



Digital Assessment and the “Machine”

Shai Olsher, Daniel Chazan, Paul Drijvers, Chris Sangwin, and Michal Yerushalmy

Contents

Introduction	2
Design of Tasks and Tools for Student Mathematical Work	5
Window 1 Interactions in the Numworx Digital Mathematics Environment	6
Window 2 Interactions in the STEP Environment	8
Interpretation and Analysis of Student Work	9
(1) Closed Choice Answers	10
(2) Algebraic Expressions as Answers	11
Window 3 Algebraic Equivalence in STACK	12
(3) Configurations of Interactive Graphical Diagrams as Answers	13
(4) Sequences of Equivalent Expressions as Answers	14
Window 4 Item Checking in MathXpert	15
(5) Free-Form Open Answer Workspace for Answers	16

S. Olsher (✉)

The University of Haifa, Haifa, Israel

e-mail: olshers@edu.haifa.ac.il

D. Chazan

University of Maryland, College Park, MD, USA

e-mail: dchazan@umd.edu

P. Drijvers

Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

e-mail: p.drijvers@uu.nl

C. Sangwin

e-mail: c.j.sangwin@ed.ac.uk

M. Yerushalmy

e-mail: michalyr@edu.haifa.ac.il

© Springer Nature Switzerland AG 2023

B. Pepin et al. (eds.), *Handbook of Digital Resources in Mathematics Education*,

Springer International Handbooks of Education,

https://doi.org/10.1007/978-3-030-95060-6_44-1

Reporting/Presentation of Results of Analysis	17
Window 5 Characterizing Student Responses in STEP	19
Window 6 Knowledge Landscape in the DME	20
Conclusion and Discussion	22
References	25

Abstract

In this chapter, we explore assessment that is performed automatically by the digital environment, what might be called assessment “through technology.” Based on experience with the design of three innovative content-specific automatic assessment platforms, the main goal of the chapter is to exemplify design considerations to map the development of digital assessments. These digital platforms use various methods to assess mathematics through student’s interaction with digital learning resources. We address the fit between an assessment’s task design and its goals, the analysis of student work by the platform, and the report that is then produced. We want tasks to offer opportunities for students to express mathematical ideas, to take advantage of the opportunities provided by automatic assessment, as well as to meet the goals of assessments. In writing tasks, designers must also take into account two other design considerations: How can the digital assessment platform interpret and analyze student work in variable and flexible ways? How can the platform “make sense” of student work, so as to be able to generate feedback or to report on learning achievements? What are the ways in which insights from this analysis will be made accessible and to whom? Taken together, examination of the design of the tasks given to students to collect data, how that data is analyzed by the machine, and how the machine reports on that analysis allows us to map current digital assessment practices. We close by emphasizing the importance of continued engagement of the mathematics education community with the design of digital assessment platforms because mathematics education stakeholders bring with them a content-specific focus on higher-level thinking in mathematics and on students’ conceptions and misconceptions.

Keywords

Digital assessment · Open ended tasks automatics assessment · Digital platforms for mathematics · Authentic assessment · Learning analytics

Introduction

Assessment is a key component of mathematics education, and using technology to support assessment has a long and interesting history. With shifts to online assessments and the inclusion of digital resources in assessments in large-scale assessments like NAEP (Johnson 1992) or TIMSS (Mullis and Martin 2017), technology’s role in influencing assessment is growing. Yet, as articulated by Drijvers et al. (2016), the phenomenon of digital assessment is varied. Digital assessments may

be paper and pencil assessments accomplished “with” machines, or they may be assessments that are offered “through” the machine, where all of students’ engagement with an assessment is mediated through technology.

Digital assessments, like any other forms of assessment, can have many goals. A digital assessment could be aimed at providing comparative summative assessment of students’ skills, or it could seek to provide support for formative assessment, carried out by either teachers or the students themselves working as individuals or in small groups. But there are other ways to capture differences in the goals of platforms. For example, many digital assessment platforms are content neutral and aim to give feedback on a range of different content (Dougiamas 2004). The design of feedback mechanisms for such platforms is often driven by efficiency considerations (Pardo et al. 2019).

As the last two sentences suggest, when students’ engagement with an assessment is through a digital platform, the machine has direct access to student work and may also automatically assess students’ work. This chapter examines that particular aspect of the phenomenon of digital assessment: the design of digital automatic assessment platforms where the “machine” analyzes student work. It is our view that machine analysis of student work, though not yet ubiquitous, is likely to have ever larger impacts over the coming years.

Rather than give this chapter the challenge of surveying the full range of assessment goals and the full panoply of digital assessment platforms, we have chosen a specific focus for this chapter; we focus on the design decisions found in platforms whose aim is to assess specific mathematical work of students during and after their work process. The platforms on which we are focusing are research-based projects of design and development that are available to use in both laboratories and in schools, in some cases internationally. The platforms we examine are designed to support a range of end users including assessment designers, the students who are being assessed, teachers who create assessments and use assessments created by others, and educational stakeholders – like students, teachers, parents, school administrators, and policymakers – who receive and could generate reports of students’ performance on the assessments at various levels ranging from a single student to a district. Specifically, we demonstrate challenges in designing automatic assessment of student work with three platforms:

- The Numworx Digital Mathematics Environment (Numworx, Drijvers 2020) is a digital mathematics platform for secondary school and university education that provides a wide range of assessment options, including
 - Autonomous tests which students can review themselves.
 - Summative tests which can serve as exams.
 - Automatic intelligent scoring and reports of student models.
- STACK (Sangwin 2013) is a computer-aided assessment (CAA) platform.
 - The design emphasis of STACK is on formative assessment.
 - The prototype interaction is that a student enters a mathematical answer in the form of an algebraic expression and STACK evaluates the student’s answers using computer algebra.

- STEP (Olsher et al. 2016) aims to facilitate guided inquiry processes in the mathematics classroom:
 - Provide information that would involve students in feedback processes.
 - Provide data for teacher’s decision-making: correctness, concept images, collective example space, student grouping.
 - Provide formative assessment processes involving different agents (teacher, student, peers).

These digital assessment platforms are used to pinpoint specific authentic mathematical activities that are assessed in order to facilitate mathematical practices in the mathematics classroom, providing teachers and learners with technological instruments to both carry out the mathematical activity and means to collect analyzable information used to create assessment insights. As a result of their focus on specific content, the design of these technological platforms for digital assessment emphasizes aspects that might be overlooked or underdeveloped in other platforms.

We will use our knowledge of these platforms to illustrate three sets of interrelated design considerations that we will use to structure the chapter:

1. The design of platforms and tasks determines what work students will submit and with what tools, how students will interact with the machine, and what information will be stored and made available for the machine to analyze. The nature of that information must fit the goals of the assessment.
2. The design of the automatic interpretation and analysis of student work is influenced by what work is available and how it is stored. Data structures influence how student work can be interpreted intelligently by the platform to fit the goals of the assessment.
3. The presentation and reporting of the results of the automatic analysis to students, teachers, and other stakeholders are influenced by the nature of the student work, how it is analyzed by the machine, and perhaps most importantly by the goals of the assessment itself. For example, is the assessment a formative assessment predominantly intended for use by the teacher, or as a summative assessment for use by other stakeholders, or as a learning resource for use by students?

Sections “[Design of Tasks and Tools for Student Mathematical Work](#)”, “[Design of Tasks and Tools for Student Mathematical Work](#),” and “[Reporting/Presentation of Results of Analysis](#)” will rely on “windows,” presenting aspects of these three platforms to support readers unfamiliar with the state of the field with respect to such automatic assessment environments. Through the windows, we seek to articulate the interrelated nature of the three design considerations outlined in those sections. In the concluding section of the chapter, we address resulting design challenges and opportunities associated with automatic assessment of student work accomplished through digital environments.

Throughout the chapter, our aim is to illustrate how automatic assessments might constitute a resource for the mathematics education community, though one that will require careful attention to design. The digital platforms we describe can be a

resource for researchers doing research on student learning of particular content or studying teachers’ mathematical knowledge of particular content. Similarly, they can be useful for educators seeking to assess the work of students to whom they teach mathematics or the work of teachers who they interact with on improving mathematics instruction.

Design of Tasks and Tools for Student Mathematical Work

When assessments are delivered in digital environments, students’ interactions with the digital platform form the basis for the assessment. When designing or choosing to use a certain digital platform, one should consider the types of interaction the platform allows and in what modalities. Closed-ended questions are simple to imagine and are limited in their authenticity. Textual inputs, whether typed as text, scribbled and parsed, or perhaps inserted as parsable mathematical syntax, provide different and possibly more elaborate information and insights about the student’s work that can be produced automatically. Interactive diagrams (Naftaliev and Yerushalmy 2017) provide a range of representations that could be dynamically linked for exploiting the unique benefits and focus that each representation provides about the different mathematical objects (an example for an interactive diagram can be found in Fig. 2).

The choice of the digital assessment platforms impacts on the goals and characteristics of the assessment. For each digital assessment platform, the goals of the assessment are connected with the characteristics of the interaction of the student with the platform. Compared to a platform that views any assessment event as a learning opportunity, an assessment platform aimed at summative assessment would probably have a more restrictive interaction with the student. When choosing to conduct a focused learning experiment, the assessment platform should support the definition of well-constrained conditions of the context in order to provide a focused assessment. Platforms could also support the assessment of the comparison of mathematical objects in various registers and modalities to assess the broadness of concept images and related concept definitions.

Student interaction with a digital platform is also greatly influenced by the nature of the tasks set for learners: What is the learner required to do? First and foremost, learners are asked to meet the requirements of the task. In many tasks, this can be done through different answer types in a digital environment, each providing different types of information: Select the correct answer/statement or type an answer mathematically or textually; show a variety of examples that could demonstrate a broad concept image; present argumentation about answer/example using various modalities; graphically identify ideas in the constructed example; construct a script describing a dynamic scenario with the constructed example; or reflect using meta-cognitive, self-reflection tools.

Finally, the end result of the learner’s interaction with the digital platform is information or data that is collected by the platform and used in the analysis to provide insights. While technology constantly pushes the boundaries of the range and quality of possibilities to collect data, the rationale for the type of information

collected is not always technological. Platform designers' choices are rooted in the goals of the assessment. Simple selection of an answer could be just the right amount of data needed to produce an in-class real-time poll. Submission of a carefully constructed solution could be the choice of platform designers that provide students with an inquiry-based sandbox or provide them with a notebook type of setting that enables saving and working on tasks and projects over multiple sessions. Platforms might also process intermediate stages in the solution process to provide a finer-grained feedback process. There are platforms that analyze examples or processes that are chosen by the student, promoting using the environment also as a notebook to save drafts and curating what will be presented to the teacher or peers. On the other hand, there are also platforms that log all of the student actions that they can record, in order to enable a holistic analysis of work processes.

In Windows 1 and 2 we present two examples of how platforms organize students' interaction with a task: one example (Window 1) describes interaction within the DME platform, and the other example (Window 2) makes use of the STEP platform. These environments allow expression of individual ideas, not only by providing open-ended tasks but also tools for writing and interaction that allow a variety of expressions.

Window 1 Interactions in the Numworx Digital Mathematics Environment

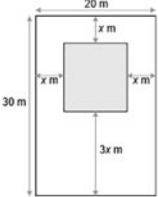
Thinking about humans interacting with tools for doing mathematics, let us first reflect on the past. Historically, paper and pencil have been the traditional environment for doing mathematics. Even today, students' experiences with doing mathematics largely concern scribbling on paper or watching a teacher doing the same with a marker on an (interactive) whiteboard. These environments offer ultimate freedom in creating mathematical representations, diagrams, tables, geometrical drawing, sketches, and in connecting them through reasoning in natural language. Together, this freedom to act, and the idiosyncratic results this may lead to, turns the paper-and-pencil environment into a strongly personalized environment, in which the user is in full control and experiences ownership. It is natural to seek to replicate the benefits of pencil and paper in a digital form. Therefore, an important requirement for digital mathematics (assessment) environments is that the interaction resembles the paper-and-pencil environment with respect to flexibility to use and freedom to work. In online activities or assessments, students should be able to do mathematics as if they are using paper and pen and in the meantime benefit from the additional affordances of the technological tools. This criterion for digital assessment environments is an ambitious one to meet but essential.

Figure 1 illustrates this criterion and shows an implementation in the form of an open answer workspace in the Numworx Digital Mathematics Environment. The left-hand side contains the task, which concerns an optimization problem. The window on the right-hand side initially is empty. It is an open answer window, in which a student can work through adding different elements. The bottom menu

Task

Building lot

Below you see the map of an imaginary building lot with a house plan in gray in the middle. The lot measures 30 by 20 meter.



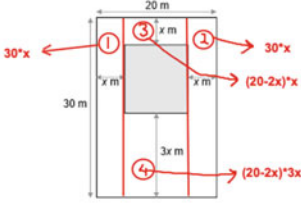
Suppose the width of the strips beside and in front of the house is x meter. The garden at the back is three times as long. As a legal requirement, the house may cover at most 20% of the building lot area.

Calculate the dimensions of the house, with a one decimal precision.

The house measures by meter

Student work

I first set up an expression for the area of the house A . The area of the building lot is $30 \cdot 20 = 600$. Next, I subtract the four strips.



$A = 30 \cdot 20 - 30x - 30x - (20 - 2x)x - (20 - 2x) \cdot 3x$
 $A = 600 - 60x - (20 - 2x) \cdot 4x$
 $A = 600 - 60x - 80x + 8x^2$
 $A = 8x^2 - 140x + 600$

As the area of the house is at most 20% of the building lot area, the biggest house area is $0,2 \cdot 600 = 120$. I can now setup an equation and solve it.

$8x^2 - 140x + 600 = 120$
 $8x^2 - 140x + 480 = 0$
 $2x^2 - 35x + 120 = 0$
 $x = \frac{35 + \sqrt{265}}{4}$ of $x = \frac{35 - \sqrt{265}}{4}$
 $x \approx 12,82$ of $x \approx 4,68$

The house width is $20 - 2x$. So $x = 12,82$ is impossible (negative width). So only a solution for $x = 4,68$.

width = $20 - 2 \cdot 4,68 = 10,64 \approx 10,6$
length = $30 - 4x = 30 - 4 \cdot 4,68 = 11,28 \approx 11,3$

Chose Text Calculator Table Graph Formula Function

Fig. 1 Open answer workspace in the Numworx Digital Mathematics Environment (<https://www.numworx.nl/en/>)

shows, as icons, the options of entering text, a calculation, a table, a graph, a formula, or an equation. The formula or equations can be entered through an equation editor or in handwriting with a stylus. In the latter case, the handwriting recognition module will display the formula in pretty print format.

In this case, the student chose to use the sketch to set up some expressions, and he/she adds some text to explain his/her thinking. Based on this, he/she sets up the equation and solves it.

Window 2 Interactions in the STEP Environment

The STEP platform is built on the belief that interactions with technology should focus on experimentation, exemplifying, conjecturing, and arguing. STEP also seeks to have learners personalize their work. Because the designers of STEP believe that example generation is a crucial aspect of inquiry, in STEP, learners personalize their work by constructing examples in an interactive diagram that they then submit in response to a task.

For example, engaging learners in generating and verifying examples of a particular mathematical concept serves two central aims of assessment: we use learner-generated examples as an indicator of the learners' understanding (conceptualization, concept images they hold), and we use example generation as a catalyst for enhancing students' understanding and expanding their example space associated with a new concept. Therefore, we design EETs (example-eliciting tasks) upon the following articulated design principles: (1) task's requirements: the example(s) are constructed and submitted demonstrating the truth for given claim(s); (2) the context of the example and its logical status is set by a given claim or by a set of constraining conditions; and (3) the example is constructed using different mathematical modalities, which are part of the design of the interactive diagram.

Figure 2 demonstrates a task where f is given by a function expression and graph: $f(x) = -2x^2 + 4x + 5$. g can be set using three independent sliders controlling the coefficients of the function expression (a , b , and c , in the applet on the left part of Fig. 2). In addition, the task lists four conditions (checkboxes on the right part of Fig. 2): two conditions of relations between two quadratic functions f and g and two

Task description

A quadratic function $f(x) = -2x^2 + 4x + 5$ is given.

Set parameters of a quadratic function $g(x)$ so that the graph satisfies as many conditions as possible.

Mark those conditions and submit appropriate examples.

The main interface shows a coordinate plane with two parabolas, $f(x)$ and $g(x)$. $f(x)$ is fixed as $-2x^2 + 4x + 5$. $g(x)$ is defined by sliders for a , b , and c , with the expression $g(x) = ax^2 + bx + c$. Below the sliders, the current values are $a=3$, $b=-6$, and $c=10$.

To the right, four conditions are listed with checkboxes:

- The graph of $g(x)$ intersects the graph of $f(x)$ in one point only.
- The same axis of symmetry of the functions .
- The graph of $g(x)$ passes through the origin of the coordinate system.
- The function $g(x)$ has a minimum.

Below the main interface, three smaller applet windows show different configurations of the sliders and the resulting graphs of $f(x)$ and $g(x)$.

Fig. 2 The quadratic multiple conditions task

conditions of properties of g . The goal is to find and submit three examples where g and f satisfy the maximum possible number of conditions. The answer consists of the marked chosen statements and the examples that demonstrate the choice. The conditions are (1) the graph of f intersects the graph of g in exactly one point; (2) the two functions have the same symmetry axes; (3) g passes through the origin $(0,0)$ of the system; and (4) the function g has a minimum. There are three possible triplets that fulfill the requirements.

To satisfy the requirements of maximum conditions on $g(x)$, there are three possible triads: 1,2,4, 1,3,4, and 1,2,3. The three submitted constructed examples in Fig. 2 are an answer submitted to meet the requirements of conditions 1, 2, and 4 (g intersects f in a single point, f and g share a symmetry line, and g is a parabola with a minimum point). Most often the work started by reading the given conditions while using the sliders to change the parameters a,b,c and observe changes of the g graph. While interactively observing the relations between the two functions, it becomes clear that pairs of conditions can be easily met, e.g., a single intersection and same symmetry line, a function with minimum that is passing through the origin. This design intends to leave space for serious exploration that leads to many partial answers. Finding the possible triads is more challenging, and finding three as different as possible examples to the same triad is leading to exploring whether the requirement of multiple examples to each of the triads can be met. Another type of logical argumentation would be required when one considers the possibility of fulfilling four conditions (a free-form explanation or free-form proof will be required in this case as providing contradicting examples will not justify a universal answer).

Finding the exact position of the graph is not easy. The design that excludes traditional direct symbolic input is a design choice made to support exploration and promote non-prototypic examples. Note that the submission presented in the middle “looks right” and might be filtered as a possible answer, whereas $g(x) = -10.5x^2 + 18.5x - 1$ is incorrect (inaccurate) and suggests that the student attended visual characteristics and did not compute or check the accuracy of the function expression.

As shown in this section, assessment platforms can offer a variety of mathematical interactions and different means to express mathematical ideas. These interactions produce various kinds of information that requires careful analysis and interpretation, so it will result in a meaningful feedback process. The next section will focus on these processes.

Interpretation and Analysis of Student Work

In Section “[Interpretation and Analysis of Student Work](#),” the main point is that assessment platforms for mathematics should offer a variety of rich interaction modes that allow students to engage in mathematical interactions and to express their individual mathematical ideas. This should be made possible through facilities to write, read, and do mathematics in a natural way, using appropriate mathematical tools, representations, and interactions. However, when students engage in a broad range of interactions with an assessment platform, new challenges arise: How can

the digital assessment platform interpret and analyze student work in variable and flexible ways? How can the platform “make sense” of student work, so as to be able to generate feedback or to report on learning achievements? This is the topic of this section.

This discussion of what a rich, interactive assessment platform makes of student work focuses on the individual item level. What is the role of the machine in assessing answers to an individual item (or part of a more complex task)? Depending on the answer type of an item, student work can vary in both substantive and superficial ways; platforms need to be able to distinguish the substantive from the superficial. To outline the challenge of interpreting and analyzing student work, we consider five different answer types, broadly sketching a spectrum of increasing sophistication of the properties established:

1. Closed choice answer items
2. Algebraic expressions as answers
3. Configurations of interactive graphical diagrams as answers
4. Sequences of equivalent expressions as answers
5. Free-form open answer workspace for answers

(1) Closed Choice Answers

This answer type appears in traditional items with a limited number of choices to make by the student, such as true/false, multiple choice, or multiple response items. These answers are easy to score and to interpret – even if the cause of a student mistake usually remains unclear. The answers can be scored with perfect reliability. These answer types can be highly sophisticated, especially when sequences are linked together, potentially with more complex question types. Meanwhile, the design of learning materials requires great skill and care. In high-stakes situations, or research applications where high validity is required, materials are typically subject to trial and refinement before use. See Mejia-Ramos et al. (2017) for a discussion of proof comprehension task development and others for concept inventory development (Carlson et al. 2010; Lane-Getaz 2013).

However, mathematics education has particular problems with this choice answer type. Some core processes in mathematics are reversible with one direction much more difficult than another (Sangwin and Jones 2017). For example, it is much easier to expand out the product $(x^2 + x + 1)(x^3 - 1)$ than to factor $x^5 + x^4 + x^3 - x^2 - x - 1$. No sensible student with an understanding of the relative difficulties of reversible processes would tackle the multiple choice question “What is the factored form of $x^5 + x^4 + x^3 - x^2 - x - 1$?” by factoring the polynomial. Instead they would expand out the options presented to them, effectively reversing the process (expand instead of factor) and subverting the intended purpose of the question. This threat to the validity of the question, by reverse engineering, is particularly problematic for mathematics.

(2) Algebraic Expressions as Answers

To avoid the validity issue of the choice answer type, automatic assessment of mathematics has accepted answers from students which are mathematical expressions and sought to establish objective properties of those expressions for over half a century. A prototype property is to establish algebraic equivalence between the student's answer and the correct answer. Next, systems establish other properties that are relevant, e.g., whether the student's answer is in the correct form.

In the example above, we wanted a factored polynomial such as $(x^2 + x + 1)(x^3 - 1)$. A string match or regular expression match, even in apparently simple situations, is completely inadequate. Here a student could also give $(x^3 - 1)(x^2 + x + 1)$ or $(1 + x + x^2)(x^3 - 1)$, and both would (probably) be acceptable as factored forms. There are many other ways of writing this polynomial as a product of powers of distinct irreducible terms, i.e., factored. An automatic digital assessment system must have some kind of computer algebra support to manipulate students' expressions and establish such properties.

The factoring example might appear rather complex and so consider this much simpler task to calculate $\left(\frac{1-i}{\sqrt{2}}\right)^{-14}$ and write the answer in the Cartesian form. The correct answer is $-i$. This item was used with year-one undergraduate students, and over a tenth of the cohort answered with expressions such as $0 - i$, $0 - 1 * i$, and $-1 * i$. Since students have just been told that the Cartesian form means they should write the complex number as $a + ib$, there is a very real potential conflict between following instructions just given in a new area of mathematics and well-established conventions that we typically do not write $+0$ and $\times 1$. If the purpose of the question is to discuss and establish norms that we do not write $+0$, even with Cartesian form, then all well and good: reject $0 - 1 * i$ as incorrect with feedback. However, if the purpose of this question is to test whether students can calculate a power using DeMoivre's theorem, and then convert the answer back into the form in which the terms in the question were stated, we should probably accept answers such as $0 - 1 * i$ as basically correct but with feedback about conventions. While we might want to accept $0 - i$, $0 - 1 * i$, and $-1 * i$, we probably don't want to accept $\frac{1}{i}$ or $\cos\cos\left(\frac{5}{2} * PI\right) - \sin\sin\left(\frac{5}{2} * PI\right) * i$ in this situation. Most teachers also want to provide feedback to students, and such feedback can only be effectively provided if the software has tools to make subtle distinctions between expressions such as $0 - i$ and $-1 * i$. In this case, we established the student's answer belonged to a particular equivalence class containing $-i$, using commutativity and associativity of addition and multiplication, together with rules for identity operators which rewrite expressions $1 \times x \rightarrow x$, etc. An automatic digital assessment system must have flexible computer algebra support which allows management of rule sets, to manipulate students' expressions and establish such properties. Such rule sets even benefit from including incorrect rules such as $(a + b)^n \rightarrow a^n + b^n$ to help establish that a student's particular answer is consistent with making a well-known error.

Many systems can establish such properties and do so effectively despite theoretical limits on what properties can be established automatically. For example, establishing the algebraic equivalence of a mathematical expression with zero is formally undecidable in general. Theoretical undecidability does not stop effective and reliable equivalence checking of simple students' answers. In Window 3 we present an item in the STACK system that demonstrates assessment of such properties.

Window 3 Algebraic Equivalence in STACK

Consider the item in the STACK system shown in Fig. 3. Clearly, there are many potential correct answers. To establish the origin lies in the plane is a simple calculation but is only one relevant property. We need \mathbf{p} to be nonzero, and we need the two direction vectors \mathbf{u} and \mathbf{v} to be linearly independent: typically in automatic assessment many separate properties are needed for completely correct answers. This item in STACK was attempted by 815 year-one undergraduate students as part of the final assessment in a linear algebra class. This question proved nontrivial. A total of 262 students (32.2%) answered "true" to the first part, i.e., they think \mathbf{p} must equal the zero vector. A total of 91 (11.2%) students gave an example of a plane with two parallel direction vectors, so not really defining a plane at all. Only

Let $\mathbf{x} = \mathbf{p} + s\mathbf{u} + t\mathbf{v}$ be the vector equation of a plane in \mathbb{R}^3 . Then the plane passes through the origin if and only if $\mathbf{p} = \mathbf{0}$? If false provide a counter example, otherwise leave the equation of a plane blank.

False **False**

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Check

✓ Correct answer, well done.
Your point $\mathbf{p} \neq \mathbf{0}$ and the origin does lie in the plane, well done!

Fig. 3 Stack item using algebraic equivalence check

329 (40.4%) were judged to have a completely correct answer, and no particular example was given by more than 25 (3.1%) of the cohort.

There are limitations on establishing many important mathematical properties. For example, it is difficult to reliably test whether a real function has a local maximum at a particular point. As with algebraic equivalence, many apparently limited situations work very reliably in practice, with students rarely actually providing answers outside the limits of reliable assessment. Where this is more likely, automatic assessment can be supplemented with a human check of the range of answers being provided and the corresponding outcomes. Indeed, in practical week-to-week teaching, it is much better to initially create a minimally working algorithm to accept correct answers and then review students’ actual answers. The temptation is to predict common mistakes or answers consistent with common misconceptions. Time spent writing checks for predicted mistakes is wasted if students never actually answer in ways which generate these outcomes. It is often much better, especially in high-stakes situations with delayed feedback, to review students’ answers before releasing outcomes. All online assessment systems record students’ interactions and generate detailed statistics for review by a teacher. The review can easily result in an updated, more reliable, assessment algorithm and a better understanding of what students really do and the frequency with which they do it. We find, especially at university, that students are genuinely interested in how the assessment process works and are typically willing to live with its limitations, provided these are rectified quickly.

In many teaching environments such limitations can be used, with students, to explain interesting mathematics. Indeed, there are very good arguments in favor of being highly explicit about the precise properties used to check if an answer is satisfactory. For example, it is very rare to see a discussion of what factored actually means in elementary algebra books.

In short, assessing whether an algebraic expression answer is precisely correct or is absolutely incorrect is relatively simple and often robust. In many situations, however, there is a range of acceptable answers and a penumbra of acceptability involving a range of issues, e.g., the conventions of written form, for which a teacher would like to provide specific feedback. So though we now have a wide range of highly sophisticated tools for automatic digital assessment, which are used extensively for formative assessment of elementary mathematical tasks, a human teacher, ultimately, remains responsible for choosing the outcomes.

(3) Configurations of Interactive Graphical Diagrams as Answers

Interactive mathematics diagrams are now well-established components in learning mathematics. Many systems enable students to manipulate the configuration of a geometric figure, and the student’s answer is the final configuration (Yerushalmy and Olsher 2020). For example, students are invited to drag a point on a predefined diagram or to construct a figure from scratch. In such items, the final diagram acts as the answer to be interpreted and analyzed.

These interpretations and analyses of interactive graphical diagrams share many issues with dealing with algebraic expressions as answers, highlighted in the previous category. In particular, the teacher remains responsible for deciding what properties are relevant and deciding what action to take when properties are (or are not) established. Example properties include deciding if two lines are perpendicular or whether a point lies on a particular line. This area of automatic assessment is less well developed than algebraic assessment but no less important and is likely to be the area of most significant development in the near future. We have already seen an example in Fig. 2, where the student's answer was the configuration of a quadratic graph. Another example is shown in Fig. 4. The response shown in this diagram is only partially correct because the range of the function is $[0,0.9]$ and not $[0,1]$ as required. Since the student drags the points A-D on-screen, the required properties of the piecewise linear graph can readily be calculated from the position of the four points.

(4) Sequences of Equivalent Expressions as Answers

Establishing objective properties of a single algebraic expression or diagram is a necessary starting point for dealing with a student's sequences of algebraic

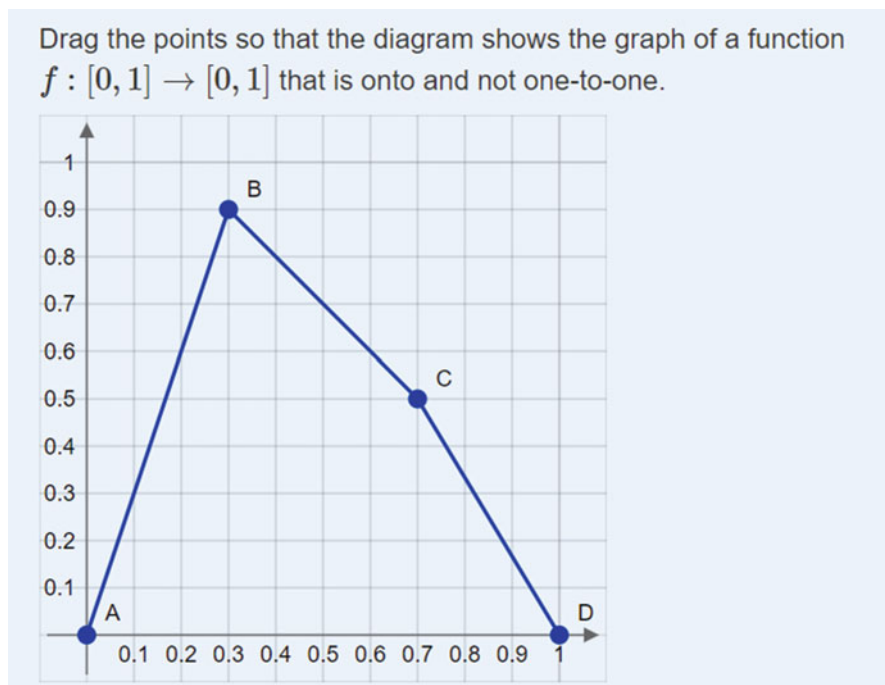


Fig. 4 Item checking students' algebraic decision-making

expressions as an answer. That is, automatically assessing students’ complete line-by-line equivalence reasoning. In school mathematics this is one of the most important forms of reasoning (Nicaud et al. 2004; Sangwin 2019). The last 10 years has seen a rapid expansion of the tools available for automatically assessing students’ line-by-line work. In Window 4 we present two examples of algebraic expressions as an answer from MathXpert and STACK. These environments provide the learner with different levels of calculations that could be performed by the system.

Window 4 Item Checking in MathXpert

A unique opportunity with systems assessing line-by-line reasoning is the ability to separate out decision-making processes from the ability to undertake calculation. Separating decision-making from calculation can be particularly useful in formative settings. An early example is MathXpert in which users are presented with a mathematical problem. Students select part (or all) of an expression, and the system suggests which operations can be performed on that selection. Having picked an option, MathXpert then performs that operation automatically, i.e., students are not also required to perform their selected calculation (Beeson 1998). In Fig. 5, the user has partially solved a calculus problem. In the last line, part of the expression has been selected, and the system has shown what rules might be applied to this part. The user chooses which rule should be applied, and the software performs the calculation.

When designing sequences of items, there are opportunities to mix between (i) specifying what students should do and (ii) actually performing the operations. This approach can be followed until students do both and reach the level of independent competence required. For example, the opposite to MathXpert would be to provide a model answer with specified operations and gaps for students to fill in until they had a complete solution. A very simple example from the STACK system

Undo Hint AutoStep ShowStep AutoFinish Finished? Graph Prev Edit Next

$$\frac{d}{dx} \left(x + \frac{1}{x} \right) \left(x - \frac{1}{x} \right)$$

$$\frac{d}{dx} \left(x^2 - \frac{1}{x^2} \right)$$

$$\frac{d}{dx} x^2 - \frac{d}{dx} \frac{1}{x^2}$$

$$\frac{d}{dx} \left(\frac{1}{v} \right) = -\frac{(dv/dx)}{v^2}$$

$$\frac{d}{dx} \left(\frac{c}{x^n} \right) = -\frac{nc}{x^{n+1}}$$

$$\frac{d}{dx} \left(\frac{u}{v} \right) = \frac{[v(du/dx) - u(dv/dx)]}{v^2}$$

evaluate derivative in one step

$$(a - b)(a + b) = a^2 - b^2$$

$$\frac{d}{dx}(u - v) = \frac{du}{dx} - \frac{dv}{dx}$$

Fig. 5 Item checking a dynamic graphical diagram in MathXpert

$$\frac{2x^3 + 6x + 1}{x^2 + 2x + 1} = \frac{(x^2 + 2x + 1) \left(\boxed{} \right) + \boxed{}}{(x^2 + 2x + 1)}$$

$$= \boxed{? + (?)(?)}$$

Fig. 6 Template item in STACK

is shown in Fig. 6. In this example, a template is provided for the students, and the last box contains a “syntax hint” to suggest the expected final form. Individual questions, such as this, have minimal value; however, *coherently organized sequences* of such materials have been found to be highly effective; see (Kinnear et al. 2022).

(5) Free-Form Open Answer Workspace for Answers

As the final and most sophisticated answer type, we now consider free-form open answer workspaces, in which students can provide complete solutions that reflect their problem-solving process, reasoning, or proof, through different mathematical representations. An example of this answer type is provided in Fig. 1 (Section “[Interpretation and Analysis of Student Work](#)”) from the Numworx environment.

Clearly, the interpretation and analysis of complete solutions, including proof, are much more difficult than assessment of a final answer. While there are prototype systems which assess proofs in specific subject areas, such as discrete mathematics, assessment of free-form proof cannot currently be done. When faced with trying to automatically assess understanding of a complete solution, including proof, the design of sequences of questions can play a valuable role. Design strategies include using faded worked examples, e.g., with judicious gaps in proofs for students to fill. Explicit assessment of separated concerns tries to ensure students’ are fully prepared for writing proofs in particular topics (Sangwin and Bickerton 2021). Lastly, reading comprehension tasks have become much more popular in recent years, such as those by Mejia-Ramos et al. (2017), but these tasks are actually quite difficult to write.

It is likely that a complete change in the way we write proofs will be required before reliable automatic checking of students’ proofs can be done. As a minimum, we need standard ways to type a proof into a machine which captures meaning and not just presentation of mathematics. Indeed, at this point students would effectively use a proof assistant, or proof checker, and this technology is right at the forefront of contemporary research (Thoma and Iannone 2021).

Taken together, this section shows that digital assessment platforms’ capacities to interpret and analyze student work are growing. Much progress has been made for the case of relatively simple answer formats, whereas more sophisticated answer

formats, such as open answer types with complete solutions, remain challenging or even impossible. These interpretation and analysis capacities determine to a large extent the opportunities for report and feedback to students on their work, which is the topic of the next section.

Reporting/Presentation of Results of Analysis

Once student work is recorded and analyzed, reporting is meant to inform students/teachers/parents/school administration/national educational systems on proceedings and state-of-the-art learning. The communication of insights to various stakeholders relies on the types of reports and presentation methods available in an automatic assessment platform. There are many important details to consider about how results of analysis are presented: What is the form that a report takes (e.g., is an analysis presented in words or in numbers?)? Do reports only look backwards or do they look forward, offering hints, or suggestive alternative strategies? Are reports given while a student is working, or are they given only after students have completed a task? Finally, what level of aggregation do reports use as their level of analysis; are they reports about individual learners, small groups of learners, whole classes, or larger units of analysis? In this section, we begin by considering these aspects of reports and then illustrate these differences with two windows.

Reporting information to the learner while working or after submitting solutions is traditionally termed as feedback, aiming to close the gap between where the student is and where the learner should be (Hattie and Timperley 2007). The reported insights intend to build awareness to learning performance in the targeted outcome content, and in some cases, these insights aim at providing general awareness to learning progress, patterns and strategies. Such personal reports are often designed to be a one-way transfer of information from an agent to the learner (Shute 2008). By way of contrast, the dialogic interaction in making sense of the reported information is considered to be an essential feedback process to enhance student’s learning. Carless (2015) described feedback in the classroom as “a dialogic process in which learners make sense of information from varied sources and use it to enhance the quality of their work or learning strategies” (p. 192).

Reports can take different forms, depending on the goal and the targeted audience. One form of commentary on work (particularly for formative practices) is didactical feedback. Didactical feedback is considered to be verbal or textual information about the student’s work and is delivered in the form of comments (told or written). There have been research and development attempts of technology-based textual reports to students. These were most often verbal reports provided, with or without technology, aimed at assessing and supporting direct acquisition of a concept or concept-related skills. Usually, these reports are promoting correct answers and the procedures leading to correct answers. Other systems seek to simulate a mathematical conversation using prepared hints or comments that address the performance of the learner compared to the envisioned or expected performance.

Forms of assessment that require adopting a comparative view of assessed learners or of different parts of the learning process, concepts, and skills have a wide choice of forms in which to present the result of the assessment, in aggregated reports commonly referred to as dashboards. While verbal comments can be compared to one another by the audience of the report, numerical summaries enable the learners to compare themselves to a required benchmark or to one another. These numerical results could also be used to provide statistical markers such as average, mean, standard deviation, and other measures that describe groups of learners. They could be incorporated as part of a dashboard, provided using a visual/graphical representation of insights or results. These dashboards could have interactive components that enable the viewers to focus on various aspects depending on their level of proficiency with the platform, offering different levels of independence from a ready-made report to fully interactive query mechanisms. Results from log data analysis, which could demonstrate, for example, learning curve analysis require careful consideration in their format of representations.

Studies of feedback (Hattie and Timperley 2007; Shute 2008) being information given to the student concern the effects of immediate vs. delayed feedback on learning outcomes or more specifically as providing online personal feedback at the pre-submission phase and most often as post-submission. Among the positive effects of immediate feedback in both phases, studies indicate helping students in their decision or motivation to practice the tasks and providing an explicit association between outcomes and causes during problem-solving. A negative effect of immediate feedback may be that it leads to dependence on information that is not available during transfer tasks, and it may lead to less care in the choice of answers and may impede metacognitive activities. On the positive side, delayed feedback may encourage learners' engagement in active cognitive and metacognitive processing, creating a sense of autonomy but may be a source for struggling and frustrating, mainly for less motivated learners. Another aspect of timing is affected by the goal of the assessment. The nature of formative assessment is in the ongoing process in which assessment informs the learning process; thus it requires frequent and smaller units of assessment. Summative assessments, on the other hand, provide the learners and other stakeholders with a snapshot of a certain point in the learning process, which requires a valid and reliable body of assessment items or events to produce meaningful insights.

In taking into consideration the insights that the analysis provided, technological platforms are not confined to presenting information only about the work that has already been completed or that is now in progress. Looking ahead, and providing suggestions for further work, or adaptation of further items for the learner to interact with is also part of the presentation that is informed by the preceding learning process. In some cases the entire process is captured in the presentation. Presentation of results connected to learning goals could appear in the form of student models or learning goal landscapes (see also Fig. 8 for an example).

Reports and feedback are often used to describe and engage with information regarding aspects of a learner's performance or understanding. These aspects may

include corrective information, an alternative strategy, information to clarify ideas, encouragement, or simply the correct answer. Generalizing the types of information provided, feedback needs to provide information specifically relating to the task or process of learning that fills a gap between what is understood and what is aimed to be understood. In some cases the reporting can also relate to more general aspects of learning that could also be used in different contexts, such as meta-cognitive skills.

In Windows 5 and 6 we present two examples of innovative personal reports, in which STEP and Numworx platforms are reporting back to students. STEP automatically points out whether or not requirements and additional characteristics are present in the student’s submission. Numworx creates an overlay mapping of student achievements and may present students a model of their achievement level to inform them on further steps.

Window 5 Characterizing Student Responses in STEP

The analysis of the submission in STEP is reported to students by means of the different types of reports that students receive at different stages of their work: pre-submission information provided during exploration and post-submission information. The post-submission report consisted of a list of task requirements and a list of additional mathematical characteristics of the submissions (exemplified in Table 1). As part of authoring a task, designers provide the platform direction about the characteristics of student submissions to be checked and associated with examples. The role of the characteristics is both to articulate mathematical ideas in a mathematical way and to introduce a competing discourse about the phenomena that the students could interact with, which they can either reject or use to refine their submissions. These characteristics can provide information that goes beyond whether students’ submissions are right or wrong and can be used to describe where in the example space each of the submitted examples resides. These characteristics have the potential to challenge the students’ current perspectives and are designed to create a dialogic feedback process. The report includes information on characteristics of each student’s constructed and submitted example. Especially when students are asked to submit multiple examples, the software has a window into what Watson and Mason (2005) call student “response spaces,” collections of learner-generated examples that fulfill a given requirement. When examining the responses of individual students, STEP can help characterize a personal example space.

In the following design (Fig. 7), we show an extract of a post-submission report that focuses on the relations between f and g . Thus the reports for two of the individual examples (out of the three submitted, due to space limitations) and for the submission as a whole are restricted to characteristics out of this category in order to produce a focused report for the learner to interact with.

Table 1 Analyzed characteristics in a student's submitted answer

Task requirements	Additional characteristics	
	Example's characteristics	Personal example space characteristics
Marked three valid triplets	<i>Function g:</i>	"Peculiar" example (extreme, degenerate)
Three different examples	<i>g</i> is a function with minimum	
Each example demonstrates all marked conditions	The graph of <i>g</i> intersects the origin	Sketch (looks right but not correct symbolically)
	<i>Relations between f and g</i>	Easy to compute coefficients
	Functions share symmetry line	Example demonstrating part of marked conditions [wide concept image]
	$f = g$ at exactly one point	
	Graphs do not intersect	Example demonstrating marked and additional conditions [narrow concept image]
	Graphs intersect in two points	
	Graphs intersect in the extremum	
	Intersection not in the extremum	
	<i>Does not meet requirements:</i>	
	Marked invalid triplet	
	Marked four conditions	
	<i>Partially meets the requirements:</i>	
	Marked two conditions or one condition	

Window 6 Knowledge Landscape in the DME

At the basis of reporting back to students on their progress lie knowledge models, representations of the targeted learning goals within a domain. These knowledge models form the "learning landscape" in which the students move. Figure 8 (left) shows an extract of such a knowledge model for the subdomain within geometry on points, lines, and intersections in a seventh-grade mathematics textbook in the Netherlands, implemented in the Numworx Digital Mathematics Environment. The yellow nodes refer to knowledge and the blue ones to skills that students are expected to master. The arrows in this oriented graph depict the dependency

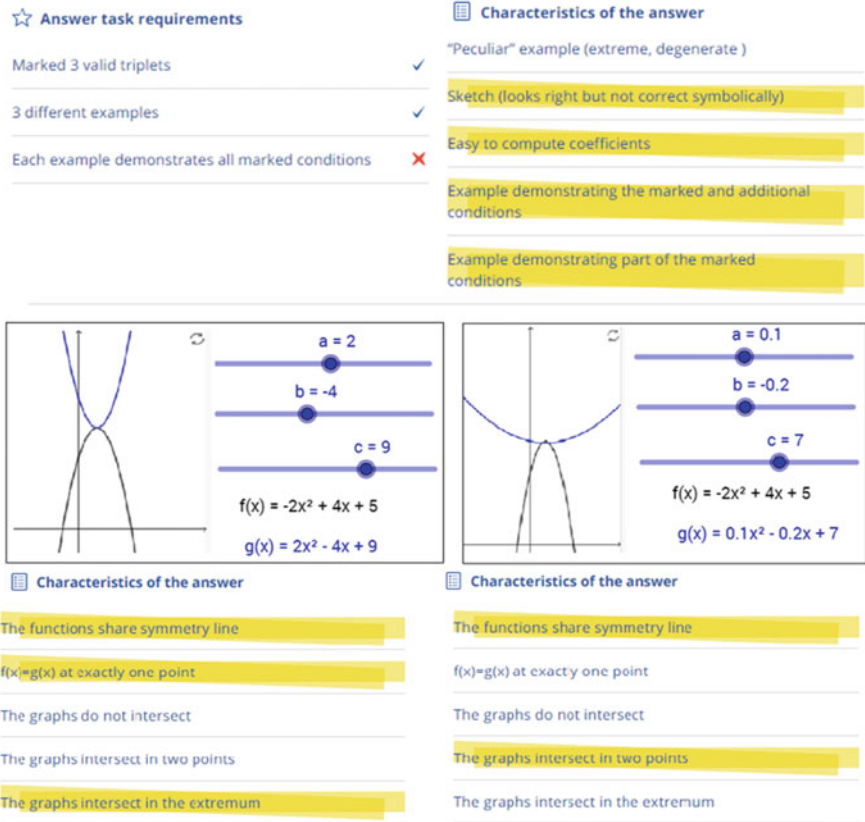


Fig. 7 Post-submission personal report characterizing a personal example space based on *requirements* (top) and additional *characteristics* (bottom)

relationships between the knowledge components: to master a node, mastering the corresponding nodes at the tails of the incoming arrows is needed. The incoming arrows, therefore, start from nodes that can be considered preliminary knowledge. Making explicit the dependency relationships between knowledge components not only is informative in offering an overview of the learning goals landscape but also allows for efficient assessment of progress: if there is evidence of a student mastering a particular knowledge component, one can assume that the “incoming nodes” will also be mastered.

After such a knowledge model is set up, the common next step is to link learning activities or tasks to specific knowledge components. Once these connections are established, student achievements on these tasks inform and load the student model according to this mapping. These “overlay models” (Brusilovsky and Millán 2007) may be presented to students as a model of their achievement level and may inform them on further steps. Figure 8 (right) provides an exemplary implementation of such an overlay student model: nodes become green if there is evidence that the

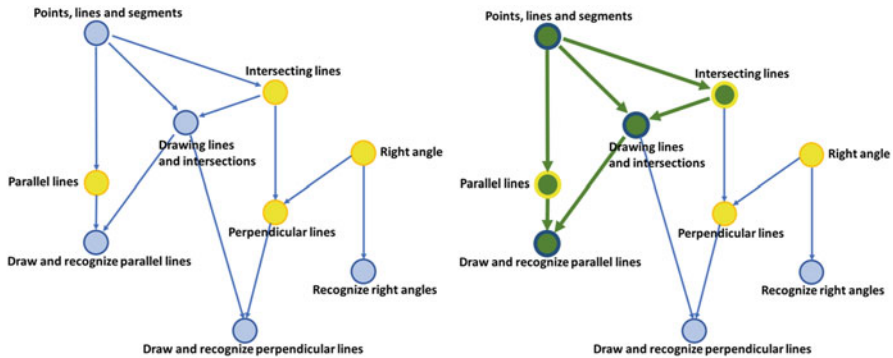


Fig. 8 A knowledge landscape including intrinsic dependencies (left) and a student model overlay (right) in the Numworx Digital Mathematics Environment (<https://www.numworx.nl/en/>)

student masters the nodes, as well as the incoming arrows and the related preliminary knowledge. Both teachers and students have access to this information. Individual student reports can be provided at the level of each of the knowledge components. This informs them on where they are in the landscape and what to do next. Thanks to the dependency relations, there is no need to assess all knowledge elements in detail, but tasks can be more complex and cover more of these elements at once.

While the focus of this chapter is reports designed for the learner, presentation of the insights is not restricted to an individual learner, and the various forms of presentation span over a wide range of audiences.

For example, the knowledge landscape in Fig. 8 concerns feedback to an individual learner. One could consider further elaboration toward more sophisticated feedback and toward other audiences. As for the former, it would be interesting to explore opportunities for including the learning setting (individual, pairs, groups) and the support level (parents, teacher, peers) in the feedback design. If the audience also includes teachers, who want to have an overview of their class' proceedings, the overlay in Fig. 8 could easily be extended to a class level. And, thinking further, why not to a teacher, school, or district level? While such reports carry a great potential to improve education, we should also take into consideration their inherent potential threat they may form in assessing teachers or teaching methods, as to implement superficial forms of “evidence-informed” educational practice.

Conclusion and Discussion

Automatic assessment can create effective means to support the implementation of mathematical practices that are widely promoted by the mathematics education community, but only if it is reliable and valid. The authors of this chapter have been designing and using automatic assessment platforms whose aim is to assess students' mathematical work to report on learning progress on specific mathematics content, as described in Sections “[Design of Tasks and Tools for Student](#)

Mathematical Work”, “Design of Tasks and Tools for Student Mathematical Work,” and “Reporting/Presentation of Results of Analysis” of this chapter. In closing this chapter, we underscore how our commitments to mathematics education influence our designs and argue that as a mathematics education community, it is important to have mathematics educators or any stakeholders that take into account the specific nature of mathematics, continue to engage with advances in automatic assessment, and continue to push such tools to support goals important to the mathematics education community. In particular, it is important that stakeholders in mathematics education engage in the design of techniques for automatic assessment. We close the chapter with four remarks on the current status of automatic assessment of students’ mathematical work and three recommendations for future directions that we see emerging and seek to encourage.

There is an irony with current online assessment; the tasks which are easiest to assess automatically are often the tasks related to calculations which the computer can readily perform, e.g., tasks involving students’ competence with symbolic manipulation, such as factoring or symbolic integration. What is the point of student fluency when computers exceed any reasonable demand of practical fluency? A risk is that digital assessment is not commonly testing competencies that match what it means to “do mathematics” in the twenty-first century.

Using contemporary terminology, it is automatic assessment that creates information for learning analytics. It is common for traditional practices to evolve and progress when they move toward data-driven practices, whether it is production processes or even the use of complex statistical formulas in sports. We see merit in such analytic processes and understand the potential benefits from the integration of learning analytics into the educational world in general and into mathematics education specifically. Yet the quality of the insights provided by data analytics depends on the fit between the information collected and the processes to be improved. When assessment practices are narrowly focused on whether students’ answers are right or wrong, they are not likely to produce insights that will improve the learning of mathematics in classrooms.

By contrast, mathematics educators leading development of technological platforms in mathematics teaching, learning, and assessment usually take either a mathematical or a didactical approach as a starting point. Led by the potential added value of these perspectives, the technological tools developed are usually novel. This approach is not necessarily aligned with the common developmental approach that is led by harnessing a tool’s existing capabilities into meaningful resources. The most trivial example is “time on task,” which is quite easy to measure within technological platforms, which led to studies of possible benefits (e.g., cheating, motivation). This is an example of the data driving the questions rather than questions driving the collection of data.

Mentioning technological developments, it is fair to say that, at the time of our writing, artificial intelligence has made almost no impact whatsoever on automatic assessment of mathematics. In practice, an assessment platform establishes only what a teacher has decided are relevant properties, typically properties decided in advance. The teacher, or task designer, ultimately remains responsible for outcomes,

feedback, and all judgments of value. On reflection, the difficulty of using artificial intelligence in mathematics assessments is hardly surprising: teachers are called upon to make many subtle judgments even in very simple-looking situations. In mathematics, a single misplaced symbol often significantly changes the meaning, and macrolevel decisions (e.g., the style of proof) can result in radically different correct solutions to a particular problem. Furthermore, the judgments required actually turn out to be surprisingly context dependent, and the judgments change over time as students progress in their mathematical career. Given the nature of mathematics as an exact science, that criteria change over time might be a surprise. The change in assessment criteria is an example of the expert reversal effect, a well-established phenomenon that what is useful for a beginner is quite different, perhaps the opposite, of what is useful for an expert; see (Kalyuga et al. 2012). A completely trivial example: early in a student's mathematical career writing fractions in lowest terms, that is to say without common factors in the numerator and denominator, is the whole point of the work. Later, in more advanced work, use of syntactic conventions (representing a rational number as a fraction in lowest terms) is important but is normally much less important relative to other issues, and at times postponing such reductions can be an important solution strategy.

Currently, adaptive testing is also not as widely used as one might expect; sequences of questions, such as faded worked example sequences or a proof comprehension task, are typically a fixed sequence of questions. Computers offer the opportunity for adaptive testing where the response to the current question, or history of previous answers, is used to select the next question. Adaptive testing is not new; adaptive testing is an idea going back to the earliest teaching machines, the history of which was told recently in Watters (2021). Indeed, adaptive testing does not require computer technology at all, and an interesting example of adaptive materials can be found in Crowder and Martin (1960). In their algebra book, students answer simple questions, and based on their answer the student moves to a specific page in a nonlinear fashion. What adaptive testing requires is good design, and the significant difficulty of good design is ultimately why so little progress appears to have been made in this direction. There are multiple examples of adaptive testing (Anderson et al. 1995; Appleby et al. 1997) that are based on mapping dependencies between different skills and models of student thinking. Yet ultimately automatic digital assessment does not make routine use of adaptive testing outside rather large specific projects.

As a first recommendation for future development, the continued development of technology has the potential to address some challenges in current assessment practices. For example, there are other considerations of design that cause developers who are members of the mathematics education community to choose specialized and uncommon technological functions or innovative solutions in mathematics assessment. Designers often seek to make the interaction between the students and tools as convenient as possible in supporting student expression of, and communicating, their ideas. Yet, convenient modalities may have drawbacks in terms of the ability of a platform to identify and analyze the mathematics in students' work. On the other hand, solutions that require severe effort from the students in

doing mathematics such as writing symbolic expressions with an equation editor present a burden that limits the expressivity of the ideas due to the efforts of communicating them. Here technological developments such as optical character recognition (OCR) may provide useful solutions.

Second, the complex thinking that is usually the focus of mathematics educators is commonly best assessed by interviews, rather than through online automatic tools. The development of technological solutions for assessment of higher-order thinking usually requires specific focus and design and is not covered by the standard mass production educational tools, which target greater audiences and potential customers. This leads to a state in which general platforms and tools are either not compatible or not connectable with other technological solutions. One possible path to extend the level of expertise available in automatic assessment is to increase the connectivity of different platforms, to harness existing technology as a baseline for further development, enable to embed additional functionalities into an existing platform, or enable the integration of a novel instrument through use of different application programming interfaces (APIs) available to developers.

When looking into the future of digital assessment, there is one concern that lingers: who has control over the assessment platforms? For example, to what extent can individual teachers write their own assessments, to what extent is data available to research, and how does innovation take place? Another tension relates to the choice of general tools and mathematics-specific tools. Both require teacher resources, and if teachers use their resources on general technology, they use less math-focused technology (Drijvers et al. 2021), a general finding that is also relevant to the case of assessment.

As researchers leading development of technological platforms in mathematics teaching, learning, and assessment, we should acknowledge that this type of development is a complex endeavor. Complex, but also possible, as demonstrated in this chapter and also essential for the evolution of the field of mathematics education. We should continue to pursue solutions that are custom-made for the specific needs of assessment of complex mathematical activity. Keeping in mind that we also tend to adopt the simple solutions that are broader in their aims and audiences, these solutions and design ideas should be kept and developed/incorporated into mainstream platforms in order to enhance their sustainability.

References

- Anderson JR, Corbett AT, Koedinger KR, Pelletier R (1995) Cognitive tutors: lessons learned. *J Learn Sci* 4(2):167–207
- Appleby J, Samuels PC, Jones TT (1997) Diagnosys: a knowledge-based diagnostic test of basic mathematical skills. *Comput Educ* 28:113–131
- Beeson M (1998) Design principles of Mathpert: software to support education in algebra and calculus. In: Kajler N (ed) *Computer-human interaction in symbolic computation*. Springer, pp 89–115. <https://doi.org/10.1007/978-3-7091-6461-7>

- Brusilovsky P, Millán E (2007) User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) *The adaptive web*, LNCS 4321. Springer, pp 3–53. https://doi.org/10.1007/978-3-540-72079-9_1
- Carless D (2015) *Excellence in university assessment: learning from award-winning practice*. Routledge, London
- Carlson M, Oehrtman M, Engelke N (2010) The precalculus concept assessment: a tool for assessing students' reasoning abilities and understandings. *Cogn Instr* 28(2):113–145. <https://doi.org/10.1080/07370001003676587>
- Crowder NA, Martin GC (1960) *Adventures in ALGEBRA*. Doubleday, Garden City
- Douglas M (2004) Moodle: A virtual learning environment for the rest of us. *The Electronic Journal for English as a Second Language*, 8(2)
- Drijvers P (2020) Digital tools in Dutch mathematics education: a dialectic relationship. In: van den Heuvel-Panhuizen M (ed) *National reflections on The Netherlands didactics of mathematics*, pp 177–196). SpringerOpen. https://link.springer.com/chapter/10.1007%2F978-3-030-33824-4_10
- Drijvers P, Ball L, Barzel B, Heid MK, Cao Y, Maschietto M (2016) *Uses of technology in lower secondary mathematics education; a concise topical survey*. Springer. <http://www.springer.com/us/book/9783319336657>
- Drijvers P, Thurm D, Vandervieren E, Klinger M, Moons F, Van der Ree H, Mol A, Barzel B, Doorman M (2021) Distance mathematics teaching in Flanders, Germany and The Netherlands during COVID-19 lockdown. *Educ Stud Math* 108:35–64. <https://doi.org/10.1007/s10649-021-10094-5>
- Hattie J, Timperley H (2007) The power of feedback. *Rev Educ Res* 77(1):81–112. <https://doi.org/10.3102/003465430298487>
- Johnson EG (1992) The design of the national assessment of educational progress. *J Educ Meas* 29(2):95–110
- Kalyuga S, Rikers R, Paas F (2012) Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educ Psychol Rev* 24(2):313–337. <https://doi.org/10.1007/s10648-012-9195-x>
- Kinnear G, Wood AK, Gratwick R (2022) Designing and evaluating an online course to support transition to university mathematics. *Int J Math Educ Sci Technol* 53(1):11–34. <https://doi.org/10.1080/0020739X.2021.1962554>
- Lane-Getaz S (2013) Development of a reliable measure of students' inferential reasoning ability. *Stat Educ Res J* 12(1):20–47. <https://doi.org/10.52041/serj.v12i1>
- Mejia-Ramos JP, Lew K, de la Torre J, Weber K (2017) Developing and validating proof comprehension tests in undergraduate mathematics. *Res Math Educ* 19(2):130–146. <https://doi.org/10.1080/14794802.2017.1325776>
- Mullis IV, Martin MO (2017) TIMSS 2019 assessment frameworks. International Association for the Evaluation of Educational Achievement
- Naftaliev E, Yerushalmy M (2017) Engagement with interactive diagrams: the role played by resources and constraints. In: Leung A, Baccaglioni-Frank A (eds) *Digital technologies in designing mathematics education tasks: potential and pitfalls*. Springer International Publishing, pp 153–173. https://doi.org/10.1007/978-3-319-43423-0_8
- Nicaud JF, Bouhineau D, Chaachoua H (2004) Mixing microworlds and CAS features in building computer systems that help students learn algebra. *Int J Comput Math Learn* 9(2):169–211. <https://doi.org/10.1023/B:IJCO.0000040890.20374.37>
- Olsher S, Yerushalmy M, Chazan D (2016) How might the use of technology in formative assessment support changes in mathematics teaching? *Learn Math* 36(3):11–18. <https://flm-journal.org/Articles/1083F0AD733094BDDDBD64476F743F.pdf>
- Pardo A, Jovanovic J, Dawson S, Gašević D, Mirriahi N (2019) Using learning analytics to scale the provision of personalised feedback. *Br J Educ Technol* 50(1):128–138. <https://doi.org/10.1111/bjet.12592>

- Sangwin CJ (2013) Computer aided assessment of mathematics. Oxford University Press. ISBN 978-0-19-966035-3
- Sangwin CJ (2019) Proof technology in mathematics research and teaching. Springer International. <https://doi.org/10.1007/978-3-030-28483-1>. 15
- Sangwin CJ, Bickerton R (2021) Practical online assessment of mathematical proof. *Int J Math Educ Sci Technol* 53:2637. <https://doi.org/10.1080/0020739X.2021.1896813>
- Sangwin CJ, Jones I (2017) Asymmetry in student achievement on multiple choice and constructed response items in reversible mathematics processes. *Educ Stud Math* 94:205–222. <https://doi.org/10.1007/s10649-016-9725-4>
- Shute VJ (2008) Focus on formative feedback. *Rev Educ Res* 78(1):153–189. <https://doi.org/10.3102/0034654307313795>
- Thoma A, Iannone P (2021) Learning about proof with the theorem prover lean: the abundant numbers task. *Int J Res Undergrad Math Educ* 8:64–93. <https://doi.org/10.1007/s40753-021-00140-1>
- Watson A, Mason J (2005) Mathematics as a constructive activity: the role of learner generated examples. Erlbaum, Mahwah
- Watters A (2021) Teaching machines: the history of personalized learning. MIT Press, Watters
- 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 – Int J Math Educ* 52:1033–1049. <https://doi.org/10.1007/s11858-020-01134-0>