Mesure et évaluation en éducation

Digital assessment of mathematics: Opportunities, issues and criteria

Paul Drijvers

Volume 41, Number 1, 2018

URI: https://id.erudit.org/iderudit/1055896ar
DOI: https://doi.org/10.7202/1055896ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et évaluation en éducation*, *41*(1), 41–66. https://doi.org/10.7202/1055896ar

Article abstract

Digital assessment of mathematics is becoming widespread, but still comes with limitations and constraints. A central question is how to design digital tests that assess mathematical knowledge in a valid way. Based on literature on validity and on assessment with and through technology, we identify arguments for and opportunities of digital assessment of mathematics, as well as its main issues. Through three case descriptions, different ways to design digital tests are explored. As a conclusion, we make a plea for assessment environments which offer rich opportunities for students to "do" mathematics and for test designers to design rich items; automated scoring also needs further development, with respect to the scoring of intermediate steps in problem-solving strategies.

Digital assessment of mathematics: Opportunities, issues and criteria

Paul Drijvers
Utrecht University

Key words: digital assessment, mathematics, automated scoring, item design

Digital assessment of mathematics is becoming widespread, but still comes with limitations and constraints. A central question is how to design digital tests that assess mathematical knowledge in a valid way. Based on literature on validity and on assessment with and through technology, we identify arguments for and opportunities of digital assessment of mathematics, as well as its main issues. Through three case descriptions, different ways to design digital tests are explored. As a conclusion, we make a plea for assessment environments which offer rich opportunities for students to "do" mathematics and for test designers to design rich items; automated scoring also needs further development, with respect to the scoring of intermediate steps in problem-solving strategies.

Mots-clés : évaluation numérique, mathématiques, correction automatique, conception d'items

Si l'évaluation des mathématiques en format numérique est de plus en plus répandue, elle n'est pas sans limites ni contraintes. Une question importante à ce sujet concerne les manières de concevoir des tests numériques qui évaluent les connaissances en mathématiques de façon valide. En nous appuyant sur la littérature sur la validité et sur l'évaluation par l'entremise de la technologie, nous présentons des arguments en faveur de l'évaluation numérique en mathématiques, et nous discutons des occasions de l'employer et des problèmes principaux de ce type d'évaluation. Par l'entremise de trois descriptions de cas, nous explorons différentes façons de concevoir des tests numériques. Pour conclure, nous encourageons les environnements d'évaluation qui offrent aux élèves des occasions de « faire » des mathématiques et aux concepteurs de concevoir des items riches. Nous avançons aussi que la correction automatique nécessite de l'amélioration, notamment en ce qui a trait aux points accordés pour les étapes intermédiaires des stratégies de résolution de problèmes.

Palavras-chave: avaliação digital, matemáticas, correção automática, conceção de itens

A avaliação das matemáticas em formato digital está cada vez mais difundida, mas, não obstante, apresenta limites e constrangimentos. Uma questão importante a este respeito tem a ver com os modos de conceber os testes digitais que avaliam o conhecimento em matemáticas de modo válido. A partir da literatura sobre a validade e a avaliação por meio da tecnologia, apresentamos argumentos a favor de uma avaliação digital em matemáticas e discutimos as oportunidades de usá-la e os principais problemas deste tipo de avaliação. Através de três descrições de casos, exploramos diferentes maneiras de conceber testes digitais. Para concluir, incentivamos os ambientes de avaliação que oferecem aos alunos oportunidades de "fazer" matemática e aos conceptores de conceber itens ricos. Sugerimos também que a correção automática requer melhorias, especialmente no que diz respeito aos pontos acordados para as etapas intermédias das estratégias de resolução de problemas.

Author's note: Correspondence related to this article may be sent to [p.drijvers@uu.nl].

We thank Sietske Tacoma and Peter Boon for their input on domain reasoners, and Ger Limpens, Pepe Palovaara, and Irene Van Stiphout for their valuable comments on earlier versions of this paper.

Introduction

Nowadays, assessment increasingly takes place with and through digital means. Different types of tests for different purposes and target groups are delivered online, or count on the availability of technological tools such as calculators and computers. In addition to initiatives at the national level in this respect, international comparative tests are already administered online (e.g., PISA since 2012)[1] or will be delivered as online tests shortly (e.g., TIMSS in 2019)[2]. For mathematics, assessment with digital technology is becoming common, and different countries adopt a variety of policies to do so (Brown, 2010; Drijvers, Monaghan, Thomas, & Trouche, 2015).

For high-stakes summative tests such as national mathematics examinations the use of digital means is under debate and many questions arise. Can we go beyond straightforward multiple-choice tasks and make students really "do mathematics" in a digital test? Can students' digital work be scored automatically with the sophistication and subtlety that is common in human scoring of paper-and-pen work? How can we avoid assessing students' ICT literacies rather than their mathematical knowledge?

These questions show that the quality of digital summative assessment in mathematics is an important topic, not in the least because of its feed-forward to the preceding teaching and learning practices. A central question in this paper, therefore, is how to design tests with and through digital technology that assess student knowledge in a valid way, and that provide them with opportunities to express themselves mathematically.

To address this question, we will identify opportunities and issues of digital assessment of mathematics, which will be underpinned by some illustrative cases. From this, the main criteria for assessment environments for mathematics will be inferred, both from a learner's perspective and a teacher or test designer's perspective. To conclude, we extrapolate these criteria to a future research and design agenda.

Theoretical perspectives

To consider digital assessment of mathematics in more detail, some theoretical perspectives may be helpful. A first, very general point of departure concerns *test validity*. Clearly, assessment needs to be valid. What do we mean by validity and how can we go beyond the initial notion of "measuring what you intend to measure"? As shown in Figure 1, Wools (2015) defines validity as a chain of inferences (see also Kane, 2013; Wools, Eggen, & Sanders, 2010). First, a student's performance is translated into a (usually numerical) score. Next this score is extrapolated to the test domain, a fictive set of all possible assignments that could reasonably be part of a test on the topic at stake. Then, we generalize the test domain into the competence domain, which refers to the competences to be assessed, for example in terms of curricular goals. The next extrapolation concerns the practice domain, in our case mathematical practices either in school or out-of-school. This chain of inferences, finally, leads to a decision, which can be a pass or a fail decision, a grade, a suggestion for the follow-up learning process, or a diagnosis. Such a chain, however, is as strong as the weakest link. In our case, the test domain, competence domain and practice domain focus on mathematics. If the performance demands a high level of mastery of the digital tool, that to a certain extent may not be related to mathematical performance, the first link may suffer from that and the validity as a whole may be affected. Even if technical and mathematics skills may not be unrelated, a careful inspection of the students' familiarity with the digital techniques needed to answer the items is needed to make sure that validity is not threatened by the digital format of the test. This is in line with the findings by Threllfall, Pool, Homer, and Swinnerton (2007), who report remarkable differences in student scores on similar items, delivered on paper and in a digital environment, respectively. The medium matters.

Validity is, of course, related to the targeted learning goals and the underlying views on the topic of those who set these goals. If appropriate problem-solving strategies are considered more important than correct outcomes of procedural work, then the test should allow for assessing these strategies, and multiple-choice items may be of limited value.

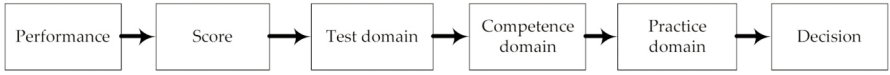


Figure 1: Validity as a chain of inferences (Wools, 2015, p. 21)

A second point that may guide the paper is the distinction of two types of technology-rich assessment, which we call assessment with digital technology and assessment through digital technology (Drijvers, Ball, Barzel, Heid, Cao, & Maschietto, 2016). *Assessment with digital technology* refers to paper-and-pen written tests, during which students have access to digital technology such as (graphing or CAS) calculators or computers. While the digital technology may be helpful for the candidate to get ideas, to visualize or explore situations, or to check answers, the results need to be written down by the student on paper. Such a model is used in final national examinations in many countries (Brown, 2010; Drijvers, 2009). While conventional paper-and-pen assessment is the point of departure here, one may wonder how the availability of the digital tools may impact on both the type of assignments and the students' solving strategies. As such, assessment with technology is more subtle than it might seem at a first glance, and test validity as mentioned above comes into play. To deal with this, policy makers in some countries have decided to have a technology-free part of the examination in which the students' paper-and-pen skills can be assessed without any interferences of digital technology.

We speak of *assessment through digital technology* when technological means are used to deliver and administer assessment. Think, for example, of online tests, in which all student responses are entered in the digital assessment environment. Examples of such environments for mathematics are well-known systems such as Maple TA™, and more specific environments such as the French PÉPITE (Grugeon-Allys, Chenevotot-Quentin, Pilet, & Prévit, 2018), and the Dutch Digital Math Environment[3]. A number of assessment software case studies can be found in Sangwin (2013). In the *through* technology case, the role of digital technology is more important than in the *with* technology case: it puts higher demands on the environment's opportunities to enter mathematical solutions and arguments and to allow students to construct their mathematical solutions. Also, it puts higher demands on the students' skills to use the environment's interface for entering all this and for using these opportunities. Again, any hindrances in this respect may endanger test validity, and this

will turn out to be one of the main themes in this article. It is our impression that the worldwide trends in assessment of mathematics can be interpreted as a movement from assessment with to assessment through technology. Therefore, in this paper we will focus on the latter.

As a third perspective, we want to focus on *automated scoring of student responses* through digital means. From the world of Intelligent Tutoring Systems (ITS), we learn that computer tutoring may be nearly as effective as human tutoring, and a plea is made for the use of ITS for assessment (VanLehn, 2008, 2011). In a study of International Baccalaureate examination assignments, Sangwin and Köcher (2016) claim that a significant proportion of assignments for 18-year-old students can be marked automatically according to marking schemes set up for human scoring, particularly if the digital scoring engine can deal with reasoning by equivalence. In the meantime, mathematics teachers seem to feel more confident about their own scoring, including the scoring of intermediate steps made by the students, or the continuation of the work after an initial mistake. Automated scoring of mathematics assignments, therefore, is not yet considered as doing justice to student work as much as human grading is. In this paper, we investigate how automated scoring of mathematics might be improved, particularly for tasks that go beyond very straightforward applications of standard procedures. This is all the more important in the light of adaptive tests, in which the student performance while doing the assessment is guiding the delivery of new items. Clearly, this can only be done if the previous items are automatically scored.

To summarize this section, we decide to focus on test validity, and on assessment through technology rather than with technology. We have a particular interest in automated scoring as a prerequisite of adaptive testing.

Arguments

Why is digital assessment becoming so widespread? What are the main reasons to challenge the traditional paper-and-pen format that dominated summative assessment practices over centuries? Following Stacey and Wiliam (2013) and Drijvers and his collaborators (2016), the following arguments for digital assessment seem to be the most important ones:

1. The rich item argument

   Digital assessment offers opportunities for rich and dynamic item types, in which film, animations, simulations, and other resources can be included; the student can interact with the material in a way that would not be possible in a paper-and-pen test.

2. The anachronism argument

   Now that digital technology is omnipresent in both daily life and professional practices, and digital tools are used more and more in education, it can be considered an anachronism to refrain to just the traditional paper-and-pen medium when (summative) assessment comes into play. To better reflect previous education and to better prepare for future demands, assessment should include the use of contemporary tools that students use outside the testing setting. Paper-and-pen testing no longer fits in the digital era that we find ourselves in.

3. The outsourcing argument

   Educational goals go beyond the mastery of basic procedural work, and include higher-order skills. Think of the attention paid to 21st century skills (P21, 2015), and, for the case of mathematics, to mathematical thinking (Devlin, 2012; Drijvers, Kodde-Buitenhuis, & Doorman, submitted). The availability of digital tools seems to trivialize part of the basic procedural work and, therefore, questions its importance. In the meantime, outsourcing basic procedural work to digital tools during assessment may save time and may offer opportunities to better address the higher-order thinking skills in tests.

4. The delivery argument

   Digital assessment allows for delivering a test simultaneously in different places. In addition, if a test is created through sampling from a large set of test items, the test can be delivered at different moments in time. Clearly, this would not be possible for paper-and-pen tests, unless the assignments would be kept secret. In general, delivery of digital tests is less time and place dependent than delivery of paper-and-pen tests.

5. The production argument

   Once an extended item database is set up, the production of new, comparable versions of a test is straightforward. In addition, if psychometric data of a large number of students for each item is available, the level of difficulty of a test can be controlled in a way that would be much harder to achieve through paper-and-pen media.

6. The feedback argument

   Digital assessment environments may generate automated feedback. Such feedback can not only be technical, but may also focus on the mathematical content. Hints, diagnostic reports, and other forms of support can be provided to scaffold the learning process. Of course, the timing of such feedback is crucial, and it applies to formative assessment goals rather than summative settings. The feedback may also be stored internally and be delivered to the student after finishing the work for diagnostic purposes.

7. The scoring argument

   The automatized scoring of student responses is an important argument, as this may not only save much time for the teacher, but also may lead to an improvement with respect to objectivity and consistency. In addition, the results of the automated scoring may be used for learning analytics.

8. The adaptation argument

   Different from paper-and-pen testing, the test items of a digital test may be adapted to the student's level on the fly, i.e., while the test is administered. For example, the level of the next assignment may be adapted to the student's results so far. In this way, the adaptive test can focus on the appropriate level for the student and hence can measure student skills more efficiently. Automated and immediate scoring, addressed in the previous point, is a prerequisite for adaptive testing.

The eight arguments of this non-exhaustive list to a certain extent hold for digital assessment in all subjects; as we will see below, they do, however, have specific repercussions for the assessment of mathematics. Furthermore, most of the above points only hold for assessment through technology, that is, for assessment through a digital assessment player. In

fact, only the first three arguments also apply to assessment with technology, such as written national examinations during which the students have access to graphing calculators.

Opportunities

From the above arguments, we can infer some important opportunities that assessment through digital technology may offer to students, test designers and test graders.

For *students*, the digital assessment environment might provide them with means to express themselves mathematically in appropriate and sophisticated ways. They can easily construct neat graphs and geometrical objects, explore properties of such objects, and change them. In short, the medium offers room for construction and expressiveness, which is in line with what was hoped for in early views on the use of digital tools in mathematics education (Noss & Hoyles, 1996). As a second opportunity, students may benefit from both the digital test's adaptivity (see argument 8), which will free them from spending too much time on items that are too hard or too easy to them, and from the feedback that may be provided to them, either on the fly for formative purposes, or after finishing a test for diagnostic means (see argument 5). With respect to the latter, Grugeon-Allys, Chenevotot-Quentin, Pilet, and Prévit (2018) provide an example for feedback on algebra, and Tacoma, Drijvers, and Boon (2017) present a diagnosis in a university-level course on statistics, based on the design and use of a so-called student model.

For *test designers*, digital assessment environments offer opportunities for the design of rich, dynamic and interactive items. Think of items in which students can manipulate objects to explore invariant properties, or can construct their own examples. A variation of item types can be used. Items that have multiple correct answers or solution strategies can be designed. Appropriate feedback design and scoring rules or marking schemes (see arguments 6 and 7) can be set up and implemented (Sangwin & Köcher, 2016).

For *test graders*, in many cases the teacher who assigns the work to their students, digital assessment environments may drastically lighten their time-consuming burden through different types of automated scoring facilities (see argument 6). First, the use of a Computer Algebra System for the interpretation of numerical and algebraic expressions and

formulas may allow for a subtle scoring of algebraic responses (Sangwin, 2013). Second, Boolean variables in Dynamic Geometry Systems offer opportunities for the automated scoring of geometrical constructions (Kovacs, Recio, & Vélez, 2017). Third, the use of so-called Domain Reasoners for a specific domain within mathematics may allow for the interpretation and scoring of intermediate steps and responses. As the example later in this paper will show, subtle automated scoring is becoming more and more adequate nowadays.

In short, there are many arguments that favor a transition from paper-and-pen to digital assessment, and such a transition in principle might offer opportunities to students, test designers and teachers. In the next section, however, the reality of digital assessment of mathematics will appear to be less favorable than we would like it to be, and we will address some issues.

Issues

Despite the above arguments for and opportunities of digital assessment, it is not self-evident that digital assessment of mathematics is a widely accepted trend towards high-quality assessment. There is an ongoing debate on the opportunities and pitfalls of digital assessment of mathematics, and we identified the following main issues that prevent it from becoming the success that one might expect it to be.

A first issue concerns the *practical demands* of digital assessment with and through technology. How can we ensure that all students have access to the digital technology? Do schools have computer work stations, or do they apply BringYourOwnDevice policies? If students bring their own device, how can we be sure that the device meets required regulations and specifications, so that, for example, Internet access is not possible in the case of computers, and that computer algebra is not available for the case of graphing calculators? How can examination rooms be set up so that students cannot look at each other's screens? What about security, privacy, possible hacks or attacks to school networks? What is the scenario if there is a network breakdown? Even though these issues should be taken seriously, we expect them to be solved in the near future and consider them as out of the scope of this paper.

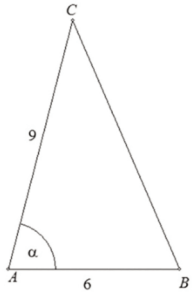
A second issue, more related to the mathematical content, concerns *the skills the students need* to use the technology appropriately. To find the zeros of a function using a graphing calculator during an examination, for example, requires some familiarity with the corresponding techniques on the device, including some notion of the syntax as well as of its limitations: depending on the type of device and on the type of function, not all zeros will be displayed. The student needs to be aware of the device's constraints and idiosyncrasies. If the learning process did not address the subtle interplay between mathematical knowledge and technical skills (cf. instrumental genesis, Trouche & Drijvers, 2010), one can wonder if the assessment is on mathematics or on technological skills. This issue may seriously question the test validity.

A third issue concerns *the limitation of the assessment environment*. Many environments used for assessment through technology provide limited means for mathematics. Equation editors, graphing tools, geometry construction tools, statistical tools are often lacking or only available in rudimentary form. This means that the students only have limited means to express themselves mathematically, for example through graphs and formulas, intertwined with natural language. In a paper-and-pen environment, however, they can sketch, write and scratch whatever they want; there are no obstacles that hinder them in showing their mathematical skills. This type of limitations in digital assessment may challenge the mathematical expressiveness that we want digital tools to offer, in learning as well as in assessment (Noss & Hoyles, 1996).

A fourth issue, finally, concerns the *limitations of automated scoring* that assessment environments offer. As will be shown in the example that follows, these limitations may lead to very artificial and non-authentic items. Such items not only may offer a strange view on mathematics, but they may also unnecessarily confuse students, which also threatens the test's validity. Also, automated scoring engines may have difficulties to identify correct further steps after an initial mistake, which would be a very natural thing to do for a human grader (VanLehn, 2008).

Given

For a triangle ABC holds $AB = 6$, $AC = 9$ and $\cos(\alpha) = \frac{1}{4}$ Also, $BC = \sqrt{k}$



Question

What is the value of k?

Given

The following system of equations with $z \neq 0$

$$\begin{cases} x - y = 0 \\ x + z = 2 \\ xy = 4 \end{cases}$$

Question

What are the solutions of the system?

Provide the answer as the product of x, y and z.

Figure 2: Two items for a Dutch test for pre-service mathematics teachers[4]

To illustrate the third and fourth issues, which mainly refer to testing through technology, Figure 2 shows two items from a Dutch online test for pre-service mathematics teachers, who will be teaching to 12-15 year-old students. In the left-hand screen, the lengths of AB and AC are provided, and the cosine of α equals ¼. The task is to find k so that the length of BC is equal to $\sqrt{k}$. As a first remark, we notice that the expressions for cos(α) and BC are not well aligned and that the parentheses around are too big. More important is that the natural question here would be to calculate BC. Why this strange detour with $\sqrt{k}$? The reason is the system's inability to deal with square root signs in student answer windows, like $3\sqrt{10}$ in this case. The students simply cannot directly enter the answer as a square root, and are asked to enter 90 instead. This limitation with respect to equation editing led the test designers to change the item into this artificial form, which clearly might puzzle candidates (Drijvers, Straat, & Wools, 2016).

In Figure 2's right-hand screen, the task is to solve a system of three equations with three unknowns. As an answer, the product of the three solutions x, y and z is requested. Again, there are some presentation issues,

such as the outlining of the condition for z being unequal to 0, and the final equation not being italicized. The most striking aspect of the problem is that the product of the solutions is asked for. First, the correct product doesn't guarantee a correct solution of the system of equations, and second, why would one want to know this? The reason for this artificial phrasing is that the assessment player and its scoring module are unable to deal with a set of three solutions. The constraint on z is added to avoid two solution sets. Once more, it is the constraints of the system that make test designers move away from what would be mathematically natural and sense making.

These examples show that the issues, if not dealt with appropriately, may lead to strange test items, in which students do not have the means to "move freely" through their mathematical knowledge, but are constrained by the limitations of the assessment environment. In such cases, we may wonder if digital assessment of mathematics is a step forward towards high-quality and valid assessment.

Three illustrative cases

The Finnish Abitti initiative

To illustrate the way in which we might deal with the above opportunities and issues, we now sketch three cases of digital assessment which we consider as relevant to its future development. The first case concerns the national assessment of mathematics for 18-year-old high-achieving students in Finland. The examination consists of two parts, one written part in which the students have no access to any form of digital technology, and a second part in which students can use (laptop) computers with different kinds of software[5]. It is this latter part that is of interest for this paper, and for which the Finnish Matriculation Examination Board took the initiative to design an electronic examination system called Abitti[6]. After starting Abitti through a USB stick, the student's computer is in a Bootable Client Lock-Down (BCLD) mode, which means that no access to software outside Abitti is possible, including Internet. Within Abitti, a player is available with the examination's task assignments and a worksheet for the candidate. In addition to this, and this is where the main point comes in, a range of mathematical software is available, offering different mathematical tools. Figure 3 contains a screen capture of the part of the Abitti menu showing the mathematical software currently available, includ-

ing GeoGebra, Maxima, Casio ClassPad and TI-Nspire. While working on a task, students can choose the software that they find most appropriate or that they are most familiar with, use it to solve (part of) the task, and paste the software's output into their examination worksheet and complete their response with reasoning and argumentation in natural language.

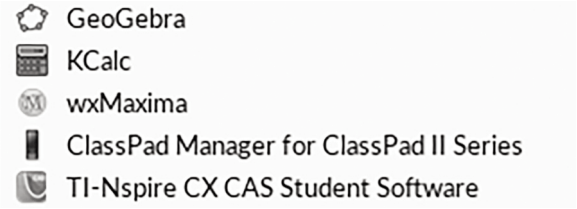


Figure 3: Mathematical tools available in the Finnish Abitti examination system[7]

This case is interesting for several reasons. First and most important, this approach allows students to use serious and sophisticated mathematical software, offering a means to express themselves mathematically and to construct mathematical objects and argumentation. The lack of expressiveness in digital assessment environments, mentioned previously as a possible limitation, is avoided here.

Second, students can during the examination use software they are familiar with, for example through its use in the teaching that precedes the examination: the components within the Abitti environment also exist as independent software, so they can be used outside of the assessment environment. In this way, the issue of computer skills challenging the test validity is solved. The wide range of software allows schools and teachers to follow their own preferences for specific tools, provided they are also available within the Abitti environment.

Third, from a policy point of view it is interesting that this was an initiative by the Finnish administration, but carried out in close collaboration with private partners and software companies. Through the integration of the software in the Abitti environment, test designers know exactly which type of tools students have access to and are no longer surprised by new features in software updates. The set of tools and their capacities are under full control of the assessment authorities.

As a downside of this approach, we acknowledge that automated scoring is not possible in this approach, as students are free to phrase their answers and to explain their reasoning. Human scoring in this format is needed, with its advantages and limitations.

To summarize this case, the Finnish model has the power of providing students with high-quality mathematical tools in a controlled environment. As such, it plays a role in the debate on future assessment strategies in the Netherlands, where a similar assessment player is under construction, as we will further explain in the next case.

The Dutch diagnostic test for 15-year-old students using Facet

The second case concerns a diagnostic test for 15-year-old students in the Netherlands. The aim of this test, an initiative by the Dutch Ministry of Education, is to provide students, teachers and parents with a diagnosis of the students' results of lower secondary mathematics education (grades 7-9) and their achievements in the light of the final national examinations, which take place at the age of 16, 17 or 18, depending on the school type. The diagnostic test is delivered through an online assessment player, called Facet, developed by the national assessment authority.

To design test items, an authoring environment called Questify Builder is used. Questify Builder is a product by Cito, a Dutch test and assessment company. After an international expert meeting (Drijvers & van Reeuwijk, 2015), the Questify-Facet chain was extended by four components: an equation editor in Facet, the computer algebra system Maxima at the background for automated scoring, parts of the Digital Math Environment[8] for tables and graphs, and GeoGebra for geometry. Figure 4 shows this configuration.

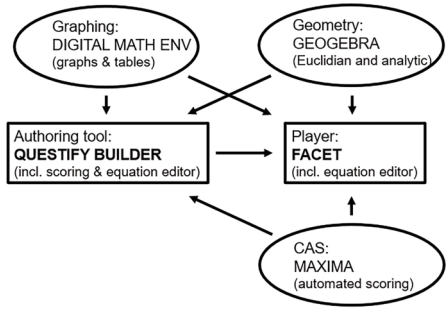


Figure 4: Configuration of mathematical tools in the authoring environment and assessment player for the Dutch diagnostic test

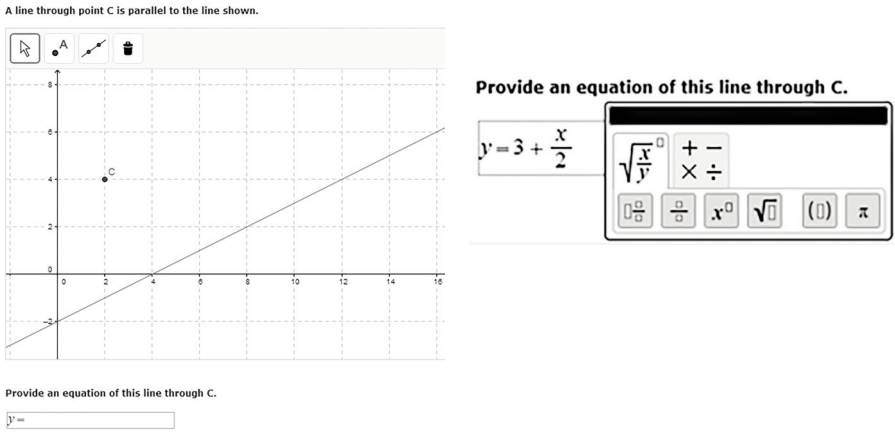


Figure 5: The parallel line item in Facet

Let us now consider two items from this test. The left-hand screen in Figure 5 shows an item in which an equation is requested of a line through point C parallel to the line that is already displayed. As a first remark, the students can use the GeoGebra screen, embedded in Facet, to explore the situation and, for example, to sketch the parallel line through C. The GeoGebra window serves as an exploration environment, as a digital scratch pad, and student work in this window is not scored in this case.

Second, as soon as the student clicks in the equation line "y = ", an equation editor pops up (see Figure 5 right-hand screen). The mathematical expression the student enters is processed by the system and evaluated by the computer algebra system Maxima. The scoring rule in this item, set by the item designer, is that the expression should be algebraically equivalent to 3 + x/2. This means that not only 3 + x/2, but also ½x + 3, x/2 + 1 + 1 + 1, as well as ½(x – 4) + 5 and ½(x – 2) + 4 are graded as correct. In this way, the – in theory infinite – number of correct solutions, each corresponding to different solution strategies, is identified. The "equally large" set of incorrect responses is also recognized, without the need for the test designer to figure out all possible student responses. The item designer can also set more strict constraints on the form of the expressions, if needed, for example by scoring both 3 + x/2 and ½x + 3 as correct, but x/2 + 1 + 1 + 1 as incorrect, because a simplified form is expected. As an aside, we note that the item designer can also decide on the GeoGebra buttons available (in this item a very limited set) and on the number of options available in the equation editor.

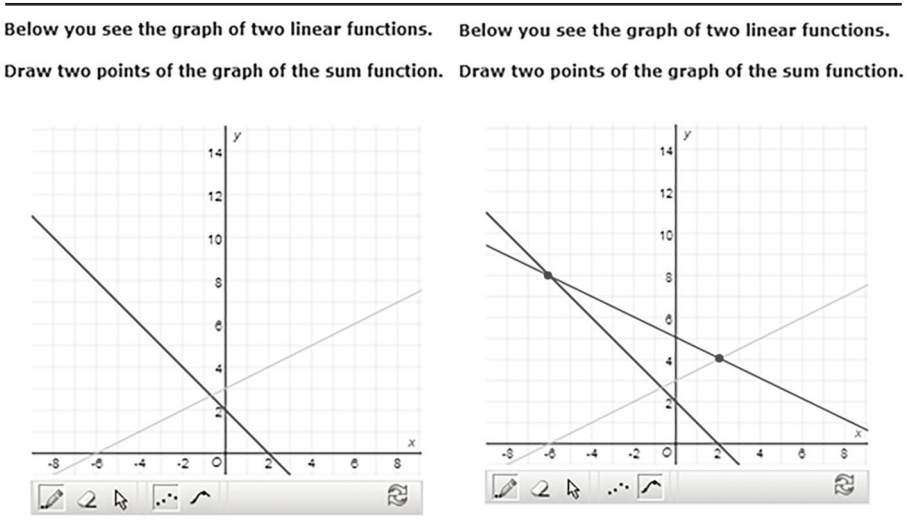


Figure 6: The sum of linear graphs item in Facet

The second item, shown in Figure 6 (left-hand screen), contains a Digital Math Environment window, in which students are requested to enter two points of the graph of the sum of the two linear functions that are already graphed. The right-hand screen shows a correct solution. But, as in the previous example, there are more correct solutions, and we want to offer to students the freedom to follow different strategies. Therefore, the scoring rule set by the item designer comes down to reading off the four co-ordinates of the two points the student defines. We don't care about the first co-ordinate of the first point, say a. The second co-ordinate of the first point, b, however, should equal to $5 - a/2$. The first co-ordinate of the second point, say c, should be different from a, and the second co-ordinate, d, should equal $5 - c/2$. In this way, the automated scoring, that might need some tolerance margins in the checks on b and d, allows students to use different strategies, but recognizes all correct responses. As an aside, we note that a somewhat more natural task might be to draw the sum graph. Figure 6 (right-hand screen) shows that students can draw the sum graph, but in this environment only by defining two points, and these points are the only student input that is evaluated. The item designers once more had to deal with the constraints of the digital environment.

To summarize, this case shows how very sophisticated mathematical tools, such as computer algebra systems, dynamic geometry systems, and graphing tools, may be needed to test quite basic mathematical knowledge in

a way that provides students with construction room and that, in the meantime, allows for automated scoring. This approach provides opportunities to avoid the issues mentioned in previous sections, also in this case where the students are younger than in the Finnish case.

If one would leave out the automated scoring criterion, as was done in the Finnish case, the Facet configuration could also be used for assessment with technology, for example by using GeoGebra within Facet as an alternative for graphing calculators. These directions are currently explored in the Netherlands.

The Digital Math Environment for intermediate feedback

The third case focuses on one of the main limitations in automated scoring: the inability to evaluate students' intermediate steps in a sensible way, like human scorers do. In mathematics education, this is very important, as many teachers want to grade the problem-solving process rather than the fluency in elementary procedures, and, therefore, do not want to punish their students too hard for a procedural mistake. Promising developments are taking place in this field. As we already mentioned above, software systems called domain reasoners have been designed, that can identify and interpret rules and "buggy" rules within a specific mathematical domain (Heeren & Jeuring, 2014). Clearly, the design of such domain-specific software integrates knowledge from computer science and mathematics didactics. Such domain reasoners can be used for automated scoring on intermediate steps, as well as for feedback and hints for sensible next steps, as it is also able to determine the student's position in a problem-solving strategy. The software identifies the steps a student makes, and can determine not only if the step is algebraically correct, but also if it brings the student closer to the solution, or rather is a detour. These features may be helpful in assessing a student's strategic skills, or, in a formative assessment scenario, may provide the student with feedback on efficient solving strategies.

As an example, Figure 7 shows the implementation of the Ideas domain reasoner for linear equations in the Digital Math Environment[9]. The student is asked to solve the linear equation that is shown in the first line. To do so, he intends to expand the brackets, and manually enters the second equation, using the fraction button in the top menu. Next, the software identifies the error in the expansion as this error type is included in the domain reasoner's set of buggy rules. It provides a red cross to indicate

the mistake, as well as sensible feedback, which is expected to help the students to correct the response. The design of the domain reasoner is primarily a computer science challenge, based on domain knowledge. The design and phrasing of the actual feedback are a didactical challenge to the educational designer of the environment.

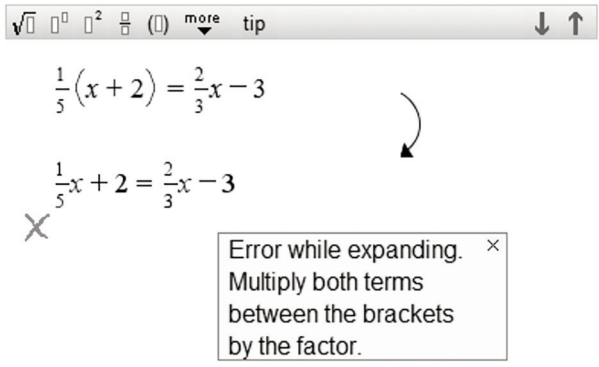


Figure 7: The Ideas domain reasoner on linear equations in the Digital Math Environment

This case shows how domain reasoners, which are developed in the field of algebra, geometry, and statistics, to mention just some, may be powerful to further fine-tune automated scoring and feedback design, particularly for feedback on intermediate results, in addition to the previously described techniques using computer algebra and Boolean variables.

Criteria for assessment environments for mathematics

What criteria for assessment environments for mathematics can we infer from the arguments, opportunities, and issues presented so far? With the three cases as inspiring examples, and building on earlier work (Drijvers, Ball, Barzel, Heid, Cao, & Maschietto, 2016), we distinguish criteria for students, which refer to the user-friendliness and mathematical options of the assessment player, and criteria for test designers, who want to use the system's authoring environment to design authentic tasks.

The student perspective

From the student perspective, it is crucial that there are mathematical means to explore, to construct, to reason, and to set up multi-step solutions. In short: students need the ability to demonstrate their capacities

for doing mathematics in a valid and authentic way, and in an environment with which they are familiar and which provides the least constraints possible. This implies that the assessment player offers sophisticated and user-friendly mathematical tools such as:

- An equation editor to enter algebraic formulas and expressions;
- A graphing tool to graph and explore functions;
- A table tool or spreadsheet tool to make tables of function values and to explore data;
- A geometry tool to carry out geometrical constructions;
- A mathematical worksheet in which the above tools can be integrated with natural language to explain and describe reasoning and problem-solving strategies.

In the three cases presented in the previous section, the exemplary items show that most of these criteria are met. The exception is the final, most challenging one, referring to integrated worksheets that resemble solutions as they are written down on paper. The Finnish case comes the closest to this, because students can write text and paste output from the mathematical tools in it. The trade-off in this case, however, lies in the need for human, non-automated scoring.

The test designer perspective

From the test designer perspective, it is important that the authoring environment provides the test designer, who usually is not a computer scientist but a mathematics educator or teacher, with user-friendly means to use a wide range of interactive and dynamic item types and to design rich items. We think, for example, of embedding video clips or applets, and more specifically of:

- Means to include the mathematical tools mentioned in the student criteria;
- Means to customize these mathematical tools for the specific item context, learning goal assessed, and target group;
- Means to set up "intelligent" and close-to-human interpretation and automatic scoring of student responses, for example through the use of CAS, Boolean variables and domain reasoners.

The three cases we presented clearly offer test designers with sets of mathematical tools to be embedded in test items. The second point on customization includes means to define the set of tools that are available, for example through limiting the menu, as was done in Figure 5 for GeoGebra. It also refers to means to set a mathematical situation for the students, as was done in Figure 6 by presenting two graphs, and to a limited extent in Figure 7 by presenting an equation. The experiences in the case of the Dutch diagnostic test reveal that setting up sophisticated automated scoring may be quite challenging for test designers.

If the above criteria for the assessment player and authoring environment are not met, there is a danger of digital assessment of mathematics remaining limited to multiple-choice items or similar item types, that do not do justice to the notion of the problem-solving process being the core skill to assess, and of mathematical skills encompassing more than the basic procedural work that is relatively easy to assess.

Towards an agenda for design and research

The central question in this paper is how to design tests with and through digital technology that assess student knowledge in a valid way, and that provide them with opportunities to express themselves mathematically. The above findings lead us to conclude that we need sophisticated mathematical tools for graphing, geometry and algebra that allow us to go beyond straightforward multiple-choice tasks and to make students really "do mathematics" in a digital test, to express themselves mathematically, to show, to produce. Test designers need means to design a wide variety of rich, interactive and dynamic tasks. In the light of adaptivity, we strive for students' digital work to be scored automatically with the sophistication and subtlety that are common in human scoring of paper-and-pen work. To a certain extent, this is realized through the use of computer algebra systems, Boolean variables and domain reasoners; further improvement is expected in the near future. In fact, the experiences in the Dutch diagnostic test development have shown that, even for relatively "easy" mathematics, we need "hard" tools such as computer algebra systems and dynamic geometry systems.

To foster test validity, it is crucial that these mathematical tools in the assessment player are similar to the tools that students use in the preceding teaching and learning. This will avoid test artifacts that relate to the user

interface of these tools, and to their limitations and constraints. Once students are familiar with these tools, we can avoid assessing students' ICT literacies rather than their mathematical knowledge. If students are not prepared for the use of the digital tools during the test, test validity is already threatened in the first step of the chain model presented in Figure 1.

Let us finish this article by extrapolating these conclusions to a future research and design agenda with respect to the digital assessment of mathematics. First, the design of user-friendly mathematical tools that can be embedded in assessment players will need to be continued. Even if a trade-off between what is feasible from a technological perspective and what is desirable from a mathematics didactics perspective will always remain, improvements in the mathematical richness of assessment players is possible and is needed.

Second, we need further improvement of automated scoring with respect to intermediate steps. In mathematics education, it is very common to value correct intermediate results with partial credit, or to value correct steps after an initial mistake, so this should be incorporated in automated scoring. The approach through domain reasoners is promising[10]. For adaptive tests, automated scoring is a prerequisite. In the meantime, we acknowledge that automated scoring should be transparent to teachers and students. Teachers and students should be able to view the student work and the scoring, to exploit the formative value of the test: test results may inform teachers about their teaching and next steps to take, and students acquire insight in the caveats in their knowledge and skills.

Third, we should consider how techniques from psychometrics can be used. We think, for example, of student models to keep track of the student's performance during the delivery of the test. Data can be logged and analyzed, for example through learning curve analysis (Tacoma, Drijvers, Boon, Jeuring, & Sosnovsky, submitted). It is a challenge to explore how the didactical, the technical and the psychometric perspectives can be integrated in digital assessment.

Finally, there is still a world to win in the alignment of teaching and learning with digital means and digital assessment. Teachers are finding their ways in teaching with different types of digital resources, and for test validity it is crucial that the types of digital tools used, the ways of use, and their role in the problem-solving process during assessment are consistent with the teaching practices in the preceding education.

Digital assessment of mathematics is a phenomenon that will play an increasingly important role in mathematics education. The challenge for teachers, designers, and researchers is to make this a success from both a mathematics didactics and a test theory perspective. The above ideas may provide guidelines for the directions to go.

Réception : 17 septembre 2017

Version finale : 26 mars 2018

Acceptation : 28 avril 2018

NOTES

1. See www.oecd.org/pisa/test-2012/form
2. See www.iea.nl/fileadmin/user_upload/General_Assembly/56th_GA/Study_presentations/eTIMSS_2019_Development_GA.pdf
3. See www.numworx.nl
4. From www.10voordeleraar.nl/toetsen/oefenmateriaal
5. In fact, the situation is somewhat more complicated and in a transition phase. Until spring 2019, exams in Finland are done with paper-and-pen in two parts: A with no calculators and B with any calculators (CAS recommended). From spring 2019 onwards, digital exams keep the same structure: part A in which the students have no access to any CAS technology, and part B in which students can use all CAS software available in the Abitti system.
6. See www.abitti.fi
7. From www.abitti.fi
8. See www.numworx.nl
9. See http://ideas.cs.uu.nl and www.numworx.nl, respectively.
10. This is addressed in an ongoing Erasmus+ project called Advise-Me (http://advise-me.ou.nl).

REFERENCES

Brown, R. G. (2010). Does the introduction of the graphics calculator into system-wide examinations lead to change in the types of mathematical skills tested? *Educational Studies in Mathematics*, *73*(2), 181-203. doi: 10.1007/s10649-009-9220-2

Devlin, K. (2012). *Introduction to mathematical thinking*. Petaluma, CA: Devlin.

Drijvers, P. (2009). Tools and tests: Technology in national final mathematics examinations. In C. Winslow (Ed.), *Nordic research on mathematics education: Proceedings from NORMA08* (pp. 225-236). Rotterdam: Sense.

Drijvers, P., Ball, L., Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). *Uses of technology in lower secondary mathematics education: A concise topical survey*. New York: Springer.

Drijvers, P., Kodde-Buitenhuis, H., & Doorman, M. (submitted). *Assessing mathematical thinking as part of curriculum reform in the Netherlands.*

Drijvers, P., Monaghan, J., Thomas, M., & Trouche, L. (2015). *Use of technology in secondary mathematics: Final report for the International Baccalaureate*. International Baccalaureate Organization. Retrieved from www.ibo.org/globalassets/publications/ib-research/technologyindpmathematicsfinalreport.pdf

Drijvers, P., Straat, H., & Wools, S. (2016). Wiskunde valide getoetst? De digitale landelijke kennistoets wiskunde van de tweedegraads lerarenopleiding vergeleken met de instituutstentamens [Mathematics assessed in a valid way? The digital national knowledge test mathematics for pre-service teachers compared with institutional examinations]. *Tijdschrift voor lerarenopleiders*, *37*(3), 27-38. Retrieved from www.ris.uu.nl/ws/files/23518131/DrijversStraatWools2016VELON.pdf

Drijvers, P. & van Reeuwijk, M. (2015). *Automatische beoordeling van wiskunde: Rapportage expertmeeting* [Automated scoring of mathematics: Report expert meeting]. Arnhem/Utrecht: Cito/College voor Toetsen en Examens.

Grugeon-Allys, B., Chenevotot-Quentin, F., Pilet, J., & Prévit, D. (2018). Online automated assessment and student learning: The PÉPITE project in elementary algebra. In L. Ball, P. Drijvers, S. Ladel, H.-S. Siller, M. Tabach, & C. Vale (Eds.), *Uses of technology in K-12 mathematics education: Tools, topics and trends* (pp. 245-266). New York: Springer.

Heeren, B. & Jeuring, J. (2014). Feedback services for stepwise exercises. *Science of Computer Programming*, *88*, 110-129. doi: 10.1016/j.scico.2014.02.021

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi: 10.1111/jedm.12000

Kovacs, Z., Recio, T., & Vélez, M. P. (2017). *GeoGebra automated reasoning tools: A tutorial*. Retrieved from http://mintlinz.pbworks.com/f/Kovacs-20160113.pdf

Noss, R. & Hoyles, C. (1996). *Windows on mathematical meanings*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Partnership for 21st Century Learning [P21] (2015). *P21 Framework Definitions*. Washington, DC: P21. Retrieved from www.p21.org/our-work/p21-framework

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford, UK: Oxford University Press.

Sangwin, C. J. & Köcher, N. (2016). Automation of mathematics examinations. *Computers & Education*, *94*, 215-227. doi: 10.1016/j.compedu.2015.11.014

Stacey, K. & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 721-751). New York: Springer.

Tacoma, S., Drijvers, P., & Boon, P. (2017, February). Using student models to generate feedback in a university course on statistical sampling. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the tenth congress of the European Society for Research in Mathematics Education* (pp. 844-851). Dublin, Ireland. Retrieved from www.ris.uu.nl/ws/files/41181049/CERME10_Proceedings_2017.pdf

Tacoma, S., Drijvers, P., Boon, P., Jeuring, J., & Sosnovsky, S. (submitted). *Student models in statistics education and their interplay with task design*.

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. Retrieved from www.learntechlib.org/p/101712

Trouche, L. & Drijvers, P. (2010). Handheld technology: Flashback into the future. *ZDM: International Journal on Mathematics Education*, *42*(7), 676-681. doi: 10.1007/s11858-010-0269-2

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113-138). Mahwah, NJ: Erlbaum.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221. Retrieved from www.public.asu.edu/~kvanlehn/Stringent/PDF/EffectivenessOfTutoring_Vanlehn.pdf

Wools, S. (2015). *All about validity: An evaluation system for the quality of educational assessment* (Doctoral dissertation). Enschede, Netherlands: Twente Univeristy.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63-82. Retrieved from www.researchgate.net/publication/273328765_Evaluation_of_Validity_and_Validation_by_Means_of_the_Argument-based_Approach