

Computer-based assessment in mathematics: Issues about validity

Anneli Dyrvold and Ida Bergvall

Department of Education, Uppsala University, Sweden

Computer-based assessments is becoming more and more common in mathematics education, and because the digital media entails other demands than paper-based tests, potential threats against validity must be considered. In this study we investigate how preparatory instructions and digital familiarity, may be of importance for test validity. 77 lower secondary students participated in the study and were divided into two groups that received different instructions about five different types of dynamic and/or interactive functions in digital mathematics items. One group received a verbal and visual instruction, whereas the other group also got the opportunity to try using the functions themselves. The students were monitored using eye-tracking equipment during their work with mathematics items with the five types of functions. The result revealed differences in how the students undertook the dynamic functions due to the students' preparatory instructions. One conclusion is that students need to be very familiar with dynamic and interactive functions in tests, if validity is to be ensured. The validity also depends on the type of dynamic function used.

ARTICLE DETAILS

LUMAT Special Issue
Vol 11 No 3 (2023), 49–76

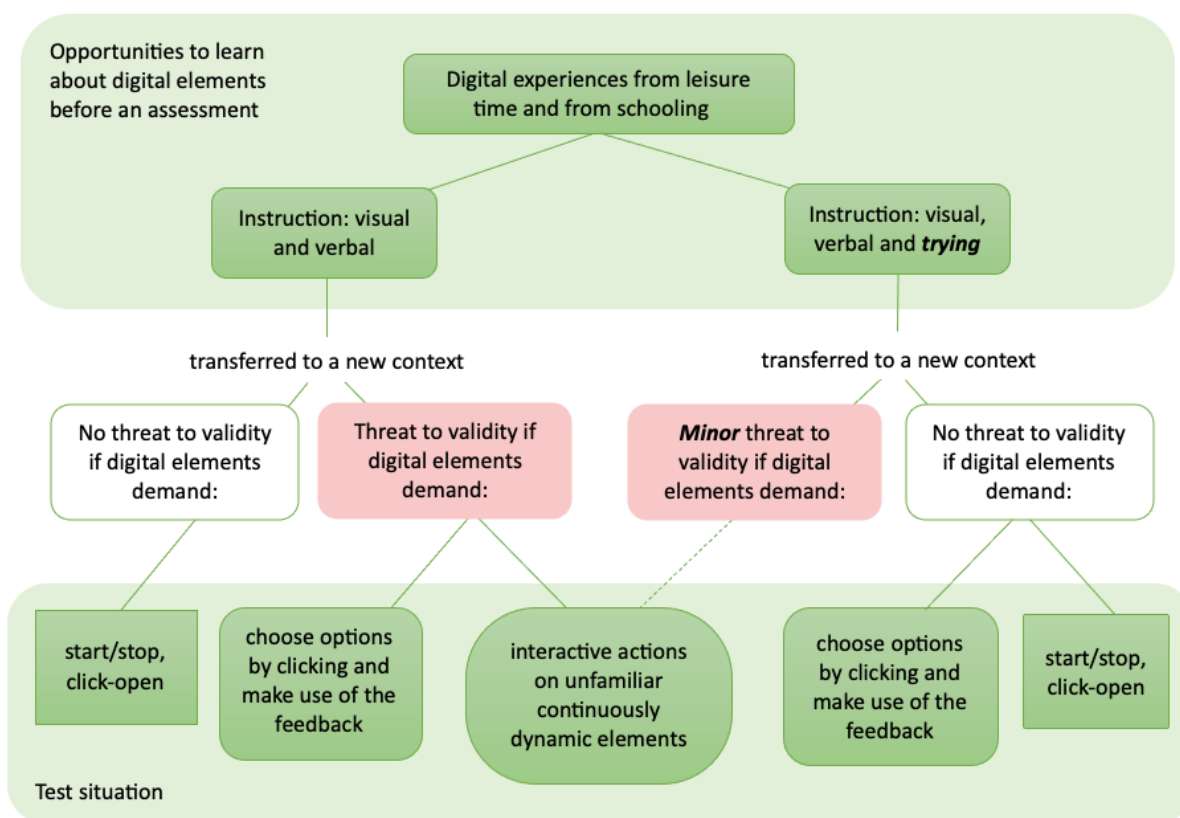
Received 27 October 2022
Accepted 14 September 2023
Published 5 October 2023

Pages: 28
References: 39

Correspondence:
a.dyrvold@gmail.com

<https://doi.org/10.31129/LUMAT.11.3.1877>

Keywords: computer-based assessment, dynamic, interactive, validity, transfer



1 Introduction

The use of computer-based assessments is continuously increasing, in mathematics education as well as in other subjects. Computers can be used in tests to simplify data collection or because computers offer practical tools in the test situation (e.g., language editing) but the purpose can also be to test aspects of digital competence. The discrepancy between merely using computers as a data collection tool or to, more or less, also assess digital competence has implications for the kind of preparation or training that is needed before a test. In both cases it is crucial that the test-taker is not disadvantaged due to misunderstandings that prevent them from using a particular digital component in the test, something that can threaten the validity of a test.

There are many reasonable arguments for digitization of tests and an increase in digital tests are expected. Bennett (2015) points out that the evolution from paper-based assessments to electronic ones is substantive. Currently, different assessments are at different stages in this evolution. In the mathematics part of PISA (Programme for International Student Assessment), digital items are included based on the argument that “a level of competency in mathematical literacy in the 21st century includes usage of computers” (OECD, 2013, p. 44). It is emphasised in the PISA framework that the digital environment provides opportunities to include more interactive and authentic items, for example with drag-and-drop and with real-world data (OECD, 2013). The benefits of a dynamic environment have also made an imprint on the mathematics assessment in TIMSS (Trends in International Mathematics and Science Study). It is argued that computer-based assessments enable the incorporation of new and better assessment methods, through the use of digital components (Mullis & Martin, 2017).

The variety of skills demanded to navigate in a digital environment risk to be underestimated, in particular if familiarity with such an environment is taken as a guarantee for a broadly applicable digital competence. The younger generations have more or less lived their whole lives using digital devices as natural tools in their everyday lives, which is why these generations are called “digital natives” (e.g., Prensky, 2001). Digital natives are perceived as possessing advanced knowledge of the use of digital equipment and technology. This generalisation has however been questioned by many researchers (e.g., Bennett et al., 2008a; Helsper & Enyon, 2010) who highlight the great variation within the generations and that there are no direct differences between digital natives and older generations, even if digital natives use digital technology to a large extent. The use of digital devices in leisure time and professionally, or in

schooling can be very different things and it cannot be taken for granted that the use of digital tools, for example in an assessment context, is problem-free.

In terms of test validity, it is crucial that students are comfortable using the digital functions offered in a test, which is not guaranteed even if they belong to a younger, “digital native” generation. It is of interest to explore what kind of familiarity with digital materials is needed for students to make use of use different kinds of digital functions in tests, because with such an understanding we can better prepare test takers, and thereby diminish threats to validity. Accordingly, this study addresses validity in relation to computer-based tests, with a particular focus on different digital functions and which kind of instructions that are sufficient before a test.

1.1 Aim and research question

The aim of this study is to contribute to understanding the role of students’ acquaintance with digital functions for how they encounter digital teaching materials and tests. The research question is: In a digital environment, are there any differences between which dynamic functions students are prone to undertake and use in the face of different instructions proposed?

2 Background

2.1 Utilising affordances of the digital media in assessments and in teaching materials

Many tests in mathematics that previously have been offered in print are successively replaced by digital counterparts; for example recent versions of PISA mathematics as well as TIMSS mathematics are offered both in paper and in digital format (OECD, 2021; Mullis & Martin, 2017). There are a variety of reasons for using computer-based assessments. For example, digital resources have the potential to enrich items in tests by inclusion of new multimodal resources such as video explanations, hints or worked out answers, or mathematical software for manipulating objects like, graphing, drawing or solving equations (e.g., see Usiskin, 2018). Digital resources also provide opportunities to organise mathematical information in new ways, for example by linking to explanations and definitions that can be shown or hidden (O’Halloran, Beezer, & Farmer, 2018). Computer-based assessments may also facilitate formative assessment (e.g., Aldon & Panero, 2020; Barana et al., 2021) or contribute benefits related

to the implementation itself, such as simplification of distribution, reduced working time for assessment and grading, automatic reporting of test results, increased usability and accessibility for students with disabilities through technical solutions (e.g., Regeringen, 2017; College Board, 2022). Results from comparisons between paper-based and computer-based mathematics tests have revealed that computer-based tests are significantly harder than the paper-based test (Bennett et al., 2008b) but there are also contradictory results. A study contrasting paper-based and computer-based tests in Korean language, mathematics, social sciences, and science revealed higher results on all paper-based tests, but only with a significant difference for Korean language. Furthermore, for female participants, the difference was significant also for mathematics (Hanho, 2014). The difference between modes (paper or computer) is also evident in students' choices of mode for responses. An analysis of which format students prefer to use when responding on tests reveal that the paper format is preferred by 71 percent (100 participants) when contrasted to digital pen or type-written response format (Davis et al., 2021). The mean results were also highest for the paper format and lowest for type-written response, but these differences were not significant. Accordingly, the occurrence of a mode effect seems to be dependent on the subject, but contextual factors such as familiarity with the hardware can also affect performance (Dadey et al., 2018).

There undoubtedly are numerous assets to computer-based tests, enabling a valid assessment of students' mathematics skills, but it is important to keep in mind that the opportunities provided by digital media are likely to place new demands on students' ability to read and navigate the digital environment and to work with the digital resources. It has been shown, for example, in a study of students' work with GeoGebra in ordinary teaching situations, that students with lower achievement levels in mathematics but used to working with GeoGebra outperformed high-performing students who were not used to working with GeoGebra (Baccaglioni-Frank, 2021). This probably applies to assessment situations as well. Even when simpler tools such as a ruler or protractor are integrated digitally, using them demands other skills than if the physical tools are used. If such tools are integrated in tests, students must be accustomed to using them to perform well (Lemmo, 2021). Harris et al. (2021) showed that students' performance on interactive tasks was lower than on static tasks in the digital environment. The reduced performance may be due to increased cognitive demands on the students as they need to work with several different mental processes simulta-

neously in interactive tasks. The vertical order in which text elements are usually arranged on ordinary sheets also works less well on a screen with its proportions, and tasks may therefore need to be structured differently, for example in columns. This restructuring may complicate reading because the student needs to coordinate information from different parts of the text in a way that they may not be used to (Lemmo, 2021).

2.2 The relation between experience with digital teaching material and performance on computer-based tests

A relation between whether digital teaching material is used in training and performance on a test in digital form is much expected, but exactly which kinds of digital features test takers need experience of is not self-evident. In a context where test takers are accustomed to digital devices both in school and at home, a proficiency in navigating in the digital environment might be expected due to the familiarity with digital devices. There are previous studies pointing to the connection between experience of being taught with digital learning materials and success in computer-based tests. For example, Hamhuis and colleagues (2020) explored potential differences in performance of Dutch primary school students on the TIMSS test, depending on whether it was paper based or conducted on a tablet. The result revealed no significant differences between the paper and the tablet tests, which can be explained by the fact that Dutch students are used to the digital format. The picture is somewhat complicated by other studies such as Smolinsky et al. (2020) showing that university students who were taught with paper-and-pencil have slightly better results on computer-based tests compared to students who were taught entirely computer-based. This result speaks against the fact that general digital habits make it easier for students when working with computer-based tests. Within another research project (Hoch et al., 2018), an interactive mathematics teaching material for the introduction of fractions was designed and evaluated. The material consisted not only of digitised text, but also of three key interactive features; interactive exercises, adaptive demands and automatic feedback. A result that emerged from the study was a negative relationship between the time students spend on the tasks and performance (Hoch et al., 2018). Hoch and colleagues suggest that an explanation to the result is that the students did not use the exercises only to acquire new knowledge, but did also spend time practising rational number concepts. Time was therefore not a good measure of proficiency, even though practising is of course a desirable activity.

Students who are used to working with digital teaching materials and various types of dynamic resources in their daily lessons are likely to have advantages when encountering new types of resources in test situations in contrast to those using print material. Research has however shown that in contemporary teaching material comprising whole courses, the utilisation of highly dynamic elements is still rather limited (e.g., see Dyrvold, 2022; Ilovan et al., 2018). Students who study with digital textbooks can therefore also meet unexpected demands in computer-based assessments.

2.3 Validity in computer-based assessments

A crucial question in relation to computer-based assessment is of course what the test aims at assessing. It can be argued that digital competence is part of a comprehensive mathematical competence (e.g., see Geraniou & Jankvist, 2019) and tests can accordingly aim at assessing also mathematical digital competence. However, such an aim is not the goal in all digital mathematical assessments, but even if that was the case, it needs to be scrutinised which kinds of digital demands are wanted in tests and not. This issue relates to Messick's (1995) argumentation concerning the essentiality of construct validity, and in the design of computer-based tests, it is decisive whether the ability to handle digital tools is part of the construct or not. Messick highlights the complexity of capturing the construct a test aims at assessing. A test measure is only one indicator of the full construct, and it is fundamental that this indicator constitutes a good representation of this construct; that is, whether task responses are solely dependent on the processes, knowledge, and strategies utilised in the performance. If the intention is not to assess digital competence, the test should not reward digital abilities, nor should the test result be affected by unwanted digital difficulties.

Another potential threat to validity is construct irrelevant variance, which can be due to construct irrelevant easiness or construct irrelevant difficulty. An example of irrelevant easiness is when the possible answers in a multiple-choice question provide 25 percent chance of answering correctly by guessing, and irrelevant difficulty might be if dynamic resources in a task are difficult to undertake, so that students fail on tasks that they would otherwise be able to solve. Computer-based assessments increase the potential sources of construct irrelevant variance because the digital environment may entail skills not demanded in paper-based tests. Crucial for validity is therefore whether digital skills needed in the digital environment, is regarded as a part of the construct, or not. In this paper, we intend to contribute to understanding

validity in computer-based tests by investigating the role of students' acquaintance with digital functions for how they encounter digital teaching materials and tests.

Regarding the mathematics content included in the construct of a test, Ripley (2009) distinguishes between computer-based tests designed based on a *migratory* strategy (keeping the test similar to the print version) or on a *transformative* strategy (inferring demands of new digital skills in the test). Depending on strategy, different validity issues arise. In assessments designed with a migratory strategy, effort is laid on ensuring the assessment is kept as similar to the print version as possible and several studies have proven a high validity in this type of assessments (Junpeng et al. 2019; Hamhuis, 2020). These studies indicate that digitalisation can be made without decreasing validity, but if many new functions are used in a computer-based test the risk of validity issues increases. Assessments based on a transformative strategy, for example, that are designed to bring *innovation* in curriculum design and learning introduces many elements that may be new to the test takers. An innovative assessment can provide rich opportunities to assess comprehensive mathematical competences that would be harder to capture in a paper-based test. The test taker can for example be offered dynamic materials (Yerushalmy & Olsher, 2020) and the tasks can contain real-world data and present students with an authentic, simulated environment (e.g., see OECD, 2013). Bennett (2015) distinguishes assessments with a transformative strategy in a *second* and *third* generation. The second generation of assessments use new item formats including for example multimedia or constructed response options and may aim at assessing new constructs. The third generation is defined as assessments that uses complex simulations and interactive features and serves both individual and institutional purposes. These kinds of assessments are integrated with instruction with repeated sampling, and accordingly new skills are assessed in more sophisticated ways. Bennett concludes that challenges in relation to the third most advanced step in this evolution of digital tests is large and that the need for a cautious development process is essential. The current study addresses features that are prominent in the second and third generation of assessments; the inclusion of multimedia as well as interactive features. The many benefits of the third generation of assessments are addressed by Bennett but some important challenges are also highlighted. In particular he points out that the most important challenge to address in research is validity and fairness of the third generation of assessments for all individuals, in especially for students at risk.

The need for awareness of potential threats to validity when using innovative assessments is essential. For example, if the functions used are novel to the test taker there is a risk for construct irrelevant difficulty and the integration of several types of skills in the construct being assessed can probably make it difficult to distinguish construct irrelevant variance. Messick (1995) exemplifies how demand of communication skills can be judged either as irrelevant in an assessment of mathematical knowledge or considered as parts of mathematical proficiency and therefore relevant to the construct. For assessments that are complex in the sense of subsuming multiple processes, awareness of construct irrelevant variance is of particular importance. The sources of variance need to be thoroughly evaluated in relation to the construct being assessed, judging how compelling the evidence for the relevance of some variance is. It is apparent that the risk for construct irrelevant variance increases as new affordances of the digital media are used but the threats to validity can also decrease if for example the option to write with digital tools increases a test taker's ability to represent their knowledge. In addition, there are many affordances of the digital media to take advantage of besides distribution, central grading and curriculum innovation. Other possibilities could be to include certain dynamic functions, such as possibilities to write more detailed answers, to use digital functions or to include, for example, GeoGebra, but without aiming at curriculum innovation. Especially if computer-based tests would be designed in accordance with digital learning materials, the potential for meaningful and valid assessments is substantial.

The current study addresses validity issues by putting the focus on how test takers are prepared for a computer-based assessment; whether barely instruction is sufficient or if the test takers might need to try features that they are expected to utilise in the test.

3 Method

The current study is part of a larger project about digital teaching material in mathematics. In this study we focus on students' encounters with different dynamic elements in mathematics items after receiving one of two kinds of instruction about the different elements. At the core of the study is differences between types of dynamic elements and between two kinds of instruction given to the students before problem solving. Thorough descriptions about the elements and the instruction are therefore given in this section. Elements are defined as a coherent part of a text that may include both words, symbols, and images, where the constituents can be static or dynamic.

Some of the elements are also interactive. The types of dynamic elements focused in this study (Table 3) is utilised in the *Facts* in an item, see Figure 2.

3.2 Data

Empirical data in this study was collected in an eye-tracking analysis of 77 grade nine students from two different schools in Sweden, during work with mathematics items. Both schools used printed mathematics textbooks for regular teaching. Each participant took a survey with questions about digital habits before working with the items. Information about the participants' grades were also collected. Data from three participants were excluded due to bad calibration or missing data from the eye-tracking. One participant rushed through all items in four minutes and data from this participant was also excluded. Data from three students were excluded because these students had not reached the first level (grade E) in Swedish as a second language. The participants were divided in two groups, referred to as *Show* and *Try* because in addition to verbal instructions one group was shown information before doing the items whereas the other group also tried dynamic features such as to start a film. Six students studied Swedish as a second language and had reached at least grade E (pass). Data from these students were also analysed because they were equally distributed between the *Show* (7%) and *Try* group (8%). All students who had reached the age of 15, gave a written consent to participate in the study (or their guardians otherwise). The students were informed that the overall purpose of the study concerned work with digital teaching materials in mathematics, and that the analysis would be carried out using eye-tracking analysis to monitor the participants' work on five mathematical items. The students were also informed that all participants are de-identified and that they could withdraw their consent at any time.

Ideally the selection of participants in the *Show* group and the *Try* group would have been matched pairs, which was not the case because during data collection the two types of information presentation were used on different occasions in two different schools. Because the study was designed ad hoc within a larger project data for the *Show* group were collected first and for fewer participants. Information about the participants experiences of digital teaching material in mathematics were gathered through a survey that teachers at the schools completed (Table 1). The survey is designed to capture a generalised view of classrooms in the two schools, and the frequency the resources were used could be ticked at three levels. A guide defined "never" as *never or at one occasion*, "seldom" as *a few times per semester*, and "often" as

several times per semester, maybe every week. The results reveal differences between the groups for questions 4, 5, and 7.

Table 1. Use of Digital Teaching Resources in Mathematics in the Show Group and the Try Group.

	Show group (<i>n</i> =24)		Try group (<i>n</i> =46)
If teachers use particular digital resources in classroom:			
1. Smart board or similar	seldom		seldom
2. Digital quizz	often		often
3. Software to dynamically present mathematics (e.g. GeoGebra)	seldom		seldom
If students use particular digital resources in math:			
4. Computer or padlet	seldom	≠	often
5. Watch short film individually	often	≠	seldom
6. Software to dynamically present mathematics (e.g. GeoGebra)	seldom		seldom
7. Mathematics apps (e.g. for repetition)	seldom	≠	often

Note. The teachers chose between alternatives: never, seldom, often.

Comments given in the survey explains the differences. Regarding question 4: In both show and the try group all student have a personal computer or padlet but computers are not used often in mathematics. The try group explains that "often" refer to the use of computers to look up solutions online or to use a particular app (Magma, see <https://www.magma.se/>). Regarding question 7: The choice "often" for the try group is explained by the use of one particular app (Magma) on a regular basis: "at least all students have access to the app". In summary the differences between the groups are explained by one group's more frequent use of short films and the other group's use of one particular app; therefore it can be concluded that the two groups' experiences of digital teaching materials are fairly similar. The results from the survey only give a general view of the participants experiences of digital teaching material, but at least the results assures that the two groups are not offered very different experiences of digital teaching materials at their schools.

Background information and information from a survey the participants completed was also used to ensure that the two groups were not too different regarding qualities that may play a role for their ability to learn from the instruction and to do a computer-based test. The experimenters followed strict protocols during the data collections and the experimenters had the same role throughout all data collection to

avoid differences else than those intended (Show vs Try). The background information about the participants is presented in [Table 2](#). The information about activities students do more than 7 hour a week is based on options in a Likert scale: “not at all”, “1-3h”, “3-6h”, “7-14h”, “more than 14h”. The participants answered the questions “On your leisure time, approximately how many hours per week do you spend” ... “on computer games?” and “on other digital activities?”. An experimenter was available to answer questions. The two options with at least 7 hours a week are included in the share presented in [Table 2](#).

One difference between the groups is that there is a larger share of girls in the *Try* group, but we have no reason to believe gender plays a large role in interpretation of information before solving the tasks. A comparison between the *Show* and the *Try* group reveals that a larger share in the *Show* group states that they spend more time on computer games and a larger share in the *Try* group does other digital activities. For the digital activities together, however, the share of students is for the two groups .74 and .77 respectively. These numbers of added fractions must be interpreted with caution because the same student can be represented in both the share who plays computer games and the share who does other digital activities.

Table 2. Information About Participants in the Two Groups of Students.

	Show group (<i>n</i> =24)	Try group (<i>n</i> =46)
Fraction of girls in group	.36	.52
Fraction who plays computer game ≥ 7 h/week	.39	.30
Fraction who does other digital activities than gaming ≥ 7 h/week	.36	.47
Mean grade in mathematics (lowest 0, highest 5)	2.43	2.71
Mean grade in Swedish (lowest 0, highest 5)	2.64	2.43

In Sweden a scale F–A is used in grading. F represents to not pass and E–A increasingly higher grades. A comparison of grades in the subjects Mathematics and Swedish revealed fairly similar mean grades between the groups. The *Show* group have slightly higher grades in Swedish and the *Try* group have slightly higher grades in mathematics.

3.3 Mathematics items with different dynamic features

Five different mathematics items were used. All students were presented these items in the same order: Item 1, 2, 3, 4, and 5. All items have some essential Facts that is needed to solve the task, at the right-hand side of the item (Figure 2). The Facts are designed in five versions for every item, based on a typology of elements designed to mirror an increasing interactivity and dynamics from type I to type V (Table 3). Accordingly there are five versions for each item. A counterbalanced combination of items in different versions was used, which ensured each student was offered all five items, and all five types of elements used in presentation of the Facts but in counterbalanced order. Thus, the same timelines with the counterbalanced order of items were used in the *Show* group and the *Try* group.

Table 3. Typology of Elements Used in Facts.

Element type	Dynamic and interactive characteristics
I	the constituents are presented similar to a printed counterpart, but on screen
II	the constituents appear after a click on a button
III	the constituents are presented in a film with a voice
IV	has constituents where the reader needs to choose options by clicking to receive a response and if needed try again and finally make use of the feedback
V	has constituents that change continuously over time when the reader drags or grab and move objects with the mouse

For all element types except the static (type I) and the film (type III) there are labels or instructions that inform the test taker about what is expected. In the static version no actions are needed by the test taker and the film is presented in a very familiar format with a triangle ► as the start button. In type II the button has the inscription “Click to open” and in type IV there were options to click on, a button “Check” and another button “Try again”. In type V there is a sign “Drag” with an arrow pointing to a coloured dot that should be dragged. Examples of the design of dynamic facts of element type IV and V are given in Figure 1.

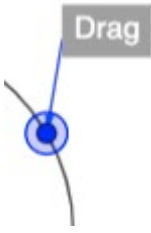
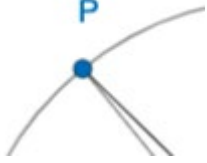
Type IV	Type V	
<p>Facts</p> <p>In this case a is equal to 1.</p> <p>This means:</p> <ul style="list-style-type: none"> <input type="radio"/> a = 1 <input type="radio"/> a = 0 <input type="radio"/> a > 1 <p><input type="button" value="Check"/></p> <p>The statement is now <input type="text"/> <input type="button" value="Try again"/></p>		<p>Drag the point P</p> 
<p>When an option is chosen and “Check” is clicked on, “True” or “False” is displayed. After clicking “Try again”, the user can start over, trying to find the correct option.</p>	<p>The blue dot can be grabbed and dragged with the mouse.</p>	

Figure 1. Examples of Facts Expressed by Dynamic Element Type IV and Type V.

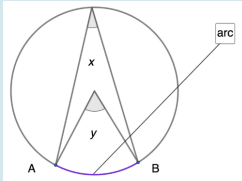
The items were chosen to be new to the participants so that the tasks could not be solved based on previous knowledge only; in particular the intention was that to solve the tasks the participants would need the information provided in the Facts. The mathematical content in five items is: 1) the inscribed angle theorem, 2) maximum and minimum of quadratic functions, 3) set theory, 4) the relation between power and roots, and 5) permutation and factorials. As an example, the item about the inscribed angle theorem and all five versions of Facts are presented in [Figure 1](#). The exact same image and wording is used in Facts type I–III but in Facts II the Facts are displayed after a click on the button and in the film the text and image with arrows appear successively accompanied by a voice reading the text.

Item with Facts type I

Inscribed angle and central angle

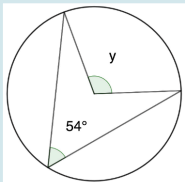
Introduction

The figure shows an inscribed angle x and a central angle y on the same arc AB.



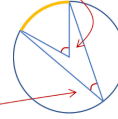
Task

Find the size of the angle y .



Facts

The *central angle*



is always twice the size of the *inscribed angle* on the same arc.

Choose correct answer:

 60°

 104°

 27°

 108°

Facts type II

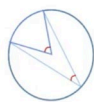
Facts

Click to open

Facts type III

Facts

Start the film

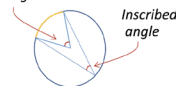


0:00 / 0:17

Facts type IV

Facts

Central angle



When the central angle and the inscribed angle on the same arc is compared it is always true that

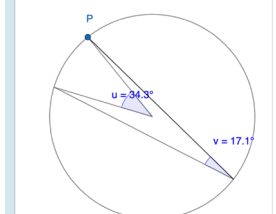
- the angles have equal size
- the inscribed angle is largest
- the central angle is twice as big as the inscribed angle

The statement is

Facts type V

Facts

Drag the point P and see how the size of the angles changes.



Note. The font of the different Facts is the same in all versions, here displayed in different sizes. In Facts V, the value for u disappears for $u > 90$

Figure 2. Example Item with Five Different Versions of Facts.

The development of the typology of elements and further examples of elements Type I–V can be found in Dyrvold (2022). This typology as well as the items were developed within a larger project and accordingly, the dynamic and interactive element types were not chosen to be demanding and test whether the participants chose to use them. In fact, the current study was developed ad hoc when participants choice to not use some Facts were identified. This omission was not expected and accordingly

the current study was deemed important in an era with expansive use of digital assessments.

3.4 Implementation: Information to test takers

To diminish distractions caused by the eye-tracking equipment and to diminish extraneous cognitive load, all participants were individually informed about the setting and about the items before the test. This information was given just before they entered the room where they worked with the items. The information about the eye-tracking was very brief and the purpose was to answer questions and to cope with potential worries. The information about the items were strictly about the visual appearance of the items and about digital functions of the items. Both student groups received the same information about the eye-tracking and about the visual appearance of the items.

All students were shown a test item ([Figure 3](#)) on a screen and were told that the particular item also would appear as the first item in the test setting, before the five items included in the study. They were also told that the test item was intended to be easy and that it only played the role of an example. While looking at the item the students were informed that all items that should be solved had the same main arrangement. Thereafter the students were informed about the main parts in the items. This information was accompanied with gestures pointing at the parts on the screen. Information was given that all items have:

- a title,
- an introduction,
- some facts that are essential to solve the task,
- a task,
- answers to choose from, and
- a button that takes the student to the next item.

Algebra

Introduction

In algebra letters a and b can represent different values.

Task

What is the value of $a + a$?

Facts

In this task use

$a = 1$

Choose the right answer: 2 4 20 To next item

Figure 3. Test Item About Algebra with Facts in Static Form.

The use of a practice item as a first item in the test setting means data from the participants first minute, when getting comfortable with being tracked, do not affect data intended for analyses. Because all items have the same parts (Figure 2) present in the same spot on the screen, the reader does not need to grapple understanding the information flow, which is likely to reduce some extraneous cognitive load. The training with a very easy test item also contributes to making the setting more similar to the use of a digital teaching material that is familiar to the reader, as in the school setting.

After receiving the first information about the main parts in the items, the students were informed that different element types were used to present the Facts in the five items and that these Facts offered information essential to solve the task. Five versions of the test item, about algebra, were used to illustrate all different element types. This information was presented differently to the two groups. The *Show* group received the visual information from a PowerPoint presentation on a laptop. The *Try* group received the visual information from html files on a laptop, and for the four items with Facts that had some dynamic element the *Try* group were asked to try the dynamic element by clicking/dragging or likewise. The spoken information was given based on

a list of information and was therefore the same for the *Show* and the *Try* group, with exception of information related to the different settings.

- For the first element type in static form, the information was the same for both groups. They were informed that these Facts were in static form, visible directly.
- For the second element type both groups were informed that one item had a button that must be clicked on to “open” the Facts. The *Try* group clicked on a button and the Facts appeared whereas the *Show* group was shown where the button would appear. For both groups, the button had the inscription “Click to open” as in the test setting.
- The third element type was in the form of a video. Both groups were informed that they could start the video and look many times if needed. They were also informed that headphones were used so no one (i.e. the experimenter) would reflect over whether they looked many times. The *Try* group was also encouraged to try and start a film.
- For the fourth element type both groups were informed that this type of presentation of the Facts was a bit like a task because they should choose an option and thereafter click “Check” to see if the option was correct. They were also informed that they should try again if they choose the wrong option. Finally, they were instructed not to think of this activity as *the task* but as contributing to make the Facts complete. They should still do the task. The *Try* group clicked on one option and then on “Check”. The *Show* group were only shown a static page displaying the layout of the options and were informed that they were supposed to try different options, check whether it was correct and retry if needed (Figure 1).
- For the fifth element type both groups were informed that there would be some content that can be dragged or moved (Figure 1). The *Try* group moved a slider and saw a number appearing whereas the *Show* group were informed that there would be a prompt to drag and were shown how it could look visually. The experimenter showed a grab-drag gesture in front of the screen visualising how to drag a slider.

The participants were asked if they had any queries and some students asked questions, for example whether they should use a mouse or a trackpad.

3.5 Eye tracking

The participants' activities while working with the practice item and the five mathematics items with Facts expressed using five different element types were captured with an eye-tracking camera. The results from the eye-tracking data are mainly used in other studies, only a minor part of the data is used in the current study. Based on participants' mouse-clicks and eye-fixations on the Facts it was coded whether a participant used the Facts or not. The static Facts was coded as *used* if the participants had fixed their gaze on the constituents in the Facts at least once. For the Facts that have a dynamic component all items were first coded (tentatively) as used if the student had clicked on any button in the facts. To ensure that the participants for which the Facts were coded as *used* actually had used the Facts, screen recordings where students' gaze were displayed were also examined for all participants. Based on the recordings, all cases where participants that had clicked on a button or dragged a dot in the Facts were coded as "used". All participants that used the Facts like this, also spent time reading the Facts, which could be seen by the gaze-path on the Facts. The difference between the measure *gaze fixation* (for Fact I) and *used* (for Facts II-V) means that the static Facts have the role of a reference.

3.6 Transfer as a theoretical frame of analysis

Understanding information and being able to use the information in new situations means that the information needs to be correctly interpreted but also later retrieved and related to a new context (i.e. transfer). In the current study oral explanations are given to all participants whereas one of the two groups also had the opportunity to try and use the dynamic features themselves. Potential differences in outcome between the two groups are therefore likely to be related to the opportunity to learn by trying the different dynamic elements. The hypothesis is that trying supports the ability to transfer information from one situation to another. Theory about transfer provides a frame for the analysis in this study.

Transfer has traditionally been defined as the application of results from prior learning in novel situations (e.g., Gass & Selinker, 1983). At the core of the historical transfer perspective is Thorndike's identical elements (Thorndike & Woodworth, 1901). Identical elements refer to overlapping features between the learning situation and the new transfer situation. This narrow focus on identical elements has been criticised. One reason is that this focus implies that abstracted rule-like processes define

successful transfer of the overlapping features and perceptual richness, for example when different forms of representation are used, is assumed to hinder transfer (e.g., Kaminiski, Sloutsky, & Hecksler, 2013). The concept of transfer has however been developed. Lobato & Hohense (2021) describes that the original, somewhat narrow, cognitivist perspective has been complemented with several other perspectives. For example, the actor-oriented transfer (AOT) perspective has been developed in response to limitations of previous conception of transfer. AOT includes the students' experience prior to the learning situation (e.g., see Lobato, 2012). Taking the actors view on transfer puts the lens on instances where prior experiences shape the students' activities in a transfer situation (Lobato, 2012). This means that this perspective is particularly useful for qualitative studies of how learners interpret situations and make connections (Lobato & Siebert, 2002). Despite the quantitative take in the current study, the AOT perspective is used to elucidate and discuss students' activities in the transfer situations and relate them to the students' previous experiences regarding gaming and other digital activities.

Furthermore, the AOT perspective puts focus on how *contextual sensitivity* can play a role for transfer (Lobato, 2012). Contextual sensitivity is defined as students' ability to utilise knowledge from previous experiences and based on the context adapt the knowledge to a variety of new transfer situations. In the current study, the transfer situation entails students' work on digital items on their own, items designed to minimise superfluous contextual factors. Thus, contextual sensitivity entails adaptation to a stringent context where the dynamic functions are the same as in the learning situation and a test situation where the students work on the digital items independently on their own.

In accordance with Nathan and Alibali (2021) and Goldstone et al. (2008) the role of perception and interactive processes are included as experiences that play roles in transfer. These processes are valuable to include because of the reciprocity between cognition and action (Nathan and Alibali, 2021). In the current study, students in two groups are offered different opportunities for perception and interactive processes when receiving instructions in the learning situation. The students receive instructions with or without the opportunity to try the dynamic functions and thereby experience rich perception and interactivity. In this way, the learning situations offer different modes to the two student groups. Due to the differences offered in the instructions in the learning situation, the ability to form and maintain connections between

the learning and transfer situation, are supposed to be dependent on different demands on sensitivity to context.

For the analysis in this study, we use the notion of transfer to understand the process when students retrieve and transfer information from one context (the examples that the students were shown or offered to try before the test situation) to another context (the test situation). In this study, information refers to the function of the dynamic elements, not to the mathematical content in the items. Thus, successful transfer is expected to show by participants use of a particular element. It is assumed that the *Try* group is advantageous because the interactive preparation offered to this group facilitates the transfer and the students' contextual sensitivity, as well as it provides advantages depending on the relation between action and cognition.

3.7 Statistical analyses

Chi square tests were used to test whether the proportion of students who use the element types differ significantly depending on the kind of instruction given beforehand. Fisher's exact test was used in analyses where at least one cell in the chi square test had an expected count less than 5 and Pearson chi square with continuity correction was used if not. The reported p-values are for two-sided tests and $p < .05$ is used as significance level. The phi coefficient is calculated to analyse effect size and Cohen's criteria for effect size is used as a rule of thumb (Cohen, 1988).

4 Results and analysis

The analyses in this study were conducted to contribute to understanding the role of students' acquaintance with digital functions for how they encounter digital teaching materials and tests. The results reveal differences in the participants' actions depending on which instruction they had received. In particular, students tend to more frequently miss elements that are both dynamic and interactive, if they are not given the opportunity to try similar elements as part of the instruction. Analyses of the participants' interactions with items including the different dynamic functions are displayed in [Table 4](#).

4.1 Results

In summary, because all participants used Facts presented using element type I (static), they appear to have understood that the role of the Facts was to offer information essential to solve the task. Very few participants missed the opportunity to use Facts presented with elements of type II-III. For Facts presented with element types IV and V on the contrary, there were many participants who did not use the Facts. There were significantly more participants in the group who only received instruction (*Show*) that missed the opportunity to use Facts presented using element type IV and the effect size was large. For Facts presented using element type V there was no significant difference between the groups; several participants in both groups did not use the Facts. Recall that element type I is in static form, element type II has a click to open function, element type III is presented as a film, element type IV is in the form of choices to click on options and retry until correct facts are presented, and element type V has a drag option that reveal continuous changes in the presented information (Figure 1).

Table 4. Whether Participants Use (No/Yes) Facts Presented With Different Element Types.

	Show group N=24		Try group ¹ N=46		Fisher's exact test	
	no	yes	no	yes	p-value	Phi-coeff.
Facts I	0 (0%)	24 (100%)	0 (0%)	46 (100%)		
Facts II	2 (8.3%)	22 (91.7%)	3 (6.5%)	43 (93.5%)	1.00	.033
Facts III	2 (8.3%)	22 (91.7%)	0 (0%)	46 (100%)	.114	.237
Facts IV	9 (37.5%)	15 (62.5%)	0 (0%)	46 (100%)	<.001	.532
Facts V	10 (41.7%)	14 (58.3%)	7 (17.9%)	32 (82.1%)	.077	.259

¹ Data from seven participants were excluded from the analysis of Facts V because they could not see these facts or because no data were collected due to technical issues.

² Pearson chi square with continuity correction is used in this test because no cells in the chi-square test have an expected count less than 5.

Based on these results it can be concluded that students are not equally equipped for a computer-based assessment if they are only presented with information in contrast to if they also are given the opportunity to try the dynamic functions that are utilised in the assessment and use them themselves. It can also be concluded that the more interactive and dynamic functions that are included in an assessment the larger is the risk of missed opportunities to fully understand the offered information.

4.2 Analysis

An analysis of the results in relation to actor-oriented transfer elucidates the role of both previous experiences and of modes in the instruction, for the ability to utilise affordances of the dynamic elements in an item. The results differ between items with elements type I-III and type IV-V. No substantial differences between the groups (*Show* and *Try*) previous digital experiences or grades have been identified and it can therefore be assumed any differences in use of the elements are related to the two types of instruction. The element type I (static) plays the role of a standard type to contrast the results against and 100 per cent of the participants used that element. Elements type II-III (click to open, and film) are very similar to elements it can be assumed all participants have experiences of using. These elements are used by almost every participant in the study (95%). This result indicates that previous experiences, presumably from both leisure time and teaching, has provided the participants with sensitivity to context in use of these elements. Because there are no differences between the groups tendency to use the dynamic elements, it is likely the instruction in the learning situation plays a minor role for the use of the element type in the transfer situation, or at least that rich perception and interaction in the instruction is not needed. From a validity perspective this implies that in computer-based tests there is no substantial difference in the students' achievement due to instructions or previous digital experiences when element type II and III are being utilised in the test. This means that in environments similar to the Nordic school context these kinds of digital elements do not threaten the accuracy by which the test distinguishes between test takers based on their mathematical proficiency.

The largest threat against validity is however identified in relation to dynamic elements type IV and type V. For element type IV there is a significant difference in the use of facts ($p < .001$, Phi coeff. = .53) between the groups and for element type V the difference is nearly significant ($p = .077$). The conclusion is made because the results indicate that a thorough instruction that includes an opportunity to try the element is essential to assure transfer related to the use of those elements. The need for instruction is apparent for these elements and even with instruction (*Show* or *Try*), as many as 20 percent of the participants miss the opportunity to use at least one of the elements (IV & V). It can be assumed that the participants intend to make connections between previous experience and the transfer situation, but because the demands of interaction with the dynamic environment is large, knowledge from previous experi-

ences may not be sufficient for successful transfer. Experiences from previous situations including dynamic and interactive elements likely differ both in features of the elements, and in the context they have been met, which can make transfer harder. Accordingly, when utilising dynamic elements that differ from those used in previous situations, the design of the instruction is eminent, and our results reveal that this is particularly important when the elements are highly dynamic and interactive as elements type IV and V.

The only element type that was used to a significantly different amount between the two instruction groups was type IV, the element where students are supposed to click, check, and potentially retry, to compose a correct mathematical statement and use that to solve the task. All participants in the *Try* group used the element in the transfer situation which means that the perceptually rich and interactive instruction were sufficient for successful transfer. On the contrary, 37.5 percent of the participants who did not get the opportunity to try elements in the instruction abandoned the option to use the dynamic element. This may be caused by differences between modes offered in the learning situation and the transfer situation, which can put too large demands on students' sensitivity to context in the target situation (the test). This result highlights a threat to validity and a reasonable source to it, namely instruction that does not provide a learning situation sufficient for the test takers to form and maintain connections between the learning situation and the transfer situation. An unwanted consequence of such instruction is construct irrelevant variance, leading to wrong inferences from an assessment.

5 Discussion

Computer based tests in mathematics are becoming more and more common as schools and teaching in general becomes more and more digitalised. Computer based tests make it possible to take advantage of features unique to the digital media, for example the inclusion of real-life data and various dynamic features (e.g., see Yerushalmy & Olsher, 2020; Ripley, 2009). Thus, innovative and computer-based tests can be used to assess other skills than those possible to test in paper-based tests; for example modelling competence may be easier to test accurately in a digital than a paper-based test. One purpose of the PISA-test for example, is to assess digital literacy, but most often the assessment is carried out with other intentions; an example is Swedish national tests where one aim is to ensure equal grading between different teachers, schools and principals. For high-stake assessments validity is, of course, of

utmost importance, and an important issue with relevance for a sustainable mathematics education is how validity can be ensured.

We can assume that it is crucial that the students master the digital environment to ensure validity in a computer-based test and there is also evidence that experience with the digital medium can lead to better test results (e.g., Baccaglioni-Franck, 2021; Dadey et al., 2018). There is also evidence that students must be accustomed to using dynamic functions to perform well (Lemmo, 2021; Harris et al., 2021). The current study contributes by highlighting the substantial risk to overlook or underestimate the need for apt instructions as preparation for a computer-based assessment. The use of a digital dynamic interface leads to an enormous increase in options about tools to use and accordingly options for the reader. The dynamic elements used in the current study had labels which informed the students about the use of the dynamic functions and the participants also received instruction before the test. Despite these instructions many participants still did not use the dynamic functions. Part of the explanation to the unwillingness to explore the digital environment can be due to unfamiliarity with the digital frame. As shown in previous research (see e.g., Dyrvold, 2022 and Ilovan et al., 2018) dynamic functions are relatively sparsely used in contemporary digital teaching materials. This means that even students who are used to working with digital textbooks during mathematics lessons, can be assumed to be inexperienced in using dynamic functions. Everyone who as a user has experienced a transition from one familiar digital platform to another, or the need to orient in a new digital administrative tool, likely recognise the frustration and lack of grit that may lead to abandoning to even try to grasp the new functions. When students choose to answer tasks on paper instead of digitally (Davis et al., 2021), the choice makes sense because the risk of misunderstanding how to present a solution using a pen is minimal, whereas missing a digital option is something that can happen. Bearing that in mind, the large share of students who missed the opportunity to use the highly dynamic and interactive elements type IV and V despite information just before the test situation, is less surprising.

As the use of computer-based assessments are spreading, we see a substantial risk for an increase in validity issues stemming from unfair demands of digital skills or of willingness to explore the media. If computer-based assessments are used as diagnosis tools or as part of formative assessments the risk is not as alarming because in a less stressful situation, it is more likely the test taker during the test develops a sensi-

tivity to context and thereby manages to benefit from using the offered dynamic elements as expected. What we learn from this study is therefore applicable predominantly for assessments used for grading or to rank participants.

In retrospect, follow up interviews with the participants who did not use the dynamic elements would have contributed to the study. More participants in the *Show* group would also have strengthened the reliability of the results and there is a chance that more participants would have explored the content and eventually used the dynamic elements if the test was part of their mathematics course. These circumstances are important to address, and it is possible the results of a follow up study would differ to some extent. Despite these development areas however, the differences in use between more or less dynamic and interactive elements and between students who get the opportunity to try the dynamic elements beforehand or not, are convincing. Ideally the study would have been designed with matched pairs in the Try and Show group. The results from the survey to teachers and to the participants (Table 1 & 2) do however reveal that the participants in the two groups have fairly similar experiences of digital teaching materials and grades in mathematics and Swedish and it is therefore likely that the differences between the groups stem from differences in instruction.

5.1 Conclusion

Based on the statistics and the analysis according to the AOT perspective two main conclusions are drawn. Firstly, if dynamic elements utilised in a test are present in different contexts in several digital devices that students have used earlier, it is likely that these experiences can be transferred to the test situation. Because the layout of a test is likely to differ from students' previous experiences it is however suggested to at least show how the elements appear in the digital environment and make room for possible questions. Secondly, there is a substantial risk to overestimate students' capability to successful transfer from previous digital experiences and their capability to be sensitive to the new digital context (a test). If the capability to correctly use a dynamic element is not part of the construct being assessed, it is therefore recommended the students are given the opportunity to use all dynamic and interactive elements in a digital environment similar to the test before the test. The first conclusion is based on results where dynamic elements where click to open (type II) and film (type III) are used, and the second conclusion is based on results for the more dynamic and interactive elements (type IV-V). It is stressed that the options and buttons used in

element type IV are very similar to response options used in online formulas or multiple-choice questions that are used very frequently in today's society. Despite that, many participants who did not try using the element in a similar context as in the test were not able to transfer previous experiences and thus missed opportunities in the test situation. This result highlights mode effect as a potential threat to the validity of an assessment, in particular when the demand for interaction is of a different kind than what is experienced by most citizens.

Acknowledgements

This work was supported by the Swedish Research Council [grant number 2019-05005].

References

- Aldon, G. & Panero, M. (2020). Can digital technology change the way mathematics skills are assessed? *ZDM*, 52(7), 1333–1348. <https://doi.org/10.1007/s11858-020-01172-8>
- Baccaglioni-Frank, A. (2021). To tell a story, you need a protagonist: How dynamic interactive mediators can fulfil this role and foster explorative participation to mathematical discourse. *Educational Studies in Mathematics*, 106(2), 291–312. <https://doi.org/10.1007/s10649-020-10009-w>
- Barana, A., Marchisio, M., & Sacchet, M. (2021). Interactive feedback for learning mathematics in a digital learning environment. *Education Sciences*, 11(6), 279–290. <https://doi.org/10.3390/educsci11060279>
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39(1), 370–407. <https://doi.org/10.3102/0091732X14554179>
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., Yan, F. (2008b). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 6(9), 1–38.
- Bennett, S., Maton, K., & Kervin, L. (2008a). The “digital natives” debate: A critical review of the evidence. *British Journal of Educational Technology*, 39(5), 775–786. <https://doi.org/10.1111/j.1467-8535.2007.00793.x>
- Cohen, J. W. (1983). *Statistical power analysis for the behavioral sciences* (2nd ed.) Lawrence Erlbaum Associates.
- College Board. (2022). Assessment framework for the digital SAT suite, version 1.0 (June 2022). College Board.
- Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, 31(1), 30–50. <https://doi.org/10.1080/08957347.2017.1391262>
- Davis, L. L., Morrison, K., Zhou-Yile Schnieders, J., & Marsh, B. (2021). Developing Authentic Digital Math Assessments. *Journal of Applied Testing Technology*, 22(1), 1–11. Retrieved from <http://jattjournal.net/index.php/atp/article/view/155879>

- Dyrvold, A. (2022). Missed opportunities in digital teaching platforms: Under-use of interactive and dynamic elements. *Journal of Computers in Mathematics and Science Teaching*, 41(2), 135–161.
- Gass, S. M., & Selinker, L. (1983). *Language transfer in language learning. Issues in second language research*. Newbury House Publishers, Inc.
- Geraniou, E., & Jankvist, U.T. (2019). Towards a definition of “mathematical digital competency”. *Educ Stud Math* 102, 29–45. <https://doi.org/10.1007/s10649-019-09893-8>
- Goldstone, R., Son, J. Y., & Landy, D. (2008). A well grounded education: The role of perception in science and mathematics. Symbols and embodiment (pp. 327–356). Oxford University Press. <http://jattjournal.net/index.php/atp/article/view/155879>
- Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: are there performance differences between timss’ paper-and-pencil test and tablet test among dutch grade-four students? *British Journal of Educational Technology*, 51(6), 2340–2358. <https://doi.org/10.1111/bjet.12914>
- Hanho, J. (2014) A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, (33)4, 410–422. <https://doi.org/10.1080/0144929X.2012.710647>
- Harris, D., Logan, T., & Lowrie, T. (2021). Unpacking mathematical-spatial relations: Problem-solving in static and interactive tasks. *Mathematics Education Research Journal*, 33(3), 495–511. <https://doi.org/10.1007/s13394-020-00316-z>
- Helsper, E. J., & Eynon, R. (2010). Digital natives: where is the evidence? *British Educational Research Journal*, 36(3), 503–520. <https://doi.org/10.1080/01411920902989227>
- Hoch, S., Reinhold, F., Werner, B., Richter-Gebert, J., & Reiss, K. (2018). Design and research potential of interactive textbooks: the case of fractions. *ZDM Mathematics Education*, 50(5), 839–848. <https://doi.org/10.1007/s11858-018-0971-z>
- Ilovan, O.-R., Buzila, S.-R., Dulama, M. E., & Buzila, L. (2018). Study on the features of geography/sciences interactive multimedia learning activities (IMLA) in a digital textbook. *Romanian Review of Geographical Education*, 7(1), 20–30. <https://doi.org/10.23741/RRGE120182>
- Junpeng, P., Krotha, J., Chanayota, K., Tang, K., & Wilson, M. (2019). Constructing progress maps of digital technology for diagnosing mathematical proficiency. *Journal of Education and Learning*, 8(6), 90–102. <https://doi.org/10.5539/jel.v8n6p90>
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2013). The cost of concreteness: The effect of nonessential information on analogical transfer. *Journal of Experimental Psychology. Applied*, 19(1), 14–29. <https://doi.org/10.1037/a0031931>
- Lemmo, A. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education*, 19(8), 1655–1675. <https://doi.org/10.1007/s10763-020-10119-0>
- Lobato, J. (2012). The actor-oriented transfer perspective and its contributions to educational research and practice. *Educational Psychologist*, 47(3), 232–247. <https://doi.org/10.1080/00461520.2012.693353>
- Lobato, J. & Hohense, (2021). Current conceptualisations of the Transfer of learning and their use in STEM education research. In C. Hohensee & J. Lobato (Eds.), *Transfer of learning: Progressive perspectives for mathematics education and related fields* (pp. 3–26). Springer.
- Lobato, J., & Siebert, D. (2002). Quantitative reasoning in a reconceived view of transfer. *The Journal of Mathematical Behavior*, 21(1), 87–116. [https://doi.org/10.1016/S0732-3123\(02\)00105-0](https://doi.org/10.1016/S0732-3123(02)00105-0)

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741–749.
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks>
- Nathan, M. J. & Alibali, M. W. (2021). An embodied theory of transfer of mathematical learning. In C. Hohensee & J. Lobato (Eds.), *Transfer of learning: Progressive perspectives for mathematics education and related fields* (pp. 27–58). Springer.
- OECD (2013), PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, OECD Publishing. <http://dx.doi.org/10.1787/9789264190511-en>
- OECD (2021). *21st-Century Readers: Developing Literacy Skills in a Digital World*. PISA, OECD Publishing. <http://dx.doi.org/10.1787/9789264190511-en>
- O'Halloran, K. L., Beezer, R. A., & Farmer, D. W. (2018). A new generation of mathematics textbook research and development. *ZDM Mathematics Education*, 50(5), 863–879. <https://doi.org/10.1007/s11858-018-0959-8>
- Prensky, M. (2001). Digital Natives, Digital Immigrants. *On the Horizon*, 9(5), 1–6.
- Regeringen (2017). Uppdrag att digitalisera de nationella proven. <https://www.regeringen.se/4a80ac/contentassets/03dee5c5cdf244afa053e26cf654a8d8/updrag-att-digitalisera-de-nationella-proven-m.m..pdf>
- Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 92–98). Office for Official Publications of the European Communities
- Smolinsky, L., Marx, B. D., Olafsson, G., & Ma, Y. A. (2020). Computer-based and paper-and-pencil tests: A study in calculus for STEM majors. *Journal of Educational Computing Research*, 58(7), 1256–1278. <https://doi.org/10.1177/0735633120930235>
- Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8, 247–261.
- Usiskin, Z. (2018). Electronic vs. paper textbook presentations of the various aspects of mathematics. *ZDM Mathematics Education*, 50(5), 849–861. <https://doi-org.ezproxy.its.uu.se/10.1007/s11858-018-0936-2>
- Yerushalmy, M., & Olsher, S. (2020). Online assessment of students' reasoning when solving example-eliciting tasks: Using conjunction and disjunction to increase the power of examples. *ZDM Mathematics Education*, 52(5), 1033–1049. <https://doi.org/10.1007/s11858-020-01134-0>