




A Department-Wide Transition to a New Mode of Computer-Aided Assessment Using STACK

Ben Davies¹  · Cosette Crisan² · Eirini Geraniou² · Teresa Smart²

Accepted: 22 July 2024
© The Author(s) 2024

Abstract

We focus on the integration of STACK—a Computer-Aided Assessment (CAA) technology—in the mathematics department of a high-ranking University in the United Kingdom. We study a department-wide project where instructors were expected to implement STACK into continuous assessment tasks for (nearly) all core modules across the first two years of undergraduate study. We present this work as a departmental case study, drawing on semi-structured interviews with six novice STACK assessment designers (and module leaders), supplemented by students' responses to an open-response feedback questionnaire, and the reflections of a co-project lead (also first author). Our thematic analysis identified four themes related to the design of STACK-based assessments by novice to STACK tutors: the process of 'STACKification', technical challenges, users' perspectives on the role of CAA, and finally, variations in assessment designers' approaches to the role of feedback. In presenting our results, we are guided by Sangwin's (2013) design principles for mathematics assessment. We consider various technical aspects of implementing STACK-based assessments as a first-time user, and the knowledge required to do so effectively and coherently. We conclude with a series of reflections on the role of CAA in undergraduate mathematics, and the ways in which such technology can be productively integrated with established practice.

Keywords Computer-aided assessment · STACK · Instructional design · Formative assessment · Undergraduate mathematics

✉ Ben Davies
ben.davies@soton.ac.uk

¹ University of Southampton, School of Education, Southampton, UK

² IOE, UCL's Faculty of Education and Society, University College London, London, UK

This paper examines the implementation of Computer-Aided Assessment (CAA) within the Department of Mathematics at a high-ranking university in the United Kingdom. We present a case study centred on a department-wide initiative where lecturers were tasked with using STACK (a System for Teaching and Assessment using a Computer algebra Kernel) for the majority of coursework. This initiative involved transitioning from the traditional method of students submitting handwritten solutions to using the STACK online platform, which automates assessment and provides predetermined, bespoke feedback coded by a combination of module leaders and graduate teaching assistants (TAs).

The Covid-19 pandemic significantly accelerated the shift to online education, including tertiary education. However, at least in this context, we understand this acceleration to have merely expedited changes that were already coming. While our research is positioned to have broader relevance beyond the context of a global health crisis, it is important to acknowledge that the data collection for this study occurred during this period. The Department of Mathematics had been considering CAA for several years prior to this study. Nonetheless, the immediate transition away from traditional handwritten assessments was largely driven by resource constraints during a difficult period, and the urgent need for remote, contactless instruction.

Our aims in this paper are two-fold. First, we seek to develop our theoretical understanding of these assessments, and the design process used to produce them in this context. Second, our research sought to contribute to the department-wide effort to develop assessment materials that will be offered to students in future iterations of the relevant modules by analysing lecturers' and students' reflections on STACK's use for and impact on mathematical learning. We view these two aims as reciprocal, in so far as the gain in theoretical knowledge is both derived from, and eventually, valuable in the development process of CAA. While we view our findings as salient for the field more broadly, we must acknowledge the context-bound nature of this work.

Our study presents findings from semi-structured interviews conducted with a group of four lecturers and two graduate TAs, specifically employed to assist in the design and integration of STACK-based assessment. Notably, all participants had been engaged in the project for less than a year, and none possessed prior experience with STACK or any other form of CAA. The first author of this paper played a key role in co-leading the department-wide STACK project, but did not take part in the data collection or primary data analysis to aid the credibility and trustworthiness of our findings.

Before presenting our findings, we review the current literature on CAAs and their increasing use in undergraduate mathematics education. We then focus on CAA's role as a formative assessment tool, followed by an overview of the framework guiding our analysis. Our subsequent research questions focus on the views and approaches of mathematics instructors who are new to CAA, and more specifically, their approaches to the design and implementation of automated feedback.

Undergraduate Mathematics and the Increasing Use of Computer-Aided Assessment

We investigate the use of CAA in tertiary mathematics and its integration with other modes of assessment across a degree programme. We note that assessment in undergraduate mathematics degrees has remained focused on closed-book written examinations (Iannone & Simpson, 2011, 2022), largely ignoring advancements in assessment methods. This leads us to question whether CAA can challenge these long-established assessment modes and embrace the potential of formative assessment.

In recent years, STACK has emerged as a prominent CAA tool for evaluating tertiary-level mathematics (Fahlgren et al., 2021). That said, while content-specific studies (e.g., Kontorovich and Locke (2023), on a particular task within elementary integral calculus) are starting to emerge, we note an absence of research addressing the implementation of STACK, or other CAAs, at scale in the tertiary mathematics education literature.

STACK utilises a computer algebra system based on open-source Maxima, to assess both numeric and algebraic aspects of students' answers. STACK can serve for summative assessment purposes, but it has great potential to support formative assessment providing tutors with evidence of student understanding and providing feedback to advance their learning (Fahlgren et al., 2021). A comprehensive exploration of STACK's capabilities is provided in Sangwin (2013).

STACK is used by educational institutions, businesses, and developers all over the world with at least 2000 active sites (https://moodle.org/plugins/qtype_stack/stats, June 4, 2024) in over 15 countries (www.stack-assessment.org, Sept 13, 2021). At present, STACK can either run as an integrated plug-in for Virtual Learning Environments (VLEs) including Moodle and ILIAS, or as a stand-alone Application Programming Interface (API) to be used in smaller, bespoke settings.

Recent advancements in STACK include the development of a fully integrated online module for introductory university mathematics (Gratwick et al., 2020; Kinnear, 2019) and an examination of task design tailored for proof-based mathematics (Bickerton & Sangwin, 2021). Kinnear (2019) notes the time- and resource-intensive process required to fully integrate the technology, but from preliminary results, concludes that these investments were worthwhile for both instructor and student. Bickerton and Sangwin (2021), on the other hand, focused on higher-level concepts associated with proof and argumentation. These authors provided a suite of design suggestions for proof comprehension tasks using STACK, including faded worked examples, reading comprehension activities and example generation tasks.

The Role of CAA in *Formative* Assessment in Undergraduate Mathematics

To understand the possible roles CAA might play, we draw on the seminal work of Black and Wiliam (2010) on formative assessment and the role of Assessment for Learning (AfL). In particular, Black and William defined assessment as:

...all those activities undertaken by teachers — and by their students in assessing themselves — that provide information to be used as feedback to modify

teaching and learning activities... such assessment becomes (our emphasis) formative assessment when the evidence is actually used to adapt the teaching to meet student needs (p. 82).

We wish to highlight the absence of discussion about the credit-bearing role of assessment here. We identify a recent, unproductive simplification of the assessment discourse in tertiary education that identifies summative assessment with credit-bearing tasks (perhaps the more reasonable simplification) and formative assessment with all non-credit-bearing/low-stakes tasks. This simplification overlooks the active intentionality required to unearth the formative capacity of any such task.

In the case of CAA, it seems that feedback has a primary role to play and must be considered an integral part of the design process (Bearman et al., 2022).

A Framework for Evaluating Computer-Aided Assessment

Our research pays particular attention to the role of CAA in the wider pedagogic design of tertiary mathematics courses. This topic has received increasing attention in recent years, with some researchers starting to explore the potential for high-stakes summative assessments to be administered using STACK (Sangwin, 2023; Sangwin & Köcher, 2016). Earlier work, like Kinnear (2019), focused on low- or no-stakes formative tasks in ancillary modules (those outside the core undergraduate mathematics degree pathway).

Design Principles for Mathematics Assessment

In order to address the issue of what would make a mathematics assessment, Sangwin (2013) presented five design principles pivotal in the development of high-quality CAA. These principles were derived from more than two decades of professional experience as a mathematics educator, and leading expert in CAA.

We have adopted this framework in light of its seminal contribution to the conceptualisation and development of contemporary CAA practice (e.g., Olsher et al., 2024; Fahlgren & Brunström, 2023; Kloosterman & Warren, 2014). The principles offered in this work help us to understand the actions of novice STACK developers and provide us with an opportunity to reflect on their salience as potential training materials for future assessment designers using this software. For clarity, the research we present here was not interventionist at this stage. The research team did not provide any formal training to participants, and to the best of our knowledge, no participant in the project was aware of Sangwin's design principles for the duration of the project.

Before presenting his five principles, Sangwin (2013) first argued that there should be an overarching principle, which states that "Assessment should reflect mathematical practice" (p.25). Then, he continued with the following:

"Principle 1 Mathematicians try to solve problems. [...]

Principle 2 Standard algorithms are both useful and important mathematical cultural artifacts in their own right. [...]

Principle 3 Mathematicians justify their solutions. The outcome of mathematics is a correct chain of reasoning, from agreed hypotheses to a conclusion. [...]

Principle 4 Accuracy is important. [...]

Principle 5 It is important to acknowledge the place of conventions which should be distinguished from arbitrary definitions or logical consequences.” (Sangwin, 2013, pp.25–27).

Regarding Principle 1, Sangwin (2013) discussed the importance of not just computations and producing a correct answer, but also reasoning. He argued that “if an assessment of students is to reflect practice, then problem-solving must be a key part of the instrument” (p.25).

For Principle 2, Sangwin (2013) argued the importance of assessing students’ understanding of when to apply any standard algorithm as well as an understanding of how the algorithm works while using it accurately and efficiently (see page 26 for further explanations).

Sangwin (2013) emphasised the significance of evaluating students’ capacity to establish coherent connections between their solution steps through reasoning, a concept he outlined as Principle 3.

Concerning Principle 4, Sangwin (2013) discussed the importance of accepting the numerous mathematical conventions and ensuring there is accuracy when using mathematical language and notation.

Finally, in relation to Principle 5, Sangwin (2013) argued that we should look for problems that allow students to showcase their deductive reasoning, but also accurate mathematical work. Students should be able to use and apply “routine mathematical techniques” relying on “mathematicians’ conventions” (p.27). So, we should use traditional word problems as they “possess many of the features of mathematical practice” and “can be used at many levels in school and at university” (p.27).

In the data we present, the CAA at stake is implemented in STACK and is used as credit-bearing assessment with the potential to play both summative and formative roles in different settings/contexts. We explore the extent to which these five principles were evident in the context we studied.

Research Questions

To further our twin aims of developing theoretical understanding for CAA design, and improvements to the particular assessment materials in use, we pursue the following research questions:

RQ1: What are novice CAA instructors’ views and approaches to implementing CAA in tertiary mathematics?

RQ2: In particular, how do these instructors approach the design and implementation of automated feedback?

These questions reflect the highly specific context from which our data are derived and the opportunistic essence of this naturalistic case study in which we seek to understand our phenomenon of interest in its real-world setting.

Methods

To answer our questions, we draw on a naturalistic case study methodology (Cohen et al., 2017) focusing on a department-wide role out of a new CAA tool across the majority of all undergraduate mathematics programmes in this university. To this end, we draw on three data sources: semi-structured interviews with six first-time users of CAA, feedback from 445 end-of-semester Student Evaluation Questionnaires (SEQs), and the personal reflections of the project co-lead and first author. Four of our six interviewees led at least one undergraduate mathematics module at the time of data collection, and we draw primarily on their experiences in these modules. The first author was also a co-lead of this department-wide project, so many reflections are made at the level of the wider context, described in more detail below.

Context

The Covid-19 pandemic brought about an urgency in changes to policy and practice rarely found in British universities. So, alongside the numerous hardships, traumas and downsides, a small collection of silver linings can be found. The urgent need to transition to a combination of online, remote and hybrid instruction practices created unprecedented opportunities for large-scale innovation with CAA.

In contrast to previous projects centred on examining the practice and consequences of redeveloping one particular module (e.g., Kinnear, 2019), this project saw the simultaneous rollout of untested STACK-based quizzes in 20+ modules across an entire department. These quizzes were summative, for-credit assessments designed as direct replacements for traditional pen-and-paper homework exercises.

Alongside the pandemic, this decision was also partially motivated by financial considerations. While the upfront investment was substantial (two full-time faculty and 10+ Teaching Assistants across 2 years), it is hoped that this investment will be returned several times over in reduced marking costs in future years. In the discussion, we explore the consequences of this resource-driven motivation and its relationship with the (future) role of CAA in the department.

Typically, such projects are developed and implemented by passionate advocates for the change that is being promoted. This department-wide project is a notable departure from this norm, in that enthusiasm varied throughout the wider group of those engaged with the implementation of STACK-based assessments. To our knowledge, this aspect of our context is (nearly) unique for CAA research in higher education. This, and other consequences of the rapid, near wholesale change for the first two years of undergraduate study are explored later in the paper.

Participants

Four lecturers (referred to as L1 – L4) and two postgraduate teaching assistants (TA1 and TA2) participated in semi-structured interviews with two members of the research team. TA1 and TA2 were members of a larger design team including two full-time faculty and six graduate students employed at different times throughout the year. Each lecturer was the leader of at least one undergraduate module and was respon-

sible for overseeing the design of their own assessments. The extent and enthusiasm with which lecturers engaged with the design team varied substantially.

We report on their first-time experiences with CAA, as they were required to automate (significant proportions of) all coursework to reduce marking workloads. In some cases, this was near-direct translations of existing tasks. In others, substantial work was required to create workable tasks in the new environment. We note that support was offered in the form of online training and assistance from two experienced STACK developers.

Materials and Procedures

Online interviews were conducted by two members of the research team. To aid the credibility and trustworthiness of the interview data, the first author was excluded from data collection as department-wide project co-leader.

These interviews were divided into two segments. Initially, participants responded to a set of questions concerning their experiences with designing and implementing assessments using STACK. Interviewers also explored the dynamics within the design team, the process of adapting existing items into CAAs using STACK, participants' satisfaction levels with their current repository of STACK-based tasks, and their aspirations for enhancing future iterations of STACK assessments. The second segment of the interview involved a stimulated reflection task. One week prior to the interview, participants were requested to identify their preferred and least preferred tasks they had contributed to. Subsequently, interviewers posed questions about each task, aiming to uncover insights into the perceived strengths and weaknesses of CAAs in general.

We also report feedback in the form of 445 responses to the Student Evaluation Questionnaires (SEQs) from eight modules across the first and second years of the undergraduate mathematics programme. SEQs were centrally administered and anonymised so it was not possible to ascertain exactly how many individuals are represented in this sample. Given the interpretivist nature of our case study, we only include responses to the open-ended 'general comments' box in the analysis presented below. Extracts from these responses are integrated into our thematic analysis.

Data Analysis

Our primary data analysis is that of the interviews conducted with four module leaders and two graduate teaching assistants who supported the wider project. These data were analysed alongside the open-text responses to the Student Evaluation Questionnaire. The latter are used to as ancillary data to provide additional context and colour to the primary data source.

The two data sources were analysed in concert, using thematic analysis (Braun & Clarke, 2006). We followed the standard, six-stage protocol, using an inductive approach to identify latent themes within the interview data. These themes are illustrated using interview extracts, and supported, in places, using brief extracts from students' written SEQ responses.

In phase one of data analysis, a member of the research team watched each interview multiple times, tidying the imperfect automated transcripts in real-time. The interviews necessarily covered descriptions of complex mathematical solutions which led to complications with the automated transcriptions.

In phase two, an inductive approach was used to ‘code interesting features ... across the entire data set’ (p. 87).

Phase three led to the series of latent themes, presented below, alongside a series of supporting extracts.

In phase four, these were iteratively reviewed by the wider research team and adapted through several passes through the data. A preliminary report was then produced, highlighting four overarching themes with supporting excerpts and commentary for review by other members of the research team.

Phases five and six saw all authors then contributing to a drafting and redrafting process leading to the final themes presented in the analysis section to follow. Analytic conversations during Phase six – ‘producing the report’ (*ibid.*) – further highlighted the role of Sangwin’s design principles. Therefore, our analysis included a discussion on how Sangwin’s design principles influenced CAA.

Consistent with our naturalistic case study approach, we present our thematic analysis blended with contextualising reflections from the project lead, and extracts from the Student Evaluation Questionnaire.

Four Emergent Themes

We identified four themes related to the design of STACK-based assessments by first-time users: (1) *the process of STACKification*, (2) *technical challenges with coding in STACK*, (3) *the role of CAA in undergraduate mathematics*, and (4) *the role of feedback*. Within each theme, we provide commentary on Sangwin’s design principles for mathematics assessment discussed earlier.

Theme 1: The Process for STACKification

As is typical in mathematics departments, module leads are responsible for curating a series of problem sheets for continuous assessment throughout the semester. All four lecturers in our study appeared to organically follow a process when transforming their existing materials into CAAs utilising STACK. We describe this process in four phases, with reference to Sangwin’s design principles.

Phase One Lecturers would review their existing question banks to determine which ones they deemed suitable for CAA. This often entailed identifying questions with straightforward or limited input requirements and those that could be programmed without extensive technical knowledge. One lecturer (L4) highlighted the necessity for effort in identifying appropriate items, stating, “*you cannot simply take an exercise sheet and immediately turn it into STACK. It requires some effort [to identify a suitable item].*” (cited in Davies et al., 2022, p. 2368). When selecting items for CAA, only one lecturer explicitly (L3) prioritised the inclusion of items critical for comprehensive coverage of content and techniques from the module. One lecturer

(L1) had taught the same two modules over a number of years and was confident that the homework sheets covered all areas of the module and could be translated into Stack. L1 prioritised those questions where the input numbers could be randomised to minimise academic integrity concerns. However, this individual felt that the process of randomisation to ensure that students had questions “*that were similar in level of difficulty in terms of the calculations*” often needed a lot of extra work in both developing and trialling/error-checking.

Others appeared to prioritise simply what could be most naturally assessed by the tool, leaving the remaining material for other modes of assessment. There is a tension here between the affordances of the CAA tool, and the desire for assessment to reflect authentic mathematical practice. We interpret this tension as a reference to Sangwin’s Principle 1, and the desire for problem-solving to play a key role in all mathematical assessment.

In many instances, the mathematical substance of existing questions remained unchanged, but adjustments were made to the required student responses to accommodate the STACK platform. One of the teaching assistants (TA2) said “*I think all of the questions that I worked on were based on problem sheets that they had used in previous years; some were kind of word for word. Some had to be adapted*”. The lecturer for the Vector Calculus course (L3) said “*As this is a methods course if they can do the method correctly, they will get to the right answer, and so it’s reasonably well aligned with what stack can do.*” In the Introductory Analysis course, the lecturer provided a concise set of proofs, proof methods, and techniques deemed essential to assess student learning. Converting the proofs and techniques into Stack questions was challenging “*The analysis questions had to be completely rewritten*” (TA2). We interpret this as an expression of Sangwin’s Principle 2, regarding standard algorithms (and techniques) as central to the assessment practices of an undergraduate mathematics module.

As STACK currently lacks the capability to assess student-generated proofs, the design team posed a set of reading comprehension tasks similar to those suggested by Bickerton and Sangwin (2021) to evaluate students’ understanding without employing a ‘prove that’-style task. Additionally, in certain cases, a set of multiple-choice items akin to those presented in Mejía-Ramos et al. (2017) proved to be suitable alternatives.

It is interesting to note that, while some students commented that the quizzes were easier than past examples, at least some students found the quizzes “*much harder than the examples and past papers*”. This diversity of perspective is to be expected in a large cohort, but it is also worthy of attention that varying the format of the exercises is likely to shuffle the perceived difficulty/value of the exercises among that cohort. On the whole, the response was positive. “*The Moodle quizzes, much like the course, were difficult but manageable if you took careful note of the lecture content. Some of the later quizzes were especially difficult and felt a bit out of line with the lecture content. However, beyond one or two of them, I felt as if the quizzes were challenging applications of the lecture content*” (student feedback).

Phase Two In collaboration with the design team, ‘preSTACKed’ documents were produced for the majority of items. These documents were predominantly formatted in LaTeX and resembled pseudocode outlining the design features that a future coder would need to implement. They included details such as the expected student inputs, the range and arrangement of random variables, and the precise wording of questions to be presented to students. In certain instances, this preSTACKing process was less formalised, consisting solely of a list of questions to be coded.

Of the four phases, this was the most uniform across modules, seemingly because of the support available from the design team. We view this as an expression of Sangwin’s Principle 4, highlighting the importance of mathematical accuracy across assessments. As noted in phase 1, it was not possible to assess many aspects of proof construction directly. So, following Bickerton and Sangwin (2021), proof constructive tasks were often replaced with a series of fill-in-the-blank style items in such a way that students could observe and respect the syntax and norms of mathematical proof-writing without being required to produce the entire proof. This approach has its limitations, discussed later, but allowed for an entire canon of assessment items to be STACKified that would otherwise have been left out.

Phase Three Among the four lecturers, three posted these preSTACKed documents on a shared workflow tracker for the design team to address. In contrast, L1 primarily handled their own coding tasks, seeking assistance from others only when encountering nuances or techniques beyond their expertise.

Phase Four Following the initial coding phase, lecturers were invited to assess each item and encouraged to verify the code for its intended functionality. Due to the relative inexperience of the design team, numerous items exhibited bugs in early iterations. Some of these bugs resulted in the incorrect marking of variations on correct answers (e.g., an answer such as $4/2$ being marked incorrect when the intended solution was the integer 2), and vice versa. While STACK provides extensive control to users regarding the assessment of such variations and can accommodate the majority of desired responses in each case, the prevalence of bugs of this nature during the project’s early stages, compounded by the rapid pace of item production, posed significant challenges for lecturers and students alike.

To some extent, we view these teething problems as an inevitable product of any time-constrained software rollout. However, with reference to Sangwin’s design principles, these repeated violations of Principle 4 (‘accuracy is important’) appeared to do meaningful damage to students’ attitudes toward, and faith in, their continuous assessment experience. This issue was particularly pronounced in one module, where at least one student observed that “quizzes almost always had errors in them”. Several others reiterated similar concerns, stating that the quizzes contained numerous errors and questions on material not yet covered, making it challenging to stay engaged.

Effective, regular communication between the lecturer and the coding team regarding the presentation of STACK quizzes to students proved to be crucial. L3

highlighted instances where discrepancies arose between the solutions entered and the notation typically taught, prompting adjustments. This involved addressing minor formatting issues and debugging. L3 would often attempt the questions beforehand, identifying, and rectifying errors pre-emptively. However, occasionally issues only surfaced when encountered by students, necessitating live debugging sessions.

It is interesting to reflect on this process in the context of Sangwin's (2013) design principles for mathematics assessment. In particular, it is interesting to note the attempts by all participants to enact a combination of Principles 1 and 2 through the process of identifying (phase one) and pseudo-coding (phase two) problems most appropriate for the format. It is clear that preserving, where possible, the problem-solving aspects from the original tasks (Principle 1) was, for some participants, a central part of the process of STACKification. Similarly, through attempts to focus attention on the most important aspects of the curriculum, we understand that Principle 2 (on the importance of standard algorithms) is also enacted here.

We expect a decline in these challenges over time, however, we emphasise their significance in this context due to their impact on attitudes toward the value of technology with respect to automated assessment.

Theme 2: Challenges in Early Implementations of New STACK Materials

Lecturers primarily directed their attention toward evaluating procedural tasks that involved students entering numeric or basic algebraic expressions. STACK possesses the capability to accommodate a diverse range of question formats, catering to various understandings and approaches. However, for our participants, implementing anything beyond numeric or algebraic equivalence tests posed a significant challenge in numerous instances. By way of a very particular example, L1 observed that when the solution to the problem included surds, STACK encountered no issues if the square root was in the numerator. However, challenges arose when the square root was in the denominator and students rationalised it. In such instances, STACK "*could not recognise this as a correct answer*". Similarly, students expressed frustrations in their feedback, with comments reflecting on instances where the system flagged their answers as incorrect: "*Was my approach incorrect? Or was my calculation incorrect? Or did I enter the answer incorrectly?*"

In reality, the answer was 'no' to all three of this student's questions, but rather, the software was not set-up in such a way that it could accurately treat algebraic variations on the expected answer. We interpret this as an explicit violation of Sangwin's accuracy principle (Principle 4), as well as an implicit violation of Principle 5, regarding the role of arbitrary definitions and logical consequences. In particular, for the student who enters a correct answer and receives an incorrect response, they may now perceive an arbitrary distinction between the CAA-endorsed distinction and their own, potentially causing further problems for this individual down the line. While this is technically a coding error on behalf of the assessment design team, our experience suggests that this is not uncommon and that only the most experienced STACK coders consistently avoid such pitfalls in the first iterations of item development.

L3 observed that when students were tasked with inputting formulas, they encountered a challenge where a single error, even a minor one like a typo, would result in all their answers being cleared. TA1 emphasised the importance of instructing students on correct formula input methods in STACK, including the insertion of Greek letters like lambda and theta, and terms with subscripts such as x_0 . A student also expressed a desire to be explicitly taught about typing certain symbols, using TeX:

I think it would be slightly easier for me if how certain symbols were typed (e.g., summation symbol sigma, inequality sign) using TeX were taught before the formal lecture contents began.

In contrast, L3 also pointed out a particularly successful instance where the coders initially faced challenges due to the existence of multiple correct answers. However, they ultimately devised an innovative coding solution to determine the correctness of the student's solution. TA1 also acknowledged that “*you kind of have to think a bit more about all the possible answers that the students could give you*” (cited in Davies et al., 2022, p. 2370) recognising that while a solution may exist, it might not always be implementable in certain scenarios. While STACK acknowledges algebraic equivalence, students can express solutions to differential equations in various forms, requiring consideration of the multitude of possible answers.

L3 emphasised the importance of ongoing professional development for both lecturers and coders to mitigate the issues discussed in this section. Nonetheless, even experienced coders encounter challenges in this domain. For those contemplating the adoption of STACK in the future, it seems necessary to establish a robust peer-review process both prior to and following implementation with student participation.

Returning to Sangwin's principles, we note evidence here regarding the importance of accuracy (Principle 4), both from the perspective of the student and the assessment designer. While Sangwin's original text focused on accuracy on behalf of the student, it is interesting to reflect on the shared responsibility for accuracy here, and the implications that technical errors can have on the students' experience of, and attitudes toward, CAAs.

Theme 3: The Role of CAA in Different Content Domains

All four lecturers initially used problem sheets that had been used as homework and assessment tasks in their previous teaching. While they acknowledged STACK's suitability for handling examples necessitating numerical or basic algebraic responses, they hesitated to explore its potential for assessing more conceptual aspects of their module curriculum. For instance, TA1 remarked on the straightforwardness of assessing calculus, “*because it involved fairly straightforward kind of mathematical methods, so we had weekly quizzes for that*” (cited in Davies et al., 2022, p. 2370). However, they emphasised the importance of clearly defined answers. Similarly, L4 expressed reservations, noting that problems involving the input of formulas could present challenges “*because formulas can be written in slightly different ways and sometimes it doesn't recognise these things as the same.*”

While questions necessitating numerical or algebraic responses could generally be adapted for STACK, assessing conceptual knowledge and proofs posed greater challenges. L1 suggested utilising a conventional Moodle quiz format for proofs, incorporating sophisticated multiple-choice questions. Similarly, L2 indicated that *“Not all [examples] were suitable because some of the questions involve some theorem or some proving which possibly could be STACKed or, if you like, but I couldn’t see a way to do that, so I concentrated on questions with numerical answers.”* (cited in Davies et al., 2022, p. 2370). Similar to excerpts from Theme 1, we interpret this as an expression of Sangwin’s first principle and the need for assessment to reflect mathematical practice.

In one module, converting the example sheets created disproportionately large challenges for both the lecturer and the coders. The lecturer worked with 2 graduate students to translate the assessment examples into STACK. In some cases, it was determined that exercises could not productively be STACKified, leaving gaps in the content coverage of the homework. It was deemed impossible to assess a proof in STACK, or to assess if the student has a robust understanding of given definitions from the course. In some cases, this was reluctantly accomplished using a series of multiple-choice questions, but it was made clear that this was not the lecturer’s preference. L4 did not accept that CAA could test student understanding. *“Maybe in 50 years’ time, artificial intelligence will develop to such a level that the computer will be able to check whether the student understands the definition of a continuous function, yes. But the moment it’s not at this stage”*. At least one student disagreed, writing *“The [quizzes] of this module actually help us understand proofs, which is quite nice and the fill in the blanks style of proofs was effective, ensuring that you knew which lemma/theorems to apply where”*. This corroborates the claims from earlier about the importance of emphasising justification and correct chains of reasoning (Sangwin’s third principle). Again, while it was not possible to use explicit proof construction tasks, it was possible for assessment designers to target and highlight the logical structure of mathematical proofs using the CAA software at hand.

Echoing concerns similar to those regarding proof-based questions, L4 questioned how a straightforward numerical or algebraic response in STACK could effectively demonstrate students’ understanding of the theory and methods they had been taught. *“[In my course] it’s not a matter of manipulating formula like in school, right? It’s a matter of showing that you understand what’s going on and it’s somehow difficult to transform it into computer-based assessment”* (cited in Davies et al., 2022, p. 2370). Moreover, L4 raised doubts about the suitability of STACK *“at a serious university... In a very good math department, you have to show that you understand, then you have to write, and explain”* (cited in Davies et al., 2022, p. 2370). However, L4 acknowledged that STACK may be more suitable for an ancillary course [for non-math majors], *“but still, it’s somehow lame [sic], even for chemists”* (cited in Davies et al., 2022, p. 2370). In contrast, L2 believed that with careful question formulation, students would need to grasp the methods and theory to arrive at the correct answer, a view also shared by L3.

These excerpts indicate significant variability in the applicability and value of STACK across different parts of the undergraduate mathematics curriculum. In this paper, we purposefully refrain from evaluating the viewpoints expressed by L4 and

others. Instead, we opt to solely present the perspectives offered by our participants and contemplate ways to enhance our provision for students in future iterations of these courses. Bickerton and Sangwin (2021) have put forth a set of alternatives for STACK-based proof assessments aimed at tackling several of the concerns raised by L4. We recognise that implementing these alternatives can be time-consuming and may not be feasible in every scenario. However, we suspect that none of our four lecturers were familiar with this recent research, and we plan to organise professional development workshops in the future. Through these workshops, we aim to expand the variety of tasks available to our students and enhance the range of conceptual understanding assessed through our STACK-based assessments.

Returning again to Sangwin (2013), it is interesting to note the challenge presented to Principle 3: Mathematicians justify their solution. This is particularly acute in the context of proof-based mathematics modules, where some participants felt that the CAA fell short of offering students the opportunity to justify their reasoning or argumentation, instead resorting to adjacent tasks in which such content could only be assessed indirectly.

Theme 4: The Role of Feedback

For most questions, the STACK feedback was intended to develop students' fluency in solving problems. In most cases, a full worked solution was provided as general feedback printed in response to all inputs. Specific hints were also given in many cases where student input was not correct. Resource constraints meant that bespoke formative feedback in the form of hints, or corrections based on the students' input, were not possible in many cases. For items with several sub-questions, the feedback was usually given after each part, one at a time, giving students an opportunity to learn from their mistakes and not get repeatedly punished for 'follow-through' errors. This type of feedback was welcomed by the students, who described it as constructive, facilitating learning while working through the questions and feedback. We interpret this adaption for follow-through errors to be a strength in CAA, in alignment with Principles 3 and 4. Regarding Principle 3, by acknowledging and explicitly rewarding solution sequences that contain minor errors but are internally consistent, this assessment design reinforces the value of a chain of reasoning from one step (or sub-question) to the next. When coupled with Principle 4 (regarding accuracy), this strength of CAA comes to the fore in contrast to traditional 'by hand' marking, in which it is often easy to miss the inherent value and internal consistency of solutions that do not yield a correct answer.

In quizzes, when making mistakes or being stuck, the students were able to attempt similar but different questions, as L1 explained that they coded STACK to "*give [the students] three tries without showing them the full solution*" before releasing the full solution. Another approach which was also thought beneficial to the students was to provide the full solution after each try, "*what I really like is for the system [STACK] to give them a different version of the question when they try again.*" (L3). All tutors considered that struggling to solve a problem and addressing mistakes were important skills for students to develop, and hence they expressed cautiousness about the

effectiveness of hints provided early on and too often, e.g., “if I give them too many hints and I’m not sure if they keep doing and keep repeating and trying.” (L1).

When STACK quizzes were used as a summative assessment task, no feedback was provided, and a worked solution would only be released once all the students completed the quiz and/or the submission deadline had passed. Lack of feedback immediately after the assessment task was a source of frustration for some students; particularly for those who had become accustomed to immediate feedback in other contexts: “They wouldn’t get the solution until after the deadline [...] so the full feedback didn’t appear until after the last student handed it in. And if anybody had an extension, that could be a whole week after it was supposed to be due, or maybe even a fortnight.” (L3). The pedagogic decision to delay feedback was left to individual lecturers and practice differed across our four participating module leads and throughout the department. We conjecture that this inconsistency may have been as much a source of the frustration as the delay itself.

Lecturers also commented on losing insight into how the cohort performed on the problem sheets in contrast with the handwritten homework tasks from previous years. Normally, lecturers would collate the feedback and comments from the marking tutors on what aspects of the taught topics students appeared to have difficulties with: “so I would ask the tutors if there’s any common problems ... then I can feed back to the class or design some extra problems to do in live classes to illustrate some common areas of problems.” (L2). This practice was not seen as achievable with STACK. As a result, lecturers felt that their teaching did not change in order to respond to problems that students encountered. We note that the Moodle-generated response logs have the potential to provide an adequate, if not dramatically improved, replacement for the collation of handwritten homework. Later, we return to this lost opportunity for these tasks to “become formative assessments” (in the sense of Black and William, 2010).

Finally, while all lecturers and coders were fully aware of the importance of students receiving detailed, ‘good quality’ feedback, they soon realised the amount of time needed to invest in coding STACK to this effect. One lecturer (L2) who tried to do so said in the interview said: “I mean that would be excellent to say and that well you’ve made a little error here you’ve differentiated this incorrectly, but that seems a very difficult task to me to do via STACK”, especially if the question has a number of steps in its solution where students get stuck. L2 did wonder if the difficulty arose from their lack of knowledge of all of STACK capabilities “maybe STACK can do [this]”. Ideally for tutors, the feedback would signal to the students if they were on the ‘wrong avenue’ in solving a problem, but they were not sure yet if this was even possible in STACK, as L2 reckoned: “I don’t really know how you do that on the STACK but there’s probably some way of doing something”.

Importantly, all lecturers interviewed shared that they had very little time available to plan for automated assessments and, in particular, to prepare what feedback was given to students. They recognised that this will require further work in the future. In L1’s words, “We did put some effort into the feedback, maybe not as much as we would have done if we’d have more time.” Using the feedback function in STACK is definitely a challenge for the future: “I should definitely start doing more feedback

so if a student gets a specific wrong answer, it's more tailored to what the student answers" (L1).

Returning one last time to Sangwin's principles, we repeat similar observations from earlier themes. In particular, first, we note the centrality of the problem-solving aspect of assessment design (Principle 1), and the importance of feedback in supporting this ambition. Second, we note the role of justification (Principle 3) and the ways in which participants used feedback to emphasise the importance of argumentation in mathematics, even if this was not always possible to assess directly, as discussed under Theme Three.

Discussion

In this section, we first address our two guiding research questions, before zooming out to the wider consequences of our work, and some recommendations for future work.

RQ1: What are novice CAA instructors' views and approaches to implementing CAA in tertiary mathematics?

To answer this question, we return to the four themes emerging from our interview data.

Theme one explored the process of 'STACKification', where conventional hand-written coursework tasks were transformed into CAAs through STACK. Despite potential opportunities for enhancement, it is noteworthy how uniformly this process was adopted by all four lecturers. Subsequent iterations of these courses are expected to involve iterative refinements of numerous items, incorporating new functionalities, addressing bugs, and providing more detailed feedback.

Theme two pertained to a primary challenge for first-time users of STACK associated with evaluating algebraic equivalence in various forms. In several cases, lecturers and members of the design team were aware that an alternative coding solution likely existed but could not execute a solution within the time constraints. Again, these concerns will likely diminish with time, and as a department, there is now an ongoing opportunity to revisit those items that did not function as expected.

Theme three reflected lecturers' perspectives on the role of CAA in tertiary mathematics more generally. All four lecturers acknowledged that STACK had the potential to contribute to at least some proportion of the undergraduate curriculum. However, these were heavily weighted toward applied mathematics, and to more procedural (rather than conceptual) tasks. Of particular note was L4's belief in the inability to assess mathematical proof using STACK or other forms of CAA. It is unclear from the data available whether these perspectives will change with time or further professional development, perhaps focused on the potential for STACK to assess a wider array of question formats. While all four lecturers acknowledged the potential of these resources in some capacity, several comments indicated discomfort with assigning course credit to students' responses. That is, several seemed uncom-

fortable with their role as summative assessments in their module. We return to this theme of formative and summative assessment later.

Theme four highlights that providing feedback specific to the problem *and* student input would help turn the STACK quizzes into a learning as well as an assessment tool. However, this relies on the lecturer using their knowledge of the type of mistakes students make using the coders to identify the mistakes that are made and code the appropriate feedback. It is not clear how often this is achievable given the current resource constraints. Perhaps future learning analytics research and/or integration with Large Language Models will prove fruitful in this area.

RQ2: In particular, how do these instructors approach the design and implementation of automated feedback?

To answer this research question, we used Sangwin's (2013) five design principles. We observed that the first three of Sangwin's five design principles were consistently observed by at least some of our participants. With respect to Principle 1, we found extracts across themes one and two referencing the importance of problem solving in assessment design.

Regarding Principle 2, we found evidence of awareness of the importance of standard algorithms and techniques in themes two and four.

Principle 3, on the importance of justification and logical reasoning, was raised in two themes as well. First, it came up in the context of the technical challenges associated with coding in STACK (theme 2). While participants demonstrated an awareness of the importance of forefronting logical justification, most expressed frustrations at the limitations of the tool in this regard. This principle was also raised in the context of feedback (theme 4) and the need to use creative (albeit potentially limited) formats to highlight the role of justification and chains of reasoning through the assessments offered.

More interestingly, perhaps, was the absence of alignment with Principles 4 and 5. Principle 4, regarding accuracy, was raised primarily in themes 1 and 2, regarding the barriers to accurate coding. It was noted that errors in the back-end coding of CAAs have potential knock-on effects for students' beliefs about and attitudes toward assessment tasks. It seems plausible that there could be further knock-on effects here by implicitly communicating to students that accuracy is not a priority for those leading the module. This was not explored in detail in our findings but could be the domain of future work in this area.

Further on Principle 4, Sangwin (2013) discussed the importance of giving students opportunities to practice and demonstrate accuracy when using mathematical language and notation. This is particularly pertinent for proof-based mathematics in which the norms of communication are often highly specific. However, the affordance of the tool seemed to limit such opportunities as the majority of proof-related questions were written in formats like 'fill-in-the-blanks', in which the convention-following aspects are primarily done for, rather than by, the student.

Finally, and somewhat similarly, evidence of alignment with Principle 5 was completely absent from our data. That is, we saw no explicit acknowledgement of the role and place of conventions in mathematical communication, and the role of sometimes

arbitrary definitions, particularly in proof-based mathematics. It is likely that this is a consequence of the affordances of the tool at hand, as argued in the previous paragraph.

Having explored the specifics of novice instructors' views and approach to CAA, and their alignment with Sangwin's (2013) design principles, we now zoom out to offer some wider reflections from the case study as a whole.

Reflections on the Role of Professional Development for CAA

Based on our interviews and our own reflections on the project, it seems that the role of CAA, and of STACK in particular, requires further examination. Many of the resources in this project were initially conceived of as replacements for long-standing coursework tasks, i.e., low-stakes summative assessments. Based on our research work so far, we learned that when conceptualised as formative resources to provide students with opportunities for (immediate) feedback and reflection on their own learning, these CAA resources have unbounded potential. However, when conceptualised as rigorous credit-bearing summative assessments, their value is less clear, and the weaknesses of the software become more prominent. We, therefore, argue that further empirical evidence is needed to effectively embed STACK in summative assessments for tertiary mathematics.

We also note an absence of evidence that lecturers are benefitting from these materials as a means to understand their students and/or tailor their future instruction. This was explicitly mentioned by at least one participant who seemed unaware that the software has the capacity to provide the exact formative feedback they identified as lost from their previous practice. This is a key benefit of formative assessment, as highlighted by Black and Wiliam (2010), and it seems eminently possible given the log files available.

Of course, like all new software, there are always implementation challenges. Theme two from the interview analysis speaks directly to this challenge and highlights the need for time and resources to be invested upfront, both in terms of task development and professional support for colleagues making the transition to a new practice. Sangwin (2013) spoke of the importance of 'accuracy' in assessment practice (Principle 4). It is conventional to think of accuracy in the context of student output, but reflections on this research question and the feedback offered by students remind us of the importance of accuracy from the designers' vantage point as well.

Further, beyond the necessary technical training, we have observed a need for practitioners to develop a clear vision for the role of their CAA in the immediate and long-term future. We now expect that the explicit use of Sangwin's (2013) design principles can add value in this regard, particularly for graduate teaching assistants who are likely coming to assessment design, of any type, for the first time.

This conjecture about the role of Sangwin's (2013) design principles is an empirical claim, and one that will require future researchers to investigate before more definite claims can be made.

Final Remarks and Future Work

This project offered a novel insight into how STACK has been used and can be used to support low-stake summative assessments across a high-ranking undergraduate mathematics programme. We have also explored the extent to which Sangwin's (2013) design principles were enacted by first-time users in this particular content and reflected on their potential value for novice users.

As the popularity of this CAA continues to expand into tertiary settings, we hope that future researchers will attend to the complex needs for professional development and investment in sustainable practice.

These resources require maintenance and ongoing improvements, as well as institutional knowledge on how best to integrate them into practice, use them effectively, and address the difficulties raised in training and recruiting colleagues. In our view, the keys to success for department-wide projects like ours lie in the investments in professional development, and ideally, in permanent full-time faculty with substantial responsibilities dedicated to CAA.

From a more insular perspective, we know that the project from which we derived our case study has a sustainable future, with the question bank continuing to evolve through ongoing use in 15+ modules. Even though this work was inspired by, and in fact launched during the Covid-19 pandemic, there is real value in supporting formative, as well as summative CAAs in tertiary education. We recognise that STACKification is not trivial. Not all 'traditional' resources make good CAA items (and vice versa) and this STACKification process can be challenging for novices. However, the assumption that it is possible to simply translate existing resources into a CAA format is one that at least warrants reflection and poses a worthy but rewarding challenge.

Despite the complexity and the challenges, CAA has the potential to make a substantial contribution to undergraduate mathematics education. To harness the potential, the research community must attend to the professional development required for supporting lecturers in their professional practice, supporting students and lecturers to reflect on teaching and learning through genuinely formative assessment practice, and in supporting the often resource-constrained assessment practices of a university department. We have explored the role and consequences of one particular department-wide implementation for low-stake summative assessment, but urge the research community to pursue further robust, large-scale, empirical investigations exploring the role and consequences of CAAs in other contexts and toward other goals.

Data availability To protect the anonymity of participants in this study, supporting data is not publicly available.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

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

- Bearman, M., Nieminen, J. H., & Ajjawi, R. (2022). Designing assessment in a digital world: An organising framework. *Assessment and Evaluation in Higher Education*, 48(3), 291–304. <https://doi.org/10.1080/02602938.2022.2069674>
- Bickerton, R., & Sangwin, C. (2021). Practical Online Assessment of Mathematical Proof. *International Journal of Mathematical Education in Science and Technology*, 53(10), 2637–2660. <https://doi.org/10.1080/0020739X.2021.1896813>
- Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 92(1), 81–90. <https://doi.org/10.1177/003172171009200119>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Cohen, L., Manion, L., & Morrison, K. (2017). *Research Methods in Education (8th ed.)*. Routledge. <https://doi.org/10.4324/9781315456539>
- Davies, B., Smart, T., Geraniou, E., & Crisan, C. (2022). STACKification: automating assessments in tertiary mathematics. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)*, 2365–2373. Free University of Bozen-Bolzano and ERME. <https://hal.science/hal-03750584>
- 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>
- Fahlgren, M., Brunström, M., Dilling, F., Bjarnheiður, K., Pinkernell, G., & Weigand, H. (2021). Technology-rich assessment in mathematics. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalová, & H. Weigand (Eds.), *Mathematics Education in the Digital Age* (pp. 69–83). Routledge. <https://doi.org/10.4324/9781003137580-5>
- Gratwick, R., Kinnear, G., & Wood, A. K. (2020). An online course promoting wider access to university mathematics. In R. Marks (Ed.), *Proceedings of the British Society for Research into Learning Mathematics*, Vol. 40. <https://bsrlm.org.uk/wp-content/uploads/2020/05/BSRLM-CP-40-1-04.pdf>
- Iannone, P., & Simpson, A. (2011). The summative assessment diet: How we assess in mathematics degrees. *Teaching Mathematics and its Applications: An International Journal of the IMA*, 30(4), 186–196. <https://doi.org/10.1093/teamat/hrr017>
- Iannone, P., & Simpson, A. (2022). How we assess mathematics degrees: The summative assessment diet a decade on. *Teaching Mathematics and its Applications: An International Journal of the IMA*, 41(1), 22–31. <https://doi.org/10.1093/teamat/hrab007>
- Kinnear, G. (2019). Delivering an online course using STACK. *Contributions to the 1st International STACK Conference 2018*. Friedrich-Alexander-Universität Erlangen-Nürnberg. <https://doi.org/10.5281/zenodo.2565969>
- Kloosterman, P., & Warren, T. L. (2014). Can technology help in mathematical assessments? A review of computer aided assessment of mathematics. *Journal for Research in Mathematics Education*, 45(4), 534–537. <https://doi.org/10.5951/jresmetheduc.45.4.0534>
- Kontorovich, I., & Locke, K. (2023). The Area enclosed by a function is not always the definite integral: Relearning through collaborative transitioning within a Learning-Support Module. *Digital Experiences in Mathematics Education*, 9(2), 255–282. <https://doi.org/10.1007/s40751-022-00116-z>
- Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, 19(2), 130–146. <https://doi.org/10.1080/14794802.2017.1325776>

- Olsher, S., Chazan, D., Drijvers, P., Sangwin, C., & Yerushalmy, M. (2024). Digital Assessment and the machine. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of Digital Resources in Mathematics Education*. Springer. Springer International Handbooks of Education https://doi.org/10.1007/978-3-031-45667-1_44
- Sangwin, C. (2013). *Computer-aided assessment of mathematics*. Oxford University Press.
- Sangwin, C. (2023). Running an online Mathematics Examination with STACK. *International Journal of Emerging Technologies in Learning*, 18(3), 192–200. <https://doi.org/10.3991/ijet.v18i03.35789>
- Sangwin, C., & Köcher, N. (2016). Automation of mathematics examinations. *Computers and Education*, 94, 215–227. <https://doi.org/10.1016/j.compedu.2015.11.014>

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