




Student approaches to generating mathematical examples: comparing e-assessment and paper-based tasks

George Kinnear¹  · Paola Iannone¹ · Ben Davies²

Accepted: 20 August 2024
© The Author(s) 2024

Abstract

Example-generation tasks have been suggested as an effective way to both promote students' learning of mathematics and assess students' understanding of concepts. E-assessment offers the potential to use example-generation tasks with large groups of students, but there has been little research on this approach so far. Across two studies, we investigate students' responses to a particular sequence of example-generation tasks, posed either on paper or through e-assessment. We identify a striking difference in students' example-generation strategies and success, for e-assessment and paper-based versions of the same tasks. This suggests the use of example-generation tasks in e-assessment may not be straightforward, and we conclude by discussing the implications for teaching and research.

Keywords Learner-generated examples · Digital tasks · Example spaces · Example-generation strategies

There is a large body of work in the mathematics education literature arguing that example-generation tasks are an effective way to develop students' understanding of mathematical concepts (e.g., Watson and Mason, 2005). However, these tasks are rarely used in practice in undergraduate assessment because it can be demanding for lecturers to judge the students' answers, particularly with large cohorts. E-assessment presents an opportunity to use example-generation tasks with large groups of students, where an automated system can check numerous student-generated examples and provide feedback on a scale that would not be practical in a manual setting.

Currently, little is known about how students approach example-generation tasks as part of an e-assessment. Previous research has developed frameworks for analysing the range of examples generated by students (Zazkis & Leikin, 2007), as well as the strategies used by experts to produce examples (Antonini, 2011). These have been successfully applied by researchers in a range of contexts (e.g., Sağlam and Dost, 2016; Edwards and Alcock, 2010; Iannone et al., 2011; Lew and Zazkis, 2019), but so far the focus has been on paper-based tasks. We know that task format can influence students' approaches (Lemmo, 2021), and that e-assessment is playing an increasingly important role in undergraduate mathematics

✉ George Kinnear
G.Kinnear@ed.ac.uk

¹ School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK

² University of Southampton, Southampton, UK

courses (Kinnear et al., 2022). Hence it is important to investigate how students respond to e-assessment versions of example-generation tasks.

In this paper, our overarching aim is to investigate students' responses to a particular sequence of example-generation tasks, posed either on paper or through e-assessment. We adopt a mixed-methods approach to examine students' example generation in both settings, in terms of the examples they generate, and the strategies used to generate them. In Study 1, we analyse a large sample of students' written and e-assessment responses, with a focus on outcomes and examples produced. In Study 2, we analyse students' responses to the same tasks during task-based interviews, to gain insight into students' example-generation strategies.

1 Literature review and theoretical framing

We begin with a review of three aspects of the literature on example generation: the theoretical grounding in variation theory; approaches to characterising the examples generated by students; and research on the strategies that students use to generate examples. In addition, we review relevant literature on the comparison between paper and digital tasks, and on students' conceptions of functions (since this is the topic addressed by our tasks).

Variation theory and example generation Variation theory (Marton et al., 2004; Runesson, 2006) forms the theoretical basis for the focus on the role of example-generation tasks in students' learning (Watson & Mason, 2005). According to variation theory, learners come to understand a particular aspect of a concept through experiencing variation of that aspect across examples. Drawing on this theoretical framing, Fukawa-Connelly and Newton (2014) analysed the range of examples of a concept (mathematical group) that was available to students in an abstract algebra class. While their analysis was grounded in variation theory, they also drew on notions described by Watson and Mason (2005), such as the *example space*: the "class of potential examples" (p. 45) that a learner associates to a given concept. Watson and Mason (2005) emphasised the situated and personal nature of these example spaces, reflecting the influence of "task, cues and environment" (p. 76). This motivates the focus of our study on the impact of varying the format of the task.

Characterising example spaces Students' example spaces can be inferred from their responses to example-generation prompts. Such inferences should be made cautiously: while a student may not give a particular example, this "does not imply that it is not within their accessible space" (Goldenberg and Mason, 2008, p. 189). A student's example space can be compared to the *conventional example space* as generally understood by mathematicians, or to the *collective and situated example space* that is established in a given classroom or other local context (Watson and Mason, 2005, p. 67). Zazkis and Leikin (2007) proposed a framework for analysing students' example spaces, in the context of clinical interviews where "participants have ample opportunity to provide or construct examples, and that multiple examples are encouraged by asking for 'another and another' and for 'something different'" (p. 19). Thus, in studies such as ours, where students are asked to respond to a fixed sequence of tasks, this type of analysis is typically not possible. Instead, students' examples can be classified based on their mathematical properties to give insight into the collective and situated example space. For instance, Popper and Yerushalmy (2021) classified students' examples of quadrilaterals into standard sub-types, considering the totality of work submitted by the

students as a collective example space for that classroom. Similarly, Fahlgren and Brunström (2023) classified students' examples of functions according to their view of important categories within the conventional example space. We use a similar approach to analyse the collective example space.

Example-generation strategies Another focus for research has been on the strategies that students use to generate examples. Drawing on task-based interviews with postgraduate students, Antonini (2006) identified three strategies:

1. *Trial and error* - "The example is sought among some recalled objects; for each example the subject only observes whether it has the requested properties or not" (p. 58),
2. *Transformation* - "An object that satisfies part of the requested properties is modified through one or more successive transformations until it is turned into a new object with all the requested characteristics." (p. 59)
3. *Analysis* - "Assuming the constructed object, and possibly assuming that it satisfies other properties added in order to simplify or restrict the search ground, further properties are deduced up to consequences that may evoke either a known object or a procedure to construct the requested one." (pp. 59-60)

Antonini's (2006) classification of strategies has proved useful in other research that sought to analyse how students respond to example-generation tasks. Iannone et al. (2011) classified undergraduate students' strategies after they solved tasks involving concepts that were new to them. They found that the trial and error strategy was dominant, accounting for about 80% of all examples generated, and suggested that students' over-reliance on this strategy could explain why the generation tasks in the study did not lead to the learning gains that the literature predicts (e.g., Watson and Mason, 2005). Other researchers have echoed the finding that undergraduate students predominantly use the trial and error strategy (Edwards & Alcock, 2010; Sağlam & Dost, 2016).

However, these studies also highlighted issues when classifying undergraduate students' strategies using Antonini's (2006) terms (or the elaboration provided by Antonini (2011)). Edwards and Alcock (2010) noted that some students appeared to use transformation or analysis strategies, but with faulty assumptions or deductions. Similarly, Iannone et al. (2011) reported that students using the trial and error strategy did not always check that their selected example had the desired properties. Sağlam and Dost (2016) highlighted two additional aspects not captured by the individual strategies. First, several students in their study began their work by writing out definitions of the concepts involved; while the authors do not interpret this a distinct strategy, it appeared to help the students to work with tasks that involved multiple concepts. Second, students often transitioned between strategies: for instance, some students generated examples through trial and error, noticed features that needed to be accounted for, then switched to a different strategy. This is a more sophisticated view of trial and error, beyond (Antonini, 2011) noting that the trials can be generated "with a precise and planned order" (p. 208).

To summarise, the current literature on example-generation strategies suggests that undergraduate students may not use all the strategies that experts use. We, therefore, draw on and develop, Antonini's (2011) framework for our analysis of students' strategies in Study 2.

Comparing paper and digital mathematics tasks The transfer of traditional pen-and-paper assessments to e-assessment can be achieved in a variety of ways. Ripley (2009) proposed a classification of transfer approaches, distinguishing between assessment migration and

assessment transformation. Assessment migration involves simply replicating a task, where the question remains unchanged except for the mode of delivery. On the other hand, assessment transformation entails implementing the task in a predominantly digital manner, such as requiring students to use applets or specific features of the digital interface to solve problems. Several studies have investigated the characteristics of each type of transfer, examining students' approaches and performance in these tasks. Of interest to our study is the finding of Lemmo (2021) that even simply migrating a paper-based task to e-assessment alters students' interaction with the task and affects their performance, showing that the migration process is not neutral. Johnson and Green (2006) and Threlfall et al. (2007) arrived at similar findings, emphasising that task characteristics and question types can influence students' working methods.

The studies discussed so far are all in the school mathematics context, and in fact, we did not find much written about the transfer of tasks from paper to digital format at the undergraduate level. One example close to the undergraduate context is the study by Yerushalmy et al. (2017), which used a task to probe school students' understanding of the definition of the tangent to a graph. The task exemplifies Ripley's (2009) notion of "assessment transformation" since the students used an interface with movable points, making the technology an integral part of the task solution. The authors found that the task provided valuable feedback to the teacher regarding students' conceptions about the tangent and that information gathered in this way could be an important tool to design feedback. Yerushalmy et al. (2017) conclude that

Designing interactive e-tasks with many correct solutions that can be worked out and expressed in different representations with mathematical tools or symbols is of significant pedagogical value. (p. 714).

It seems plausible to think that such important pedagogical value would exist also for similar tasks designed for undergraduate students.

Function concepts The tasks that we use in our studies are based on the concepts of injectivity and surjectivity of functions. We chose this topic pragmatically since it was studied in the course where we collected data. Moreover, there has been relatively little research published about students' understanding of these concepts. Bansilal et al. (2017) found that a majority of pre-service teachers relied on faulty reasoning to conclude that a given function was surjective, such as being many-to-one. Similarly, Thoma and Nardi (2019) noted that students would often use an inaccurately recalled definition to decide whether a given function was injective or surjective (based on a commognitive analysis of exam scripts). In a study focused on more sophisticated concepts from abstract algebra, Uscanga and Cook (2022) noted that students' conceptions of injectivity and surjectivity warrant further study, based on an analysis of textbook examples and interviews with instructors.

Students' conceptions of piecewise functions are also relevant to our tasks. Previous research has found that students are reluctant to work with piecewise functions (e.g., Kontorovich and Li, 2022). In some cases, this is due to the conception that functions are defined by a single rule, so that a piecewise function is therefore really two (or more) functions (Mirin, 2022). Other difficulties may stem from a conception that function graphs should be "reasonable" (Vinner, 1983, p. 303), while piecewise functions may have graphs that have sudden changes in behaviour.

2 Study 1: quantitative analysis of assessment responses

We developed a sequence of example-generation prompts that students could respond to either in traditional written format or through e-assessment. To gather responses from students, we used a quasi-experimental design with successive cohorts of students taking an undergraduate course in pure mathematics. This was a pragmatic approach, as we were able to embed the tasks in relevant assessments taking place in the course, with the two different formats used in successive years (we return to the limitations of this design in the Discussion). The set of responses to the two formats sheds light on the collective example space (Watson & Mason, 2005) of the two cohorts of students, and also gives insight into which examples are most accessible.

For this study, we pose two research questions:

RQ1: How did students' performance compare across the two formats?

RQ2: How did the examples that were produced (i.e., the collective example space) compare across the two formats?

2.1 Methods

Materials We devised the sequence of four example-generation tasks shown in Fig. 1, drawing on advice from Watson and Mason (2005, p. 131) to prompt students to “make up an example with some constraints” and to “add constraints sequentially”. We refer to the four prompts as Task 1 to Task 4.

The e-assessment version of the task sequence was implemented using STACK (Sangwin, 2013), with interactive graphics provided by the JSXGraph plugin.¹ For each task, the software showed a grid representing $[0, 1] \times [0, 1]$, with four movable points generating a tripartite piecewise linear function. Figure 2 shows a successful attempt at Task 1. A task designed in this way is in the category of assessment transformation (Ripley, 2009) as it involves students using an interface (in this case the grid with the movable points) to solve the given task.

In this case, the e-assessment design was driven by the need to provide students with an intuitive interface for producing function graphs, while also producing output that can be interpreted by the Maxima computer algebra system used by STACK to evaluate students' answers. For instance, a different e-assessment design could have provided students with a digital “paintbrush” to draw their graph,² but this would have required a more complicated interface (e.g., to allow for erasing and tweaking the drawing), and it would be difficult to assess the resulting sketch automatically. In our design, the four movable points provide the facility to express a suitable variety of function behaviours, and the properties of the example can be readily inferred from the coordinates of the four points. Thus, the tasks were not strictly identical across the two formats: the e-assessment version provided students with some scaffolding by providing the axes and a prototype function graph, while limiting students to piecewise linear examples (a point we return to in the Discussion). Nevertheless, we argue that since the example-generation prompts were identical in both formats, and both formats allowed for a range of responses, it is reasonable to compare students' responses across the two formats.

¹ further examples of this functionality are given by Hooper and Jones (2023)

² an example of such an applet, that is constrained to drawing polynomials of a particular degree, is available at <https://jsxgraph.uni-bayreuth.de/share/example/sketch-polynomial>

In each case, draw the graph of a function with the given properties, and label important points:

- $f_1 : [0, 1] \rightarrow [0, 1]$ has image $[0, \frac{1}{2}]$.
- $f_2 : [0, 1] \rightarrow [0, 1]$ has image $[0, \frac{1}{2}]$ and is not injective.
- $f_3 : [0, 1] \rightarrow [0, 1]$ is surjective and not injective.
- $f_4 : [0, 1] \rightarrow [0, 1]$ is injective, not surjective, and passes through $(0.2, 0.8)$ and $(0.5, 0.5)$.

Fig. 1 The written version of the task sequence

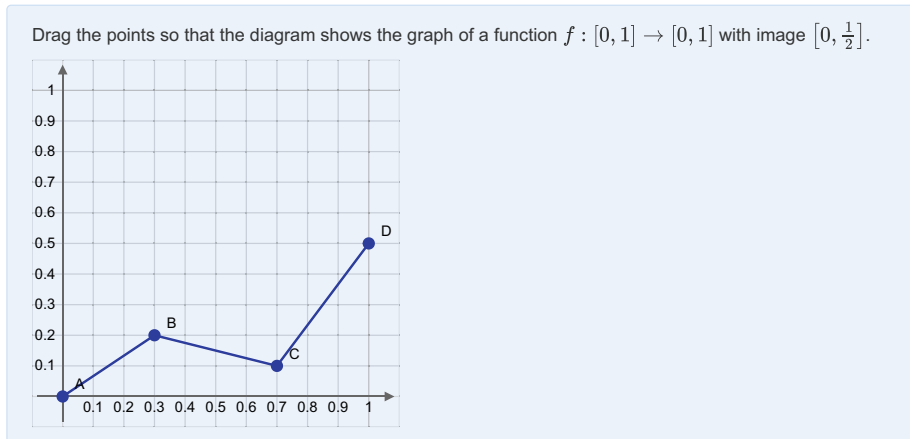


Fig. 2 The e-assessment version of Task 1. The points A, B, C and D are constrained to the grid. Note that before manipulation, all four points default to the x -axis

Participants and procedure Participants were students at a large, research-intensive UK university, in a first-year introduction-to-proof module that follows Liebeck's (2018) textbook. Most students were taking the course as a compulsory part of the first year of an undergraduate mathematics degree, although approximately a third of the students were taking the course as an option on another programme (e.g., physics or computer science). The course curriculum was the same for both the 2019/20 and 2020/21 cohorts, with an overall design based on a weekly schedule of assigned reading, lectures, small-group problem-solving classes, and assessed homework (though aspects of the course design were adjusted to mitigate the Covid-19 pandemic; for instance, lectures took place online in 2020/21). Ethics approval for the study was obtained through the School of Mathematics at The University of Edinburgh, which granted administrative consent for the research team to analyse students' anonymised responses during the normal running of the course.

The e-assessment tasks were completed by 322 of 377 students from the 2019/20 cohort. The tasks were embedded in the week 8 "reading quiz" that students were asked to complete online at the start of the week. Students completed the quiz unsupervised in their own time, with a time limit of 1 h once the quiz was opened (the mean time spent on the quiz overall was 28 min). Students received automatically generated feedback from STACK only after the deadline had passed. The best 8 of 10 weekly quiz scores contributed 5% to the course result.

The written tasks were completed by 333 of 451 students from the 2020/21 cohort. The tasks were included in the week 8 written assignment; the best 8 of 10 written assignments

contributed 25% to the course result. The written assignments were also completed unsupervised in students' own time, within a 1-week period. Students received written feedback from their tutors 1 week after submission. Students in the course are encouraged to work with peers, and make use of online resources, but there is a clear expectation that work submitted for assessment should reflect the student's own understanding.

2.2 Data analysis

We reviewed all of the responses to judge their correctness and the types of examples that were generated. For the written responses, each author independently analysed 125 students' responses. For each example, we noted whether it was fully correct and classified the type of mathematical function (e.g., linear, quadratic, piecewise) in order to describe the collective example space. These judgements were subjective since students did not always write an algebraic expression; to provide a check of reliability, we distributed the coding so that 42 students' work was analysed by at least two authors. We achieved a high level of inter-coder reliability on the sample that was double-coded, with a Krippendorff's alpha value of 0.897 for correctness and 0.857 for example type (Krippendorff, 2004).

For the e-assessment responses, correctness was evaluated using the existing STACK code that checks properties of the student answer. We also partially automated the identification of example types, using an R script (available at osf.io/sa24w) to identify patterns of common responses. For example, if the four points did not all have distinct x -coordinates, we categorised the example as "not a function" since it would necessarily include a vertical segment.

2.3 Results

Our first research question concerned the relative performance of the written and e-assessment groups. Figure 3 shows that, in aggregate, students from the written group performed substantially better than the e-assessment group on all four tasks. Moreover, there was a strikingly similar pattern in performance across the four tasks in the two groups, with Task 4 being the most difficult.

Our second research question sought details of the types of examples generated by each group, to shed light on their collective example spaces. Here we focus on the correct examples (for analysis of incorrect examples, see Kinnear et al., 2023). Our analysis identified common groups of examples; these are summarised in Fig. 4 and we give further details for each task in turn.

Task 1: image $[0, \frac{1}{2}]$ For the written group, the vast majority of students (90%) gave the straight line $y = \frac{1}{2}x$ as their example.³ The e-assessment group gave a more diverse array of responses in this case, most frequently choosing monotonically increasing piecewise functions. The $y = \frac{1}{2}x$ example was also common, and relied on students realising that the points could be dragged horizontally as well as vertically.

Task 2: image $[0, \frac{1}{2}]$ and not injective To satisfy the non-injectivity condition, 54% of responses from the written group used a parabola (either concave down, as shown in Fig. 4,

³ Students typically did not give an explicit algebraic formula with their response. These formulae were not a main feature of our analysis and are only included here for brevity.

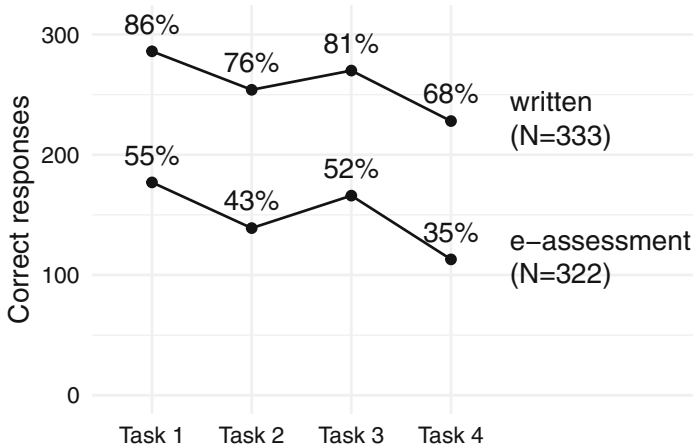


Fig. 3 Number and proportion of correct responses to each task from each group

or concave up). Piecewise linear graphs were also common, sometimes explicitly specified as $y = |x - \frac{1}{2}|$. The remaining responses were made up of cubic curves (31 examples) and various other more complicated curves (26 examples); for instance, one student drew $y = |\frac{1}{2} \cos(2\pi x)|$.

Responses from the e-assessment group were dominated by the “hump” shape shown in Fig. 4, with 43 students giving the example as shown and a further 11 giving the concave-up version. Since it was not possible to represent a parabola using the e-assessment interface, this is perhaps the students’ best approximation. The next most common type of response was nondecreasing, with at least one flat segment. Other common responses were a “zig-zag” shape with segments alternately increasing and decreasing (28 examples) and variations of the hump shape where the segment between the middle two points was not flat (24 examples)

Task 3: surjective and not injective Among the written responses, we observed a similar mix of functions as in the previous task, with parabolas (58%) and piecewise linear functions (20%) dominating. The correct e-assessment responses were quite varied, but most often the non-injectivity was due to a change from increasing to decreasing (or vice versa) resulting in a “peak” (or “valley”) shape. Among these are 22 examples of the “hump” shape from Task 2. Another common group of answers were monotonic with a flat segment, which were predominantly nondecreasing (48 of the 50 examples). As with Task 2, a further common response was a “zig-zag” shape (39 responses).

Task 4: injective, not surjective, and passes through (0.2, 0.8) and (0.5, 0.5) This task was designed to prompt students to generate a bespoke example, since standard linear/quadratic functions cannot satisfy all the constraints. For the written group, most correct answers were piecewise linear (72%) in contrast to the previous tasks where smooth curves were predominant. Still, many students in the written group drew smooth curves for this task too (62 examples). For the e-assessment group, essentially the only choice was whether to restrict the image at the top or the bottom. The largest group of examples omitted values at

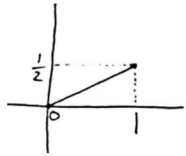
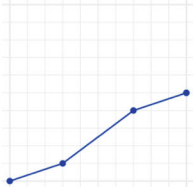
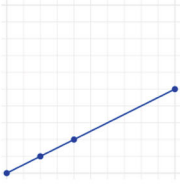
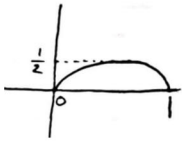
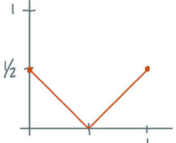
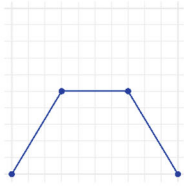

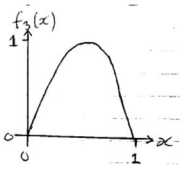
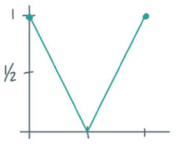
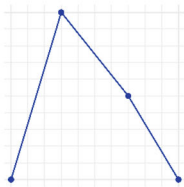
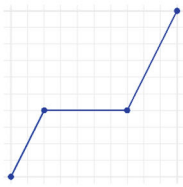
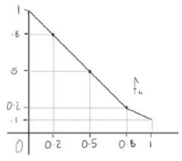
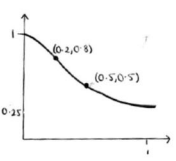


	Written		E-assessment	
Task 1	286 correct examples		177 correct examples	
	258 (90%) linear		84 (47%) increasing	53 (30%) linear
				
Task 2	254 correct examples		139 correct examples	
	137 (54%) quadratic	57 (22%) piecewise linear	54 (39%) hump	33 (24%) nondecreasing
				
Task 3	270 correct examples		166 correct examples	
	156 (58%) quadratic	55 (20%) piecewise linear	73 (44%) peak/valley	50 (30%) monotonic
				
Task 4	228 correct examples		113 correct examples	
	164 (72%) piecewise linear	62 (27%) other	55 (49%) top end	28 (24%) both ends
				

Fig. 4 Overview of the collective example spaces, showing the most common types of correct examples generated by students for each task

the top end (55 examples), but omitting at both ends (28 examples) and only at the bottom (24 examples) was also common.

2.4 Discussion

We found that performance on the e-assessment version of the tasks was substantially lower than on the written version. This large difference in performance — even on simple tasks like Task 1 — was surprising. Our analysis of the types of (correct) examples used by students revealed clear differences in the collective example spaces elicited by the two formats. For the written version, students predominantly used familiar types such as linear, quadratic and cubic functions where possible. For the e-assessment version, where students did not have access to these familiar types of function, the most common answers often appeared to approximate the more familiar examples (e.g., the “hump” example for Task 2 approximating a parabola).

We offer three plausible explanations for the large difference in performance. The first is that the e-assessment version requires students to produce examples consisting of piecewise linear functions. Our analysis of the written examples showed that students rarely used such functions. Moreover, there is some evidence in the literature (Hohensee, 2006; Mirin, 2022) that students at this stage have difficulties working with piecewise functions. A second possible explanation is that the students found the e-assessment interface difficult to use. While we cannot entirely rule this out based on the available data, we note that most students were able to produce a response. Finally, it could be the case that students tackled the tasks — in written form and online — differently, for example by using different example-generation strategies.

We also note two limitations of our longitudinal (in particular, repeated cross-sectional) design. First, it is possible that the two groups of students in this study had different levels of motivation toward the tasks. Both versions were part of assessments that contributed to course grades, but the written assignment had a higher weight than the e-assessment version. However, in both years, student performance on both forms of assessment was broadly similar, with average scores across the class of around 75% each week. Second, the written assessment took place toward the end of the week of study, while the e-assessment tasks were set near the start (as part of a quiz on the assigned reading). Thus, the higher performance on the written version could have been due to greater familiarity with the concepts.

The key finding from this study, of a stark difference in performance between the two groups, provides empirical evidence of the potential for task format to influence student performance. We designed a second study to develop a better understanding of the possible explanations for differences between the two groups. In what follows we describe the study and its outcomes.

3 Study 2: Qualitative analysis of task-based interviews

Our second study focused on students’ example-generation strategies as a possible explanation for the differences between the written and e-assessment tasks. We carried out a series of interviews, in person and online, during which students were asked to solve tasks 1 to 4. For this study, we asked the two research questions:

RQ1: Is there a difference in example-generation strategy use between students attempting tasks in written and e-assessment formats?

RQ2: What student difficulties contribute to unsuccessful attempts at example generation?

3.1 Methods

Participants were students enrolled in the same introduction to proof module as in the first study, at the same university, but in the 2021/22 academic year. An email was sent to all 536 students enrolled on the course, with an invitation to either an in-person interview or an online one. Nineteen students replied, 11 for the in-person interviews and 8 for the online ones. The interviews lasted between 15 and 30 min and were video recorded. Ethics approval for the study was obtained from the Ethics Committee of the School of Science at Loughborough University (UK) where the second author worked at the time of data collection.

The interviewers (the first author for the in-person ones, the second author for the online ones) used the same interview schedule, following a task-based model (Maher & Sigley, 2020). We described the project aims, then invited students to solve the same four tasks as in the previous study: the written version for the in-person interviews and the e-assessment version for the online ones. The students had access to reference materials from the course, including the definitions of functions, domain, range, injectivity and surjectivity (see osf.io/rp7a3). The interviewer intervened either when the student seemed to pause (e.g., “what are you thinking about?”), or to ask for clarification regarding what the student was doing (e.g., “how did you come up with that?”). When the student offered a solution, the interviewer asked whether it was correct and why (e.g., “can you explain what’s going on?”, as in Fig. 5). This sometimes led to the student noticing an error and revising their solution to


Peter		
Task 2: Draw the graph of a function $f_2: [0, 1] \rightarrow [0, 1]$ that has image $[0, \frac{1}{2}]$ and is not injective		
Transcript with time stamp (I=Interviewer, S=student)	Example Generated	Analytical summaries
<p>[4:09] [reading task]</p> <p>[4:22] S: So injective... that means that, er. So each of the points that are mapped, they have to have at most one output, I suppose.</p> <p><i>[interviewer offers reference materials and student consults them]</i></p> <p>[6:43] <i>[starts drawing axes and annotating points on the y-axis]</i></p> <p>[7:14] <i>[draws the curve]</i></p> <p>[7:24] S: I guess, something like this?</p> <p>I: OK, can you explain what’s going on?</p> <p>S: Er, not every element in the codomain is mapped, so it doesn’t go... the graph doesn’t span the whole codomain <i>[gesturing at $y=1$]</i>. But a point here <i>[gesturing at the horizontal line $y=1/2$]</i> could be mapped onto a point, say $\frac{1}{2}$ there.</p>	<p>Synthesis, incorrect</p>  <p>Prompted validation</p>	<p>Student seems to confuse injective and surjective, and this remains the case even after consulting the reference materials</p> <p>The strategy seems to be Synthesis because of the apparent attention to the properties, even though the student’s understanding of those was flawed.</p>

Fig. 5 Analysis of Peter’s attempt at Task 2. On the left-hand side is a (reduced) transcript of the interaction with time stamps; in the middle, a picture of the student’s example is annotated with the strategy assigned by the coder and an indication of whether it is correct; on the right is the analytical commentary

produce a new example. In order to align the interviewing styles of the two interviewers, the second author observed the first interview of the first author before starting their interviews and a debriefing meeting was held afterward.

3.2 Data analysis

We began the analysis by investigating the students' example-generation attempts in terms of the strategies described by Antonini (2011). Our unit of analysis was an attempt to solve one of the tasks, thus there were 76 units of analysis in total. For each unit of analysis, we produced a transcript and an analytical summary of the content (for example, commenting on the types of difficulties the student encountered while solving the task), as well as coding the strategy used by the student. An example of this is shown in Fig. 5.

Our coding of example-generation strategies proceeded by triangulation analysis (Mok & Clarke, 2015), in three stages:

1. We each selected at random one interview to analyse independently. We returned to the original videos — either recordings of the online interviews or the video-recorded in-person interviews — to code the example-generation strategies according to those identified by Antonini (2011). We discussed these interviews and our coding to arrive at a common understanding of the different strategies. From this discussion, we identified two main difficulties.

First, some students produced a chain of examples before stopping, and could therefore use a mix of different strategies. To address this, we agreed to note details of each example generated by a student and to code the strategy used in each case.

The second difficulty that we encountered was that some students employed a strategy that did not fit exactly under any of those described by Antonini (2006, 2011). In particular, we noted several cases where students re-read the definitions relevant to the task and tried to make sense of them before using them to construct an example. We named this strategy *Synthesis* to indicate an approach based on elaboration of the relevant definitions before starting to construct the given example according to these definitions. The *Synthesis* and *Analysis* strategies are related, in that both are based on reasoning about mathematical properties. However, in the *Analysis* strategy, the experts proceed deductively starting from the hypothesis that an example (the required mathematical object) exists and deducing further properties until they recognise such an object among those in their example space or they can construct one. Our students, on the other hand, seemed to proceed inductively, starting from the relevant definitions and elaborating on them (e.g., interpreting algebraic constraints graphically), before looking to find or construct an example that may fit them. We therefore chose the name *Synthesis* to indicate that students' reasoning was proceeding in the opposite direction⁴ to the experts in Antonini's (2006) study. Crucially, *Synthesis* goes beyond simply reading the definitions and then

⁴ Antonini (2011) notes that his use of "Analysis" is because the strategy is "similar to the method used by ancient Greeks for both geometrical construction and search for proofs" (p. 208). In the context of problem-solving, Polya (1945) highlighted a reciprocal relationship between analysis and synthesis in the writings of Pappus, noting that "analysis is invention, synthesis, execution; analysis is devising a plan, synthesis carrying through the plan" (p. 146). Our use of "Synthesis" likewise alludes to the reciprocity with Antonini's Analysis: while both strategies depend on reasoning, the deductions flow in opposite directions.

employing one of the Trial and Error or Transformation strategies, as some of the students in Sağlam and Dost's (2016) study did. In the case of the example-generation attempts coded as Synthesis in our sample, students elaborated on the definitions to establish constraints that guided their generation of an example.

2. We each coded one more interview independently, using the shared understanding of the strategies that we had developed. We shared the coding and discussed how the strategies were used in the interviews coded so far. At this stage, we refined the definitions of the strategies, as summarised in Table 1.
3. We independently coded the rest of the interviews. We each coded a similar number of interviews, ensuring that each author coded both in-person and online interviews to avoid bias.

After the first round of coding was complete with the analytical commentary for the 76 units of analysis, the second author carried out a second round of coding focused on the students' difficulties (RQ2). The types of difficulties across the interviews were classified by pattern coding (Saldaña, 2021), which allowed grouping of types of difficulties, as well as investigating the distribution of strategy usage and effectiveness across the written and e-assessment tasks.

We note that 32 of the 76 example-generation instances consisted of a sequence of examples. In these instances, the student generated an example and then realised, either by themselves or following a prompt by the interviewer, that the example did not fulfil the given requirements and then produced another one (and at times two or three more). To enable comparison with results from Study 1, we scored each example-generation task with 1 for a correct answer and 0 for an incorrect one. For sequences of examples, we scored the last example generated by the student before any prompting by the interviewer (for instance, Sofia's attempt in Fig. 6 was scored 1 because the final example was correct and arrived at without prompting from the interviewer; Porsia's attempt in Fig. 9 was scored 0 based on the incorrect first example, since the second example was only produced after prompting).

Table 1 Definitions of the example-generation strategies

Strategy	Description	Example
Trial and Error (T&E)	The student tries examples from their example space to see whether they may fit the object requested	Fig. 8
Transformation	The student modifies an example which does not yet have all the required properties	Fig. 6
Synthesis	The student recalls or revises the relevant theory (e.g., the definition of injective function) and uses their elaboration of these definitions to guide their approach to the task	Fig. 5, Fig. 7
Analysis	The student assumes the required object exists and uses the properties to deduce consequences that either evoke a known example or a procedure to construct one	None

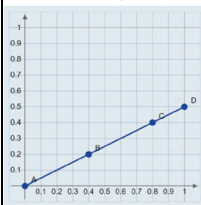
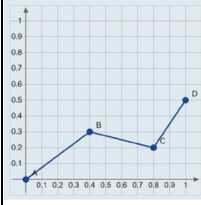
Sofia		
Task 2: Draw the graph of a function $f_2: [0, 1] \rightarrow [0, 1]$ that has image $[0, \frac{1}{2}]$ and is not injective		
Transcript with time stamp (I=Interviewer, S=student)	Example Generated	Analytical summaries
<p>[9:20] <i>[reading question]</i></p> <p>[9:30] S: It's the same question as last time, but the function needs to not be injective. Which means, eh... injective means that there's only one value in the domain that maps to each value in the codomain. So you can't have... if it's not injective, then there needs to be two... two values in the domain that give the same in the codomain.</p> <p>[10:10] I: Exactly. So how would you realise such a function?</p> <p>S: Well, I'm just thinking about my last one and whether that was injective or not. Because that was just a straight line [...] I think my last one was injective.</p> <p>[10:50] I'm just going to put in my last one, just to see if that helps me think about it. <i>[moves points into place]</i></p> <p>[11:11] S: So, the same function... isn't going to work for this. It's not injective, then it needs to... there needs to be a point on the graph where you can draw a horizontal line across it and it'll hit the graph twice [...] it needs to change direction at some point, basically.</p>	<p>Transformation, incorrect</p>  <p>Unprompted validation</p>	<p>There are some signs of Synthesis here, but the student's strategy seems to be mainly based on whether the previous example will suffice or in what ways it might fail to work.</p>
<p>[11:46] <i>[moves point C down, then point B up]</i></p> <p>[11:55] S: I would want it to look something like this.</p> <p>I: Yeah. Why do you think this would work?</p> <p>S: Because, I was just saying, as long as there's somewhere on the graph that we can get the same y value multiple times for different x values, then it's not injective.</p>	<p>Transformation, correct</p>  <p>Prompted validation</p>	<p>The interface may have heightened the sense of transformation here, as the student moved points around to achieve a desired effect (described at 11:11).</p>

Fig. 6 Sofia's attempt at Task 2, consisting of a sequence of two example-generation instances, both of which came from the Transformation strategy

3.3 Findings

3.3.1 Scoring the example-generation attempts

The scoring of the example-generation tasks produced the grouping in Table 2. We observe

Table 2 Scores for each student, reflecting the number of example-generation tasks they completed successfully without prompting from the interviewer

Score	Students	
	Written	E-assessment
0	Paul	Sally
1	Phoebe, Prue	Serena, Steve, Sergio
2	Pamela, Peter, Philip	Sofia, Stefan, Siobhan
3	Pat, Pier, Porsia, Patrick	Sam
4	Padme	

Note that the names are pseudonyms, that do not necessarily reflect the gender of the participants

that the scores reflect the findings from Study 1: students completing the written version performed better than students completing the e-assessment version. For the e-assessment tasks, only one student achieved a score of 3 and none a score of 4.

3.3.2 Strategy usage across attempts

To analyse the strategies used by students, and their effectiveness, we consider separately the cases where students generated a single example and where they generated a chain of examples.

Single-example instances There were 44 example-generation instances that consisted only of one example. These were correct 29 times; for the remaining 15 incorrect examples, the student either abandoned the task or was not able to explain their answer and the interviewer moved to the following task. Table 3 shows the distribution of strategy usage and success across the two task formats.

The Trial and Error strategy was the most used among the single-example instances — with 18 instances, 12 of which gave a correct example — but not by much. In fact, the Synthesis strategy was used 16 times in this sample and mostly on the written version of the tasks, with only two students using this strategy in the e-assessment version (both times producing a correct example). While all students were given reference material from the lectures, it was almost exclusively those in the written group who used this material to guide them (albeit not always successfully) to the construction of the required example. When the Synthesis strategy was not successful, this was because the student constructed an example starting from inaccurate premises or drew inaccurate inferences. One example of this is the solution to Task 2 proposed by Peter (Fig. 5). During the interview, Peter confused the definitions of injective and surjective, even after consulting the lecture notes. From this misunderstanding, they constructed an example of an injective function whose image is $[0, \frac{1}{2}]$ and not $[0, 1]$ as the task requested.

Similarly, the Transformation strategy was used predominantly by students in the written group (successfully in most cases). This was surprising as we anticipated that the STACK interface would encourage students to transform examples by moving points on the grid.

Chains of examples There were 32 instances of chains of examples. These were more common among the e-assessment group (18 out of 32, compared with 14 out of 44 for the written group). Only six of the 32 chains were coded correct, while 16 chains ended with a correct example only after prompts by the interviewer. This low number of correct instances is not surprising: a chain typically begins when the student's first example is not correct (we did observe one case where the student spontaneously changed their own correct first example).

Table 3 Distribution of example-generation strategies and outcomes for instances when a single example was generated

	Trial and error		Transformation		Synthesis	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Written	7	2	6	1	8	6
E-assessment	5	4	1	2	2	

The most common way for a chain to end was with a final example obtained through the Transformation strategy, with 22 occurrences out of the 32 chains. In four of these cases, students generated a correct example at the end of that chain without prompts, while in a further 13 cases, the final example was correct but produced only after prompts by the interviewer (such as in Fig. 6). The remaining five cases ended with an incorrect solution.

Another chain of strategies adopted in four instances consisted of a series of T&E attempts, with the generated examples being quite disconnected from each other. All four instances were in the e-assessment group, and only one of the attempts was eventually successful (following prompting from the interviewer). This repeated use of T&E may have indicated that the student was confused as to what the exercise required.

We notice that nine out of the 32 sequences started with a Synthesis strategy. This strategy was implemented in these cases with incorrect premises, and may indicate that the student did not have a clear grasp of the definitions involved in the task. For instance, Philip consulted the definition of injectivity and tried to interpret it graphically, but confused the roles of x and y (which led to drawing the S-shaped first example in Fig. 7).

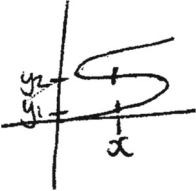
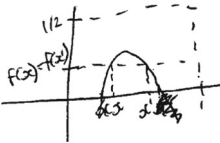
Philip		
Task 2: Draw the graph of a function $f_2: [0, 1] \rightarrow [0, 1]$ that has image $[0, \frac{1}{2}]$ and is not injective		
Transcript with time stamp (I=Interviewer, S=student)	Example Generated	Analytical summaries
<p>[4:08] <i>(Starts off looking at definition of injectivity; seems to get muddled with x's and y's.)</i></p> <p>[4:41] S: I couldn't have a graph that had multiple... x... multiple... y values with different values of x.</p> <p>[4:59] S: So I couldn't have something like this <i>[draws S shape]</i> because then I would have multiple y values for the same value of x here. I'd have y_1 and y_2 are the same <i>[annotates these points]</i>.</p> <p>[5:17] S: No, wait, hang on. No, y is the same. <i>[draws new set of axes]</i>. So I suppose this one here <i>[indicates example of $y=x/2$ from Task 1]</i> is injective, is 1-1, the first graph.</p> <p><i>[conversation with interviewer about the properties of the Task 1 example]</i></p> <p>[6:20] I: OK, so the one that you've drawn for the first task was injective [S: yes] but for this task we're looking for one that's <i>not</i> injective.</p> <p>[6:27] S: Ah, right. <i>[Student re-reads the question and consults the definitions again]</i></p> <p>[6:38] S: So I have $f(x)$ equalling $f(x')$ but I have x not equalling x' <i>[while writing this down]</i></p> <p>[6:50] S: So I have $f(x)$ and $f(x')$ here; x here and x' here <i>[annotates these on the axes]</i></p> <p>[7:10] S: So, we have something like that <i>[draws parabola shape]</i>, in the interval 1 and a half. So $f(x)$ equals $f(x')$, and that is x and that is x'. And that one's not injective.</p> <p>[7:35] I: OK, so that's the "not injective" part. Are you happy with everything else?</p> <p>[7:44] S: So we're still in the same interval, and we've still got the same image. So as long as I keep under a half <i>[adds dotted line at $y=1/2$ and annotates it]</i> and 1 <i>[adds dotted line at $x=1$]</i>, I think I'm still complying with the rest.</p>	<p>Synthesis, incorrect</p>  <p>Unprompted validation</p> <p>Synthesis, incorrect</p>  <p>Unprompted validation of "not injective".</p> <p>Prompted to consider other properties; student says they are satisfied</p>	<p>The first example here is not a function, but the student corrects themselves quickly.</p> <p>The student reasons based on the previous example, but returns to the definition when producing the next one.</p> <p>The student seems to view the domain and image requirements as constraints that don't necessarily have to be met exactly ("as long as I keep under a half").</p>

Fig. 7 Analysis of Philip's attempt at Task 2

3.3.3 Difficulties with example-generation tasks

The second round of analysis identified a number of difficulties that students encountered in solving the tasks. Many of the attempts failed because the student did not attend to all of the requirements of the task. For example, for task 3, students were supposed to find a function (first requirement) from $[0, 1]$ to $[0, 1]$ (second requirement regarding the domain and codomain of the function) which was surjective (third requirement), and not injective (fourth requirement). Some students attended to only one or two of the requirements, as we can see in the case of Sally’s answer to Task 2 (Fig. 8).

Another common difficulty was with notation; for instance, many students confused the interval notation $[0, 1]$ with the notation for a point $(0, 1)$ or had difficulties in coordinating the abstract representations of functions that appeared in the lecture notes (for the material students were given during the interview see osf.io/rp7a3). Finally, some students misunderstood what the task asked for, for example looking for injective functions when the tasks asked for non-injective functions. The last type of difficulty is a logic one: three students equated non-injective to surjective (or non-surjective to injective). We observe however that none of these difficulties could be predominantly seen in one for the online or written groups, with all students but one showing some of these difficulties some of the time.

This last stage of the analysis confirms that the two groups of students displayed similar difficulties with the tasks, so the marked difference in achievement observed in both studies may come from characteristics of the task format. Indeed, the only difficulty that was unique to one group concerned the use of the e-assessment interface. While some students said that they had used this type of graphical interface before, for others it took some time and experimentation before they could see how to move the points around to create piecewise linear functions.

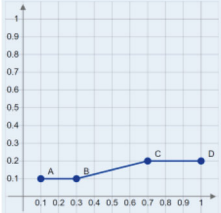
Sally		
Task 2: Draw the graph of a function $f_2: [0, 1] \rightarrow [0, 1]$ that has image $[0, \frac{1}{2}]$ and is not injective		
Transcript with time stamp (I=Interviewer, S=student)	Example Generated	Analytical summaries
<p>[7:30] S: it is a non injective. Injective – is that every different value of x has a different value of y? I: Yeah, and this needs to be <i>not</i> injective.</p> <p>[7:43] S: Yeah, so a value of x... [moves points A, B, C, D into position]</p> <p>[7:52] S: maybe like this? I: Yeah, so let’s see all the... the function needs to go from $[0, 1]$ to $[0, 1]$ with image $[0, \frac{1}{2}]$, and it needs to be not injective.</p> <p>[8:05] S: Yeah, so... if values of x are 0.1 and 0.3, we have the same value of y which is 0.1 [I: yeah] and similarly the case for C and D.</p>	<p>Trial and error, incorrect</p>  <p>Prompted validation</p>	<p>When prompted to validate, the student focuses only on the condition of non-injective</p>

Fig. 8 Analysis of Sally’s attempt at Task 2. In her final answer, Sally has attended to the non-injective requirement but not to the image requirement which was supposed to be $[0, \frac{1}{2}]$

3.4 Discussion

The results of our task-based interviews with 19 students were consistent with the pattern we observed in Study 1: students achieved much more success with the written versions of the tasks compared with the e-assessment versions. This study was also designed to examine reasons behind the difference in student performance between the two versions, in terms of students' use of example-generation strategies (RQ1) and the difficulties they encountered with the tasks (RQ2).

Regarding RQ1, we found that while our participants did use the Trial and Error and Transformation strategies, they did not use the Analysis strategy in the same way as the experts in Antonini's (2006) study. Instead, we observed students using a strategy that we named Synthesis. The absence of Antonini's (2006) Analysis strategy may be related to the well-documented difficulty that students have with abstraction. One of the coping mechanisms observed in previous research is that students reduce the level of abstraction of problems in order to be able to make sense of them (Hazzan, 1999). Employing the Analysis strategy entails assuming that an abstract object with the required properties exists, and then deducing more properties until the actual object is found. This level of abstraction may just be too much for students starting their mathematics degree. Our findings showed that the example-generation strategies of experts did not transfer exactly to describing novices' approaches. A key contribution from our study is the identification of a new strategy — Synthesis — used extensively by students, that may help us to understand novices' example generation.

For our analysis of students' example-generation strategies, we distinguished between cases where students generated a single example or a chain of examples. For the single-example cases, we found that the Trial and Error and Synthesis strategies were used a similar amount (18 and 16 times respectively), while in previous studies, the Trial and Error strategy was predominant (e.g., Iannone et al., 2011; Antonini, 2011). The frequent use of the Synthesis strategy that we observed may indicate that the students' example spaces were not yet sufficiently rich to support Trial and Error, or that they were not yet familiar with the definitions and their implications. For the cases where students generated a chain of examples, we noted that these were most common among the e-assessment group and only rarely successful. The low success rate is perhaps explained by the widely observed phenomenon that students find it difficult to review their own work (e.g., Selden and Selden, 2003; Kontorovich, 2019).

RQ2 concerned the factors affecting student performance in tasks, and we observed that the students' difficulties aligned with existing literature. The students had difficulties with notation (see also Selden, 2012), they often did not attend to all requirements of the task, and in some cases they drew incorrect logical inferences (such as equating non-injective to surjective, see also Weber, 2001). However, these difficulties were observed in both groups, making it unlikely that they alone explain the difference in outcomes.

We note three aspects of the design of the study that could affect comparisons between the two groups. First, we did not seek to ensure the groups were balanced in terms of prior mathematics attainment, so it could be that the differences we observed were due to the makeup of the groups. However, we had no selection criteria for the groups, and we observed similar difficulties related to the mathematics concepts involved in the task between the two groups. Second, the interviews for the two groups were conducted by different interviewers, and it could be that our interviewing technique was so different that it yielded different outcomes. We believe this is unlikely as the interviewers designed the interviews together and great care was taken to achieve a consistent interviewing style, although, of course, it would be impossible to guarantee that no difference between interviews was present. Third,

the lower performance of the e-assessment group could be due to difficulties with using the interface, as noted in the discussion of Study 1. However, the students had time during the interviews to familiarise themselves with the interface, and all students managed to produce examples.

We suggest two more plausible accounts for the differences in performance between the two groups in this study. First, as noted in the discussion of Study 1, the written group was able to use familiar examples (e.g., parabolas) while the e-assessment group was forced to operate with a restricted example space of piecewise linear functions. Students were perhaps unable to make the transfer from a known example to an example with the same characteristics but represented by a piecewise linear function. A second explanation for the differences in performance lies in the types of example-generation strategies used by the two groups. We found that the e-assessment group made much less use of the Synthesis strategy than the written group. This could reflect a general tendency for students to engage with e-assessment tasks in a different way to written work. Previous research has found that undergraduate students often engage passively with e-assessment tasks, relying on the software to check answers (Dorko, 2020); as one student explained, “When you do it on paper you do it more properly” (Rønning, 2017, p. 101).

4 General discussion

Across two studies, we investigated how students responded to the same sequence of example-generation prompts, in either written or e-assessment format. Our main findings were that students were much more successful with the written version of the tasks than the e-assessment version and that they employed different strategies in each case.

Our studies had some limitations concerning their longitudinal nature and the timing of the data collection. However, the similar pattern of results across both studies increases our confidence that the difference in outcomes is best explained by the different modalities of the tasks. A further limitation is that our studies were based on a single-task sequence. The nature of the task can influence the example-generation strategies that may be employed; for instance, we designed Tasks 2 and 3 to invite Transformation. It may also be that our tasks were too simple to prompt the use of the Analysis strategy. Future research could investigate different topics and tasks, to explore how the design of tasks affects the range of strategies employed to solve them.

Our findings about students’ example-generation strategies raise a question for undergraduate teaching: to what extent should students learn the Analysis strategy? Iannone et al. (2011) suggested students’ reliance on Trial and Error might explain the unexpectedly low learning gains from generating examples; students may only benefit when employing more expert-like strategies. On the other hand, this could be an instance where students lack “the knowledge, experience, or ability to engage in the activities that are productive for mathematicians” (Weber et al., 2020, p. 10), and therefore encouraging students to use other approaches (such as the Synthesis strategy that we identified) may be more worthwhile.

It seems that the promise of example-generation tasks as a way to help students learn has yet to be fully realised. Closer consideration of the way that students approach these tasks (across a range of topics) may help to unlock that potential. Our findings also demonstrate that the transfer of paper-based example-generation tasks to e-assessment is not neutral, with students adopting different strategies in each case. Further work is needed to explore how e-assessment can be used most effectively to enrich students’ example spaces.

Appendix: Further examples of student work

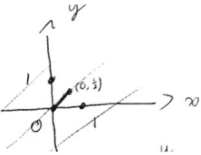
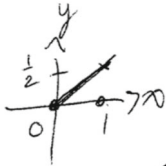
Porsia		
Task 1: Draw the graph of a function $f_1: [0, 1] \rightarrow [0, 1/2]$ with image $[0, 1/2]$		
Transcript with time stamp (I=Interviewer, S=student)	Example Generated	Analytical summaries
<p>[2:15] (Student reads the task and draws the Cartesian axes; spends some time trying to recall the definition of image and consulting slides from the lectures).</p> <p>[4:43] S: OK, so of course it's going to be smooth, however, so... (draws the segment in the first picture) it will kind of, just, end at $1/2$ however. (Student asks if the function must be linear; interviewer notes the task only asks for "a function").</p> <p>S: OK, let's go for a nice linear function, it's quite easy. And so, yeah, because the image... yes because the image is... I guess you could then contract the scales a bit, just so it ends at $1/2$, but yes, it would kind of end at the point $(0, 1/2)$ (annotates the point $(0, 1/2)$).</p> <p>[5:29] I: Ok so do you want to talk me through all the parts of the question ...explain why you are happy with that.</p> <p>S: Ok so we are given this function and we are given that it has domain $[0, 1]$ and we know that maps to this subset of $[0, 1]$ erm... oh my, I did it the other way [...] so a function that maps $[0, 1]$ to codomain $[0, 1]$ but our image is $[0, 1/2]$... so at least I think of the image as [...] what has been realised in our graph.</p> <p>[6:33] S: Like, for example when I am thinking of a function like x^2 (writes $f(x)=x^2$ and sketches its graph) so our codomain can be all the real numbers but our image is the positive reals, so the image is what's actually being realised by the graph.</p> <p>[6:55] I: Ok so can you point that out to me in your graph? Where you're seeing that?</p>	<p>Trial and Error, incorrect</p> 	<p>Student draws a graph straight away – after some thinking they consult the notes and slides.</p> <p>The first confusion is that the student thinks of a smooth (continuous?) function.</p>
<p>[7:00] S: Where I'm seeing that? OK... I have to change our graph. [...] So actually, maybe what would be better... is, instead of having this weird half-interval – if I may start over again...</p> <p>[7:26] S: (crosses out the first graph – draws new Cartesian axes and labels the x- and y-axes) this is 0, this is 1, this will be $1/2$. Then I think something like this (draws the second graph of the straight line) makes actually more sense.</p> <p>[7:36] S: Because then you actually... what I even forgot is, every function must map all of its domain, otherwise you don't include stuff in the domain. So this (pointing at the first example) is already wrong, I can see, because (pointing at 1 on the x-axis in the first example) what does 1 map to? Right, so? Here 1 actually maps to something. And that is $1/2$. And we have 0 maps to something 0. Eh, yeah. I think that this one is right.</p>	<p>The interviewer prompts a check by asking the student to check the definition against the graph produced.</p> <p>Transformation previous graph</p>  <p>Correct example, but only following a prompt from the interviewer</p>	<p>After the conversation about images, the student changes their mind and finds a better example, one that fulfils the requirements of the task.</p> <p>Issue with the domain in the previous example, only noticed after attending to the image.</p>

Fig. 9 Analysis of Porsia's attempt at Task 1, which included a chain of two examples being produced. The second example was produced after prompting from the interviewer, so this attempt was assigned a score of 0 based on the example that Portia generated before prompting from the interviewer

Data availability Supporting data and code can be found at <https://doi.org/10.17605/OSF.IO/XDU3H>.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Antonini, S. (2006). Graduate students' processes in generating examples of mathematical objects. In Novotná, J., Moraová, H., Krátká, M., & Stehliková, N. (Eds.), *Proceedings of the 30th conference of the international group for the psychology of mathematics education* (pp. 57–64). Prague: Charles University. Retrieved from <https://files.eric.ed.gov/fulltext/ED496932.pdf#page=65>
- Antonini, S. (2011). Generating examples: Focus on processes. *ZDM - Mathematics Education*, 43(2), 205–217. <https://doi.org/10.1007/S11858-011-0317-6>
- Bansilal, S., Brijlall, D., & Trigueros, M. (2017). An APOS study on pre-service teachers' understanding of injections and surjections. *The Journal of Mathematical Behavior*, 48, 22–37. <https://doi.org/10.1016/j.jmathb.2017.08.002>
- Dorko, A. (2020). Red X's and green checks: A model of how students engage with online homework. *International Journal of Research in Undergraduate Mathematics Education*, 6(3), 446–474. <https://doi.org/10.1007/s40753-020-00113-w>
- Edwards, A., & Alcock, L. (2010). How do undergraduate students navigate their example spaces? Proceedings of the 13th Conference on Research in Undergraduate Mathematics Education. Retrieved from <http://sigmaa.maa.org/rume/crume2010/Archive/Edwards.pdf>
- Fahlgren, M., & Brunström, M. (2023). Designing example-generating tasks for a technology-rich mathematical environment. *International Journal of Mathematical Education in Science and Technology*, 1–17. <https://doi.org/10.1080/0020739X.2023.2255188>
- Fukawa-Connelly, T. P., & Newton, C. (2014). Analyzing the teaching of advanced mathematics courses via the enacted example space. *Educational Studies in Mathematics*, 87, 323–349.
- Goldenberg, P., & Mason, J. (2008). Shedding light on and with example spaces. *Educational Studies in Mathematics*, 69(2), 183–194. <https://doi.org/10.1007/s10649-008-9143-3>
- Hazzan, O. (1999). Reducing abstraction level when learning abstract algebra concepts. *Educational Studies in Mathematics*, 40, 71–90.
- Hohensee, C. (2006). Students' thinking about domains of piecewise functions. Proc. 28th annual meeting of the North American Chapter of the Int. Group for the Psychology of Mathematics Education, 2, 586–593.
- Hooper, C., & Jones, I. (2023). Conceptual Statistical Assessment Using JSXGraph. *International Journal of Emerging Technologies in Learning*, 18(1), 269–278. <https://doi.org/10.3991/ijet.v18i01.36529>
- Iannone, P., Inglis, M., Mejía-Ramos, J. P., Simpson, A., & Weber, K. (2011). Does generating examples aid proof production? *Educational Studies in Mathematics*, 77(1), 1–14. <https://doi.org/10.1007/s10649-011-9299-0>
- Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment*, 4(5), n5.
- Kinnear, G., Iannone, P., & Davies, B. (2023). Insights about functions from examplegeneration tasks: Combining e-assessment and written responses. In Drijvers, P., Csapodi, C., Palmér, H., Gosztonyi, K., & Kónya, E. (Eds.), *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13)* (pp. 2399–2406). Alfréd Rényi Institute of Mathematics and ERME.

- Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., & Wong, T. (2022). A collaboratively-derived research agenda for e-assessment in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education*. <https://doi.org/10.1007/s40753-022-00189-6>
- Kontorovich, I. (2019). Why do students not check their solutions to mathematical problems? A field-based hypothesis on epistemological status. *International Journal of Mathematical Education in Science and Technology*, 50(7), 1050–1062. <https://doi.org/10.1080/0020739X.2019.1650304>
- Kontorovich, I., & Li, T. (2022). Not as straightforward as it appears: Undergraduates leverage areas to find definite integrals. *International Journal of Science and Mathematics Education*. <https://doi.org/10.1007/s10763-022-10339-6>
- Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, 30(3), 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Lemmo, A. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education*, 19(8), 1655–1675. <https://doi.org/10.1007/s10763-020-10119-0>
- Lew, K., & Zazkis, D. (2019). Undergraduate mathematics students' at-home exploration of a prove-or-disprove task. *The Journal of Mathematical Behavior*, 54, 100674. <https://doi.org/10.1016/J.JMATHB.2018.09.003>
- Liebeck, M. (2018). A concise introduction to pure mathematics (Fourth Edition). Chapman and Hall/CRC.
- Maher, C. A., & Sigley, R. (2020). Task-based interviews in mathematics education. *Encyclopedia of mathematics education*, 821–824. https://doi.org/10.1007/978-3-030-15789-0_147
- Marton, F., Tsui, A. B., Chik, P. P., Ko, P. Y., & Lo, M. L. (2004). *Classroom discourse and the space of learning*. Routledge.
- Mirin, A. (2022). Where we see one function, they see two. *International Journal of Mathematical Education in Science and Technology*, 1–29. <https://doi.org/10.1080/0020739X.2022.2123285>
- Mok, I. A. C., & Clarke, D. J. (2015). The contemporary importance of triangulation in a post-positivist world: Examples from the learner's perspective study. In *Approaches to qualitative research in mathematics education* (pp. 403–425). https://doi.org/10.1007/978-94-017-9181-6_15
- Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- Popper, P., & Yerushalmy, M. (2021). Online example-based assessment as a resource for teaching about quadrilaterals. *Educational Studies in Mathematics*, 110(1), 83–100. <https://doi.org/10.1007/s10649-021-10109-1>
- Ripley, M. (2009). Transformational computer-based testing. The transition to computerbased assessment, 92.
- Rønning, F. (2017). Influence of computer-aided assessment on ways of working with mathematics. *Teaching Mathematics and its Applications: An International Journal of the IMA*, 36(2), 94–107. <https://doi.org/10.1093/teamat/hrx001>
- Runeson, U. (2006). What is it possible to learn? On variation as a necessary condition for learning. *Scandinavian journal of educational research*, 50(4), 397–410.
- Sağlam, Y., & Dost, Ş. (2016). A qualitative research on example generation capabilities of university students. *International Journal of Science and Mathematics Education*, 14(5), 979–996. <https://doi.org/10.1007/s10763-015-9624-7>
- Saldaña, J. (2021). The coding manual for qualitative researchers. The coding manual for qualitative researchers, 1–440.
- Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford University Press.
- Selden, A. (2012). Transitions and proof and proving at tertiary level. Proof and proving in mathematics education: The 19th ICMI study, 391–420.
- Selden, A., & Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, 34(1), 4–36. <https://doi.org/10.2307/30034698>
- Thoma, A., & Nardi, E. (2019). Recall and substantiation routines in exam scripts: Injective and surjective functions. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Eleventh Congress of the European Society for Research in Mathematics Education*. (Vol. TWG14). Utrecht, Netherlands: Utrecht University.
- 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. <https://doi.org/10.1007/s10649-006-9078-5>
- Uscanga, R., & Cook, J. P. (2022). Analyzing the structure of the non-examples in the instructional example space for function in abstract algebra. *International Journal of Research in Undergraduate Mathematics Education*, 1–27. <https://doi.org/10.1007/s40753-022-00166-z>

- Vinner, S. (1983). Concept definition, concept image and the notion of function. *International Journal of Mathematical Education in Science and Technology*, 14(3), 293–305. <https://doi.org/10.1080/0020739830140305>
- Watson, A., & Mason, J. (2005). *Mathematics as a constructive activity*. <https://doi.org/10.4324/9781410613714>
- Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational studies in mathematics*, 48, 101–119.
- Weber, K., Dawkins, P., & Mejía-Ramos, J. P. (2020). The relationship between mathematical practice and mathematics pedagogy in mathematics education research. *ZDM - Mathematics Education*, 1, 3. <https://doi.org/10.1007/s11858-020-01173-7>
- Yerushalmy, M., Nagari-Haddif, G., & Olsher, S. (2017). Design of tasks for online assessment that supports understanding of students' conceptions. *ZDM - Mathematics Education*, 49(5), 701–716. <https://doi.org/10.1007/s11858-017-0871-7>
- Zazkis, R., & Leikin, R. (2007). Generating examples: From pedagogical tool to a research tool. *For the Learning of Mathematics*, 27(2), 15–21.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.