Educ Stud Math (2007) 66:335–348
DOI 10.1007/s10649-006-9078-5

Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer

John Threlfall · Peter Pool · Matthew Homer · Bronwen Swinnerton

Published online: 22 May 2007

Abstract This article explores the effect on assessment of 'translating' paper and pencil test items into their computer equivalents. Computer versions of a set of mathematics questions derived from the paper-based end of key stage 2 and 3 assessments in England were administered to age appropriate pupil samples, and the outcomes compared. Although in most cases the change to the different medium seems to make little difference, for some items the affordances of the computer profoundly affect how the question is attempted, and therefore what is being assessed when the item is used in a test. These differences are considered in terms of validity and legitimacy, that is whether the means used to answer a question in a particular medium are appropriate to the assessment intention. The conclusion is not only that translating paper and pencil items into the computer format sometimes undermines their validity as assessments, it is also that some paper and pencil items are less valid as assessments than their computer equivalents would be.

Key words Assessment · Computer-based · Key stage tests · Mathematics assessment · Paper and pencil · Validity

1 Introduction

Assessment of mathematics through the medium of paper and pencil, as exemplified in the end of key stage tests in England, is a well-established part of an officially sanctioned formal procedure. At the relevant ages, this *is* the assessment that counts: it is seen to embody many of the approved educational values in mathematics and is a strong determinant of how teachers set about their work of teaching mathematics in mainstream classrooms. The medium of pencil and paper is currently an inseparable part of the

J. Threlfall (✉) · P. Pool · M. Homer · B. Swinnerton
Assessment and Evaluation Unit, School of Education, University of Leeds, 11-14 Blenheim Terrace, Woodhouse Lane, Leeds LS2 9HX, UK
e-mail: J.Threlfall@education.leeds.ac.uk

assessment, and a change to the medium of presentation threatens that highly invested arrangement, and seems to risk losing some of what is valued.

Nevertheless, Raikes and Harding (2003) give a number of reasons for making a computer-based version of a paper and pencil test, including:

- Increased efficiency and consequently lower costs;
- Greater flexibility regarding administration (such as tests on demand);
- Instant scores and/or feedback;
- Fewer errors in marking.

This article explores some aspects of what may be lost and gained by undertaking mathematics assessment on computer. Russell et al. (2003) report studies which compare equivalent paper-based and computer-based assessments, saying that the studies indicate that using computers as a medium for assessment can have significant effects. However, McGuire and Youngson (2002) found no evidence of any differences due to the medium itself, i.e. familiarity with computers and attitudes such as fear of computers are not an issue for assessment. What differences do occur seem to be a result of how the change to the computer format changes the demands of the questions in relation to the marks received. Part of this difference is the availability of 'partial credit' on paper but not on computer (Beevers et al. 1999). However, this article is concerned not with marking issues but with performance issues, with how in some questions the change to a computer format alters what pupils do, and therefore what those items assess.

The concept of "affordance" is commonly used to highlight the effects of the interaction between a pupil and the computer interface on a pupil's response to a task. Affordances are possibilities for action that relate to the capabilities of the actor (Gibson 1979) but this definition is not without ambiguity, and a variety of uses and meanings of the term is detailed in Brown et al. (2004). For example, as Watson (2003) observes, there is also the matter of "participants' perceptions" (page 104), and this means that affordances do not determine activity. In the Human–Computer Interaction (HCI) literature there is a persistent debate about whether affordances are qualities of the computer interface or merely a matter of perception in the user (McGrenere and Ho 2000). For the purposes of this article, however, the perspective of Greeno (1998) seems best. He refers to affordances as "qualities of systems that can support interactions and therefore present possible interactions for an individual to participate in" (page 9). In the specific context of computer interfaces, Gaver (1991) introduces the notion of sequential affordances as a way to refer to situations where afforded exploratory action leads to discovery of further affordances.

One way to consider how different possibilities for action may affect performance is in terms of demand on working memory, and Sweller's (1994) cognitive load theory, although developed as an account of learning, can be applied to the demands in assessment tasks. Cognitive load theory deals with the difficulty of a task in terms of the interactivity between elements of the task. Interactivity is the degree to which elements have to be dealt with simultaneously, and if elements can be dealt with successively rather than simultaneously, cognitive load will be lower. While Sweller (1994) acknowledges that element interactivity depends on the knowledge of the individual, and is not a function of the task as such (p. 306), it is reasonable to compare tasks across different mediums of presentation, looking at differences in the potential for successive and simultaneous treatment in the affordances of the context. In addition, while accepting that this may not apply for all individuals, it seems likely that the differences will matter for at least some individuals, and will therefore be of importance for assessment.

2 Materials and methods

The study investigated the effects on both the material and the performance of pupils of moving paper-based assessment material onto computer. This was achieved by devising paper-based questions that were similar to the end of key stage assessments used nationally in England at the end of key stage 2 (age 11) and key stage 3 (age 14), and making a computer-based version of each of them. The two kinds of assessment were administered to samples of pupils in years 6 and 9, respectively.

For this exercise, an attempt was made to replicate as far as possible the question that originated in the paper and pencil format. As a result, the full range of opportunities that the computer offers for different forms of presentation of the information in the question (animation, hidden information, simulated practical activity, dynamic interaction and so on) were not utilised. As far as possible, the information was presented on screen in the same way as in the paper version. (For a discussion of the implications of the effects of utilising a wider range of possibilities, see Threlfall and Pool 2004).

The work was based on the 2003 published end of key stage material at KS2 and KS3. Modified questions were developed from this material, typically by altering numerical values and changing the contexts of the published questions. The purpose of this was to minimise the effect of any prior experience of the pupils with the actual test questions for 2003, which may have been used for practice in preparation for the end of key stage tests.

At each key stage, 24 questions were selected, aimed at the middle ability range of pupils. These items were then modified as described above and prepared as 24 paper questions. These 24 questions were then programmed (with necessary presentational changes) into a screen format for use on a computer. The material was configured into four different tests of 12 questions, with each question appearing in two tests. Each pupil in the sample took a paper test with its complementary computer test (i.e. the computer test containing the 12 questions not in that paper test).

In small scale studies like this one, any consideration of the effect of the medium on performance is subject to the risks of either practice or sample effects. If the same pupils take both the paper and computer versions of the same material there is a risk that what they learn when completing the question the first time will affect their performance on the second occasion. However, if one sample takes the paper test and a different one the computer test, any differences that are found may be due to the sample taking each test rather than to the medium in which it was presented. The design used to try to eradicate these effects is a modification of the 'single-group design' for situations where 'two tests to be linked are given to the same group of examinees' (Hambleton et al. 1991, p.128) – a commonly used design for equating tests. In this situation, it was the items that were to be linked, rather than whole tests, as they were the focus of interest.

The model identified four groups of pupils, who each took one test on paper and one on computer, such that every item was taken by all pupils on either computer or on paper, but in different combinations, (see Table 1).

Across the model, each set of items and each group of pupils is linked to the others by chains of shared experience. For example, in the paper and pencil items, group 1 took items A–F, which creates a link to group 3 who also took these items, and group 3 took items M–R, which creates a link with group 2. Group 2 also took items S–X, which creates a link with group 4, and group 4 took items G–L, which creates a link back to group 1. A similar chain operates for the computer items, and the computer and paper mediums are linked by the same pupils answering questions in both media.

Table 1 Distribution of questions in the four test pairings

	On paper	On computer
Group 1	ABCDEF GHIJKL	MNOPQR STUVWX
Group 2	MNOPQR STUVWX	ABCDEF GHIJKL
Group 3	ABCDEF MNOPQR	GHIJKL STUVWX
Group 4	GHIJKL STUVWX	ABCDEF MNOPQR

This model allows linkage of the items without any pupil taking both the paper-based and computer-based version of the same item, which removes the risk of a practice effect. By applying Item Response Theory (IRT) (Hambleton and Swaminathan 1985) within this model, the differences between samples could be compensated for.

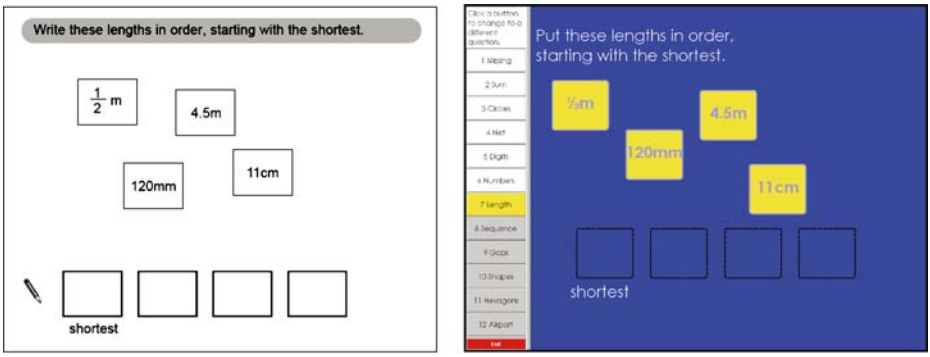
Altogether, each question was tried by equal numbers of pupils (approximately 400) in each format. Paper mark schemes were developed in the style of the end of key stage mark schemes, with computer mark schemes replicating these as far as possible. The questions were marked to give a facility (percentage correct in the sample), with the data collated into one dataset. The dataset was subjected to an IRT parameter estimation analysis known as 'concurrent estimation' which places the parameter estimates for each item on the same scale, made possible by the item linkage described above. The analysis results in estimates of the difficulty and the discrimination of each item and also estimates the ability of each pupil in the sample. Using the item parameter estimates and the ability measures provided by the concurrent estimation, the probability of a correct response on any item in either medium by any pupil in the sample can be computed, regardless of whether they actually took that item in that medium on their test. The 'predicted' facility, which derives from this, is the computed facility of any item in a given medium by all pupils in the sample, not just those who actually took that item in that medium. It is an estimate of performance which takes samples into account, and therefore offers a greater confidence when comparing facilities. It is these 'predicted' facilities which are reported throughout this article.

3 Results

Overall the performance was comparable in each medium. At key stage 2 the pupils' scores (amended to allow for sample effects) were 3% better overall on computer than on paper. At key stage 3 the pupils' scores were 5% better overall on paper than on computer. Of course this overall comparability may mask greater differences for individuals in how they

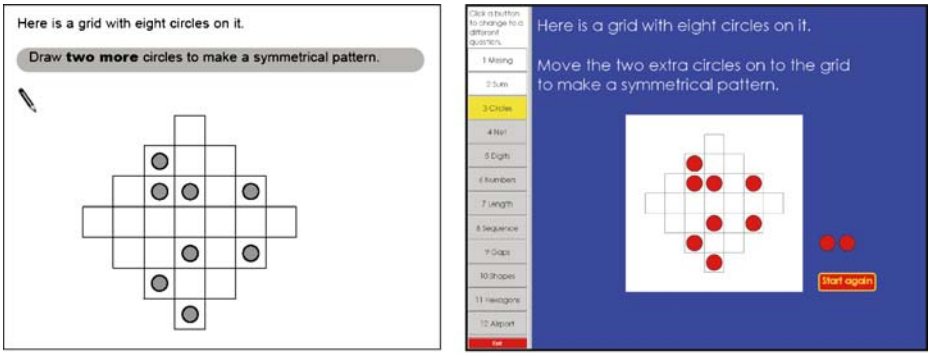
Table 2 Items showing a large difference in facility between paper and computer

	Comparison of facilities	Item	Paper facility (%)	Computer facility (%)	Difference(%)
KS2	Higher on Computer	Length	39.5	52.7	13.2
		Circles	64.5	88.1	23.6
		Sum	77.1	94.9	17.8
		Diagonals	35.9	67.3	31.4
	Higher on Paper	Blocks	69.0	56.8	12.2
KS3	Higher on Computer	Calculation	21.2	32.9	11.7
	Higher on Paper	Shapes	49.0	14.6	34.4



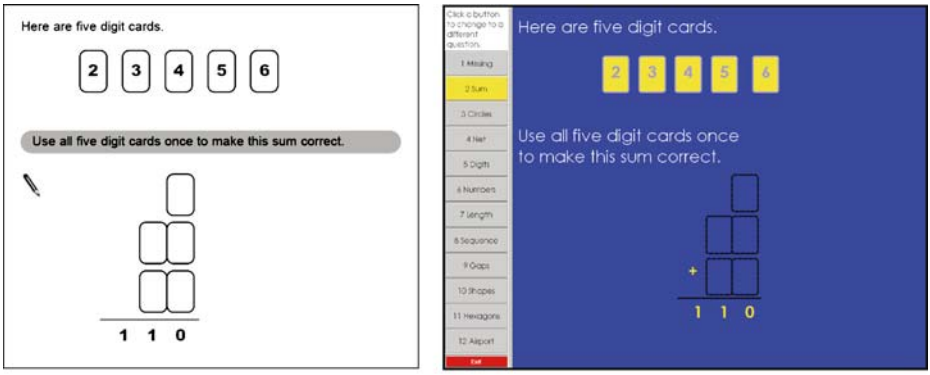
Facility on paper: 39.5% Facility on computer: 52.7%

Fig. 1 The paper and computer versions of "Length"



Facility on paper: 64.5% Facility on computer 88.1%

Fig. 2 The paper and computer versions of "Circles"



Facility on paper: 77.1% Facility on computer: 94.9%

Fig. 3 The paper and computer versions of "Sum"

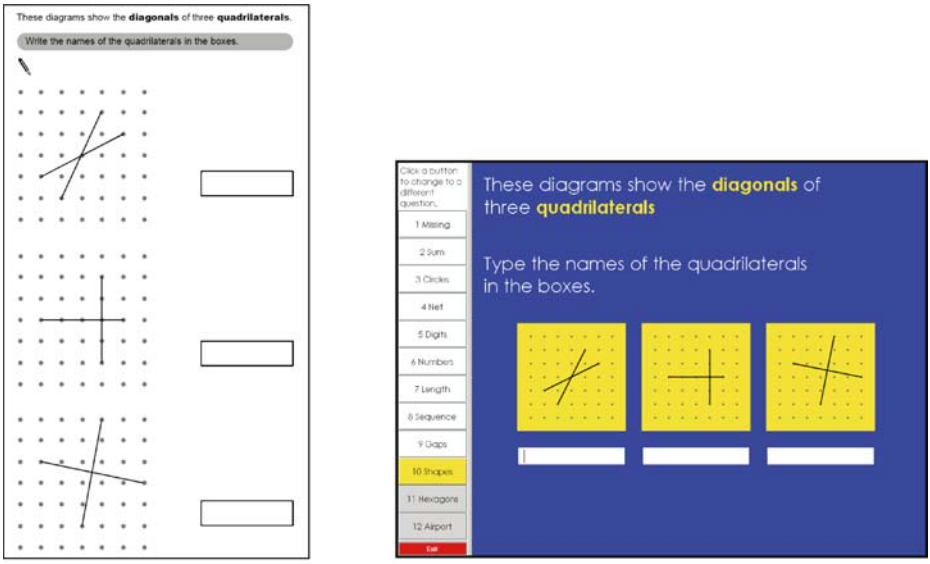


Fig. 4 The paper and computer versions of "Diagonals"

responded to the computer medium compared to the paper medium, but as the pupils took different questions in each medium, it is not possible to measure this effect.

A difference of only 5% or less in performance cannot be said to be indicative of an underlying effect. However, at the individual question level a number of questions showed significantly larger differences in facility, sometimes with more pupils answering the computer question correctly, and sometimes with more pupils answering the paper question correctly. Seven such questions, summarised in Table 2 and subsequently shown in Figs. 1, 2, 3, 4, 5, 6, 7, are the focus of this paper.

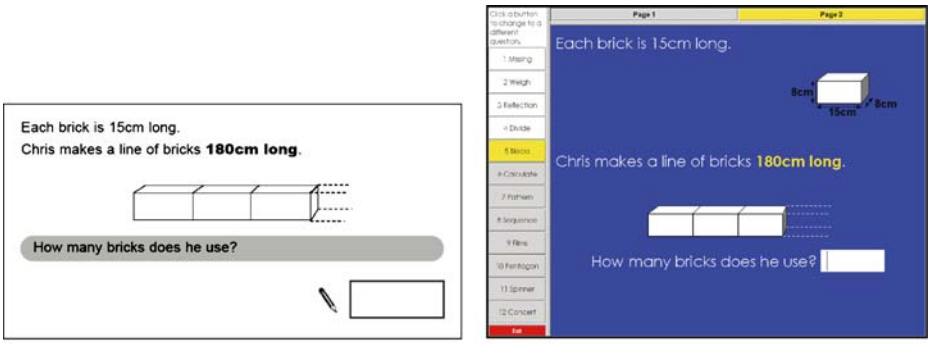


Fig. 5 The paper and computer versions of "Blocks" (second part)

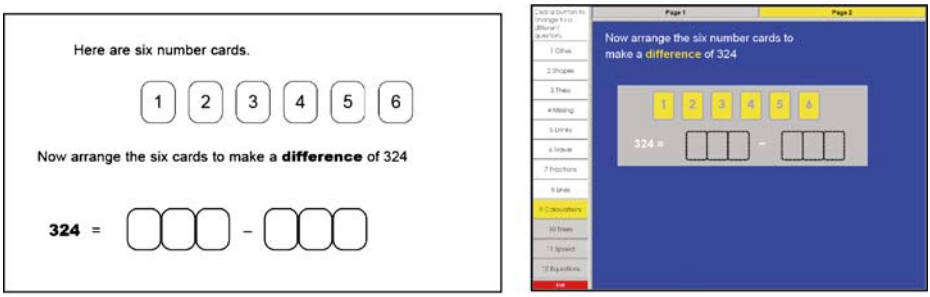


Fig. 6 The paper and computer versions of "Calculation" (third part)

4 Discussion

When considering the reasons that particular computer questions may be answered more readily than paper questions or vice versa, it is common to begin by considering imagined general features of the two media, such as the computer being more appealing, or paper being more familiar. In this case such general features, which relate ultimately either to motivation and attitude or to skills specific to one medium, are unlikely to be the source of the differences on particular questions, since the differences apply in both directions.

The next consideration is of how these questions differ from the other questions that did not show such marked performance variations across the two formats. What do the questions have in common with each other but not with the rest?

In four of the five cases in which the computer performance is better, a common feature of the item is that they involve elements that have to be arranged to give a solution. In *Length* for example, there are four lengths to put in order, in *Circles* two circles have to be added to a shape to make a symmetrical arrangement, and in *Sum* and *Calculation* a set of digits have to be arranged to make a sum or a difference. This is not the case for any of the

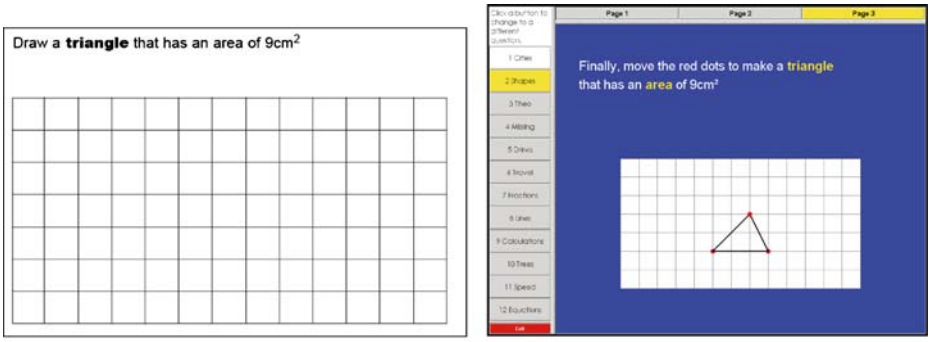


Fig. 7 The paper and computer versions of "Shapes" (third part)

other 44 items trialled, and this suggests strongly that the computer presentation brings a benefit to such questions, by way of its affordances.

Across these four test questions, one set of possible interactions that the computer supports (Greeno's definition of affordance – see Greeno 1998) are exploratory trials, which are an example of a 'sequential affordance,' as highlighted by Gaver (1991). The pupil answering any of these questions can move objects or tokens into position, see how they look, and then move them back again or into a different position – and exploratory activity of this kind was often observed as pupils completed the computer tests. This is not an affordance of the paper and pencil medium for these questions, and so represents a *relative* affordance of the computer medium in this context. It is an affordance that can account for the difference in performance, since, as a result of it, the pupil has a way of getting the question right when using the computer that is not available in the paper and pencil form. A higher facility could then be taken to suggest that such an opportunity had been taken up by a number of pupils. This matters for assessment, because how the question will be answered must always be considered. If a correct answer by an individual to a question is to count as evidence of mathematical competence in that individual, how they got the answer has to be either observed or presumed. For example, guessing and copying are not legitimate ways of getting an answer, and always undermine the assessment. If either is likely, the assessment is considered a poor one. But what about moving the items into position to see how they look? Is that a legitimate action in relation to what is being assessed? That surely depends on what is being assessed, and so must be evaluated on a case by case basis.

A typical strategy for *Length* (Fig. 1) is to begin by trying to determine the shortest. On paper this requires up to six comparisons, which must all be made mentally, to decide the shortest, which can then be written in position. On computer, however, any of the options can be placed in the 'shortest' position as a starting hypothesis, which can be tested out by comparisons one at a time, making changes as required. The placing of items eases the task of comparison, relative to the paper and pencil question, and more pupils succeed. The computer affordance – which is sequential in nature (Gaver 1991) – reveals by contrast that an implicit requirement of what is assessed in the paper question is the co-ordination of a number of mental comparisons – a need to deal with elements simultaneously, which brings an increased cognitive load (Sweller 1994) and may explain the poorer performance.

Yet such a demand seems irrelevant to the assessment, a hindrance to the pupil demonstrating their competence. It might be felt that ordering skills are not in themselves part of what is important when assessing pupils' ability to compare measurements with different units, and that ordering skills should be assessed separately, when not bound up with demands on knowledge of units of length. The exploratory placing of items would then seem legitimate in an assessment of ordering measurements, and the computer version would be the more valid assessment item. Under those circumstances, the higher facility on computer suggests that, in the paper version, some pupils with the competence are not succeeding, an example of a 'false negative' in assessment.

Of course, and on the other hand, if what is intended in the assessment is indeed the confluence of measurement comparisons with ordering skills, then the affordance of the computer would be undermining the assessment. The higher facility could then be taken as evidence of some 'false positives' on computer, with success obtained by non-legitimate means when the sought competence is absent. In this way, the affordances of the medium with regard to any particular assessment item can be related directly to the validity of the item.

In *Circles* (Fig. 2), two circles have to be located so as to make the overall design symmetrical. On paper, the circles cannot actually be drawn until after a decision has been made about where they should go, because of the mess resulting from a change of mind.

The pupil needs to decide that it will look right without being able to try it properly, so has either to be able to visualise, or be analytic – for example by matching pairs across possible lines of symmetry. Either of these involves dealing with elements simultaneously, with high demands on working memory. On computer, the pupil can put the two circles on and make a judgement by recognition – does this arrangement look symmetrical? If not, he or she can move them elsewhere (or, if he or she cannot remember which circles were placed and which were already there, can "start again"). The sequential affordance of the computer medium brings a smaller 'cognitive load' and enables easier success – by recognition of symmetry rather than through visualisation or by analysis. If pupils are willing to try things out, the question is assessing whether they recognise symmetry when they see it.

The implicit aspect of the paper assessment is that a desirable understanding of symmetry is more than just the ability to recognise it when one sees it, but also should incorporate elements of visualisation and/or analysis. If that is accepted, then the activity afforded by the computer is not legitimate for the assessment, and the computer question is less valid as an assessment item. The higher facility on computer would then imply that some pupils who do not have the competence are nevertheless succeeding, an example of 'false positives.'

An alternative perspective to this is that the visualisation and analysis needed to answer the paper version arise from the constraints of the medium, but are not in fact part of what is important in the understanding of symmetry. If recognition is closer to what is wanted, then the computer assessment is more valid, and the paper assessment is flawed, having the possibility of triggering 'false negatives.'

In both *Sum* (Fig. 3) and *Calculation* (Fig. 6) digit cards have to be placed so as to give a specified answer. Again the benefit of being able on computer to place items is suggested by the improved facility in both cases. There is also an additional benefit in the computer medium on these particular questions of being able to keep track of what has been used. This prevents errors from repeating digits or using digits that are not in the provided set, of which there were a number of examples in the paper and pencil question – although not enough to make a significant impact on facilities.

The affordance of moving numbers into position without cost in turn affords the strategy of trial and error, another clear example of a sequential affordance, with the opportunity for dealing with elements successively bringing low cognitive demand. Pupils can put different combinations of numbers down until they find a set that 'works.' On paper this is usually done mentally, and it is implicit in the paper and pencil assessments that a desirable skill in number work is to be able to organise and manipulate numbers in the abstract, i.e. to have the ability to add and subtract numbers that are arranged and rearranged mentally – even though this is a more cognitively demanding activity. As the computer versions do not require it, the affordance of the computer may be said to lead to less valid assessment items. Again, however, there is an alternative perspective, which is that the mental manipulation of digits is not part of what is wanted in the assessment, so that the paper version is less valid than the computer version.

The four items described above were all examples of the same computer affordance improving pupil performance, but in the different cases this could be said to lead to both more valid and less valid assessments – depending on assumptions about which qualities and skills are the intended focus of the assessment. The other cases of contrasting performance on the computer and paper versions of questions can also be considered in terms of the affordances that may have given rise to the difference in performance, and the effect of that on the quality of assessment.

Examining the performance of pupils on the remaining questions that have contrasting facilities across the two media reveals that two of them showed extensive evidence of working on the paper and pencil scripts. This is an affordance of the paper and pencil medium that is not available on the computer screen itself, and although the pupils were allowed to use a 'working booklet' alongside the computer, very few of the Key Stage 2 pupils took up the opportunity to do so – and both of the examples are from Key Stage 2.

The two cases show contrasting features. On *Blocks* (Fig. 5) the written working on the paper scripts was almost exclusively a calculation to work out 180 divided by 15, often as a standard division 'sum.' As it was not intended that this be calculated mentally, which would increase cognitive load too much, the paper-based affordance is legitimate, and the assessment on computer is likely to be giving 'false negatives' – pupils who have the required competence nevertheless failing the assessment because irrelevant aspects of the task bring too great a burden onto working memory. This would continue to be the case for computer based assessment of arithmetic unless and until pupils do make use of paper and pencil alongside the computer, or 'scratchpads' on the computer itself, to perform calculations using algorithms.

On *Diagonals* (Fig. 4) the written working consisted of drawing in the shapes that are 'made' by the diagonals. However, the resulting inaccuracies, evident in the pupil scripts, seem to have led many pupils astray, as they drew shapes that they then did not recognise. Here the affordance of the paper and pencil medium prevented an accurate assessment of the pupils' shape knowledge. There is an implicit acceptance of drawing as a legitimate activity in paper and pencil assessment – but its legitimacy can be questioned in this case. The apparent negative effect of drawing on performance is interesting, but is perhaps not the real issue. The more crucial question is whether the item is a more valid assessment if drawing is not allowed. To re-visit the issue raised in *Circles* (above): Is it desirable in learning about shape and space to have elements of visualisation and/or analysis, rather than it being just a matter of recognition? Drawing can enable shape and space questions to be answered on the basis of recognition, so if visualisation and analysis is wished for, drawing is an affordance of the paper and pencil medium that may undermine the assessment of shape and space. In most questions about shape it seems likely that drawing would do so by improving the performance of pupils on an item, as it is a sequential affordance that enables elements to be dealt with successively rather than simultaneously, and thereby reduces the cognitive load. Even though in this case performance is worsened rather than improved, the computer assessment would still be the more valid one.

The final question that showed a significant difference in performance across the two media is the third part of *Shapes* (Fig. 7) – a question for 13 year old pupils asking them to create a triangle with an area of 9 cm^2. Prima facie, it might be thought that the computer version would be easier for pupils than the paper version. Part of the reason for this is that the computer version removes the error of drawing the wrong shape: on paper, a number of pupils drew rectangles or squares with an area of 9 cm^2, but on computer only triangles could be made. However, the main benefit of the computer medium might be expected to arise from the computer affordance of exploratory action – pupils can try out different shapes 'for size' and by that means arrive at a correct solution on a trial and error basis. However, in practice the computer performance is considerably worse than on paper. It seems that the computer affordance to enable exploratory action was not as useful as might be supposed.

The reason for this may relate to the nature of the possible exploratory activity. Under the constraints of the software being used, it was not possible to programme the computer to simulate drawing on screen, i.e. drawing lines successively, as may be done in a drawing

'package.' The way in which the computer question was programmed was to put a triangle on screen whose shape and size could be changed by dragging any of the three vertices. In this way different triangles can be made, including the same range of different triangles with an area of 9 cm^2 that might be drawn on paper. The dynamic interaction of the computer question does enable exploratory activity, creating any number of triangles that can be evaluated for their area, but it does so in a different way than on paper, and that seems to be critical.

On the paper and pencil question there were two observed elements to the typical approach to answering this question. The first was that most pupils began by drawing a horizontal line, and then building a triangle up from it. The second was that many pupils evaluated size by counting squares. These actions represent what Greeno (1998) calls an 'attunement' to the affordance of the medium, an example of "well co-ordinated patterns of participating .... using artefacts that provide resources for practices" (page 9). In other words, the pupils used learned ways of drawing shapes on squared paper that have particular areas. However, there was not evidence of a parallel attunement in the computer question. That may have been more likely to occur if the response format of the question had been accommodating to beginning with a horizontal line, but the fact that moving any of the vertices changed two lines at once was a new situation, for which the pupils had not developed a 'well co-ordinated pattern of participating'. So the exploratory affordance led to pupils making a succession of triangles, many of which were difficult to check for size by counting squares because of awkward part square measurements, leading to poorer performance on the question.

The computer version of *Shapes* seems to require a more analytic and strategic approach to the problem than the paper version does. On paper, a piecemeal approach is a feasible attunement to the affordances of the medium, starting with a plausible line, then seeing what it leads to. In the absence of a similar attunement to the computer affordances – for example to begin by extending the horizontal base – pupils probably had to consider the problem in terms of the formula for the area of a triangle, and understand what was needed before changing the shape. The exploratory affordance offered by the computer on this question was not a sequential affordance, as it did not lead on to other possibilities for successful action, but the paper and pencil affordance of drawing was.

What are the implications of this for assessment? Here there is some ambiguity, because expectations for working out the areas of shapes change with the age and level of the pupil. For older pupils, the purpose of the assessment tends to be to ascertain whether pupils can use the formula for the area of a triangle. It is also often assumed that if a question can be done successfully using a desired method then success is evidence that it has been done using that desired method. There is an assumed logic that moves from "If a pupil knows the mathematics they will get it right" to "If a pupil gets it right then they know the mathematics." In this case the assumption of an assessment about the areas of shapes may rely on the inference that since a pupil can use the formula to get the answer, then each pupil who got the right answer did so by using the formula.

However, the relationship between the computer and paper versions of this question suggest that this is not necessarily the case. The paper question could be done using an attunement that does not rely at all on the formula, so the assumption of assessment that to get a correct answer the pupil must have used the formula is more likely to be correct on the computer version – making it the more valid assessment.

However, at another level, for example with younger pupils, the approaches that may be used in answering the paper question are also considered legitimate, and valued as worthwhile. Under those circumstances the computer version would be inferior, as it would prevent those methods from being used.

5 Conclusion

Different affordances of the two media, computer and paper, can have effects on performance, but the issue of which offers the more valid assessment is not directly related to whether performance is improved or not. As the examples have shown, the difference in medium sometimes leads to better performance, and sometimes to better assessment, but not necessarily at the same time. Among the most affecting aspects of computer questions seem to be the opportunities that are often afforded to explore variables within the problem and to try out solutions. As a result of these, some of the demands of paper questions are avoided in the computer versions of the items, and the pupils gain greater success than on paper. However, whether or not that leads to more valid assessment needs to be considered on a question by question basis. In some cases the demands of the paper and pencil question that are finessed by the affordances of the computer question are part of what is important about the mathematics of the assessed aspect, and so the computer version is less valid as an assessment. In other cases the demands of the paper version are not relevant to what is important in the mathematics, and so the computer question is better.

There can also be affordances of the paper questions, such as the opportunity to draw or 'jot', which are largely absent from the computer medium, and which can result in the paper question being answered more successfully – although it should be noted that in this study the opportunity to draw and jot was in practice available to pupils. That they did not make use of it may be taken as an example of the effect of 'participants' perceptions' (Watson 2003).

In some cases the methods that are used in the paper questions are legitimate means to answer the question in terms of the mathematics that is being assessed, and the computer question is less appropriate as a means to assess that mathematics. In other cases, the paper version allows pupils to answer successfully using means that are not legitimate for the assessment, and the computer version is more valid. Examples of all these four types have been examined above, and are summarised in Table 3.

In *Diagonals*, there was also an example of a case where the apparent affordance (on paper) led to worse performance as a result of inaccuracies in drawing. One can readily imagine questions where through the different affordances on paper compared to computer the change to the medium has affected how the question is done, but it has not affected the overall level of performance, as the affordances of the two media have had equivalent effects.

As a result it is reasonable to ask of *any* question whether the computer and paper affordances are legitimate in relation to the assessment, and which version is therefore the more valid assessment.

Table 3 Examples of performance effects and assessment legitimacy

	The computer affordance leads to better performance	The paper affordance leads to better performance
The affordance is legitimate to the assessment	Length (computer assessment more valid)	Blocks (paper assessment more valid)
The affordance is not legitimate to the assessment	Circles, Sum, Calculate (paper assessment more valid)	Shapes (computer assessment more valid)

There is risk in computer assessment that the affordances and/or constraints that are built in to the question by how the question is realised bring effects on the nature of the mathematical activity involved in completing the question – which in some cases can be sufficiently dramatic to suggest that the nature of the mathematics has changed (see Threlfall and Pool 2004).

If the usual assumptions are made, the consequential effects of these changes on performance could lead to inaccurate assessment. Making a computer item as much like the paper item as possible may be thought of as the best approach to overcoming this problem, to try to preserve the assessment in the form that has become familiar. However, this treats as unproblematic the affordances of the paper medium, which, as has been seen in some cases above, could already be affecting performance in a way that distorts the assessment, and where the computer medium contains an opportunity to improve the assessment.

Analysis in terms of affordances and attunements and their effects on 'cognitive load' can highlight some of the assumptions of conventional paper and pencil assessment, and draw attention to the possibility of certain kinds of effects when paper and pencil tests are being developed as computer-based tests. It can also suggest where to focus attention when deciding whether assessment is improved as a result. However, improvement to assessment is not determined by the framework of analysis, but has to be addressed on a case by case basis by mathematics practitioners and the mathematics community. It requires clarity in what is wanted from assessment – What behaviours are valued? Which approaches are legitimate? With these questions clarified, the analysis can proceed to suggest whether a computer-based assessment item is more or less likely to assess the mathematics fairly, compared to its paper-based equivalent.

Acknowledgement The work reported in this article was derived from a project funded by the Qualifications and Curriculum Authority.

References

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

Brown, J., Stillman, G., & Herbert, S. (2004). Can the notion of affordances be of use in the design of a technology enriched mathematics curriculum. In I. Putt, R. Faragher, & M. McLean (Eds.), *Proceedings of the 27th annual conference of the mathematics education research group of Australasia,* vol. 1, pp.119–126.

Gaver, W. W. (1991). Technology affordances. In *Proceedings of the Special Interest Group on Computer–Human Interaction (SOGHCI) conference on human factors in computing systems.* New Orleans, pp. 79–84.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Houghton Mifflin.

Greeno, J. G. (1998). The situativity of knowing, learning and research. *American Psychologist, 53*(1), 5–26.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and application.* Dordrecht: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

McGuire, G. R., & Youngson, M. A. (2002). Assessing ICT Assessment in Mathematics. *Maths CAA Series*, *Maths, Stats and OR network*. http://ltsn.mathstore.ac.uk/articles/maths-caa-series/mar2002/index.htm Accessed 11/11/2005.

McGrenere, J., & Ho, W. (2000). Affordances: Clarifying and evolving a concept. In *Proceedings of the Graphics Interface 2000.* Canadian Human–Computer Communications Society (pp. 179–186), Toronto.

Raikes, N., & Harding, R. (2003). The horseless carriage stage: Replacing conventional measures. *Assessment in Education, 10*(3), 267–277.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education, 10*(3), 278–293.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312.

Threlfall, J., & Pool, P. (2004). How might the assessment of mathematics through dynamic interactive computer items be different from that in conventional tests?. *Paper presented at the 10th International Conference on Mathematics Education* (ICME 10), Copenhagen, Denmark, July 4–11.

Watson, A. (2003). Affordances, constraints and attunements in mathematical activity. In *Proceedings of the British Society for Research into Learning Mathematics, 23*(2), pp. 103–108.