Providing individual student feedback at scale for mathematical disciplines

Stanyon, Robert; Tomlinson, Austin; Kainth, Manjinder; Wilkin, Nicola

DOI:
10.1145/3491140.3528313

Document Version
Peer reviewed version

Citation for published version (Harvard):
Stanyon, R, Tomlinson, A, Kainth, M & Wilkin, N 2022, Providing individual student feedback at scale for mathematical disciplines. in *L@S 2022 - Proceedings of the 9th ACM Conference on Learning @ Scale: Proceedings of the Ninth ACM Conference on Learning @ Scale.* vol. 2022, L@S 2022 - Proceedings of the 9th ACM Conference on Learning @ Scale, Association for Computing Machinery (ACM), New York, pp. 400-404, L@S '22, New York, New York, United States, 1/06/22. https://doi.org/10.1145/3491140.3528313

Providing Individual Student Feedback at Scale for Mathematical Disciplines

Robert Stanyon
rjs217@bham.ac.uk
Theoretical Physics Research Group
School of Physics and Astronomy
University of Birmingham
Birmingham, UK

Austin A. Tomlinson
a.a.tomlinson@bham.ac.uk
School of Metallurgy and Materials
University of Birmingham
Birmingham, UK

Manjinder Kainth
manjinder@6bit.co.uk
6 Bit Education Ltd
Birmingham, UK

Nicola K. Wilkin
n.k.wilkin@bham.ac.uk
Theoretical Physics Research Group
School of Physics and Astronomy
University of Birmingham
Birmingham, UK

ABSTRACT

This WIP discusses the preliminary results of a paid-for-pilot, at the University of Birmingham, of a new assessment and feedback platform — Graide. Graide uses machine learning and AI to assist educators in the grading process. It has been shown to increase both the detail of feedback for individual students and consistency of feedback across the cohort. Graide enables increased oversight of the assessment process whilst providing opportunities for continuous training of markers, whilst also reducing the time taken to grade work by up to 89%.

ACM Reference Format:
Robert Stanyon, Austin A. Tomlinson, Manjinder Kainth, and Nicola K. Wilkin . 2022. Providing Individual Student Feedback at Scale for Mathematical Disciplines. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3491140.3528313

1 INTRODUCTION

In order to be assured of progress in a discipline, a learner needs to measure their understanding. Particularly in disciplines where the learning is hierarchical, as is the case in mathematical disciplines, timely and personalised feedback is key to the students' progress and confidence in the discipline.

Hence, learning at scale will be most effective if the assessment processes scale accordingly. The optimal solution from a student's perspective [4] of rapid, individual, and detailed feedback clearly cannot be delivered with existing systems as the ratio of teacher to student moves into the hundreds. As we enable learning for all, particularly for financially impoverished education systems or learners who do not have a peer-study group, or access to office hours, we are left in a position where excellent educational material is available but without the teacher feedback, meaning students are unable to achieve as well as their better funded, equivalently-able peers.

Further, extreme teaching ratios, as are enabled with learning at scale, result in the teacher being less immediately aware of the progress of the student cohorts. Aiming for the middle of an 'in session' cohort's progress can be achieved through quick-fire quizzes. However, one relies on the assumption that the current cohort on a programme is the same in terms of their prior learning, and technical expertise. Holes in student learning and technical expertise are a particular issue after the disruptions in education caused by the pandemic. They have not been seen before in the university setting within which this work has been undertaken. Continuous assessment and feedback is used to both identify and address these issues, but doing this at scale poses a significant burden in time and resources.

Automated grading solutions already exist and can operate at scale. The most basic and highly widespread solution is multiple choice assessment[1]. It benefits from automatic response and if written well with good distractors, one can deduce the types of challenges a student is encountering in their understanding [7]. The major drawback with this is that students are presented with the correct answer which poses similar problems to those associated with simultaneous and sequential line-ups for eyewitnesses of crimes: 'Are Suggestiveness-Induced Hits and Guesses True Hits?' [14], where the psychology of recall with and without a reference is tested. By analogy, we are concerned that, at the level of understanding, critical thinking is not necessarily measured by successfully answering a multiple choice question [11]. A significant concern is that by being unable to reward 'method' or 'technique' marks in a multiple choice format, one does not reward and therefore incentivise students who have made a final mistake (often algebraic) in reaching their final decision on which item to select. Additionally, there is extensive evidence that multiple choice testing of reversible

L@S '22, June 1–3, 2022, New York City, NY, USA.

ACM ISBN 978-1-4503-9158-0/22/06…$15.00
https://doi.org/10.1145/3491140.3528313

[1] a challenge: find an online form that does not rely on a multiple choice at some point!

processes (common in mathematical disciplines) leads to students verifying the answers rather than directly solving the problem, often not testing the correct method [10].

The alternatives therefore are Computer Aided Assessment (CAA) platforms, which have been in development since the 1960s. Two of the most frequently used systems today are STACK[9] and Möbius[2]. Both enable more detailed mathematical understanding to be assessed, but with a cost in terms of preparing the students (and educators) to use the system. In particular, they require that questions are modified to be deliverable by the computer system.

STACK has a "reasoning by equivalence engine" which enables line-by-line comparison of mathematical working. This requires the educator to anticipate the steps (and errors) the student will make and is sensitive to small errors in the syntax that students use in their answer.

Möbius [6] is a digital assessment platform with a "grading code" feature which enables educators to randomise their questions so students can repeat assignments with a different set of questions, enhancing the learning. Randomised questions need to be carefully programmed and students need to learn how to use symbolic input to access the assessments.

Gradescope [5] is an aid to assessment which has a slick interface which offers significant improvements over traditional marking systems. It also is capable of basic grouping of identical responses so they can be marked in bulk.

To date, we are able to use the above systems to reduce the time required for grading and to scale to meet the demands of large cohorts. These systems fall into the **2nd GEN** category, as described by Bennett in 1998 [1]. Bennett classified CAA platforms into three different "generations", based on how the user interacted with them:

1st GEN platforms enable direct conversion between a traditional assessment into an electronic format. For example, a written assignment that requires students to submit a word processed essay.

2nd GEN platforms blend digital learning and assessment by utilising digital multimedia as part of the process.

3rd GEN platforms would be radically different and utilise cutting-edge technologies to aid learning an assessment in novel and unexplored ways.

STACK and Möbius were created as **2nd GEN** systems without the foundation of a well built **1st GEN** platform. As a result significant pedagogical changes are required when using these systems. This creates a barrier to entry due to the changing of the medium and the pedagogy of the assessment at the same time.

This WIP describes a new **1st GEN** system: Graide [3], the concept and algorithms of which form the basis of Stanyon's thesis [12][2]. Conceptually, Graide is an assistant to the grader, rather than attempting to replace the grader. It is now in paid-for pilot at the University of XXXXXXX. We report here early results from the study and how the system mitigates the pedagogical hurdles and short-falls of previously existing systems.

In the rest of this WIP we discuss issues with current assessment feedback processes, what the downstream consequences of these processes are, and demonstrating how Graide addresses these issues.

The key takeaways from this report are that Graide has:

(1) Made the grading and feedback process faster for teachers,
(2) Increased the quality of feedback for students and made it more consistent,
(3) Improved visibility of the assessment and feedback process, allowing for continuous training of teaching assistants which did not exist before.

2 CURRENT ASSESSMENT AND FEEDBACK PROCESSES AND THEIR CONSEQUENCES

There are three main issues with current assessment and feedback processes: they are fragmented, lack oversight, and take a significant amount of time at scale.

The first issue concerns the fragmentation of assessment and feedback tools. The main stages of delivering feedback are:

(1) Creating an assignment.
(2) Delivering that assignment to students.
(3) Students attempting and submitting their assignment.
(4) Educators grading and giving feedback to those assignments.
(5) Students receiving that feedback.

Issues arise because these stages are often managed in different systems. For example, many lecturers create assignments in word processors or LaTeX editors and print them out for students. Students then attempt the questions on a few sheets of paper and submit them in collection boxes, which are sorted and distributed (often to multiple graders), graded, collated for data entry, and finally returned back to students. The entire process is reliant on each stage being efficient and can break down if one or more encounters difficulty.

Where multiple graders assess the same questions, but for a different set of students, consistency of marks and feedback can be brought into question. Each grader may have a particular bias for awarding marks for different steps of a problem, which is especially troublesome if the rubric is vague or graders have not calibrated themselves with each other in advance [8].

It is important to note that during the pandemic, the delivery, submission, and feedback tools all had to be managed by institute learning management systems (LMSs). While this simplifies the process in principle, module leads and administrators often have to separately manage deadlines, a team of graders, and releasing feedback once moderation is complete. LMSs are fundamentally not designed to manage each stage holistically and so digitisation of assessment is not, on its own, a better solution than traditional methods.

The next issue is the lack of oversight in the process. With multiple graders involved, it is easy for issues in the quality and consistency of marks to be missed. With often hundreds of students involved in the process, quality control is required, but there are often no resources to check every graded script for consistency aside from sampling which will likely only highlight the most obvious of issues.

Finally, and most importantly, assessment and feedback takes a significant amount of time. A typical university year group can

[2]Figures are closely based upon those in the thesis

have over 200 students. At a conservative 10 minutes per script this would take over 33 hours to grade.

Where defragmented scale solutions exist, there are additional issues that arise around the cognitive load on students. Becoming familiar with CAA systems provides an additional challenge to the student and can be detrimental to their achievement. This has been recently shown at scale in an intervention in Germany amongst secondary students where the same assessment was taken via a pencil and paper set up and online. Grades overall were reduced in the online version, but even more so for the students who were already weaker in the material being assessed. [13]. One can therefore postulate that a system that requires little learning from the individual students to become proficient users will be representative of the student's competence in the disciplinary material that is being assessed.

3 HOW GRAIDE WORKS

Graide is an end-to-end assessment and feedback platform which allows educators to create, deliver, and grade assignments all in one place. It allows educators to grade work *significantly* faster, improve the feedback students receive, and provide opportunities for continuous training. It does this in three ways:

(1) A streamlined assignment creation, deliver, and attempting workflow with a state-of-the-art editor which uses machine learning.
(2) An enhanced grading interface which is visible to all parties involved in grading.
(3) Machine learning assistance while grading to learn how educators give feedback in order to eliminate repetition.

3.1 Streamlined assignment creation, delivery, and answering/responding

Since Graide is a **1st GEN** system, traditional questions can be copied over and posed to students. Students may upload their handwritten solutions directly as scanned files or may opt to input their work digitally where competence precedes. Written scripts can also be uploaded in bulk meaning students can sit a formally invigilated examination and the graders can still benefit from the enhanced workflow of the grading engine. This may be required due to institutional regulations which aim to eliminate collusion and other forms of cheating.

Work can then be either digitised using state-of-the-art Optical Character Recognition (OCR) and on device handwriting recognition (both of which use machine learning) or kept as a scan depending on which mode is chosen for grading work: AI Grading or PDF Mode.

3.2 An enhanced grading interface

Graide's interface allows for distributed grading, grading per question (which is not practical with hundreds of paper based assignments), has a feedback centred rubric which is shared amongst all graders, and allows for location-based feedback. The interface for grading PDFs (PDF Mode) is depicted in figure 1.

Currently the AI does not function in PDF Mode. Therefore, this is used when the nature of the assessment and expected responses is unlikely to benefit from the AI engine. For example, scientific reports with no expected similarity between students' submissions. Note that as this information is stored, it is expected to be used to further enhance the AI grading capability in the short-term future.

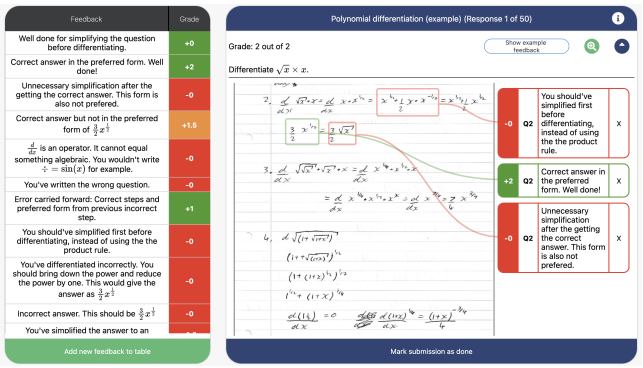

Figure 1: PDF grading interface with location-based feedback (right) and shared rubric (left).

3.3 Machine learning assistance while grading

AI grading is powered by an engine that rapidly learns how an educator is grading. By collating a large set of submissions into a simplified response tree, depicted in figure 2, it then suggests feedback in subsequent scripts where the student has followed a similar method to those already graded. The grader only has to check the suggested feedback on following scripts, occasionally teaching the AI as edge-cases appear, and confirm the feedback to move on. This empowers educators to give precise and detailed feedback as their efforts are automatically applied to all relevant cases.

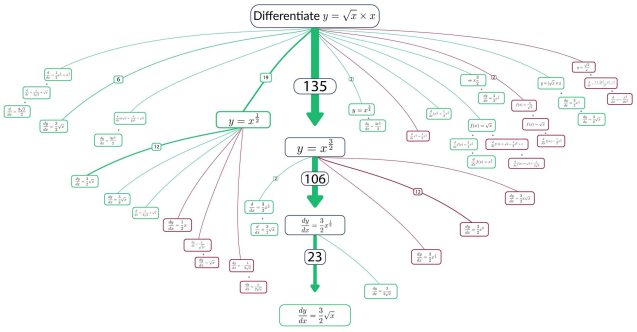

Figure 2: Simplified response tree collated by the AI engine.

4 HOW USING GRAIDE HAS IMPROVED PROCESSES AT THE UNIVERSITY OF XXXXXXX

The University of XXXXXXX has piloted Graide since the beginning of the 2021-2022 academic session. Module leads within the College of Engineering and Physical Sciences have used it over a breadth of different assignment styles, subjects, and levels. There have been benefits across the spectrum and are detailed in the following subsections.

4.1 Delegation and collation of grading tasks

For a mathematics module with over 500 students and several academics involved in the teaching, a written examination was submitted digitally by students and exported to be graded on Graide. Traditionally, physical scripts would have had to be passed around graders as they assessed specific questions from their parts of the course which requires collation and coordination that relies on everyone keeping to time to allow scripts to move around. By distributing workload by sub-parts of each question on Graide, 15 graders (academics and TAs) were able to grade scripts in parallel without any additional need to coordinate.

Since graders focused on a single question, they had a reduced cognitive load and were more likely to provide a consistent response to all students.

4.2 Continuous training opportunities

Where grading is delegated to a team of graders, often postgraduate teaching assistants, the feedback scheme provides module leads an oversight of the nature and minutia of feedback given to students. Module leads can, for example, edit inappropriate feedback so it updates on all scripts it was applied to or tweak marks for different steps. This means module leads can advise their team of graders on better practice while easily correcting issues in the feedback scheme.

4.3 Increased quality and consistency of feedback

A shared bank of feedback between all graders ensured the consistency of feedback to students. Additionally, the opportunity to write once and use many times incentivised an increase in feedback given. In a study comparing historical grading to AI feedback in Graide we found an increase in feedback of 7.2 times (the average amount of feedback increased from 23 to 166 words). When surveying users 76% of users said it was easier or significantly easier to give consistent feedback.

4.4 Faster grading and feedback

When used in PDF mode, 88% of users said Graide was faster or significantly faster than other grading systems they had used (on paper / digital ink), the remaining 12% said it was the same speed.

When used in AI mode to grade historical work we found Graide was 89.7% faster than on paper. The average amount of time to grade a script went from 11.2 minutes to 68.6 seconds. In addition to this Graide scaled remarkably quickly as shown by figure 3. Note also that this scale is seen in as few as 10 scripts and the trend only improves from there. This implies that Graide will be effective for small cohorts as well as large.

5 CONCLUSION

In summary - the Graide platform is a system that requires no technical expertise from the student, beyond uploading their work. The grading undertaken via Graide has significantly improved consistency, resulting in detailed feedback for method marks. Due to the machine learning engine, the grading accelerates as more scripts are graded. To enable learning at scale, one must be able to assess and provide feedback at scale and give granular, individualised feedback to the student. Our preliminary studies indicate that Graide is able to fulfill this role in mathematical disciplines, and will soon be in trial for short answer text response questions.

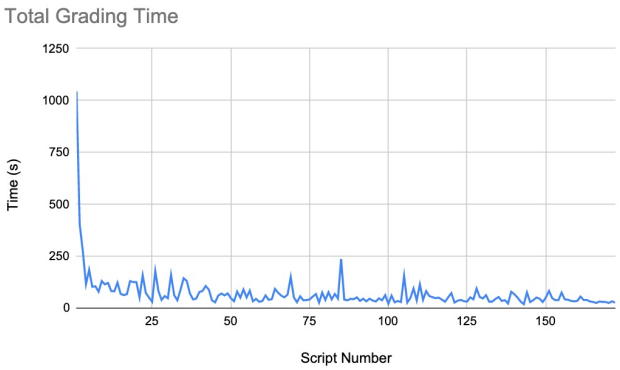

Figure 3: Time taken to grade script as a function of script number. Initial time is required to set up rubrics, and a dramatic reduction in time occurs subsequently.

Results and discussion in this WIP were originally presented in Stanyon's PhD thesis [12].

ACKNOWLEDGMENTS

This research was funded in whole or in part by the Funder [Grant number EP/N509590/1]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

REFERENCES

[1] RE Bennett. 1998. Reinventing assessment. *Princeton, NJ: Educational Testing Service* January 1998 (1998). http://www.aacompcenter.org/sites/default/files/resource/imported/PICREINVENT.pdf

[2] Digital Ed. 2020. Mobius. Retrieved Feb 28, 2022 from https://www.digitaled.com/mobius

[3] 6 Bit Education. 2021. Graide. Retrieved Feb 28, 2022 from https://www.6bit.co.uk/

[4] Graham Gibbs and Claire Simpson. 2005. Conditions Under Which Assessment Supports Students' Learning. *Learning and teaching in higher education* 1 (2005), 3–31.

[5] Gradescope. 2022. Gradescope. Retrieved Feb 28, 2022 from https://www.gradescope.com/

[6] Maplesoft. 2022. Maplesoft - Software for Mathematics, Online Learning, Engineering. Retrieved Feb 28, 2022 from https://www.maplesoft.com/

[7] Michael E. Martinez. 2010. Cognition and the question of test item format. *http://dx.doi.org/10.1207/s15326985ep3404_2* 34, 4 (2010), 207–218. https://doi.org/10.1207/S15326985EP3404{_}2

[8] D. Royce Sadler. 2013. Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice* 20, 1 (2013), 5–19. https://doi.org/10.1080/0969594X.2012.714742 arXiv:https://doi.org/10.1080/0969594X.2012.714742

[9] Christopher James Sangwin. 2013. *Computer aided assessment of Mathematics.* Vol. 15. Oxford University Press. 583–605 pages. https://doi.org/10.1093/acprof:oso/9780199660353.003.0001

[10] Christopher J. Sangwin and Ian Jones. 2017. Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics* 94, 2 (2017), 205–222. https://doi.org/10.1007/s10649-016-9725-4

[11] Kathrin F. Stanger-Hall. 2012. Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE Life Sciences Education* 11, 3 (9 2012), 294–306. https://doi.org/10.1187/CBE.11-11-0100/ASSET/IMAGES/LARGE/294FIG4.JPEG

[12] Stanyon. Robert. 2022. *Improving student assessment and feedback through the application of computer algebra and machine learning*. Ph.D. Dissertation. University of Birmingham, UK.

[13] Inga Wagner, Philipp Loesche, and Steven Bißantz. 2021. Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing. *European Journal of Psychology of Education* (2021). https://doi.org/10.1007/S10212-021-00532-6

[14] Gary L Wells, Nancy K Steblay, and Jennifer E Dysart. 2012. Eyewitness Identification Reforms: Are Suggestiveness-Induced Hits and Guesses True Hits? *Perspectives on psychological science : a journal of the Association for Psychological Science* 7, 3 (5 2012), 264–71. https://doi.org/10.1177/1745691612443368