European Journal of Science and Mathematics Education, 2023, **11**(2), 360-375
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Establishing a physics concept inventory using computer marked free-response questions

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

Holly Hedgeland [1,2]
0000-0003-3703-7942

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

Nicholas St. J. Braithwaite [1]
0000-0002-1586-3736

[1] School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[2] Clare Hall, University of Cambridge, Herschel Road, Cambridge, CB3 9AL, UK
[*] Corresponding author: mark.parker@open.ac.uk

Citation: Parker, M. A. J., Hedgeland, H., Jordan, S. E., & Braithwaite, N. St. J. (2023). Establishing a physics concept inventory using computer marked free-response questions. *European Journal of Science and Mathematics Education, 11*(2), 360-375. https://doi.org/10.30935/scimath/12680

ARTICLE INFO

Received: 13 Sep 2022
Accepted: 18 Nov 2022

ABSTRACT

The study covers the development and testing of the alternative mechanics survey (AMS), a modified force concept inventory (FCI), which used automatically marked free-response questions. Data were collected over a period of three academic years from 611 participants who were taking physics classes at high school and university level. A total of 8,091 question responses were gathered to develop and test the AMS. The AMS questions were tested for reliability using classical test theory (CTT). The AMS computer marking rules were tested for reliability using inter-rater reliability (IRR). Findings from the CTT and IRR studies demonstrated that the AMS questions and marking rules were overall reliable. Therefore, the AMS was established as a physics concept inventory which uses automatically-marked, free-response questions. The approach used to develop and test the AMS could be used in further attempts to develop concept inventories which make use of automatically-marked, free-response questions.

Keywords: computer-marked assessment, automated marking, concept inventories, free-response questions, physics education

INTRODUCTION

Concept inventories are used within science education research, mostly to test the effectiveness of teaching approaches (Porter et al., 2014). Examples of concept inventories deployed across a variety of disciplines, are 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).

The first concept inventory was the force concept inventory (FCI) (Hestenes et al., 1992), which was designed to test conceptual understanding of Newtonian mechanics. The FCI contains 30 multiple-choice questions, with distractor options chosen to correspond with student misconceptions. The FCI is regularly used to collect physics education data and has been the subject of much discussion in literature, as summarized by Eaton (2021) and Yasuda et al. (2021).

Most concept inventories make use of multiple-choice questions, which are quick and easy to administer (Lee et al., 2021). Problems with multiple-choice questions have been identified in literature (Nicol, 2007; Zhang & VanLehn, 2021). In particular, students select answers from a pre-constructed list of options which were written by another person (Simon & Snowdon, 2014), which makes confident mapping of conceptual misunderstanding difficult. In contrast, students are required to write their own answers to free-response questions (Mitchell et al., 2003). For concept inventories, free-response questions would provide more detailed information about student thinking.

Rebello and Zollman (2004) investigated the use of free-response questions in a concept inventory. They gave students free-response versions to four FCI questions and compared the free-response answers to the distractors from the corresponding multiple-choice questions. They found that there were cases where the student answers were similar to the distractor options, but there were also cases where students wrote answers which contained different ideas. This showed that the distractors provided were insufficient to cover all student thought processes pertaining to the questions.

More information is available about student knowledge and understanding when free-response questions are used. However, free-response questions take a long time to manually mark, especially when class sizes are large. For large-scale deployment of free-response questions, it is preferable to automate the marking process. This can be achieved through the Pattern Match question type of the Moodle question engine (Hunt, 2012). Pattern Match takes an algorithm-based approach by matching words as a string of letters, with the capacity for word order and proximity, negation, and truncation to be considered, along with two methods of handling student spelling mistakes. The software is capable of good marking accuracy provided that student responses are used as training data to develop the answer-matching rules. The Pattern Match question type forms the basis of the current study.

Through Pattern Match, it is possible to author and automatically mark short free-response answers, facilitating the development of a physics concept inventory which makes use of short answer free-response questions. The FCI has been widely used and its questions have been validated. Hence, the FCI was the logical choice for an investigation of free-response questions in concept inventories. The purpose of the current work is to outline the development of a version of the FCI, which makes use of automatically-marked free-response questions. This work describes quantitative investigations into the performance of the questions and the accuracy of the corresponding automated marking rules. The study was guided by the two following research questions:

1. **RQ1:** To what extent is a free-response modified version of the FCI reliable?
2. **RQ2:** How reliable are automated marking schemes when used to mark free-response versions of FCI questions?

METHODS

Background

The study was conducted at a university that specializes in distance learning. A completely free-response version of the FCI had previously been distributed to undergraduate physics students at two other institutions. Responses were marked by hand and were used to develop marking rules for the questions using pattern match. It was found that some of the questions were unsuitable for use in free-response format. These questions required students to describe a trajectory, which is not easy to describe in words. Some questions were split into two parts. Question wording was kept as close as possible to the original FCI, in respect of the large-scale use and validation of the FCI questions. A 33-question instrument was proposed, with a mixture of free-response and multiple-choice questions. These 33 questions were assembled into an online concept inventory, called the alternative mechanics survey (AMS).

Data Collection

Approval for the work was obtained from the relevant ethics committee. The response data set was collected by uploading the AMS to two online physics teaching platforms. Potential participants were contacted and provided with a link to the AMS. The participants were drawn from a mixture of undergraduate

and high school students. Data were gathered over a period of three academic years. A total of 8091 responses were gathered to the AMS questions. This data was used to test and develop the AMS using an iterative process.

Data Analysis–Classical Test Theory

Classical test theory (CTT) was used to analyze the performance of the AMS questions. CTT is explained in the pioneering work of Crocker and Algina (1986) and is described in work such as Ding and Beichner (2009). CTT outlines statistics used to analyze various aspects of test functionality. Some CTT statistics are calculated for individual test items, while others are calculated for the entire test. As such, data from complete AMS attempts were retained for the calculation of the CTT statistics.

The *difficulty* of a test item is the proportion of test-takers who answered the test item correctly. It is calculated using the formula:

$$P = \frac{N_1}{N}, \tag{1}$$

where P is difficulty, N_1 is the number of correct responses, and N is the number of completed tests. A larger difficulty value corresponds to an easier question. The difficulty takes a value between 0 and 1, with the acceptable range of values being from 0.3 to 0.9 (Doran, 1980). It is desirable to have a range of different difficulty values across the test, as this helps to differentiate between higher and lower performing test takers.

The *discrimination* is defined as the ability of the test item to distinguish between higher-performing test-takers and lower-performing test-takers. It is calculated using the formula:

$$D = \frac{N_H - N_L}{N/4}, \tag{2}$$

where D is discrimination, N_H is the number of correct responses given by test takers in the upper quartile of the overall test score, N_L is the number of correct responses given by test takers in the lower quartile of the overall test score, and N is the total number of test takers. The discrimination has a value between 0 and 1, with the acceptable range of values being from 0.3 to 1 (Doran, 1980).

A statistic related to the discrimination is the *point biserial coefficient*. This measures the correlation between the scores on an item and the total scores for the entire test. This means that it measures how well each item tests material that is consistent with the rest of the test. It is calculated using the formula:

$$r_{pbi} = \frac{\overline{X_1} - \overline{X_0}}{\sigma_x}\sqrt{P(1-P)}, \tag{3}$$

where $\overline{X_1}$ is the mean total test score of the test takers who answered the item correctly, $\overline{X_0}$ is the mean total test score of the test takers who answered the item incorrectly, σ_x is the standard deviation of all of the scores, and P is the difficulty of the item. The point biserial coefficient takes a value between 0 and 1, with the acceptable range of values being from 0.2 to 1 (Kline, 1986).

The entire test is reliable when it is consistent. Hence if the same test takers repeated the same test without learning from the experience, they would be expected to get the same scores. This is an unfeasible experiment; thus, reliability is treated in a different way. When the items test similar material, test takers would be expected to give similar responses to these related items. The *Kuder-Richardson reliability* measures the extent to which an entire test is constructed using questions that test similar material. It is calculated using the equation:

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum P_i(1-P_i)}{\sigma_x^2}\right), \tag{4}$$

where K is the number of items on the test, P_i is the difficulty of the i^{th} item, and σ_x^2 is the standard deviation of the total score. A value of r_{test} of 0.7 or above shows that the test is reliable overall.

The Kuder-Richardson reliability expands the idea of testing the reliability of individual test items to testing the reliability of the entire test. In a similar manner, *Ferguson's delta* (δ) expands the idea of assessing the discrimination of individual test items to assessing the discrimination of the entire test. It is calculated using the equation:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - N^2/(K+1)}, \tag{5}$$

where N is the number of test takers, f_i is the number of test takers who score i on the test, and K is the number of items on the test. A δ value of 0.9 or above shows that the overall test has good discriminatory capabilities.

Data Analysis–Inter-Rater Reliability

Inter-rater reliability (IRR) was used to analyze performance of AMS marking rules. Various IRR statistics are outlined by Artstein and Poesio (2008), as well as by Zwick (1988). IRR statistics are used to test the agreement of different raters when classifying subjects. The appropriate statistic must be chosen based on the properties and assumptions of the situation. Data from all non-blank responses to AMS questions were retained for the calculation of the IRR statistics.

The most basic IRR statistic is the *percentage agreement*, which measures the proportion of cases where the raters agree upon the classification of subjects. It is calculated as follows.

The *agreement* value agr_i for the subjects i is defined as:

1, if the two raters assign i to the same category,

0, if the two raters assign i to different categories.

Hence, percentage agreement A_0 over all of the subjects *i* is:

$$A_0 = \frac{1}{n}\sum agr_i, \tag{6}$$

where n is the total number of classified subjects. Percentage agreement has a value from zero to one, with values equal to or greater than 0.95 indicating a good level of agreement (Jordan, 2012). A high percentage agreement value alone is insufficient to show that agreement is genuinely at a high level. This is because percentage agreement does not account for sample size or chance agreement. As its name implies, chance agreement arises when raters assign a subject to the same category by random chance. Advanced IRR statistics are designed to account for this random agreement.

A_0 is the value of the percentage agreement. A_e is defined as the agreement that is expected to arise by chance. The value of $1 - A_e$ gives the maximum amount of true agreement (not by chance) that is possible to attain, and the value of $A_0 - A_e$ gives the amount of true agreement that is observed. Dividing $A_0 - A_e$ by $1 - A_e$ gives the proportion of true agreement, which was observed, accounting for agreement that rises by chance. Therefore, the agreement statistic when chance agreement is accounted for is defined as:

$$A = \frac{A_0 - A_e}{1 - A_e}. \tag{7}$$

Cohen's kappa (κ) (Cohen, 1960) is an advanced IRR statistic, which is calculated using this formula for *A*. The way in which A_e is calculated is based upon assumptions about the way that raters classify subjects. Cohen's kappa assumes that raters would produce different distributions if they did their classifications by chance. This is a realistic scenario for the current study, as raters would not be expected to perform their classifications in exactly the same way, even if they were putting objects into categories at random. Mathematically, this means that if the raters use categories labelled *k*, then n_{c_1k} is the number of times the first rater assigns an object to category *k*, and n_{c_2k} is the number of times the second rater assigns an object to category *k*, and *j* is the total number of object classified, then the probability P_{c_1k} of the first rater assigning an arbitrary object to category *k* is:

$$P_{c_1k} = \frac{n_{c_1k}}{j}. \tag{8}$$

Similarly, the probability P_{c_2k} of the second rater assigning an arbitrary object to category *k* is given by:

$$P_{c_2k} = \frac{n_{c_2k}}{j}. \tag{9}$$

Hence, the probability P_k of both raters assigning an arbitrary object to category *k* is given by:

$$P_k = P_{c_1k} \times P_{c_2k} = \frac{n_{c_1k}}{j} \times \frac{n_{c_2k}}{j} = \frac{n_{c_1k}n_{c_2k}}{j^2}. \tag{10}$$

Summing over the *K* different classifications gives the A_e value for Cohen's kappa:

$$A_e^{\kappa} = \sum \frac{n_{c_1k}n_{c_2k}}{j^2} = \frac{1}{j^2}\sum n_{c_1k}\, n_{c_2k}. \tag{11}$$

Putting A_e from equation (11) into equation (7) gives the corresponding Cohen's kappa statistic, which takes a value between 0 and 1. Values for Cohen's kappa which are 0.8 or above illustrate good rater agreement (Artstein & Poesio, 2008). Cohen's kappa was chosen for use in this study because its assumption that different markers will have different marking distributions, even when marking at random, matches with the expected behavior of both human and computer marking.

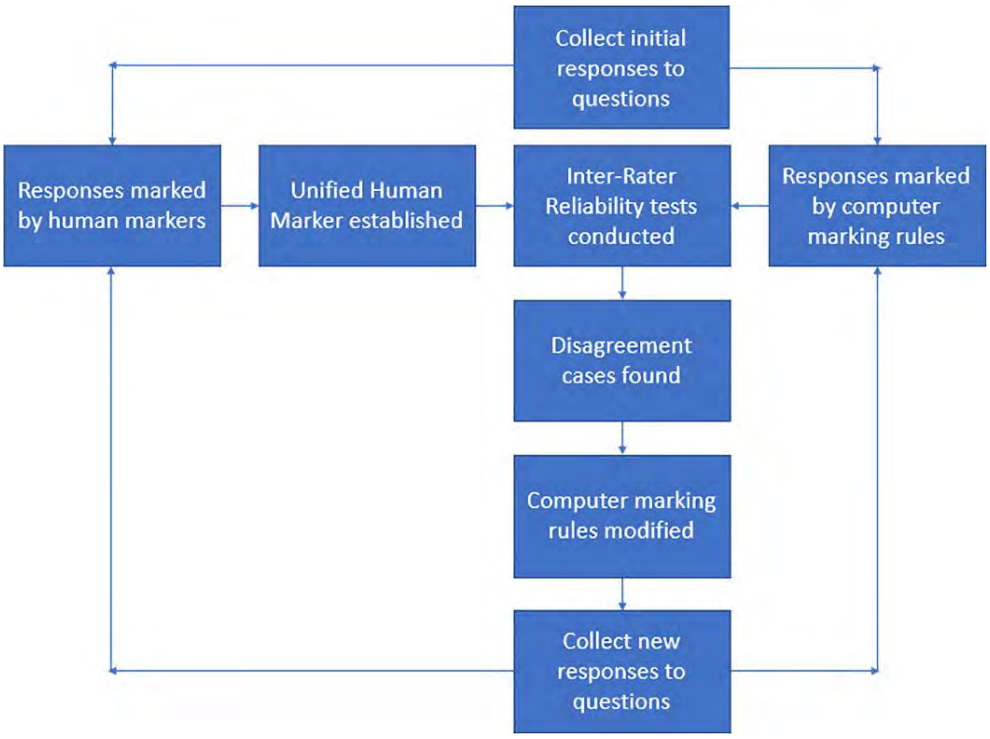

Figure 1. Flowchart showing IRR-based process used to test and develop AMS computer marking rules (Source: Authors' own elaboration)

For the IRR calculations, human marking of the responses was required to compare the computer marking against. Humans do not mark in a consistent way, even when provided with a mark scheme (Butcher & Jordan, 2010), so more than one human marker was needed. Five human markers were recruited for this task. All of the markers had a background in physics, meaning that the subject matter of the AMS was straightforward for them to understand and mark.

For the responses collected for the AMS, a standardized set of marking guidelines was given to the markers. The markers were instructed to award a mark of one for answers which they deemed to be *correct*, and a mark of zero for answers which they deemed to be *incorrect*. No partial credit was given. The marked responses for each question from each marker were collected together to establish a *unified human marker* (UHM) by examining how the majority of markers chose to mark each of the responses. For example, if a response was marked as *correct* by four of the markers, and *incorrect* by one of the markers, then the UHM would award a mark of one for this response. Cases where the response was marked one way by three of the markers and the other way by two of the markers were deemed to be *borderline*. These cases were examined by members of the author team, who made a final decision on whether the response should be marked as correct or incorrect. Because it was built from a consensus of experts, the UHM was designed to provide the most reliable marking for each response. The UHM was treated as the *master mark scheme* in his study.

The IRR statistics were used to develop and test the computer marking as follows. For each free-response question, Cohen's kappa values were calculated for the UHM against the computer marker. These values were used to identify problematic cases where the computer marking rules were not functioning at the required level, and to identify generic difficulties in the computer marking. The number of times the UHM disagreed with the computer marking on each question was noted, and the number of cases which were false positives (where the UHM marked the response as *incorrect* and the computer marker marked the response as *correct*) and false negatives (where the UHM marked the response as *correct* and the computer marker marked the response as *incorrect*) were also noted. The disagreement cases on each question were used to improve the marking rules by adding suitable negation rules to counter the false positives, and by using the false negative cases to add new rules to cover alternative wordings for correct answers. To check for consistency, these new marking rules were also back-tested against the responses used to develop them, as well as against other previous responses sets where applicable. This process for developing and testing the marking rules is illustrated in **Figure 1**.

Table 1. Academic years in which the AMS data was gathered

Version of the AMS	Academic year data gathered	Version of the AMS iterated into
Version 1	2017-2018	Version 2
Version 2	2018-2019	Version 3
Version 3	2019-2020	Final version

Table 2. Overall CTT statistics for version 1 and version 2 of the AMS

Classical test theory statistic	Value (AMS version 1)	Value (AMS version 2)	Desired values
Mean difficulty	0.65	0.55	[0.3, 0.9]
Mean discrimination	0.53	0.61	≥0.3
Mean point biserial coefficient	0.52	0.44	≥0.2
Kuder-Richardson reliability	0.92	0.87	≥0.7
Ferguson's delta	0.98	0.98	≥0.9

RESULTS AND DISCUSSION

The AMS went through three rounds of iterative development and testing. A new version of the AMS was produced at the end of each round. Version 1 of the AMS was attempted by 328 test-takers, and data collected were analyzed using CTT and IRR in order to iterate version 1 into version 2 of the AMS. Version 2 of the AMS was attempted by 81 test-takers, and these data were analyzed using CTT and IRR to iterate version 2 into version 3 of the AMS. Version 3 of the AMS was attempted by 201 test-takers, and the collected data were analyzed using IRR to iterate version 3 into the final version of the AMS. A total of 8,091 question responses were used to develop and test the AMS. Details about the academic years over which the data were gathered is given in **Table 1**.

In what follows, the findings from the CTT and the IRR analyses are presented under separate headings. Asterisks in the results indicate that the calculation provided a result which was outside the acceptable range of values.

Findings Related to the Classical Test Theory Analysis

For consideration of overall test function of the AMS, the mean values of the difficulty, discrimination, and point biserial coefficient were calculated. These are given in **Table 2**, together with the Kuder-Richardson reliability and Ferguson's delta values for the entire test.

The mean value of the difficulties of the individual questions on version 1 of the AMS was 0.65, and the mean value of the difficulties of the individual questions on version 1 of the AMS was 0.55. Both of these values were within the acceptable range for difficulty, which implied that both version 1 and version 2 of the AMS were overall functioning in the desired way in terms of difficulty.

The mean value of the discrimination of the individual questions on version 1 of the AMS was 0.53, which was within the acceptable range for the discrimination value. Similarly, the mean of the discrimination values of the individual questions on version 2 of the AMS was 0.61, which was also within the acceptable range for the discrimination value. Taken together, these calculations implied that version 1 and version 2 of the AMS were both overall capable of distinguishing between the higher-performing and lower-performing students.

The mean value of the point biserial coefficients of the individual questions on version 1 of the AMS was 0.52, and this was also within the acceptable range for the point biserial coefficient value. For version 2 of the AMS, the mean value of the point biserial coefficients of the individual questions was 0.44, and this was within the acceptable range of values for the point biserial coefficient. This implied that the version 1 and version 2 of the AMS overall contained questions that assessed similar topics. The difficulty, discrimination, and point biserial results provided important evidence for the overall functionality of the AMS questions.

Using the previously calculated difficulty values for each of the items, the standard deviation of the total scores, and K=33 for the 33 AMS items, a Kuder-Richardson reliability of 0.92 was calculated for version 1 of the AMS. Likewise, a Kuder-Richardson reliability value of 0.87 was calculated for version 2 of the AMS by using the previously calculated difficulty values for each of the items, the standard deviation of the total scores, and K=32 for the 32 AMS items. In both the cases of version 1 and version 2 of the AMS, the calculated value was above the threshold value for Kuder-Richardson reliability of 0.7, which showed that both version

Table 3. Cohen's kappa values for the version 1 UHM against the version 1 computer marking

Question	Number of responses	Cohen's kappa (AMS version 1)
Q1	328	0.92
Q2	307	0.79*
Q3	305	0.38*
Q4	304	0.28*
Q5	301	0.93
Q11	280	0.76*
Q13	277	0.91
Q17	276	0.95
Q19	275	0.81
Q20	275	0.71*
Q22	275	0.83
Q23	275	0.72*
Q25	255	0.94
Q27	255	0.89
Q29	255	0.80
Q30	254	0.38*
Q31	254	0.98
Q32	254	0.73*

1 and version 2 of the AMS were reliable overall. The findings from these Kuder-Richardson reliability calculations corroborate with what was found in the analysis of the mean difficulty and point biserial coefficients for version 1 and version 2 of the AMS.

Next, taking the values of the frequency for each possible score, N=254 as the number of test-takers, and K=33 for the number of items, a Ferguson's delta value of 0.98 was found for version 1 of the AMS. Similarly, a Ferguson's delta value of 0.98 was found for version 2 of the AMS, by taking the values of the frequency for each of the possible scores, N=60 for the total number of test-takers, and K=32 for the number of test items. In each of the cases of version 1 and version 2 of the AMS, the value calculated was above the threshold value for Ferguson's delta of 0.9. This showed that the overall version 1 and version 2 of the AMS could discriminate between lower and higher scoring students. The results of these Ferguson's delta calculations agreed with what was concluded from the analysis of the mean discrimination values for version 1 and version 2 of AMS.

The CTT analysis conducted above demonstrated that the questions on both version 1 and version 2 of the AMS were functioning at an acceptable level. The CTT statistics were stable between version 1 and version 2 of the AMS, which showed that the AMS questions had reached a sufficient level of development. This outcome may be expected, since the AMS questions were taken from the FCI, and the FCI questions have previously been developed and tested extensively, as summarized by Scott and Schumayer (2017). As a result, the CTT analysis did not need to be repeated for further iterations of the AMS. Instead, these further iterations of the AMS placed focus on improving the accuracy of the AMS computer marking rules.

Findings Related to the Inter-Rater Reliability Analysis

For consideration of the effectiveness of the AMS marking rules, the Cohen's kappa values were calculated for each version of the AMS. To test for consistency, additional back-testing calculations were also carried out where applicable. The results are given in the following. This starts with **Table 3**, which gives the results of the Cohen's kappa calculations conducted using data gathered from version 1 of the AMS. Note that some questions on each version of the AMS were not in free-response format. These questions were mostly based on trajectories, which are difficult to describe in words. Cohen's kappa values were not calculated for these questions, leading to the discontinuous question numbers.

The acceptable range of values for Cohen's kappa are [0.8, 1]. From **Table 3**, Q1, Q5, Q13, Q17, Q19, Q22, Q25, Q27, Q29, and Q31 had acceptable values for Cohen's kappa, meaning that the marking rules were functioning well for these ten questions. For the other eight questions, Q2, Q3, Q4, Q11, Q20, Q23, Q30, and Q32, the Cohen's kappa values were outside the acceptable range of values, which implied that the marking rules for these questions required further development.

The version 2 marking rules were developed by adapting the version 1 marking rules using the strategy outlined in the *Data Analysis–Inter-Rater Reliability* subsection. In addition, version 2 of the AMS increased its

Table 4. Cohen's kappa values for the version 2 UHM against the version 2 computer marking

Question	Number of responses	Cohen's kappa (AMS version 2)
Q1	81	1.00
Q2	75	1.00
Q3	74	0.32*
Q4	74	0.86
Q5	73	0.93
Q7	71	0.67*
Q11	66	0.69*
Q12	66	0.79*
Q13	66	0.68*
Q15	66	0.22*
Q17	64	1.00
Q18	64	0.55*
Q19	64	0.44*
Q20	64	0.92
Q21	64	0.59*
Q22	60	0.93
Q23	60	0.82
Q25	60	0.89
Q27	60	0.92
Q28	60	0.75*
Q29	60	0.83
Q30	60	0.49*
Q31	60	0.97
Q32	60	0.77*
Q33	60	0.25*

free-response scope by converting seven multiple-choice questions from version 1 into free-response format. The seven questions converted were Q7, Q12, Q15, Q18, Q21, Q28, and Q33. These questions required marking rules, and these were inherited from corresponding free-response questions from version 1 of the AMS, which tested similar content and concepts. **Table 4** shows the results of the Cohen's kappa calculations conducted on data collected from version 2 of the AMS.

From **Table 4**, Q1, Q2, Q4, Q5, Q17, Q20, Q22, Q23, Q25, Q27, Q29, and Q31 all had Cohen's kappa values which were within the acceptable range of [0.8, 1]. It follows that the marking rules were functioning well for these twelve questions. For the other thirteen questions, Q3, Q7, Q11, Q12, Q13, Q15, Q18, Q19, Q21, Q28, Q30, Q32, and Q33, the Cohen's kappa values were outside the acceptable range of values. This implied that the marking rules required further development in these cases. It is of note that seven of these questions, Q7, Q12, Q15, Q18, Q21, Q28, and Q33, were being ran in free-response format for the first time. The marking rules would not be expected to function at the required level for these questions. To check for consistency, the version 2 marking rules were back-tested against the response data gathered from version 1 of the AMS. The results are given in **Table 5**.

15 questions in **Table 5**, Q1, Q2, Q4, Q5, Q13, Q17, Q19, Q20, Q22, Q23, Q25, Q27, Q29, Q31, and Q32, have Cohen's kappa values, which are within the acceptable range of values. This illustrates that the marking rules for these fifteen questions had stabilized after two rounds of iterative development. **Table 5** identifies three questions, Q3, Q11, and Q30, which have values outside the acceptable range of values. These results are consistent with the above findings given in **Table 3** and **Table 4**. The calculations presented in Tables 3, 4 and 5 all identified that the marking rules for Q3, Q11, and Q30 were not functioning at the required level, indicating that the iteration from version 1 to version 2 of the AMS was insufficient to have reached the acceptable Cohen's kappa value for these questions. This finding flagged these questions as possible problematic cases.

The version 3 marking rules were developed by adapting the version 2 marking rules using the strategy outlined in the *Data Analysis–Inter-Rater Reliability* subsection. To facilitate with the data collection process, version 3 of the AMS was split into three sub-tests, version 3A, 3B, and 3C, for administration. The three sub-tests were designed to be of similar lengths, and to test a similar balance of concepts and content. This approach has previously been used in a study conducted by partitioning the traditional multiple-choice FCI

Table 5. Cohen's kappa values for the version 1 UHM against the version 2 computer marking

Question	Number of responses	Cohen's kappa
Q1	328	0.99
Q2	307	1.00
Q3	305	0.73*
Q4	304	0.91
Q5	301	0.98
Q11	280	0.77*
Q13	277	0.93
Q17	276	0.96
Q19	275	0.90
Q20	275	0.96
Q22	275	0.97
Q23	275	0.94
Q25	255	1.00
Q27	255	0.96
Q29	255	0.98
Q30	254	0.39*
Q31	254	1.00
Q32	254	0.92

Table 6. Cohen's kappa values for the version 3 UHM against the version 3 computer marking

Question	AMS version 3 question set	Number of responses	Cohen's kappa (AMS version 3)
Q1	Test 3B	47	1.00
Q2	Test 3B	47	0.95
Q3	Test 3A	118	0.90
Q4	Test 3A	118	0.93
Q5	Test 3A	118	0.98
Q7	Test 3A	118	1.00
Q11	Test 3B	47	1.00
Q12	Test 3B	47	1.00
Q13	Test 3B	47	1.00
Q15	Test 3B	47	0.84
Q17	Test 3B	47	1.00
Q18	Test 3C	36	1.00
Q19	Test 3B	47	1.00
Q20	Test 3C	36	1.00
Q21	Test 3A	118	0.88
Q22	Test 3C	36	0.93
Q23	Test 3C	36	0.73*
Q25	Test 3C	36	1.00
Q27	Test 3C	36	1.00
Q28	Test 3A	118	0.88
Q29	Test 3A	118	1.00
Q30	Test 3A	118	0.79*
Q31	Test 3A	118	1.00
Q32	Test 3C	36	1.00
Q33	Test 3C	36	0.82

into smaller parts (Han et al., 2015, 2016). Results from the Cohen's kappa calculations conducted on data collected from version 3 of the AMS are given in **Table 6**.

From **Table 6**, 23 questions, Q1, Q2, Q3, Q4, Q5, Q11, Q13, Q17, Q19, Q20, Q22, Q25, Q27, Q29, Q31, and Q32, each have values for Cohen's kappa, which are within the acceptable range of values. This shows that the marking rules are functioning at the required level for these twenty-three questions. Only two questions, Q23 and Q30, have values for Cohen's kappa which are outside the acceptable range of values. Hence, the marking rules for version 3 of the AMS were found to be highly functional in the vast majority of cases.

As a check for consistency, the version 3 marking rules were back-tested against the response data from both version 1 and version 2 of the AMS. The results are given in **Table 7**. The blank entries in **Table 7** correspond to questions which were not free-response format in version 1, meaning that there was no response data to make a comparison with the version 3 marking rules on these questions.

Table 7. Cohen's kappa values for version 1 UHM and version 2 UHM against version 3 computer marking

Question	Number of responses	Cohen's kappa (AMS version 1)	Number of responses	Cohen's kappa (AMS version 2)
Q1	328	0.99	81	1.00
Q2	307	0.99	75	1.00
Q3	305	0.70*	74	0.86
Q4	304	0.91	74	0.89
Q5	301	0.98	73	0.96
Q7	-	-	71	1.00
Q11	280	0.92	66	1.00
Q12	-	-	66	0.94
Q13	277	0.94	66	0.86
Q15	-	-	66	0.85
Q17	276	0.96	64	1.00
Q18	-	-	64	0.82
Q19	275	0.89	64	0.84
Q20	275	0.96	64	0.96
Q21	-	-	64	0.77*
Q22	275	0.97	60	0.97
Q23	275	0.85	60	0.96
Q25	255	0.96	60	1.00
Q27	255	0.96	60	0.96
Q28	-	-	60	1.00
Q29	255	0.99	60	0.91
Q30	254	0.76*	60	0.96
Q31	254	0.99	60	1.00
Q32	254	0.91	60	0.81
Q33	-	-	60	0.84

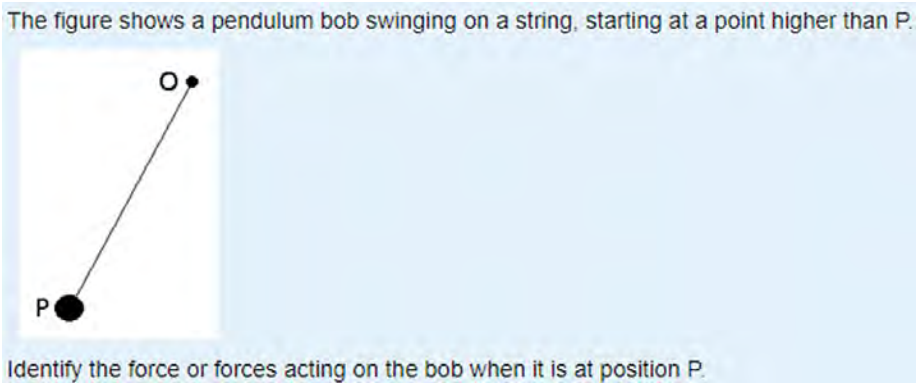

Figure 2. Updated version of Q21 found on the final version of the AMS (Source: Authors' own elaboration)

From **Table 7**, the back-testing of the version 3 marking rules highlights two problem cases when compared with marking data from version 1, and one problem case when compared with the marking data from version 2. These cases were Q3 and Q30 for the version 1 back-testing, and Q21 for the version 2 back-testing. The actions taken for each of these cases were, as follows.

In the case of Q3, the question was flagged as problematic because almost every test-taker who attempted the question provided a correct answer to it. Q3 asks test-takers to identify which forces act on a stone while it is in flight, with the correct answer being 'weight' (or equivalent). Q3 was a new question added to the AMS, meaning that it did not appear in the original multiple-choice FCI. For these reasons, it was decided to remove Q3 from the next iteration of the AMS.

For Q21, it was found that the diagram provided did not correspond to the situation being asked about in the question. Q21 asks test-takers to identify the forces acting on a boy when he is on a swing, with the sought correct answers being 'weight' and 'tension in the rope' (or equivalent). However, the original diagram corresponding to Q21 was misleading, as additional reaction forces could be identified acting between the boy and the swing seat. To rectify this, it was decided to modify the diagram and the question, but to keep Q21 in free-response format in the next iteration of the AMS. The updated Q21 is show below in **Figure 2**. The original Q21 is shown below in **Figure 3**.

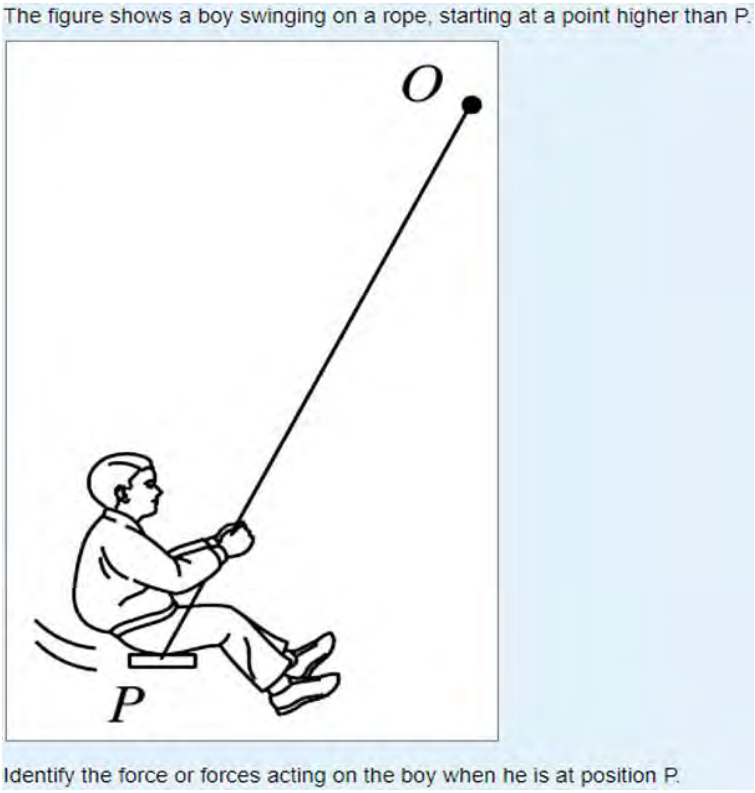

Figure 3. Original version of Q21 found on version 1, version 2, and version 3 of the AMS (Source: Authors' own elaboration)

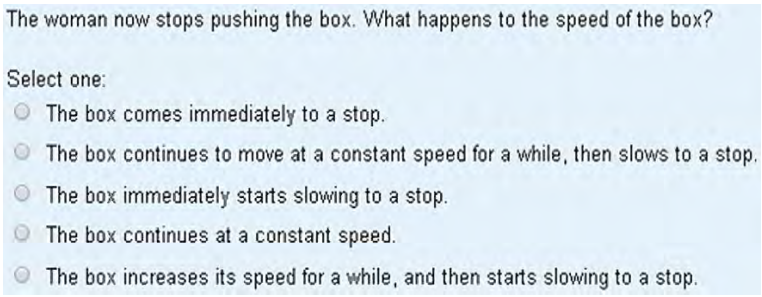

Figure 4. Multiple-choice version of Q30 found on the final version of the AMS (Source: Authors' own elaboration)

A different issue was uncovered with Q30. For Q30, the test-taker was required to state what happens to the speed of a moving box after the constant force being applied it is removed. The correct answer to this question required two parts, 'slows down' and 'stops'. For example, the answer 'the box slows down and stops' would be considered as correct, whereas the answer 'the box stops' would be considered as incorrect. It was found that the marking rules were incapable of differentiating between correct and incorrect answers because of the level of description required in order to provide a correct answer to the question. Despite three rounds of testing using a total of 432 responses, this problem had persisted throughout the AMS development process. To resolve this issue, Q30 was returned to multiple-choice format for the next iteration of the AMS. The multiple-choice version of Q30 is shown in **Figure 4**.

Using the strategy outlined in the *Data Analysis–Inter-Rater Reliability* subsection, the version 3 marking rules were developed into the final version marking rules, which was the terminal iteration in the AMS development process. The wordings of questions Q11, Q21, and Q23 were modified for the final version, and questions Q3 and Q19 were removed. Since this was the terminal iteration of the AMS, no new responses were collected to further develop and test the AMS. To test for consistency, the marking rules for the questions

Table 8. Cohen's kappa values for version 1 UHM and version 2 UHM against final version computer marking

Question	Number of responses	Cohen's kappa (AMS version 1)	Number of responses	Cohen's kappa (AMS version 2)
Q1	328	0.99	81	1.00
Q2	307	0.99	75	1.00
Q3	305	-	74	-
Q4	304	0.91	74	0.89
Q5	301	0.98	73	0.96
Q7	-	-	71	1.00
Q11	280	-	66	-
Q12	-	-	66	0.94
Q13	277	0.94	66	0.86
Q15	-	-	66	0.85
Q17	276	0.96	64	1.00
Q18	-	-	64	0.86
Q19	275	-	64	-
Q20	275	0.96	64	0.96
Q21	-	-	64	-
Q22	275	0.98	60	1.00
Q23	275	-	60	-
Q25	255	0.96	60	1.00
Q27	255	0.96	60	0.96
Q28	-	-	60	1.00
Q29	255	0.99	60	0.91
Q30	254	-	60	-
Q31	254	0.99	60	1.00
Q32	254	0.91	60	0.81
Q33	-	-	60	-

on the final version were back-tested against responses from version 1 and version 2 of the AMS, where applicable. The results of this back-testing are given in **Table 8**.

The blank entries in **Table 8** are explained as follows. In the cases of Q17, Q12, Q15, Q18, and Q28, the questions were not free-response in version 1, meaning that there was no response data to compare the final version marking rules with. For the cases of Q3 and Q19, these questions were removed from the final version of the AMS, meaning that back-testing calculations did not need to be run for these questions. In the cases of Q11, Q21 and Q23, the wording of the questions was changed for the final version of the AMS. The responses gathered to the version 1 and version 2 equivalents of these questions did not fully correspond to the Final Version wording of the questions, meaning that back-testing calculations would not have produced meaningful statistical results. In the cases of Q30 and Q33, the questions were converted to multiple-choice format for the final version of the AMS, meaning that back-testing calculations were not required.

The final version of the AMS contains 31 questions, including 21 free-response questions and 10 multiple-choice questions. From **Table 8**, the back-testing of the final version marking rules had no problem cases when compared with marking data from both version 1 and version 2 of the AMS. This indicated that the marking rules had stabilized for the 19 AMS questions with back-testing data. Since the AMS marking rules had gone through four iterations, this outcome would be expected from the final version of the AMS marking rules. For reference, changes made between different versions of the AMS are catalogued in **Table 9**.

The main research output of this work was to investigate the reliability of the AMS, a concept inventory which was constructed using computer-marked free-response questions. The findings from the CTT analysis showed that the AMS questions were reliable, and the findings from the IRR analysis showed that the AMS marking rules were reliable. Future work aims to use the AMS system on a larger scale in order to collect further data which can be used to investigate student understanding of physics concepts.

LIMITATIONS

The AMS The AMS questions were marked using a discrete marking metric. This means that responses were marked as either correct or incorrect, with no option for partial credit. It follows that the corresponding

Table 9. Changes made between different versions of the AMS

Iteration of the AMS	Changes made to questions	Changes made to marking rules
Version 1 to version 2	Q7, Q12, Q15, Q18, Q21, Q28, & Q33 converted from multiple-choice to free-response questions.	Marking rules for all free-response questions updated.
Version 2 to version 3	AMS question set split into three sub-tests, version 3A, 3B, & 3C, to facilitate with data collection.	Marking rules for all free-response questions updated.
Version 3 to final version	Wordings of Q11, Q21, & Q23 were modified. Q3 & Q19 were removed. Q30 & Q33 converted from free-response to multiple-choice questions.	Marking rules unchanged for free-response questions.

CTT, and IRR calculations gave single numbers as outputs, with no additional measure of confidence given for the result. This is acknowledged as a limitation of the study. For similar studies in the future, this effect could be mitigated by using marking schemes which use continuous marking metrics, as these would allow for partial credit to be given. Further, the AMS could be extended to give test-takers the option to explain their answers and to rate the confidence of their answers, as advocated by the work of Smith and Tanner (2010). Such approaches were beyond the scope of the current work.

Version 1 of the AMS had *N*=254 competed attempts, whereas version 2 of the AMS had *N*=60 completed attempts. This shows that a smaller sample size was used in the version 2 study. Because CTT is grounded in the objective of testing whether tests and their questions are reliable, it follows that it is preferable to have more completed attempts to analyze. This implies that on a qualitative level, the CTT calculations for version 1 of the AMS would be expected to be more reliable than those for version 2 of the AMS. For the IRR statistics, computer marking with higher effectiveness can be developed when more responses are used to develop the marking rules (Butcher & Jordan, 2010). This point was highlighted by the underperformance of some of the marking rules on version 2 of the AMS, as these were developed using a smaller number of responses.

The collection of responses was limited by the availability and suitability of student cohorts at contacted high schools and universities. Collecting more responses would have been useful for further development and testing of the AMS. In particular, further testing could be of benefit for the cases of Q11, Q21, and Q23, as the wording of these questions was modified for the final version of the AMS. Further investigation of Q30 and Q33, which were found to be unsuitable to be run in free-response format, could prove beneficial for future attempts to establish other concept inventories using free-response questions. Considerations about question wording and the suitability of questions to be run in free-response format could serve as avenues for future work.

The iterative approach used to develop and test the AMS marking rules has limitations. For the strategy of adding new marking rules to negate cases of false positive marking, there exists the possibility of correct answers being caught by the new rules and marked as incorrect (Butcher & Jordan, 2010). This can lead to new false negative cases arising. Striking a balance between negating false positives and false negatives is not straightforward in such cases. Different concerns arise when using false negative cases to add new marking rules to include a wider range of correct answers. Through the open-ended free-response format, there are many different ways which students can give a correct answer. This makes it difficult to develop marking rules to cover all possible correct answers (Sychev et al., 2020). Furthermore, developing marking rules to account for every answer wording could make the marking rules specific to the answer set used to develop them, which raises the possibility of an over-fitting concern (Zehner et al., 2016). Further responses would be required to investigate the limitations of the rule development process outlined above.

CONCLUSIONS

The study investigated two research questions:

1. **RQ1:** To what extent is a free-response modified version of the FCI reliable?
2. **RQ2:** How reliable are automated marking schemes when used to mark free-response versions of FCI questions?

The first aim of the study was to investigate the reliability of FCI questions when posed in free-response format (RQ1); the second aim of the study was to investigate the reliability of automated marking schemes when used to mark these free-response versions of FCI questions (RQ2). Data were collected for the study by

having 610 participants work through the AMS. A total of 8,091 question responses were gathered to develop and test the AMS.

To answer RQ1, response data for version 1 and version 2 of the AMS were collected from students during the academic years 2017-2018 and 2018-2019. To test the AMS questions for reliability, CTT statistics were calculated for the version 1 and version 2 responses data sets. The CTT analysis found that the questions on version 1 and version 2 of the AMS functioned well, meaning that the free-response AMS questions were overall reliable. Furthermore, the CTT calculations were shown to be consistent for both version 1 and version 2 of the AMS, which demonstrated that the AMS questions were stable and did not require further development.

To answer RQ2, response data for version 1, version 2, and version 3 of the AMS were collected from students during the academic years 2017-2018, 2018-2019, and 2019-2020. In order to test the reliability of the marking rules for the free-response AMS questions, IRR statistics were calculated for the response data collected from the free-response questions on version 1, version 2, and version 3 of the AMS. The IRR calculations found that the marking rules progressively improved in effectiveness throughout the iterative development process of the AMS, and the marking rules were found to be reliable overall for the free-response questions on the final version of the AMS. In addition, successful IRR back-testing against responses from previous academic years illustrated that the marking rules on the final version of the AMS were consistent.

Taken together, the findings from the CTT and IRR studies found that the AMS questions and marking rules were overall reliable. Thus, the final version of the AMS was established as a physics concept inventory which contains automatically-marked, free-response questions. Investigating the reliability of the AMS was the main research output of this work and has established the principle of implementing questions of this type within concept inventories, in addition to producing an inventory suitable for practical use. Future work aims to use the AMS system on a larger scale in order to collect further data which can be used to investigate student understanding of physics concepts. The approach used to successfully develop and test the AMS could be used to guide future attempts to develop other concept inventories which make use of automatically-marked, free-response questions.

Author contributions: MAJP: designed the research method, conducted the research, and wrote the article body text & **HH, SEJ, & NSJB:** supervised the research and provided corrections and comments to improve the article body text. All authors approve final version of the article.

Funding: This article was supported by the UK Open University.

Acknowledgements: The authors would like to thank David Sands, Ross Galloway, and Christine Leach for their work, which was important for the initial setting up of the concept inventory used during the study. The authors also would like to thank the participants who answered the questions on the various different versions of the AMS, as their participation was essential to the success of the research.

Ethics declaration: This study was approved by The Open University Human Research Ethics Committee (the approval code: HREC/2017/2629/Parker/1) on July 31, 2017.

Declaration of interest: Authors declare no competing interest.

Data availability: The data underpinning the contents of this article are publicly available in the PhD thesis "Establishing physics concept inventories using free-response questions", which can be found at: https://oro.open.ac.uk/73254/.

REFERENCES

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers and Education, 55*, 489-499. https://doi.org/10.1016/j.compedu.2010.02.012

Cohen, J. (1960). A coefficient for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. https://doi.org/10.1177/001316446002000104

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Wadsworth Group/Thompson Learning.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*, 020103. https://doi.org/10.1103/PhysRevSTPER.5.020103

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

Doran, R. (1980). *Basic measurement and evaluation of science instruction*. NSTA.

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

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

Han, J., Bao, L., Chen, L., Cai, T., Pi, Y., Zhou, S., Tu, Y., & Koenig, K. (2015). Dividing the force concept inventory into two equivalent half-length tests. *Physical Review Special Topics-Physics Education Research, 11*, 010112. https://doi.org/10.1103/PhysRevSTPER.11.010112

Han, J., Koenig, K., Cui, L., Fritchman, J., Li, D., Sun, W., Fu, Z., & Bao, L. (2016). Experimental validation of the half-length force concept inventory. *Physical Review Special Topics-Physics Education Research, 12*, 020122. https://doi.org/10.1103/PhysRevPhysEducRes.12.020122

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. In *Proceedings of Computer Assisted Assessment 2012 International Conference*.

Jordan, S. (2012). Short-answer e-assessment questions: Five years on. In *Proceedings of the 2012 International Computer Assisted Assessment Conference*.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. Methuen.

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

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

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

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

Scott, T. F., & Schumayer, D. (2017). Conceptual coherence of non-Newtonian worldviews in force concept inventory data. *Physical Review Physics Education Research, 13*, 010126. https://doi.org/10.1103/PhysRevPhysEducRes.13.010126

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

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

Sychev, O., Anikin, A., & Prokudin, A. (2020) Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research, 59*, 264-272. https://doi.org/10.1016/j.cogsys.2019.09.025

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

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

Zehner, F., Salzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement, 76*(2), 280-303. https://doi.org/10.1177/0013164415590022

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

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*(3), 374-378. https://doi.org/10.1037/0033-2909.103.3.374