of communication between two people and therefore I consider that, although automatic systems might support students in developing the relevant skills, a human should be involved in the marking of the final piece of work.

I join the call for variety in assessment (Main 2022, Section 6.3.4). Variety supports student diversity and enables appropriate methods to be used for the assessment of different learning outcomes. Computer-marked assessment can motivate and build the confidence of students and provide them with information about their learning. At the same time, it has tremendous potential to free human markers from the drudgery of marking relatively straightforward questions, something that is particularly significant when class sizes are large Teacher time is then freed to help students to interpret the information that the computer has provided, and to deliver effective teaching and the types of assessment that only humans have the skills to do.

References

Ashburn R 1938 An experiment in the essay-type question. *J. Experiment. Educ.* **7** 1-3

Bangert-Drowns R L, Kulik C L C, Kulik J A and Morgan M 1991 The instructional effect of feedback in test-like events *Rev. Educ. Res.* **61** 213-238

Black P and Wiliam D 2009 Developing the theory of formative assessment *Educ. Assess. Eval. Acc.*. **21** 5-31

Bloom B S, Engelhart M D, Furst E J, Hill W H and Krathwohl DR 1956 *Taxonomy of Educational Objectives: The Classification of Educational Goals* (New York: McKay)

Bloxham S, den-Outer B, Hudson J and Price M 2016 Let's stop the pretence of consistent marking *Assess. Eval. High. Educ*. **41** 466-481

Boitshwarelo B, Reedy A K and Billany T 2017 Envisioning the use of online tests in assessing twenty-first century learning: a literature review *Res. Pract. Tech. Enhanc. Learn*. **12** 1-16

Bridgeman B 1992 A comparison of quantitative questions in open-ended and multiple-choice formats *J. Educ. Measure.* **29** 253-271

Brown G 2010 The validity of examination essays in higher education *High. Educ. Quart*. **64** 276-291

Bull J and Dyson M 2004 *Computer-Aided Assessment* (York: LTSN Generic Centre)

Butcher P and Jordan S 2010 A comparison of human and computer marking of short free-text student responses *Comp. Educ.* **55** 489-499

Cassady J and Grindley B 2005 The effects of online formative and summative assessment on test anxiety and performance *J. Tech. Learn. Assess.* **4** Article 1

Dawkins H, Hedgeland H and Jordan S 2017 Impact of scaffolding and question structure on the gender gap *Phys. Rev. Phys. Educ. Res.* **13** 020117.

Dawson P, Sutherland-Smith W and Ricksen M 2020 Can software improve marker accuracy at detecting contract cheating? *Assess. Eval. High. Educ*. **45** 473-482

de Oliveira C, Sobral S, Ferreira M and Moreira F 2021 How does learning analytics contribute to prevent students' dropout in higher education *Big Data Cog. Comp.* **5** Article 64

Dichev and Dicheva 2017 Gamifying education *Int. J. Educ. Tech. High. Educ.* **14** Article 9

Draper S 2009 Catalytic assessment *Brit. J. Educ. Tech.* **40** 285-293

Gardner-Medwin A R 2019 *Innovative Assessment in HE: a handbook for Academic Practice* 2nd edn, ed C Bryan and K Clegg (New York: Routledge) Chapter 12 (Certainty-based Marking: stimulating thinking and improving objective tests)

Gibbs G and Simpson C 2005 Conditions under which assessment supports students' learning. *Learn. Teach. High. Educ.* **1** 3-31.

Hestenes D, Wells M and Swackhamer G 1992 Force concept inventory *Phys. Teach.* **30** 141-158

Holmes N 2015 Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module *Assess. Eval. High. Educ*. **40** 1-14

JISC 2006 *e-Assessment Glossary (short)* (Bristol: JISC) Online: https://www.yumpu.com/en/document/read/23670459/e-assessment-glossary-short-version-jisc

Jordan S 2011 Using interactive computer-based assessment to support beginning distance learners of science *Open Learn.* **26** 147-164

Jordan S 2013 E-assessment: Past, present and future *New Dir. Teach. Phys. Sci.* **9** 87-106.

Jordan S and Bolton J 2023 Student engagement with a novel online assessment strategy *Int. J. Assess. Tools Educ.*

Kay A E, Hardy J and Galloway, R K 2020 Student use of PeerWise: A multi-institutional, multidisciplinary evaluation *Brit. J. Educ. Tech*. **51** 23-35

Kitto K and Knight S 2019 Practical ethics for building learning analytics *Brit. J. Educ. Tech.* **50** 2855-2870.

Lobb R and Harlow J 2016 Coderunner: A tool for assessing computer programming skills *ACM Inroads* **7** 47-51

Main P 2022 *Assessment in University Physics Education* (Bristol: IOP Publishing)

Miller T 2009 Formative computer-based assessment in higher education *Assess. Eval. High. Educ.* **34** 181-192.

Mitchell T, Aldridge N, Williamson W and Broomhead P 2003 Computer based testing of medical knowledge *7th Int. Comp. Assist. Assess. Conf.*

Montenegro-Rueda M, Luque-de la Rosa A, Sarasola Sánchez-Serrano J L and Fernández-Cerero J 2021 Assessment in higher education during the COVID-19 pandemic: A systematic review *Sustainability* **13** 10509

Nicol D and Macfarlane-Dick 2006 Formative assessment and self-regulated learning *Stud. High. Educ.* **31** 199-218

Or C and Chapman E 2022 Development and acceptance of online assessment in higher education *J. App. Learn. Teach.* **5** 10-26

Parker M, Hedgeland H, Jordan S and Braithwaite N 2023 Establishing a physics concept inventory using computer marked free-response questions *Eur. J. Sci. Math. Educ*. **11** 360-375

Perelman L 2008 Information illiteracy and mass market writing assessments *College Composition Comm.* **60** 128-141

Prisacari A A 2015 The testing effect in general chemistry *MSci Thesis* Iowa State University

Riegel K and Evans T 2021 Student achievement emotions: Examining the role of frequent online assessment *Aust. J. Educ. Tech.* **37** 75-87

Ridgway J, McCusker S and Pead D 2004 *Literature Review of E-assessment* (Bristol: Futurelab)

Roediger III H L and Karpicke J D 2006 The power of testing memory *Persp. Psych. Sci* **1** 181-210.

Sadler I, Reimann N and Sambell K 2022 Feedforward practices: a systematic review of the literature *Assess. Eval. High. Educ.* 1-16 Online: https://doi.org/10.1080/02602938.2022.2073434

Sands D, Parker M, Hedgeland H, Jordan S and Galloway R 2018 Using concept inventories to measure understanding *High. Educ. Ped.* **3** 173-182

Sangwin C 2013 *Computer Aided Assessment of Mathematics* (Oxford: Oxford Univ. P.)

Sangwin C and Harjula M 2017 Online assessment of dimensional numerical answers using STACK in science *Eur. J. Phys.* **38** 035701.

Shermis M D and Burstein J 2013 *Handbook of Automated Essay Evaluation* (New York: Routledge)

Süzen N, Gorban A N, Levesley J and Mirkes E M 2020 Automatic short answer grading and feedback using text mining methods. *Proc. Comp. Sci.* **169** 726-743

Van Gaal F and De Ridder A 2013 The impact of assessment tasks on subsequent examination performance *Active Learn. High. Educ.* **14** 213-225

Winstone N and Carless D 2020 *Designing Effective Feedback Processes in Higher Education* (Abingdon: Routledge)