

From paper and pencil-to computer-based assessment: Some issues raised in the comparison

Alice Lemmo, Maria Alessandra Mariotti

▶ **To cite this version:**

Alice Lemmo, Maria Alessandra Mariotti. From paper and pencil-to computer-based assessment: Some issues raised in the comparison. CERME 10, Feb 2017, Dublin, Ireland. hal-01949268

HAL Id: hal-01949268
https://hal.science/hal-01949268

Submitted on 9 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From paper and pencil- to computer-based assessment: Some issues raised in the comparison

Alice Lemmo[1], Maria Alessandra Mariotti[2]

[1]University of Palermo, Italy; alice.lemmo@unipa.it

[2]University of Siena, Italy; mariotti21@unisi.it

Comparative studies on pen-and-paper and computer-based test principally focus on statistical analysis of students' performances. In educational assessment, comparing students' performance (in terms of right or wrong results) does not imply a comparison between the solving processes followed by students. In this paper we present an example of task analysis that allows to highlight how students' solving processes could change in switching from paper to computer format and how these changes could be affected by the use of one environment rather than another. The aim of our study lies in identifying possible consequences that specific changes in task formulation have, in terms of students' solution processes.

Keywords: Computer based assessment, comparative study, task analysis.

Introduction

Computer-based assessment is an actual issue. The increasing use of tests administered in the digital environment allows research in mathematics education to develop new fields of study. On the one hand research in computer based tests concerns the validity of these tests, on the other it focuses on their comparability with existing paper tests. In these two perspectives, large-scale surveys were conducted; they involve students from different educational levels, from primary to secondary instruction (Drasgow, 2015; Way, Davis, & Fitzpatrick, 2005).

Computer-based tests mainly involve institutions in large scale assessment (OECD-PISA, OECD-PIAAC, NAEP, …); one of the major interests of these institutions is to anchor every new test with ones from the previous years rather the diachronic study of students' performances in the different surveys. For this reason, some studies focus on the test-mode effect comparing performances of students taking on computer and paper-based tests.

Literature on these topics shows very mixed results; there is empirical evidence that paper-based and computer-based tests will not return the same results. On the one hand, some studies show equivalence in students' performances; on the other hand, different researches highlight a significant discrepancy on scores. For example, Kim and Huynh (2007), as many other researchers (e.g. Kapoor & Welch, 2011; Lottridge et al., 2008), show that there is no statistical evidence suggesting that the administration modality changes the coherence and consistency of computer-based tests. On the contrary Clariana and Wallace (2002) point out to empirical evidence suggesting that students involved in paper-based and digital-based tests will not obtain the same results. At a more general level, in a meta-analysis of computer versus paper-based cognitive ability tests, Mead and Drasgow (1993) found that on average, paper-based test scores were very slightly higher than computer-based test scores.

The main characteristic of all these studies, which involve comparative analysis of outcomes using quantitative and statistical methods, is that they show the comparability between tests administered

on paper and pencil and in a computer environment. Such comparison is developed contrasting students' performances and it is grounded in the implicit assumption that students who achieve the same performance implement the same solution processes. Threlfall et al. (2007) propose a more accurate analysis; they focus on students' solution processes and explore the effect on students' attitudes when they are involved in paper and pencil tests migrated into a digital environment. As shown by Threlfall et al., in some cases changing to a different environment seems to make little differences in the solution process. However, for some particular tasks the computer environment deeply affects how students approach the tasks. An important issue arises: task comparability cannot be measured only in terms of students' outcomes but it is also established by the comparison between the solving strategies that they use.

These diversified results suggest that task comparability needs a deeper analysis. In particular, the comparison between students implies the problem of how and when two tasks could be considered equivalent. Ripley (2009) proposes a possible solution to this question. He distinguishes two main approaches to the use of digital devices in order to enhance assessment: *migratory* and *transformative* approach. He defines the migratory approach to be the use of technological support as a tool of administration; it consists in a transition in digital format of tasks conceived for paper format. The transformative approach involves the transformation of original paper tests integrating new technological devices which support interactive tools (graphs, applets, …) that enhance new affordances. There are no specific studies comparing these two approaches; a possible reasonable hypothesis is that migratory could be a suitable approach to construct what in the literature is called an equivalent task. By definition, the migratory approach has the aim to maintain most of the task features unvaried in the translation process but this transition to a new environment cannot be completely unbiased. The migration from an environment to another one is not neutral because it depends on intrinsic properties of the environments. The adoption of migratory approaches is undervalued; the assumption that the translation process causes few changes on the task formulation and that these changes do not cause significant alterations on the solution processes is not to be neglected.

The purpose of this study is to examine whether the migratory approach may have effects on students' solving processes. In this perspective, the issue of test validity arises; in other words, does the use of a migratory approach maintain the validity of the original test? Below, we present part of a wider study that has the aim to analyse possible changes in students' solution procedures related to the migration from a pen and paper to a digital environment. In Ripley's words, we consider tasks that could be defined migratory or other authors could call equivalent tasks. In particular, we present one example of analysis that compares a task in his migration from paper to computer, highlighting the impact that the changes could have on students' solutions.

Word problem in a migration process

In many tests, especially in large scale assessment, knowledge and skills are assessed through units consisting of a stimulus (e.g. text, table, chart, figures, etc.) followed by a certain number of tasks associated with this common stimulus. These particular features connect these kinds of tasks with word problems. In a wide perspective, the term *mathematical word problem* refers to any mathematical task where significant background information is presented through a verbal text rather

than in mathematical notation. As word problems often involve a narrative of some sort, they are occasionally also referred to as *story problems* (Verschaffel , Greer, & De Corte, 2000).

Mathematical word problems have an important role in teaching; for many decades researchers in mathematics education have focused on the possible difficulties that students encounter when they solve them. Verschaffel et al. (2000) highlight the fact that many of the difficulties met by students lie in the preliminary phase of understanding the problem situation. Interpreting students' attitudes in solving word problems is complex because it involves multiple interacting factors, both cognitive and metacognitive: stereotypes of standard problems, implicit and explicit rules that regulate mathematical activity, students' beliefs, etc. (Verschaffel , Greer, & De Corte, 2000).

Considering word problem texts (in particular, its formulation features) introduces the important issue of *representation*. Goldin and Kaput (1996) describe two distinct meanings of the term *representation*. On the one hand, the *external representations* refer to "physically embodied, observable configurations such as words, graphs, pictures, equations, or computer microworlds" (ibid., 400); on the other hand, the *internal representations* concern "possible mental configurations of individuals, such as learners or problem solvers" (ibid., 399). In the case of word problems, the solver interacts with the *external representation* presented and produces a personal *internal representation* linked with the one that she already has. Obviously, being internal, such configurations are not directly observable but they could be inferred through the solution process that the solver employs. For this reason it is possible to confirm that a change in the *external representation* could influence the construction of the internal representation and so the adoption of the solving process. Therefore, it is possible to suppose that the formulation of mathematical word problem influences both cognitive and metacognitive factors that are involved in word problem solution. Goldin (1982) highlights that small differences in some features of word problems can deeply affect the process of solution. In particular, Mayer (1982) and later De Corte & Verschaffel (1985) observe that the difficulties noticed within problem solving activities may come from an inadequate interpretation of the text.

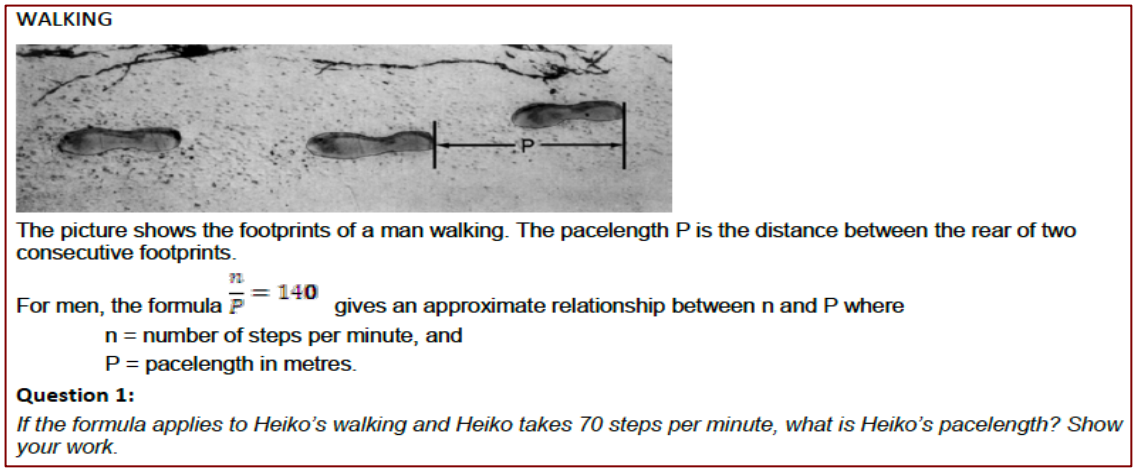
Thus, in the perspective of comparison, it is necessary to analyse the differences between tasks to determine the possible differences that occur in students' solution processes. Identifying possible changes in a mathematical word problem requires to consider many text features. For this reason, the task is simplified by dividing the word problem into simpler elements. Gerofsky (1996) describes word problems in terms of three main components: the *set-up component* which establishes the characters and location of the story; the *information component* which encompasses the information needed to solve the problem; and finally, the *question component* which expresses the request and focuses on goal and aim.

Our purpose is to analyse tasks through specific variables that might influence the behaviour of students in the solving process. Obviously, checking these differences is a general issue that could be presented whether or not there is a migration process in a new environment; possible changes could happen even just in the paper environment.

Analyses of a migrated word problem

We present the analysis of one of the items presented in the Draft 2015 PISA Mathematics framework (OECD, 2013). Figures 1 and 2 show the two versions of the famous task: "Walking", administered

in PISA 2003 survey. The text of the item has not changed; therefore, narrative or linguistic differences are not recognized in the *set-up* component.



WALKING

The picture shows the footprints of a man walking. The pacelength P is the distance between the rear of two consecutive footprints.

For men, the formula $\frac{n}{P} = 140$ gives an approximate relationship between n and P where

n = number of steps per minute, and

P = pacelength in metres.

Question 1:

If the formula applies to Heiko's walking and Heiko takes 70 steps per minute, what is Heiko's pacelength? Show your work.

Figure 1: "Walking" paper version, administered in PISA 2003 survey

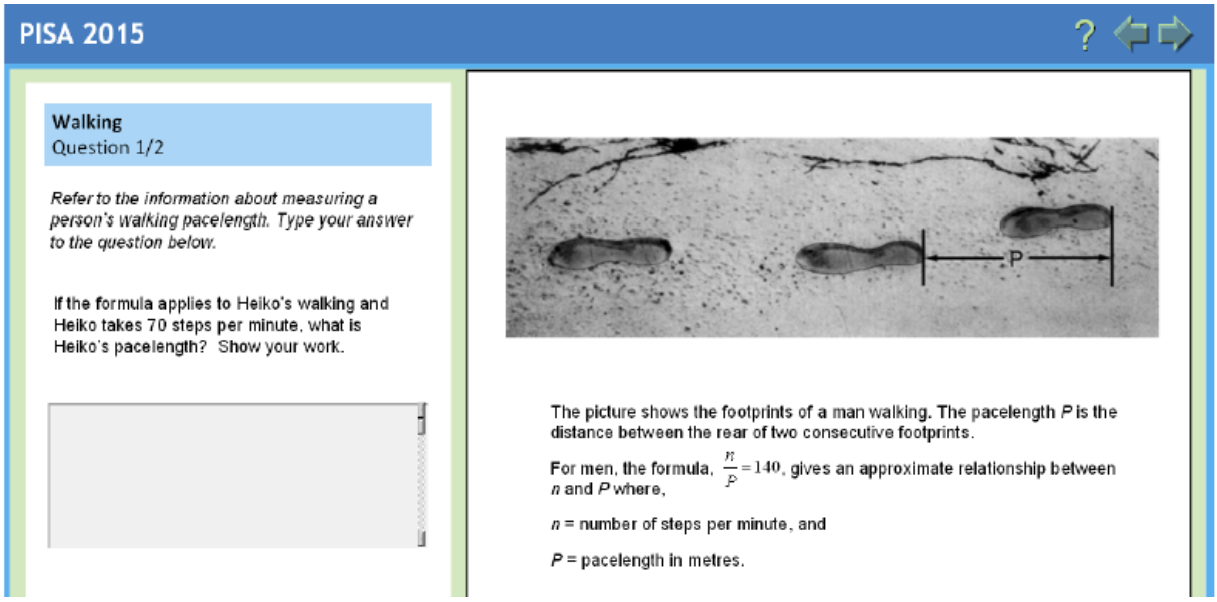


Figure 2: "Walking" computer version, shown in PISA 2015 Framework Draft

First of all, there is a difference in the editing of the text. In the paper version the task is presented in a compact way: *set-up* and *information* components are given in the same text and the *question* is presented under this text. In the digital format, the task is divided into two main sections. On the right there are the *set-up* and *information* components: they consist of an image and a description of the situation, both in words and algebraic formulas. On the left there is the question. The difference in editing seems to complicate the task; in the digital format, the text is presented in two separate columns. Therefore, the change of the question position could create variances in solution processes: the solver has to coordinate the interpretations of the different parts in which the text is divided. In other cases this change could affect the solver's comprehension. For instance, Thevenot et al. (2007) show that putting the question before a word problem (rather than classically presenting it at the end) conditions problem solution in young students and in particular it facilitates students in engaging a correct solution process. In Fig 2, the question is presented in the bottom-left of the screen; in this case the solver probably reads the question before reading the *set-up* and *information* components.

According to Thevenot et al, this fact suggests that in the digital format the interpretation of *set-up* and *information* components could be affected by the previous reading of the question.

Concerning the component *question*, there is another notable difference. In the digital format (Fig. 2), the first part of the *question* text shows the instructions for answering to the task ("Type ... below") and how to coordinate the information presented in the context ("Refer to … pacelenght"). This aspect enriches the *question* and the length of the text that students have to comprehend and interpret (in the paper version there is not any kind of instruction).

The test item format is changed; a text box in the digital version replaces the free space presented in the paper. The test item format has a strong impact on students' solution process. Kazemi (2001) investigates children's mathematical performance on test items focussing on the typology of the questions. In his study, Kazemi uses multiple-choice questions and juxtaposes them to other open-ended problems. He highlights that the typology of questions affects students' thinking in designing and interpreting problems. This impact is emphasized when there is a change of environment and so a change of tools available to the solver. Concerning computer and paper and pencil based tests, Russell and Haney (1997) describe a comparative study in terms of students' performances. They show that there are differences in performance related to the type of test item formats; substantial changes are not found in the case of tasks with multiple choice questions but there are relevant differences in the case of open response items. Moreover, assuming that the student is familiar with the writing tools available (for example the keyboard) it is reasonable to suppose that this change would not result in significant differences in the solution process. However, in using the free space in paper format, the solver has a different freedom of expression with respect to the case of the text box: in paper and pencil, the solver can produce sketches, calculate and write text both in natural and in symbolic language. These actions are not allowed in a simple text box in which one can only enter the characters on the keyboard or otherwise perform the actions allowed by the available writing tool, depending on the software used.

Finally, in both tasks the same picture is presented; nevertheless, it is possible to notice that in the digital version the picture is presented on the screen with all the strengths and limitations of the software that supports it. For example, it might be difficult (or impossible) to analyse the image through common and simple manipulation action such as turning the paper, complete the picture by drawing lines, highlight points, etc.; these actions are possible only in paper and pencil environment.

Conclusion

In the previous example, the migration process could at first appear accurate but a deep analysis shows the opposite. At a first glance, the highlighted little differences might appear superfluous; however they are crucial to analyse and interpret students' behaviour in the solving process. The literature described in the first part of the paper indicates that each difference observed in the example may affect the solver. For instance, the change in the task editing could simplify the text comprehension if it is presented in a linear way; on the contrary, the reading could be difficult if the verbal description is fragmented in several parts. Furthermore, the position of the *question* may encourage the solver to develop a correct solution process or it could complicate the *set-up* comprehension because the solver has to coordinate its interpretation with the information presented in the text components. These little differences hide important consequences for assessment, especially if the purpose of the migration process is to ensure continuity between the paper and the

computer administration. The comparison studies presented in the literature assume that the tasks administered in the two environments are equivalent. However, our analysis shows that this starting assumption should be changed. The equivalence between performances (in terms of right or wrong results) does not imply an equivalence between the processes adopted by the students. Therefore, the analysis of the results collected in the two environments probably is not equivalent in terms of educational assessment. The answers produced by students in the digital environment seem hardly comparable with what they do on paper. Thus, there is a substantial difference in terms of the assessment; it cannot be ignored especially by national or international large scale assessment.

In addition, the change in type of test item format is crucial because it strictly depends on the intrinsic feature of the environment and on the familiarity that the solver has with the tool available. We recall also that there are cases where it is impossible to translate a task from paper to digital format through the migratory approach; for example, the tasks that require the use of physical tools and measuring instruments as ruler, compass, or other. A special case is the one of items that have the goal of assessing students' drawing abilities. In this case, the item may request to draw a figure starting from a given one, or from given measurement, or from written verbal instructions. In these cases, it is possible to introduce an ad hoc software or applet that simulates the use of drawing tools. Moreover, the issue of students' familiarity with these software or applets arises (Bennett, Persky, Weiss, & Jenkins, 2010). In the case of lack of familiarity with the use of the instrument, the digital device could be largely useless; students that use digital tools may be disadvantaged comparing them to students that use paper and pencil and physical tools. This example highlights a very serious and complex issue. Further research is needed to define criteria of control that allow to check and to compare all the little differences that occur in the migration process.

In our wider study we define a specific instrument to monitor such differences. In particular, we identify specific variables that might influence the students' behaviour in solving a certain task. We organize such variables into a table that we call *comparison tool*. Such tool is constituted by a system of indices related to the structure of word problems described before. In particular, we identify five different indices that represent possible changes that may occur in the migration process:

- **Story** refers to the *narrative dimension* (Zan, 2011) of the task (for example: characters, background, narration, etc);
- **Linguistic form** that indicates also the number and the length of sentences (for example: syntactic, organization of the sentence, lexical, etc);
- **Type of item formats** concerns the types of possible responses (for example: the question has constructed-response, selected-response, etc);
- **Format and editing** refers to layout features and position of the different components (for example: paragraphs, font, underlines, spaces, etc);
- **Data representation** is related to semiotic register used for representing information (Duval, 1993) (for example: verbal register, iconic register, math register, etc).

Each index is related with studies concerning word problem formulation and its impact on students' solution process. For example, many authors show that the narrative dimension attached to mathematical tasks is relevant to students in terms of the their availability to solve the task (Sowder, 1989). Other studies draw attention to the importance of language in student performance on

assessments (Abedi, Lord, & Plummer, 1995). Moreover, many authors pay attention to the role of representation in the teaching and learning process (Duval, 1993).

Considering the example above, the comparison tool highlights differences related to three of the five indices. In particular, tasks are different in terms of *linguistic form*, *type of item format* and *format and editing*. In fact, in the digital version there are more sentences than in the paper task. Furthermore, even if both questions are open-ended, in the computer task there is the restriction caused by the text box where type. Finally, there is a strong difference in editing: in the two task versions, the components are presented in different parts of the page/screen. The comparison tool highlights a certain number of differences; such differences could confirm our hypothesis in terms of possible differences in students' behaviour.

References

Abedi, J., Lord, C., & Plummer, J. (1995). *Language Background as a Variable in NAEP Mathematics Performance: NAEP TRP Task 3D: Language Background Study.* Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *The Journal of Technology, Learning and Assessment , 8* (8).

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Technology, 33* (5), 593-602.

De Corte, E., & Verschaffel, L. (1985). Beginning first graders' initial representation of arithmetic word problems. *The Journal of Mathematical Behavior , 4*, 3-21.

Drasgow, F. (Ed.). (2015). *Technology and Testing: Improving Educational and Psychological Measurement*. Abingdon, Oxon, United Kingdom: Routledge.

Duval, R. (1993). Registres de représentation sémiotique et fonctionnement cognitif de la pensée. *Annales de didactique et de sciences cognitives, 5*, 37-65.

Gerofsky, S. (1996). A linguistic and narrative view of word problems in mathematics education. *For The Learning of Mathematics, 16* (2), 36-45.

Goldin, G. A. (1982). The measure of problem-solving outcomes. In J. Garofalo, & F. Lester (Eds), *Mathematical problem solving: Issues in research* (pp. 87-101). Philadelphia: Franklin Institute Press.

Goldin, G. A., & Kaput, J. J. (1996). A joint perspective on the idea of representation in learning and doing mathematics. In *Theories of mathematical learning* (pp. 397-430).

Kapoor, S., & Welch, C. (2011). Comparability of Paper and Computer Administrations in Terms of Proficiency Interpretations. *Annual meeting of the National Council on Measurement in Education.* New Orleans: LA.

Kazemi, E. (2001). Exploring test performance in mathematics: the questions children's answers raise. *Journal of Mathematical Behavior, 21*, 203-224.

Kim, D. H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *The Journal of Technology, Learning and Assessment, 6* (4).

Lottridge, S., Nicewander, A., Schulz, M., & Mitzel, H. (2008). *Comparability of Paper-based and Computer-based Tests: A Review of the Methodology.* Monterey: Pacific Metrics Corporation.

Mayer, R. (1982). The psychology of mathematical problem solving. In F. L. Lester, & J. Garofalo (Eds.), *Mathematical problem solving. Issues in research* (pp. 1-13). Philadelphia: The Franklin Institute Press.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449.

OECD. (2013). *Draft PISA 2015 Mathematics Framework*. OECD publishing.

Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 92-98). Luxembourg: Office for Official Publications of the European Communities.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education policy analysis archives, 5* (3), 1-18.

Sowder, J. (1989). Searching for affect in the solution of story problems in mathematics. In D. B. Adams (Ed.), *Affect and mathematical problem solving:· A new perspective* (pp. 104 - 113). New York: Springer-Verlag.

Thevenot, C., Devidal, M., Barrouillet, P., & Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. *The Quarterly Journal of Experimental Psychology, 60* (1), 43-56.

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66* (3), 335-348.

Verschaffel , L., Greer, B., & De Corte, E. (2000). *Making sense of word problems.* Lisse, The Netherlands: Swets & Zeitlinger.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2005). *Practical questions in introducing computerized adaptive testing for K-12 assessments.* (pp. 1-20). Pearson Educational Measurement.

Zan, R. (2011). The crucial role of narrative thought in understanding story problem. In K. Kislenko (Ed.), *Current state of research on mathematical beliefs. Proceeding of 16th MAVI Conference* (pp. 331-348). Tallinn.