CAA 2012

Short-answer e-assessment questions: five years on

Sally Jordan

The Open University

sally.jordan@open.ac.uk

Abstract

Computer-marked short-answer questions, with tailored feedback, have been in regular use in the Open University Science Faculty for 5 years. High marking accuracy has been retained following a move to algorithmically-based answer matching. However, short-answer free-text e-assessment remains an underused technology. This paper suggests some reasons for this, in particular the need for a substantial number of marked student responses for use in developing the answer matching.

Short-answer e-assessment questions, with tailored feedback, have been in regular use on two high-population Open University Science Faculty modules since 2007. About 24 of these questions are currently in use, alongside other question types, with around 5000 individual student users per year.

Short-answer questions have the advantage of requiring students to construct a response for themselves, rather than selecting from a number of predetermined options (Jordan & Mitchell, 2009). Evaluation has shown that our answer matching is accurate (Butcher & Jordan, 2010) and that students engage well with the questions and the feedback provided (Jordan, 2012). But yet, short-answer free-text e-assessment remains an underused technology, at the OU and elsewhere. This paper's aim is to explore some of the reasons for this and to challenge some misconceptions about short-answer free-text questions.

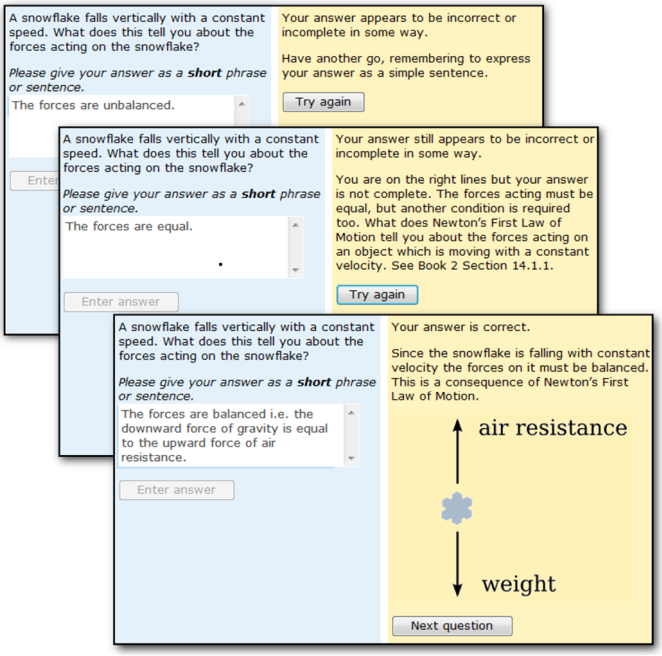
What is a short-answer question?

Our 'short-answer' questions are intended to generate answers of around a sentence in length. In early trials, a few students were found to submit very long answers (up to around 100 words), which were difficult to mark because they frequently included both correct and incorrect aspects, acknowledged by others (e.g. Mitchell et al, 2002) as a problem area for automated marking. For this reason, since 2009, answers have been restricted to no more than 20 words (Jordan, 2012).

Questions of this type, and the technology required to mark them, are characteristically different from those where answers are expected in the form of essays. Essay marking software such as E-rater (Attali & Burstein, 2006) marks writing style, whilst the focus in short-answer questions is on the content of the answer.

Software for marking short-answer questions includes C-rater (Leacock & Chodorow, 2003), Intelligent Assessment Technologies (IAT) Free Text Author system (Jordan & Mitchell, 2009) and the system developed by Sukkarieh, Pulman & Raikes (2003, 2004). These systems, reviewed by Siddiqi & Harrison (2008), are all based to some extent on computational linguistics. The IAT software, used at the OU from 2006-2009, draws on the natural language processing (NLP) techniques of information extraction, but provides an authoring tool that can be used by a question author with no knowledge of NLP.

The IAT software sat within OpenMark[1], allowing multiple attempts at each question, with an increasing amount of feedback provided after each unsuccessful attempt (Figure 1).



A snowflake falls vertically with a constant speed. What does this tell you about the forces acting on the snowflake?

*Please give your answer as a **short** phrase or sentence.*

The forces are unbalanced.

Your answer appears to be incorrect or incomplete in some way.

Have another go, remembering to express your answer as a simple sentence.

A snowflake falls vertically with a constant speed. What does this tell you about the forces acting on the snowflake?

*Please give your answer as a **short** phrase or sentence.*

The forces are equal.

Your answer still appears to be incorrect or incomplete in some way.

You are on the right lines but your answer is not complete. The forces acting must be equal, but another condition is required too. What does Newton's First Law of Motion tell you about the forces acting on an object which is moving with a constant velocity. See Book 2 Section 14.1.1.

A snowflake falls vertically with a constant speed. What does this tell you about the forces acting on the snowflake?

*Please give your answer as a **short** phrase or sentence.*

The forces are balanced i.e. the downward force of gravity is equal to the upward force of air resistance.

Your answer is correct.

Since the snowflake is falling with constant velocity the forces on it must be balanced. This is a consequence of Newton's First Law of Motion.

Figure 1. Increasing feedback (top left to bottom right) on a typical short-answer free-text question.

[1] OpenMark Examples, http://www.open.ac.uk/openmarkexamples/ (Accessed 7th April 2012)

In 2009, the OU swapped to OpenMark's 'PMatch' for marking short-answer questions. PMatch does not rely on computational linguistics, but rather on matching simple sequences of words. For example, one of the rules for the 'Metamorphic' question (see Appendix A for all the questions discussed in this paper) might be expressed in everyday language as

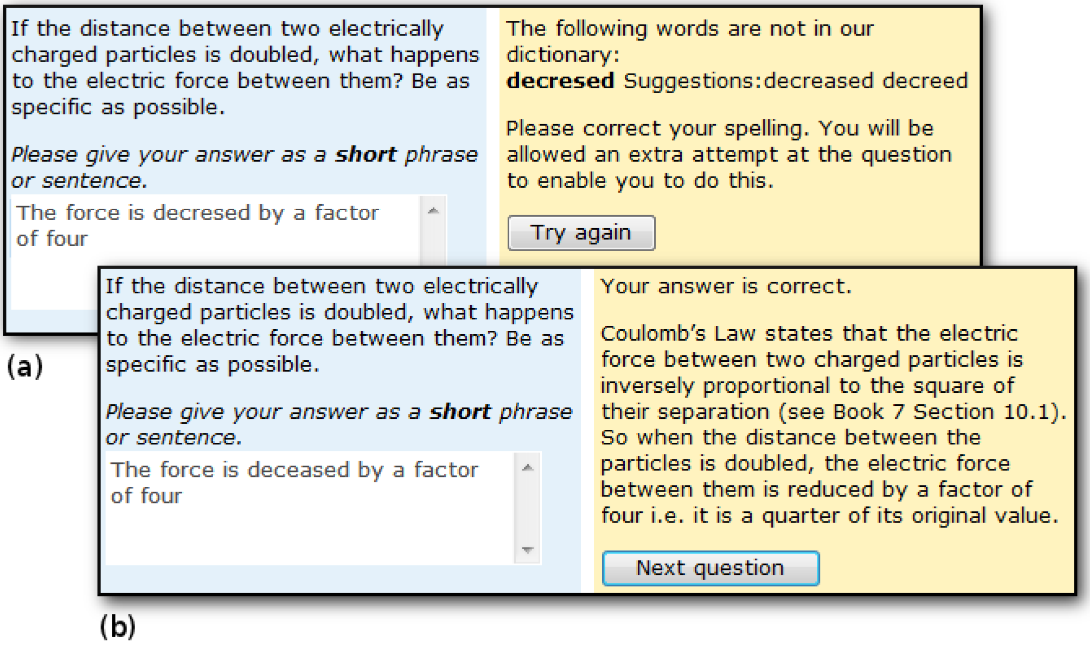
> *Accept answers that include the words 'high', 'pressure' and 'temperature' or synonyms, separated by no more than three words.*

or in PMatch's Java as

> *m.match("mowp3","high|higher|extreme|inc&|immense_press&|compres&|[deep_burial]_temp&|heat&|[hundred|100_degrees]")*

The 'Metamorphic' question has 10 rules of this type, used to mark students' responses and to generate feedback. Other questions have between 2 and 23 rules for marking plus up to 6 rules used purely to generate feedback.

Within words of more than three letters, single incorrect, transposed, missing or extra letters can be allowed (Figure 2b). In addition, a spell-checker informs students when a word they have used is not recognised by the spell-checker's dictionary (to which scientific words can be added if necessary), and offers suggestions for correct spelling (Figure 2a).



If the distance between two electrically charged particles is doubled, what happens to the electric force between them? Be as specific as possible.

Please give your answer as a ***short*** *phrase or sentence.*

The force is decresed by a factor of four

The following words are not in our dictionary:
decresed Suggestions:decreased decreed

Please correct your spelling. You will be allowed an extra attempt at the question to enable you to do this.

If the distance between two electrically charged particles is doubled, what happens to the electric force between them? Be as specific as possible.

Please give your answer as a ***short*** *phrase or sentence.*

The force is deceased by a factor of four

Your answer is correct.

Coulomb's Law states that the electric force between two charged particles is inversely proportional to the square of their separation (see Book 7 Section 10.1). So when the distance between the particles is doubled, the electric force between them is reduced by a factor of four i.e. it is a quarter of its original value.

Figure 2. Two methods for dealing with incorrect spelling. (a) A spell-checker identifies words not in a dictionary. (b) Similar but incorrect words can be accepted (in this case "deceased" instead of "decreased").

A pattern matching question type 'Pattern match', based on OpenMark's PMatch, has recently been released into the Moodle 2.1 quiz engine[2].

[2] OUeAssessment_1.0, http://labspace.open.ac.uk/course/view.php?id=3484 (Accessed 7th April 2012)

Analyses of marking accuracy

Human-computer marking comparison

Surely computers can never mark questions of this type as accurately as human-markers?

In 2007, the IAT marking of between 92 and 248 student responses to each of seven short-answer questions was compared with that of six course tutors. The results are reported in detail in Butcher & Jordan (2010) and Table 1 includes a summary of the results for three questions that were also included in the 2012 analysis described below. For each question, the marking agreement with the question author is expressed as a percentage of responses for which the same score was given and also as a kappa (□) inter-rater statistic (Cohen, 1960). The kappa statistic is calculated from the formula

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \qquad \text{(Equation 1)}$$

where *P(a)* is the proportion of times the marker agrees with the question author and *P(e)* is the proportion of times they would be expected to agree by chance. A Kappa statistic of greater than 0.8 is usually taken to indicate excellent agreement.

Table 1. Summary of previously reported results from a human-computer and a computer-computer marking comparison, updated with 2012 data.

Question	**Previous results (Butcher & Jordan, 2010)**				**2012 results**	
	Number of responses	**% agreement with author (□ inter-rater statistic)**			**Number of responses**	**% agreement with human 'expert'** **(□ inter-rater statistic)**
		Range for 6 human markers	**IAT**	**PMatch**		
Direction	189	97.4 to 100 (0.92 to 1)	99.5 (0.98)	99.5 (0.98)	1767	99.3 (0.97)
Intrusive	92	92.4 to 97.8 (0.79 to 0.95)	98.9 (0.97)	96.7 (0.91)	1616	98.2 (0.84)
Snowflake	248	83.9 to 97.2 (0.75 to 0.94)	97.6 (0.95)	98.8 (0.97)	2218	98.4 (0.97)

The results shown in Table 1 are typical of those for all the questions included in the 2007 human-computer marking analysis. The computer marking was always at least as accurate as the worst of the human markers and sometimes it was more accurate than all the markers. Although this result was initially surprising, it is in line with frequently reported concerns about the accuracy of human marking (e.g. Garner, 2010). Disagreement between the human markers and the question author was found to arise from uncertainty over how to mark 'borderline' responses (in this study, all questions were simply marked as either right or wrong), errors in marking caused by insufficient subject knowledge, and slips. However, the most startling finding was the variation between individual markers. For one question, 64% of all student responses were marked as correct by at least one human marker and incorrect by another. Conole & Warburton (2005) identified consistency as an advantage of computer marking.

PMatch's marking accuracy

Surely you need sophisticated technology?

Table 1 also includes data from an earlier comparison of the marking accuracy of IAT and PMatch. The startling finding that answer matching developed in PMatch was as accurate as that developed in IAT was the primary reason for our decision to swap from IAT to PMatch answer matching in 2009.

The purpose of the 2012 analysis was to check that PMatch answer matching was still accurate, on a range of questions in regular use. Eleven short-answer free-text questions were selected, on the basis of their use by large numbers of students from the module *Exploring science* during 2011. For each of the questions, between 1591 and 2218 responses were marked by a single human marker with good knowledge of the question author's intentions. It is acknowledged that the human marker will have made some mistakes, but this is ignored in the current study. The human marker indicated when he was unsure about his grading of a response; this uncertainty gives some indication (probably an underestimate) of the proportion of responses that different markers might mark differently or indeed of responses that one marker might mark differently on a different occasion. The human 'expert's' marking was compared with that of PMatch.

Some data for the three questions that were also included in the earlier analyses is included in Table 1 to ease comparison. More detailed results for all the questions are given in Table 2.

Although there is some variation between questions, the PMatch marking accuracy remained good, and there was always a higher proportion of responses over which the human marker was in doubt than responses where PMatch marked differently from the human marker. PMatch's least well performing questions, especially 'Diffraction grating' and 'Ripple tank', tended to be those where the human marker was most likely to be in doubt. Many of the false positives (responses that PMatch marked as correct but the human considered to be incorrect) and false negatives (responses that PMatch marked as incorrect but the human considered to be correct) were responses where the human marker was actually in doubt. For example, for 'Ripple tank', the marker was in doubt in 21 of the reported 27 false positives and 15 of the 27 false negatives. In the most extreme case, 'Sandstone', all of the 9 false positives and 8 false negatives were flagged as uncertain by the human marker.

The kappa inter-rater statistic was greater than 0.9 for all but one question. The kappa statistic for 'Intrusive' (□ = 0.84) was lower than would be expected on the basis of percentage agreement, because the question was very well answered by students (95% of students got the question right at first attempt) so the computer and human were more likely to agree by chance. This is reflected in the kappa statistic.

Table 2. Comparison of the PMatch marking of 11 questions with that of a human 'expert'.

Question	**Number of responses**	**% agreement with human 'expert'**	**□ inter-rater statistic**	**% of all responses which were false positives**	**% of all responses which were false negatives**	**% of all responses where human marker was unsure**
Ball	2092	98.9	0.97	0.7	0.4	3.6
Diffraction grating	1938	97.6	0.94	1.7	0.7	7.2
Direction	1767	99.3	0.97	0.6	0.06	5.0
Electric force	1591	98.4	0.96	0.6	1.0	3.4
Intrusive	1616	98.2	0.84	0.4	1.4	4.7
Kettle	1725	97.9	0.94	0.2	1.9	3.7
Metamorphic	1727	99.1	0.92	0.2	0.7	1.8
Ripple tank	1706	96.8	0.92	1.6	1.6	6.4
Sandstone	1872	99.1	0.98	0.5	0.4	3.3
Slide	1862	98.6	0.97	0.6	0.8	3.5
Snowflake	2218	98.4	0.97	1.0	0.6	3.9

Examples of responses that may cause difficulty for computer marking

Responses which are 'borderline', in particular where a response includes aspects of both a correct and an incorrect response, have already been identified as problematic for human and computer markers alike, though a human perhaps has the advantage of being able to identify an incorrect statement that shows that the student does not understand the topic. For example, in answer to 'Ripple tank', several responses such as the following were received:

> *The spread of the dots increases as the wavelength is increased and the spacing decreases.*

Whilst it is true that the diffraction pattern will decrease in the scenario described in the question, this response appears to relate to the diffraction of light by a diffraction grating rather than the diffraction of water in a ripple tank!

Some similarly bizarre answers had already been accounted for following earlier analyses of responses. For example, in answer to 'Slide', responses which described motion on a swing e.g.

> *When the speed is greatest – which is at the lowest point in the swing cycle.*

were marked as incorrect, with feedback including the following:

> Note that this question is about a child on a slide not a swing.

The answer matching for all the questions had been developed iteratively, using responses from students taking the module *Exploring science* or its predecessor. This reduced the number of cases where a response was inaccurately matched simply because a synonym or different way of expressing an answer had not been seen in earlier student responses. However, there were a few correct responses to 'Kettle' which were missed for this reason, for example

> *It remains static.*
>
> *It does not rise anymore.*

For the future, responses such as these can easily be matched by adding 'static' as a synonym for 'same, constant, steady, stable' (all of which were already matched) and 'It does not rise' as another way of saying 'It does not increase' or 'It does not change' (already marked correct).

Which PMatch rules were used in matching the student responses

Although the answer matching for some of the questions included a large number of rules, in every case some rules were found to be much more effective than others. Figure 3 shows the number of responses matched by the most frequently triggered rules for 'Ball' and 'Snowflake'. Green colouring indicates a rule that marked a response as correct; red colouring indicates a rule that marked a response as incorrect.

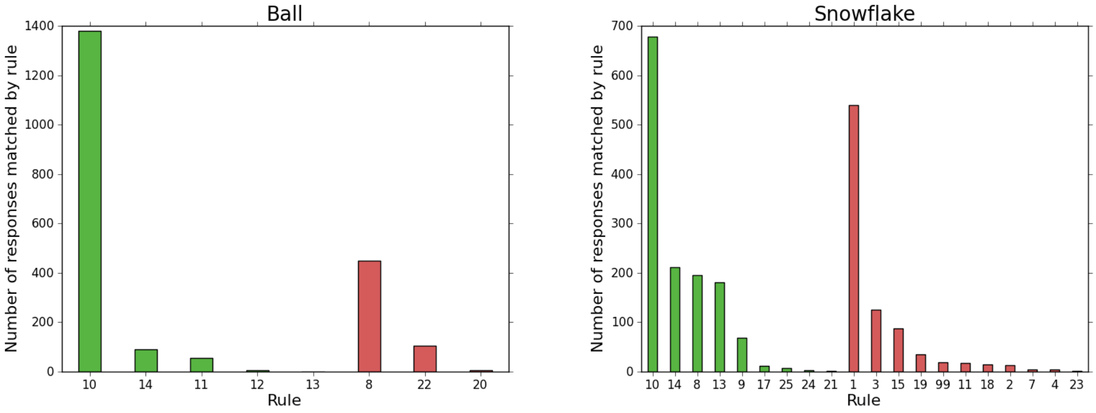


Figure 3. The most commonly triggered rules for 'Ball' (which had 7 rules to match correct responses and 2 rules to match incorrect responses) and 'Snowflake' (which had 10 rules to match correct responses and 13 rules to match incorrect responses).

Discussion

What makes a good question?

In principle, answer matching can be written for any question that has distinct correct and incorrect answers, although when there are many ways of expressing a correct (or incorrect) response, developing the answer matching rules becomes tedious. The 23 rules in 'Snowflake' is undoubtedly a lot, but if we had been prepared to accept a slightly less accurate overall match, around 10 rules would probably have been adequate.

There is perception that short-answer questions are more difficult for students than other question types. Figure 4 shows the proportion of questions that students got right at first, second and third attempt, for all short-answer questions in one presentation of *Exploring science*, in contrast with all multiple-choice type questions and all questions requiring a numerical answer. It may be that our short-answer questions are just too easy: they do not appear to present our students with much difficulty.

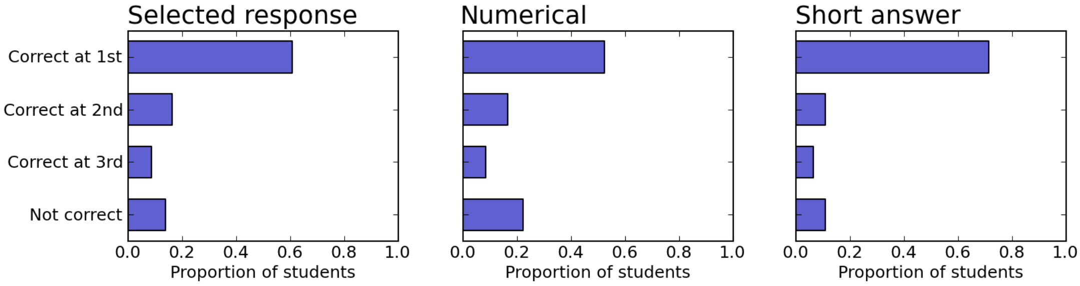


Figure 4. The proportion of students getting questions right at first, second or third attempt, or not at all, for different question types.

It is sometimes argued that short-answer questions can assess only low-level content-related learning outcomes. Although answers must be short, it is possible to write imaginative questions which require students to interpret information that has been given to them (e.g. in 'Electric force', 'Sandstone' and 'Snowflake') rather than simply reproducing content taught in the module.

Figure 4 shows that short-answer questions are well answered by students, and Tables 1 and 2 show that the answer matching is accurate. However, it is also important that students have confidence in the computer's marking. This means that it is important to strive for accurate answer-matching (even when the question is in formative-only use) and when a response is marked incorrect because it *is* incorrect, the student needs to understand this, rather than assuming that the computer has 'got it wrong'. In practice, this highlights the importance of providing students with targeted feedback (Jordan, 2012).

Although PMatch's answer matching is based on simple rules, it is important to include rules to cope with negation and, in some cases, word order. In 'Snowflake', answers of 'There are no unbalanced forces acting' and 'The forces are balanced' must both be marked correct, whilst 'The forces are unbalanced' and 'The forces are not balanced' must both be marked as incorrect. In 'Ball', 'Kinetic energy is converted to gravitational energy' is correct, but 'Gravitational energy is converted to kinetic energy' is incorrect.

What are the real barriers to wider-take up?

How long does it take to develop the answer matching?

The original answer matching for each of the questions described in this paper was written in a matter of hours. However, the answer matching was then refined in response to answers from real students (initially those studying the predecessor module). It is vitally important to use real student responses – students express their answers in different ways from lecturing staff.

However, the iterative development of answer matching adds to the time required, and makes this difficult to quantify. There is a tension between accepting that a question's answer matching is 'good enough' – and no doubt matches the vast majority of student responses – and continuing to strive towards perfection. In practice, the questions analysed in this paper appear to be in good enough shape to be used without further intervention for the life of the module (which has two presentations of around 2000 students each year, and would normally be expected to run for 10 years). So the considerable time that has gone into their development can be justified. This may be more difficult to justify for modules with smaller student numbers.

The answer matching for the questions described in this paper was written by the author (who has no experience of computer programming but has a background in mathematics and physical science) and the OU's E-assessment Adviser. Workshops have been held to demonstrate PMatch to others, but as yet the software has not be much used. It is likely that some academics will find it difficult to write sufficiently rigorous answer matching, especially given a general reluctance to move beyond multiple-choice questions (Hunt, 2012). This raises questions about who should be authoring sophisticated e-assessment items of this type.

Machine learning offers potential for removing the drudgery from the development of answer matching. However, responses still need to be marked by a human marker in the first instance, to provide a dataset for the computer to use in generating rules. Other users of short-answer free-text questions (e.g. Mitchell et al, 2003) have used human-marked responses from different assessment tasks, e.g. examinations.

How many student responses do you need?

Mitchell et al. (2003) used paper-based marking guidelines and approximately 50 marked student scripts in developing their answer matching whilst Sukkarieh et al. (2003) used approximately 200 marked student answers per question for training, and approximately 60 answers per question for testing. Our experience it that the number of responses required to develop sufficiently robust answer matching varies from question to question, but is usually measured in hundreds of responses. A larger training set will increase the accuracy of subsequent marking (Butcher & Jordan, 2010). Even for answer matching based on the inspection of thousands of responses, there will be occasional answers that are incorrectly matched (for example, the two correct responses to 'Kettle' given above) simply because they are expressed in a way that has not been seen before.

For modules with large students numbers, long lifetimes, and some way of capturing student responses for use in question development, answer matching can be iteratively refined for as long as the question author has the time and motivation to do so. The issue is one of deciding how far it is appropriate to go.

Modules with small student numbers face a more serious problem. We contend that accurate answer matching cannot be developed without a substantial corpus of real student responses. If you are prepared to collect and human mark responses for a number of years prior to developing your answer matching rules for subsequent automatic use or to develop answer matching iteratively, accepting that your initial computer marking may be quite inaccurate, there may be a way forward. However, large student numbers provide both the student responses necessary for question development, and the justification, in terms of future marking time saved, for the time spent in development.

Conclusion

In their paper 'The unreasonable effectiveness of data', Halevy, Norvig & Pereira (2009) refer to datasets measured in trillions, used by Google, for example in translating text. The datasets described in the current paper are only measured in thousands, however it is these relatively large numbers of student responses that have led to the remarkable marking accuracy achieved by both IAT and PMatch.

The title of Halevy et al's paper is a deliberate reference to an earlier paper by Eugene Wigner 'The unreasonable effectiveness of mathematics…' (1960) in which Wigner reflects on the fact that much of physics can be explained with simple mathematical formulae such as $F = ma$. We agree with Halevy et al that human behaviour cannot be so neatly described. However provided you 'follow the data' (Halevy et al., p.12), simple algorithmically-based answer matching can provide robust and effective matching for short-answer free-text e-assessment questions.

Acknowledgements

The work described in this paper was partially funded by the Centre for Open, Learning of Mathematics, Science, Computing and Technology and this funding is gratefully acknowledged, as is the invaluable assistance of Barbara Brockbank, Phil Butcher, Tom Mitchell, Richard Jordan and the six course tutors who participated in the human-computer marking comparison.

References

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

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

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational & psychological measurement*, *20*(1), 37-46.

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

Garner, R. (2010) Don't trust exam results, says marking expert. *The Independent*, 22nd April 2010.

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

Hunt, T. (2012) Computer-marked assessment in Moodle: past, present and future. In *Proceedings of the International Computer Assisted Assessment (CAA) Conference, 10th-11th July 2012*. Southampton: University of Southampton.

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

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

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

Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference, 9th-10th July 2002*. Loughborough: Loughborough University.

Mitchell, T., Aldridge, N., Williamson, W. & Broomhead, P. (2003). Computer based testing of medical knowledge. In *Proceedings of the 7th International Computer Assisted Assessment (CAA) Conference, 7th-8th July 2003*. Loughborough: Loughborough University.

Siddiqi, R. & Harrison, C.J. (2008). On the automated assessment of short free-text responses. Paper presented at the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge, UK.

Sukkarieh, J.Z., Pulman, S.G. & Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free-text responses. Paper presented at the 29th International Association for Educational Assessment (IAEA) Annual Conference, Manchester, UK.

Sukkarieh, J.Z., Pulman, S.G. & Raikes, N. (2004). Auto-marking 2: using computational linguistics to score short, free-text responses. Paper presented at the 30th International Association for Educational Assessment (IAEA) Annual Conference, Philadephia.

Wigner, E.P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure & Applied Mathematics*, *13*(1), 1-14.

Appendix A The questions discussed in the paper

Ball

A ball is thrown vertically upwards into the air. What is the main energy conversion that takes place as the ball rises towards its maximum height?

Note: you should only consider the energy conversions that take place after the ball has been thrown and before it reaches its maximum height.

Diffraction grating

A red laser beam is shone through a diffraction grating (as shown in the *Making Waves* video sequence in Book 3 Activity 9.1.). What would be the effect on the observed diffraction pattern of replacing the diffraction grating with one in which the lines are closer together (i.e. *d* is smaller)?

Direction

What does an object’s velocity tell you that its speed does not?

Electric force

If the distance between two electrically charged particles is doubled, what happens to the electric force between them? Be as specific as possible.

Intrusive

Why do intrusive igneous rocks have larger crystals than extrusive ones?

Kettle

Water is heated in an electric kettle until it boils. When the water is boiling, what happens to its temperature?

Metamorphic

Metamorphic rocks are existing rocks that have 'changed form' (metamorphosed) in a solid state. What conditions are necessary in order for this change to take place?

Ripple tank

Water waves in a ripple tank are diffracted by a narrow aperture, as shown in the *Making Waves* video sequence in Book 3 Activity 9.1. What would be the effect of increasing the width of the aperture?

Sandstone

A sandstone observed in the field contains well-sorted, well-rounded, fine pitted and reddened grains. What does this tell you about the process that led to the deposition of this rock and the environment in which it formed?

Slide

A boy climbs slowly to the top of a slide and then slides down it. At which point will his kinetic energy be a maximum?
Note: Your answer should ignore the effects of friction.

Snowflake

A snowflake falls vertically with a constant speed. What does this tell you about the forces acting on the snowflake?