Student Reaction to a Modified Force Concept Inventory: The Impact of Free-Response Questions and Feedback

Journal Item

European Journal of Science and Mathematics Education, 2022, **10**(3), 310-323
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Student Reaction to a Modified Force Concept Inventory: The Impact of Free-Response Questions and Feedback

Mark A. J. Parker [1*]
0000-0002-1984-9624

Holly Hedgeland [2]
0000-0003-3703-7942

Nicholas Braithwaite [1]
0000-0002-1586-3736

Sally Jordan [1]
0000-0003-0770-1443

[1] The Open University, UK
[2] University of Cambridge, UK
[*] Corresponding author: mark.parker@open.ac.uk

Citation: Parker, M. A. J., Hedgeland, H., Braithwaite, N., & Jordan, S. (2022). Student Reaction to a Modified Force Concept Inventory: The Impact of Free-Response Questions and Feedback. *European Journal of Science and Mathematics Education, 10*(3), 310-323. https://doi.org/10.30935/scimath/11882

ARTICLE INFO

Received: 21 Jan 2022
Accepted: 1 Mar 2022

ABSTRACT

The study investigated student reaction to the alternative mechanics survey (AMS), a modified force concept inventory, which used automatically marked free-response questions and offered limited feedback to students after their answers had been submitted. Eight participants were observed in completing the AMS, and they were interviewed to gain insight into what had been observed; the resultant data set was analyzed by thematic analysis. This revealed six key themes: "use of free-response questions supported deep learning", "interpretation of the AMS instructions affected answer length", "the idea of being marked by a computer did not affect answer structure", "participant reaction to the usability of the AMS was mostly positive", "reactions to the AMS depended upon what participants thought it was for", and "limited feedback was a useful addition to the AMS". Participants gave answers of differing length, being guided by the question wording as well as by their own preferences. It was found that participants valued being given feedback on their performance. Participants reacted positively to the free-response questions and could see potential for the use of this question type, opening up possibilities for the use of automatically marked free-response questions in concept inventories in the future.

Keywords: computer-marked assessment, concept inventories, feedback, free-response questions, physics education, student reaction

INTRODUCTION

Concept inventories have been used in science education for decades. The objective of this use varies (Smith & Tanner, 2010), but most are designed with the notion of testing different teaching approaches by measuring the learning gain (Porter et al., 2014). Typically, this is done by having students attempt the concept inventory as a pre-test before instruction on a topic, and then have them attempt the same concept inventory again as a post-test after instruction on the topic. Scores from the pre-test and the post-test are then compared to find the learning gain, and hence gauge the effectiveness of the teaching methods (Bailey et al., 2012).

The first concept inventory was the force concept inventory (FCI) (Hestenes et al., 1992). This contains thirty multiple-choice questions and was designed to test understanding of the physics concept of Newtonian mechanics. The questions have minimal mathematical content and have distractor options based on common student misconceptions. Since its introduction, the FCI has been widely used and the subject of much discussion within the research community (Eaton, 2021; Lasry et al., 2011; Yasuda et al., 2021).

Many other concept inventories, including the brief electricity and magnetism evaluation (Ding et al., 2006), the force and motion conceptual evaluation (Thornton & Sokoloff, 1998), the biology concept inventory (Garvin-Doxas et al., 2007), and the astronomy diagnostics test (Hufnagel, 2002; Zeilik, 2003) make use of multiple-choice questions. Multiple-choice assessments are generally considered to be quicker and easier to administer than their free-response counterparts, and thus to be preferable when assessing large classes (Lee et al., 2021). However, problems arising from an over-reliance on multiple-choice questions have been identified (Nicol, 2007; Zhang & VanLehn, 2021). When answering a multiple-choice question, students are selecting from a list of ideas that were constructed by somebody else, whilst in answering a free-response question, they are required to construct their own answer (Mitchell et al., 2003; Simon & Snowdon, 2014). In the context of concept inventories, this means that free-response questions provide more information about student thinking than their multiple-choice counterparts.

Rebello and Zollman (2004) previously investigated the idea of using free-response questions in a concept inventory. They found that free-response questions provided further insight into student thought processes, making them a viable alternative to multiple-choice questions. However, human marking of free-response questions is time consuming. In order for free-response questions to be a viable alternative, the making process needs to be automated. This is possible with the pattern match question type within the Moodle question engine (Hunt, 2012), which takes an algorithm-based approach to mark responses using a computer. The pattern match question type forms the basis of the current study.

Since it has been shown to be possible to author and automatically mark short free-response answers, it should be possible to develop a concept inventory that makes use of free-response questions instead of the usual multiple-choice questions. Since the FCI has been widely used, and its questions have been validated, it was the logical concept inventory on which to base an investigation. In this work, we outline the early-stage development of a version of the FCI that includes some free-response questions and describe an investigation into student reaction to the modified instrument.

Given that concept inventories are generally designed to measure learning gain and hence to assess teaching methods, they do not usually give feedback to students, though this has been done occasionally with the aim of increasing student self-efficacy (Chen et al., 2004; Lawrie et al., 2013). These studies also investigated student reaction to limited feedback, provided by the modified FCI, after students had submitted their answers.

The purpose of the current work was to investigate student reaction to feedback and free-response questions, which may be useful in gaining deeper insight into student learning. The study's two research questions were:

1. RQ1: How do students react to free-response modified versions of FCI questions?
2. RQ2: How do students react to being given feedback on multiple-choice & free-response FCI questions?

METHODS

Background

The study was conducted at a non-traditional university that delivers its teaching at a distance. A free-response version of the complete FCI had previously been given to undergraduate physics students at two other universities, and the responses were marked by hand. These marked responses were used to develop marking rules using pattern match. At this stage it was discovered that, in their current form, some of the questions were unsuitable for use in the free-response format. These questions were mainly those that required students to describe a trajectory, as this is not easy to articulate. It also proved necessary to split some of the questions from the original FCI into two parts. It was decided to retain question wording that was

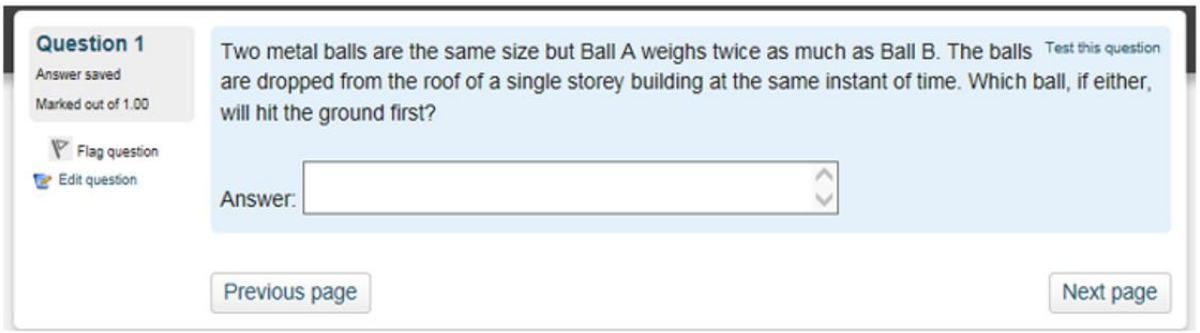

Figure 1. An example of a question on the AMS

as close as possible to the original FCI, respecting the large-scale use and validation of the original FCI questions.

A 33-question instrument was proposed, with 18 free-response questions and 15 multiple-choice questions. These 33 questions were then put together to form an online concept inventory, dubbed the *alternative mechanics survey* (AMS).

Data Collection

The investigation utilized the methodology of usability testing (Barnum, 2010). The study was conducted in a usability laboratory, which features a human computer interaction lab and a live observation room. The human computer interaction lab contains a computer on which participants can trial new software, as well as a webcam and audio link which allows the participants' actions and commentary to be both streamed and recorded. The live observation room allows viewers to watch the participant and their screen in real-time via the webcam and audio link.

As a precursor to the main usability testing study, the AMS was trialed with four subject experts in a think-aloud setting, and it was found that the AMS questions were interpreted in the desired way. After obtaining approval for the work from the relevant ethics committee, eight participants were selected for the study. The opportunity to be involved in the investigation was offered to second year physics students and first-year PhD students at the institution. An email message was sent to potential participants, and this gave details about the logistics of the usability testing, but did not give any details about the background of investigating use of free-response questions in concept inventories. Four undergraduate students and four postgraduate students were selected, being the first eight participants to respond. Participants were not paid for their involvement, but at the end of the work were given an *Amazon* voucher worth £20 as a token of appreciation for their involvement.

Figure 1 shows a typical AMS question. Participants could receive feedback when they had completed all the questions, by clicking the *submit all and finish* button at the bottom of the final screen.

During the study, participants worked through the AMS while being recorded, and were watched remotely from the live observation room. Participants were free to think aloud if they wished, but they were not given explicit instructions to do so, because this can affect the process being observed (Dockter & Mestre, 2014). After completion of the AMS, a semi-structured interview was conducted, which was also recorded. No time limit was placed on the AMS, but participants were told that it would probably take between 30 minutes and 1 hour to complete.

Feedback was given to participants after they had completed the AMS. The feedback provided was limited to knowledge of whether the answer was correct or incorrect, plus the correct answer for multiple-choice questions. In addition, in order to give direction to the semi-structured interviews that followed the usability testing, some verbal feedback was provided to the students on their AMS answers. This prevented the interviews from being derailed by participants wanting to check their answers, which allowed the focus to be placed on a discussion of whether the questions had been interpreted and interacted with in the intended way from a usability testing standpoint. Similar approaches have previously been used to add structure to interviews in the context of evaluating the effectiveness of remote teaching laboratories by Nickerson et al. (2007) and Scanlon et al. (2004).

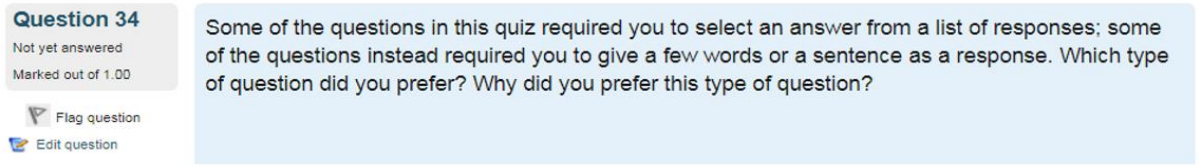

Figure 2. Q34 of the AMS

Table 1. Codes associated with the "use of free-response questions supported deep learning" theme

Code	Number of times coded
Free-response questions required thought (C1)	7
Free-response questions encouraged creativity (C2)	6
Multiple-choice distractor options misleading (C3)	9
Selected-response made multiple-choice easier (C4)	12
Identified benefits of both multiple-choice and free-response questions (C5)	10

There were two components to the data from these trials. The first was the responses given by the eight usability laboratory participants to the interview questions; the second were responses to Q34 of the AMS gathered from the large-scale administration of the AMS. Q34 was a qualitative question added to the end of the AMS, asking how the students found the different question types on the AMS; a screenshot of the question can be found in **Figure 2**. The large-scale administration of the AMS was conducted for the quantitative development and testing of the AMS, and was undertaken by distributing the AMS to high school and university students electronically, with a sample size of approximately 254 (Further detail of the quantitative development is not the focus of the current article).

Data Analysis

Thematic analysis (Braun & Clarke, 2006; Braun et al., 2014) was used to analyze the data. This is an appropriate approach when a small number of participants produce a rich qualitative data set. The aim of thematic analysis is to reduce this data set into an interpretable form, and these are the eponymous themes of the method. Themes emerge from the investigator's assimilation of the data set, which prevents arbitrary conclusions from being drawn. Thematic analysis is a robust and widely used methodology which has been used in a range of contexts, including the analysis of forum posts (Smedley & Coulson, 2017), focus group data (Garcia & Hamilton-Giachritsis, 2017; Varagona & Hold, 2019), blog entries (Castro & Andrews, 2018), open-ended question responses (Grogan & Jayne, 2017), interview data (Robertson et al., 2018), and literature (Filia et al., 2018). The version of thematic analysis used in the current study is outlined by the University of Auckland (2017). Where appropriate, responses to Q34 taken from a wider range of test-takers were used for triangulation.

RESULTS AND DISCUSSION

For the interview data, thematic analysis identified 17 codes, which grouped into 6 themes. These themes were "use of free-response questions supported deep learning", "interpretation of the AMS instructions affected answer length", "the idea of being marked by a computer did not affect answer structure", "participant reaction to the usability of the AMS was mostly positive", "reactions to the AMS depended upon what participants thought it was for", and "limited feedback was a useful addition to the AMS". Each of these are discussed below, and the findings from the Q34 responses are triangulated with the interview data where this is relevant.

Findings Related to the "Use of Free-Response Questions Supported Deep Learning" Theme

There were 5 different codes associated with the "use of free-response questions supported deep learning" theme, and this theme was coded 44 times overall. This theme covers participants' reactions to answering the free-response questions, and their comparisons of these with other question types. The codes associated with the theme are presented in **Table 1**. Note that unless otherwise stated, the occurrence of the codes was more or less equal between the eight participants.

Table 2. Reasons given by the students from the large-scale administration of the AMS for preferring the free-response question type

Reasons for preferring FRQs	Tally	Meaning
Allowed you to think/write own words	27	Refers to student being able to think for themselves, and to write down their own answer to the question.
Tests understanding	8	Refers to students identifying that free-response questions provide a better test of student understanding than multiple-choice counterparts.
MCQ makes them doubt themselves	3	Refers to idea that options given by multiple-choice questions can cause the student to doubt that the answer that they have given is correct.
Quicker to answer	2	Refers to the idea that free-response questions are quicker to answer than their multiple-choice counterparts.

Table 3. Reasons given by the students from the large-scale administration of the AMS for preferring the multiple-choice question type

Reasons for preferring MCQs	Tally	Meaning
Answer unambiguous	14	Refers to student being able to select from a definite list of responses.
Easier/quicker to answer	41	Refers to the multiple-choice questions being easier and quicker than their free-response counterparts.
Made them think	4	Refers to the students finding multiple-choice questions to be challenging to answer.
Provides guidance	38	Refers to the scaffolding provided in the multiple-choice questions to facilitate with giving an answer.
Writing own answers is hard	14	Refers to the student having difficulties when asked to write their own answer to the free-response questions.
Issues with the free-response question interface	11	Refers to the students having difficulties using the interface when answering the free-response questions.

Findings from the Q34 responses

Overall, 229 students who participated in the large-scale administration of the AMS gave responses to Q34. Within these responses, it was found that 44 preferred free-response questions, 150 preferred the multiple-choice questions, and 35 preferred neither question type. The reasons given by the students for these preferences are given in **Table 2** and **Table 3**.

Combined discussion of AMS usability testing and Q34 response findings

From codes C1 and C2, it appears that the usability laboratory participants found that free-response format made them think about the questions, which allowed them to be creative when constructing their own answers. In this way, free-response questions were identified as being useful for students by making them think more deeply. These findings were backed up by the responses to Q34 given by students in the large-scale administration of the AMS, where free-response questions were identified as being useful for finding out about student understanding. Taken together, the above reflections showed that the students were capable of seeing the educational value of the free-response questions on the AMS. This finding illustrates that there is potential demand from students for free-response questions in other concept inventories.

Codes C3 and C4 show that participants recognized that multiple-choice distractor options can lead test-takers to select an incorrect answer, making multiple-choice questions misleading; similar observations have previously been reported by Woodford and Bancroft (2005). In addition, some of the participants admitted to making use of *eliminate and guess* methods (Sangwin, 2013) and other strategic techniques when answering multiple-choice questions. Similar concerns were also raised in the responses to Q34, where multiple-choice questions were found to be confusing by some students; whereas other students thought that multiple-choice questions were too easy because of the guidance that is inherently built-in to them. Crisp (2007) pointed out that these factors make it difficult to draw conclusions about student understanding from multiple-choice questions, and this is one of the main motivations for making use of free-response questions in the AMS.

However, some participants noted as a positive factor that guidance was available in the question wording and answer options of multiple-choice questions, which could help them to answer the question when they were unsure. This contrasts with the above findings, where the same guidance was perceived as a negative aspect of the question type; this split was also present in the Q34 responses. It is of note that students who were less sure of their answers were typically in favor of the guidance offered by multiple-choice questions, whereas students who wanted to challenge themselves by doing the questions tended to be against the guidance offered by multiple-choice questions. This indicates that the students had different perceptions of the multiple-choice questions based on what they wanted to achieve by working through the AMS, and this could also be related to the level of attainment of the students. Further data would be required to investigate this observation in a more general setting.

Somewhat surprisingly, evidence provided in **Table 2** suggests that some students took longer to answer the multiple-choice questions than the free-response questions. It is possible that these students read through all of the possible multiple-choice options, which led them to do more on-screen reading in order to answer these questions; such complaints about computer-based assessment have previously been raised in the work of Nardi and Ranieri (2019). However, it is also possible that these students were reading through each of the options and evaluating the degree to which each option was wrong, rather than looking for the only option that could feasibly be correct. This is the reverse of the *eliminate and guess* strategies outlined previously, and instead illustrates how students could make use of the questions for learning purposes.

In the case of code C5, usability lab participants and responses to the Q34 questions identified that multiple-choice questions were easier to answer, whereas free-response questions make students consider the problems in more detail. Thus, multiple-choice questions were perceived as being good by students, since they could answer these to build confidence; on the other hand, the free-response questions were perceived as being useful for markers, because the answers better reflected students' understandings and misunderstandings. These findings highlight a contrast between what students and educators may perceive as positive aspects of question design. The multiple-choice questions can be perceived as positive by students because it makes their task easier; such a perception may not be shared by educators. Similarly, free-response questions can be perceived as positive by educators because they provide insight about students' conceptual understanding; this perception might not be shared by students. In this way, both free-response and multiple-choice questions have their merits for students and educators, and these need to be considered when designing questions of either type.

Findings Related to "Interpretations of the AMS Instructions Affected Answer Length" and the "The Idea of Being Marked by a Computer Did Not Affect Answer Structure" Themes

There were two codes related to the "Interpretations of the AMS instructions affected answer length" theme: "information given about answer length resulted in shorter answers" (C6) was coded 11 times while "information given about answer length resulted in longer answers" (C7) was coded 14 times. To facilitate the discussion of this theme, the mean answer length in words for each of the participants in the study is shown in **Figure 3**.

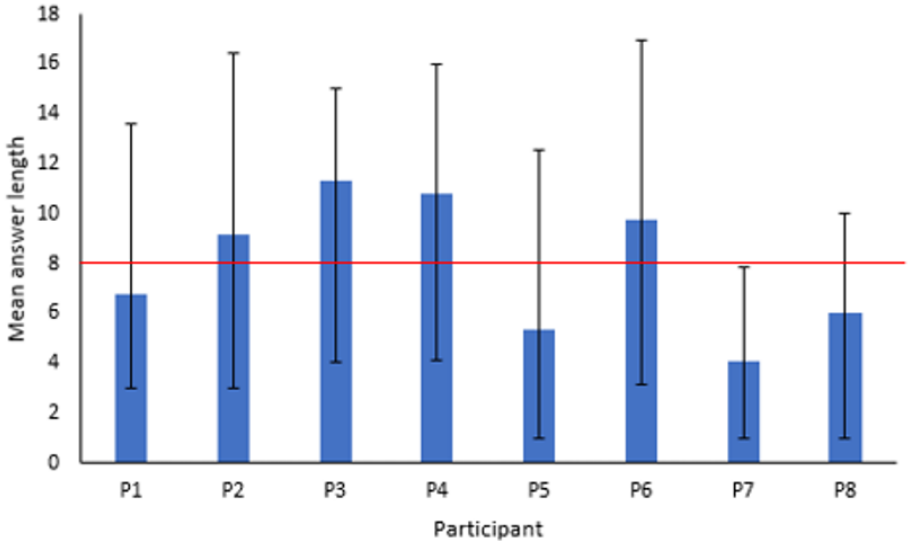

Figure 3. Graph showing the mean answer length in words by participant (blue bars) and the standard deviation adjusted for skew of each participant's answer lengths (error bars). The red horizontal line represents the mean of these mean answer lengths

The length of participants' answers to the free-response questions were found to vary between participants. The data and responses suggested that none of the participants struggled with the length of their answers, but some were consciously aware of it. In addition, there was evidence that the participants used the guidance provided to determine what amount of detail was appropriate to present in their answers. In the cases highlighted by code C6, the participants made their answers shorter, as the guidance had told them that answers of a few words would be sufficient; whilst in the cases identified by code C7, the participants made their answers longer, alluding to the 20-word limit that they had been alerted to. It is possible that the participants' reactions to guidance were examples of the regular exam techniques and study skills.

There was only one code associated with the "the idea of being marked by a computer did not affect answer structure" theme, and this code was "speculated about free-response question marking" (C8). From code C8, some of the participants thought about how the marking system worked for the free-response questions on the AMS but there is no evidence that any of the participants tried to use this sort of consideration in an attempt to *beat the system* while working through the AMS. In relation to this, one participant postulated that the computer was searching for keywords to mark the answers against, whereas another participant noted that it would be easier to setup effective automated marking for one-word answers than for sentence answers.

No responses given by students to Q34 referred to using *beat the system* approaches to answer the free-response AMS questions. This partially alleviates concerns about students trying to employ such approaches in a large-scale administration of the AMS. However, it is worth noting that none of the usability laboratory participants or student respondents to Q34 had any reason to be particularly invested in achieving a high score on the AMS, since it did not count towards any course grade. Further, data would be required to investigate the prevalence of this approach to answering free-response questions more generally.

The reaction of the participants towards the computer marking of their answers was indicative of the *computers as social actors* framework (Reeves & Nass, 1996), which postulates that people can respond to a computer in the same way that they would respond to another person. Even in formative use, test takers expect to be given a score which accurately reflects their level of understanding, and accurate marking is important in order to retain student confidence. In addition, particular care needs to be taken to ensure that alternative and incorrect spelling is accounted for, so as not to disadvantage users with dyslexia or for whom English is not their first language. These considerations highlight that free-response questions need to be carefully designed and tested in order to be deployed both inclusively and effectively in an educational setting.

Table 4. Codes associated with "participant reaction to the usability of the AMS was mostly positive" theme

Code	Number of times coded
AMS was easy to use (C9)	27
Issues encountered with the AMS interface (C10)	12
Against adding a time limit to the AMS (C11)	18

Findings Related to "Participant Reaction to the Usability of the AMS Was Mostly Positive" and "Reactions to the AMS Depended Upon What Participants Thought It Was for" Themes

The "participant reaction to the usability of the AMS was mostly positive" theme was coded 57 times overall, and it pertains to the observation that, most of the time, the participants managed to work through the AMS without encountering issues. There were three codes associated with this theme, and these are outlined in **Table 4**.

Occurrences of code C9 highlighted participants were able to work through the AMS interface without issues most of the time. This was also found in the large-scale administration of the AMS, as students managed to enter responses to the questions, and there were no complaints about the usability of the AMS in the responses to Q34. When issues did arise, these were picked up by code C10; the issues highlighted by this code were mostly minor and question specific, meaning that they were easily resolved. Taking the findings from these codes together implied that the AMS was presented in a format that could easily be worked through by test-takers. This was a positive outcome from the usability testing, as it meant that the overall AMS did not need to be fundamentally re-designed for later versions.

Code C11 captured participants' ideas about adding a time limit to the AMS. Participants were not placed under any time limit when they worked through the AMS, and they each took different amounts of time to complete it. In general, the participants were not strongly in favor of adding a time limit, as it was not perceived to add anything useful to the AMS. However, some participants were against the idea of adding a time limit. These participants reasoned that if the AMS was designed to learn about student understanding, then adding a time limit would add unnecessary pressure to test-takers, which could lead to their levels of understanding being misrepresented. These reflections highlight the importance of considering what the purpose of the AMS is before trying to develop it further, since making unnecessary additions to the AMS could cause it to diverge from its original goals.

There were two codes under the "reactions to the AMS depended upon what participants thought it was for" theme: "AMS was compared to familiar types of assessments" (C12) was coded 19 times and "AMS identified as a conceptual evaluation tool" (C13) was coded 25 times. These codes relate to ideas that participants had about what the purpose of the AMS was, and what it could potentially be used to do.

Participants had various ideas about how the AMS could be used to assess students, and these were collected together by codes C12 and C13. Whether the AMS would be used in a high-stakes summative context or a low-stakes formative context was a key point raised by these codes. In code C12, it was postulated that the AMS could be developed to test more advanced topics, or be used as part of a summative assessment. Through code C13, one idea was that the AMS could be used to check that understanding of Newtonian mechanics had reached a certain level, which would make the AMS into a formative diagnostics test with similar objectives to the *mechanics diagnostics test* (MDT) (Halloun & Hestenes, 1985) upon which the FCI was originally based. Beyond this, a further idea was that the AMS could be used as a way of testing the effectiveness of different teaching methods, which aligns with the traditional use of concept inventories (Porter et al., 2014). These reflections indicate that the AMS is perceived to be a versatile tool, with the potential for it to be used in both assessment and teaching contexts.

Findings Related to the "Limited Feedback Was a Useful Addition to the AMS" Theme

The "limited feedback was a useful addition to the AMS" theme was coded 183 times overall, making it the theme that was coded most frequently. The theme refers to the limited feedback to the AMS questions that was given to the participants, and their reactions to this feedback. Note that the feedback is referred to as *limited* here since it only told the participants whether their answers were marked as correct or incorrect, with

Table 5. Codes associated with "limited feedback was a useful addition to the AMS" theme

Code	Number of times coded
Limited feedback provided by AMS was used to reflect upon performance (C14)	97
Felt as if the greater level of feedback should be provided to AMS questions (C15)	31
Felt that a lower level of feedback should be provided to the AMS questions (C16)	7
Responded positively to receiving feedback on AMS performance (C17)	48

the correct answer also given to the multiple-choice questions. The theme consists of 4four codes, and these are shown in **Table 5**. Note that codes C14 and C17 were referred to particularly frequently by the participants.

Discussion of giving feedback to the AMS questions

When the participants did the AMS, the instructions did not tell them that they were going to get feedback on their performance, or what the detail of this feedback would be. All but one of the participants received the feedback, and the behavior and interview responses of these participants indicated that they were not surprised to get some sort of feedback after completing the AMS. Participants who received the feedback were found to be interested in their own performance on the AMS, as captured by code C14. All the participants who received the feedback were observed to scroll through their answers to check where they were right and where they had gone wrong; in general, participants paid more attention to the instances where they had been wrong. In these cases, participants analyzed their own answers and asked questions to the interviewers about why their line of reasoning was incorrect. As a result, these participants were interested in using the AMS usability testing experience to better their understanding of the concepts being assessed. This sort of self-regulated learning (Nicol, 2021) gives students responsibility for their own learning. It is a well-established principle in formative assessment, but worthy of further investigation in the context of concept inventories; this is because concept inventories do not usually give feedback to students, although this has been done occasionally with the aim of increasing student self-efficacy (Chen et al, 2004; Lawrie et al., 2013).

Participants had different ideas about the level of the feedback that should be given by the AMS, and these were encapsulated within codes C15 and C16. At this point, it is important to note that levels of feedback on assessed tasks can be classified in several different ways. Carless (2006) considers whether the purpose of the feedback is advice for improving the current assessment, advice for future assessments, a means of explaining or justifying a grade, or a ritual. Shute (2008) instead classifies the different types of feedback that can be given as verification of response accuracy, explanation of the correct answer, hints, and worked examples. Jensen et al. (2021) describes feedback using conceptual metaphors, which include feedback being used for coaching and as a learner tool.

In line with Weaver (2006), some of the participants felt as if being told whether they were right or wrong was sufficient, because they could tell what they had done wrong from this low-level feedback, and then take necessary steps to improve without further prompting. However, in a concept inventory, this approach assumes that students can work out the nature of their conceptual misunderstanding without further guidance. To counter this possible issue, other participants felt as if a model answer should be given, which would serve to highlight where their line of reasoning had gone awry. However, a potential disadvantage of this approach is that students could memorize the answers for future use, rather than building up their physics knowledge and understanding. Such behavior has been reported previously (Bull & McKenna, 2004).

One participant reasoned that the level of the topic being tested could be used to determine the level of the feedback provided. This participant felt that the level of feedback given by the AMS was appropriate, but a higher level of feedback would be required for more advanced topics. Related to this, another participant pointed out that the level of feedback required was related to the purpose of the AMS. As an example, the participant postulated that if the AMS was meant as a diagnostic test, then summary feedback with a list of topics requiring attention would be helpful. These participants' ideas about different levels of feedback are comparable to Nyquist's (2003) distinction of *weaker feedback*, in which students are just told about their score, as compared with *stronger formative assessment*, which gives information about correct answers, explanation of the answers, and activities to undertake to improve.

Getting some feedback was perceived as an important part of the process by most participants, and this was highlighted by code C17. The feedback was found to be useful, even though it was limited to knowledge of whether their responses were right or wrong, with the correct answer also given for some of the questions. This is in agreement with the findings from the literature that students like receiving feedback, even if they do not properly make use of it (Brown & Glover, 2006) or if the feedback intervention is unhelpful (Kluger & DeNisi, 1996). In relation to these findings, Henderson et al. (2019) place emphasis on the importance of designing effective feedback interventions.

For some participants, feedback was seen as positive because it told them how well they had done. When making use of limited feedback in this way, students may simply be checking that they are making reasonable progress, in line with the findings of Dawson et al. (2018), Draper (2009), and Scott (2014). At an even lower level of feedback, Millar (2005) found that students are interested in knowing their score, even if this does not contribute in any way to their course grade, as was the case for the AMS. For another participant, feedback was welcomed because it showed them where they had gone wrong, allowing them to refer back to the course material to improve their own understanding. As was the case for this participant, the presence of computer-generated feedback has previously been shown to deter students from using a *trial and error* approach to answering the questions (Walker et al., 2008), which aligns with the aims of using concept inventories to investigate student understanding.

Whatever their reasons for wanting feedback, and whatever use they saw that it had, the participants in general saw feedback as a good thing. Giving feedback to students enables them to take responsibility for their own learning and allows them to gain independence (Boud & Soler, 2016; Carless, 2020). Meanwhile, the more conventional use of concept inventories to provide feedback to teachers is in line with the recently recognized field of *learning analytics* (Clow, 2013; Sedrakyan et al., 2020; Zilvinskis et al., 2017). The provision of feedback to both students and teachers marks a welcome move towards a more personalized type of teaching and learning, where students' needs are responded to in a way that is based upon their own strengths, weaknesses and willingness to engage. In the context of the work reported here, this could be an area for further investigation in the future.

LIMITATIONS

The qualitative study used eight participants. Testing and interviewing a greater number of participants would have been useful, but in order to maintain the project schedule, the usability tests were limited to these eight participants. This is acknowledged as a limitation of the study, as it does not allow for ready subdivision of qualitative data by specific characteristics of the participants, such as study experience or demographic. It is of note that the lead author was the only one who went through the data and coded it, which is another limitation of the study. To counteract this, the codes were enumerated and gathered together under themes which exhibited broader patterns of behavior. This approach was taken in order to mitigate the effect of annotator bias, and to prevent arbitrary conclusions from being drawn from the data. To make the meaning of the themes clear, the statement of each theme gave the main conclusion from the underlying codes associated with it, and this resulted in themes which were longer than usual.

CONCLUSIONS

The study investigated two research questions:

1. RQ1: How do students react to free-response modified versions of FCI questions?
2. RQ2: How do students react to being given feedback on multiple-choice & free-response FCI questions?

The first aim of the study was to investigate how students reacted to free-response concept inventory questions (RQ1); the second aim of the study was to investigate how students reacted to being given feedback on concept inventory questions (RQ2). Data were collected for the study by having eight participants work through the AMS in a usability testing setting, and conducting interviews with the participants about their experience. Further data were collected from the large-scale administration of the AMS in the form of qualitative responses to the feedback question Q34, and the findings from the different data sets were triangulated where relevant.

In the context of RQ1, participants generally reacted positively to being asked AMS questions in the free-response format. It was found that free-response questions made participants think more deeply about the questions, which encouraged them to be creative when writing their answers. Participants also noted that answers to free-response questions provide more information about student understanding to the marker; in this way, participants could see the educational value of using free-response questions instead of multiple-choice. Taken together, this suggests that it is feasible to use free-response AMS questions in place of the multiple-choice FCI counterparts, which validates their use in the AMS.

In the context of RQ2, participants viewed getting feedback as an important part of the process of working through the AMS, and they responded well to receiving it. Feedback was found to be useful by participants because they were interested in finding out about how they had done on the AMS. The feedback was limited in detail, and participants had different ideas about the level of feedback that should be given by the AMS; these were often related to what the participants thought that the purpose of the AMS could be. Taken together, this suggests that there is an opportunity to make use of formative concept inventories that give feedback as a tool for guiding more independent, student-driven learning.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approve final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Acknowledgements: The authors acknowledge the previous work done by David Sands, Ross Galloway, and Christine Leach, which was important for initial setting up of concept inventory used during the study. The authors also owe a debt of gratitude for the eight usability testing participants, without whom the research could not have been completed.

Ethics declaration: Ethical approval for this study was obtained from The Open University Human Research Ethics Committee with approval code HREC/2017/2629/Parker/1 on 31 July 2017.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Bailey, J. M., Johnson, B., Prather, E. E., & Slater, T. F. (2012). Development and validation of the star properties concept inventory. *International Journal of Science Education, 34*(14), 2257-2286. https://doi.org/10.1080/09500693.2011.589869

Barnum, C. B. (2010). *Usability testing essentials: Ready, set, test*. Morgan Kaufmann Publishers.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education, 41*(3), 400-413. https://doi.org/10.1080/02602938.2015.1018133

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., Clarke, V., & Terry, G. (2014). Thematic analysis. *Qualitative Research in Clinical Health Psychology, 24,* 95-114. https://doi.org/10.1007/978-1-137-29105-9_7

Brown, E., & Glover, C. (2006). Evaluating written feedback. In C. Bryan, & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 81-91). Routledge.

Bull, J., & McKenna, C. (2004). *Blueprint for computer-aided assessment*. Routledge. https://doi.org/10.4324/9780203464687

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219-233. https://doi.org/10.1080/03075070600572132

Carless, D. (2020). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*. https://doi.org/10.1177/1469787420945845

Castro, A., & Andrews, G. (2018). Nursing lives in the blogosphere: A thematic analysis of anonymous online nursing narratives. *Journal of Advanced Nursing, 74*(2), 329-338. https://doi.org/10.1111/jan.13411

Chen, J. C., Kadlowec, J., & Whittinghill, D. (2004). Work in progress: Combining concept inventories with rapid feedback to enhance learning. In *Proceedings of the 34th Annual Frontiers in Education*. https://doi.org/10.1109/FIE.2004.1408580

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683-695. https://doi.org/10.1080/13562517.2013.827653

Crisp, G. (2007). *The e-assessment handbook.* Continuum.

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2018). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*(1), 25-36. https://doi.org/10.1080/02602938.2018.1467877

Ding, L., Chaby, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics-Physics Education Research, 2,* 010105. https://doi.org/10.1103/PhysRevSTPER.2.010105

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics-Physics Education Research, 10,* 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Draper, S. (2009). What are learners actually regulating when given feedback. *British Journal of Educational Technology, 40*(2), 306-315. https://doi.org/10.1111/j.1467-8535.2008.00930.x

Eaton, P. (2021). Evidence of measurement invariance across gender for the force concept inventory. *Physical Review Physics Education Research, 17,* 010130. https://doi.org/10.1103/PhysRevPhysEducRes.17.010130

Filia, K. M., Jackson, H. J., Cotton, S. M., Gardner, A., Killackey, E. J., & Cook, J. A. (2018). What is social inclusion? A thematic analysis of professional opinion. *Psychiatric Rehabilitation Journal, 41*(3), 183-195. https://doi.org/10.1037/prj0000304

Garcia, M., & Hamilton-Giachritsis, C. (2017). Getting involved: A thematic analysis of caregivers' perspectives in Chilean residential children's homes. *Journal of Social and Personal Relationships, 34*(3), 356-375. https://doi.org/10.1177/0265407516637838

Garvin-Doxas, K., Klymkowsky, M., & Elrod, S. (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: Report on a National Science Foundation-sponsored conference on the construction of concept inventories in the biological sciences. *CBE Life Sciences Education, 6*(4), 277-282. https://doi.org/10.1187/cbe.07-05-0031

Grogan, S., & Jayne, M. (2017). Body image after mastectomy: A thematic analysis of younger women's written accounts. *Journal of Health Psychology, 22*(11), 1480-1490. https://doi.org/10.1177/1359105316630137

Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college students. *American Journal of Physics, 53*, 1043-1056. https://doi.org/10.1119/1.14030

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development, 38*(7), 1401-1416. https://doi.org/10.1080/07294360.2019.1657807

Hestenes, D., Wells, M., & Swackhamer, G., (1992). Force concept inventory. *The Physics Teacher, 30,* 141-158. https://doi.org/10.1119/1.2343497

Hufnagel, B. (2002). Development of the astronomy diagnostic test. *Astronomy Education Review, 1*(1), 47-51. https://doi.org/10.3847/AER2001004

Hunt, T. (2012). Computer-marked assessment in Moodle: Past, present and future. *Digital Education Research Network*. https://dern.acer.org/dern/ict-research/page/computer-marked-assessment-in-moodle-past-present-and-future

Jensen, L. X., Bearman, M., & Boud, D. (2021). Understanding feedback in online learning–A critical review and metaphor analysis. *Computers & Education, 173,* 104271.https://doi.org/10.1016/j.compedu.2021.104271

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical view, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the force concept inventory. *American Journal of Physics, 79*(9), 909-914. https://doi.org/10.1119/1.3602073

Lawrie, G., Wright, A., Schultz, M., Dargaville, T., O'Brien, G., Bedford, S., Williams, M., Tasker, R., Dickson, H., & Thompson, C. (2013). Using formative feedback to identify and support first-year chemistry students with missing or misconceptions. A practice report. *International Journal of the First Year in Higher Education, 4*(2), 111-116. https://doi.org/10.5204/intjfyhe.v4i2.179

Lee, N. W., Shamsuddin, W. N. F. W, Wei, L. C., Anuardi, M. N. A. M., Heng, C. S., & Abdullah, A. N. (2021). Using online multiple choice questions with multiple attempts: A case for self-directed learning among tertiary students. *International Journal of Evaluation and Research in Education, 10*(2), 553-568. https://doi.org/10.11591/ijere.v10i2.21008

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

Mitchell, T., Aldridge, N., Williamson, W., & Broomhead, P. (2003). Computer based testing of medical knowledge. In *Proceedings of the 7th International CAA Conference* (pp. 249-267).

Nardi, A., & Ranieri, M. (2019). Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology*, *50*(3), 1495-1506. https://doi.org/10.1111/bjet.12644

Nickerson, J., Corter, J., Esche, S., & Chassapis, C. (2007). A model for evaluation the effectiveness of remote engineering laboratories and simulations in education. *Computers and Education, 49*(3), 708-725. https://doi.org/10.1016/j.compedu.2005.11.019

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and higher Education, 31*(1), 53-64. https://doi.org/10.1080/03098770601167922

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756-778. https://doi.org/10.1080/02602938.2020.1823314

Nyquist, J. B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Vanderbilt University Press.

Porter, L., Taylor, C., & Webb, K. (2014). Leveraging open source principles for flexible concept inventory development. In *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education* (pp. 243-248). https://doi.org/10.1145/2591708.2591722

Rebello, N., & Zollman, D. (2004). The effect of distractors on student performance on the force concept inventory. *American Journal of Physics, 72,* 116. https://doi.org/10.1119/1.1629091

Reeves, B., & Nass, C. (1996). *The media equation*. Stanford University Press.

Robertson, A. E., Stanfield, A. C., Watt, J., Barry, F., Day, M., Cormack, M., & Melville, C. (2018). The experience and impact of anxiety in autistic adults: A thematic analysis. *Research in Autism Spectrum Disorders, 46*, 8-18. https://doi.org/10.1016/j.rasd.2017.11.006

Sangwin, C. J. (2013). *Computer aided assessment of mathematics.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

Scanlon, E., Colwell, C., Cooper, M., & Di Paolo, T. (2004). Remote experiments, re-versioning and re-thinking science learning. *Computers and Education, 43*(1-2), 153-163. https://doi.org/10.1016/j.compedu.2003.12.010

Scott, S. V. (2014). Practising what we preach: Towards a student-centred definition of feedback. *Teaching in Higher Education*, *19*(1), 49-57. https://doi.org/10.1080/13562517.2013.827639

Sedrakyan, G., Malmberg, J., Verbert, K., Jarvela, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior, 107,* 105512. https://doi.org/10.1016/j.chb.2018.05.004

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Simon, & Snowdon, S. (2014). Multiple-choice vs free-text code-explaining examination questions. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research* (pp. 91-97). https://doi.org/10.1145/2674683.2674701

Smedley, R. M., & Coulson, N. S. (2017). A thematic analysis of messages posted by moderators within health-related asynchronous online support forums. *Patient Education and Counseling, 100*(9), 1688-1693. https://doi.org/10.1016/j.pec.2017.04.008

Smith, J. I., & Tanner, K. (2010). The problem of revealing how students think: Concept inventories and beyond. *CBE Life Sciences Education, 9*(1), 1-5. https://doi.org/10.1187/cbe.09-12-0094

Thornton, R., & Sokoloff, D. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*, 338. https://doi.org/10.1119/1.18863

University of Auckland. (2017). *Thematic analysis*. https://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis/about-thematic-analysis.html

Varagona, L. M., & Hold, J. L. (2019). Nursing students' perceptions of faculty trustworthiness: Thematic analysis of a longitudinal study. *Nurse Education Today*, *72*, 27-31. https://doi.org/10.1016/j.nedt.2018.10.008

Walker, D. J., Topping, K., & Rodrigues, S. (2008). Student reflections on formative e-assessment: Expectations and perceptions. *Learning, Media and Technology*, *33*(3), 221-234. https://doi.org/10.1080/17439880802324178

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment and Evaluation in Higher Education*, *31*(3), 379-394. https://doi.org/10.1080/02602930500353061

Woodford, K., & Bancroft, P. (2005). Multiple choice questions not considered harmful. In *Proceedings of the Seventh Australasian Computing Education Conference* (pp. 109-115).

Yasuda, J., Mae, N., Hull, M. M., & Taniguchi, M. (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical Review Physics Education Research, 17,* 010115. https://doi.org/10.1103/PhysRevPhysEducRes.17.010115

Zeilik, M. (2003). Birth of the astronomy diagnostic test: Prototest evolution. *Astronomy Education Review, 1*(2), 46-52. https://doi.org/10.3847/AER2002005

Zhang, L., & VanLehn, K. (2021). Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments, 29*(6), 1019-1036. https://doi.org/10.1080/10494820.2019.1619586

Zilvinskis, J., Willis, J. III., & Bordon, V. M. H. (2017). An overview of learning analytics. *New Directions for Higher Education, 179*, 9-17. https://doi.org/10.1002/he.20239