Progressay AI Augmented Marking pilot

NCFE Functional Skills English Language Reading and Writing Level 2 and Level 3

We're NCFE. We're shaping smarter learning.

We're NCFE: an educational charity and leader in vocational and technical learning. **We combine over 170 years of education experience** with deep insight, working with a network of expert collaborators to shape smarter solutions around the greatest learning needs. In doing this, we're working for a fairer education system for all learners to power inclusivity and choice.

In 1848, we were born from the belief that no learner should be left behind. Today, we're taking up that cause with fresh energy. Our vision and goals will be achieved through:

- Creating education for a fairer world
- Moving towards a smarter education eco-system
- Using our influence to shape real change
- Promoting the idea that potential is personal.

progressay

Progressay is an award winning edtech start-up providing online based learning assessment. Their AI driven automated essay marking platform is designed to make marking more meaningful and accelerate learning.

For more information about this report please contact the NCFE Assessment Innovation Team by emailing aif@ncfe.org.uk

Contents

Executive Summary

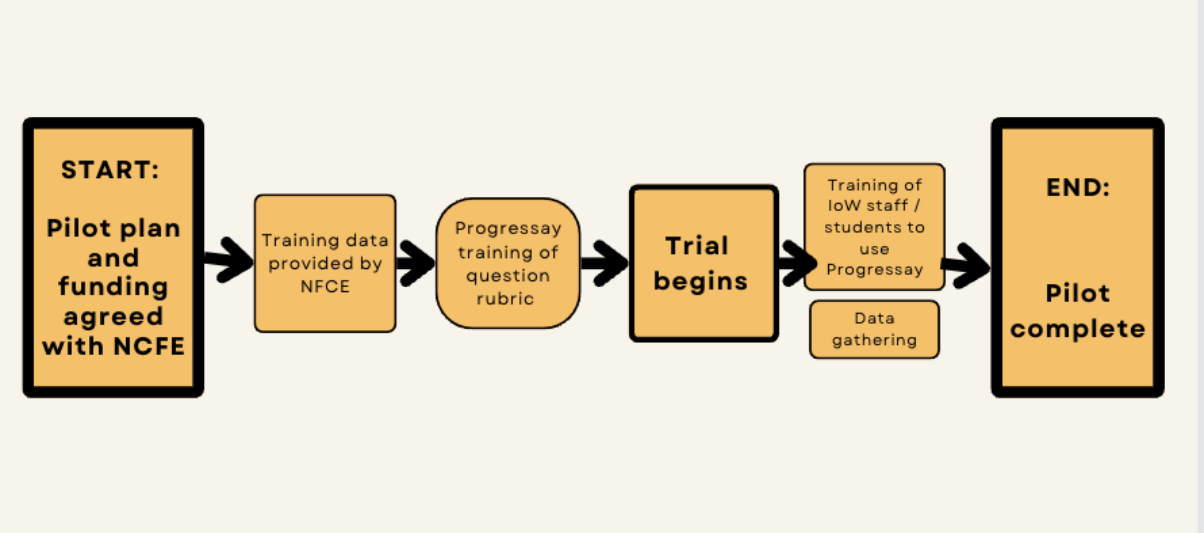
Project Summary

- Aim of pilot
- Target group
- Basic delivery info
- Who delivered the intervention
- Number of students/areas
- Brief description of evaluation design
- Brief details of developers and any funders other than NCFE
- Brief description of qualitative work undertaken
- Dates when the pilot started and finished

Aim of Pilot

This pilot evaluated an automated AI Augmented real time feedback process to determine its effectiveness in enhancing student attainment and self-efficacy. Its design was informed by research undertaken by Hattie and Timperley (2017) that indicated students who received swift feedback on low stakes testing exhibited the biggest improvements in attainment. Additionally, Opitz et al (2011) findings were influential, demonstrating that participants who were given immediate feedback showed a significantly larger increase in performance than those who received delayed feedback.

The pilot was structured as follows:



This research pilot had two objectives:

1. Attainment:

- To what extent does real-time, computer-generated/ AI Augmented, diagnostic feedback impact upon student attainment?

2. Learner Efficacy:

- To what extent does real-time, computer-generated/ AI Augmented, diagnostic feedback impact upon fostering learner efficacy?

Importantly, it should be noted at this stage that the Progressay AI tool does not use OpenAI to produce any of the marks and comments. Progressay uses a closed data setting to train its AI specifically for the purposes of student Data Privacy - this ensures it remains in line with Dept of Education guidelines regarding the use of student work to train AI systems (as outlined in 'GenAI in Education Policy Paper' Oct 2023).

Delivery Information:

119 students aged between 14-16, in full time education at the Isle of Wight School, were given access to the Progressay online platform containing the Level 2 Functional Skills Reading and Writing course. Students had access to the platform for a duration of one month, during which they engaged in 10 lessons. The pilot

required that students used the Progressay designed online NCFE learning platform for the entire process.

They began by taking a pre-trial test to measure their baseline performance in English Reading and Writing Level 2. Students were then required to work their way through the Progressay designed online course content. This had been built from scratch using NCFE Past Papers and involved specially created content to develop student skills. (While the complete design and build of the interactive courses was not necessarily part of the original project plan, this became essential in order to run the pilot as NCFE did not have an accessible platform on which Progressay were allowed to 'plug-in' their marking tool. The online courses were therefore an essential element in order for students to submit their work and engage with the AI augmented marking and feedback).

As they worked their way through the online course students were presented with low stakes exam style tests to complete. They submitted their responses and the Progressay platform gave AI augmented feedback in real time about their work. This included both formative and summative feedback comments, taking the form of both positives (What Went Well) and negatives (Even Better If). It was presented to students using a mixture of numerical data and descriptive narrative (e.g. You made X spelling errors meaning your spelling was in line with the class average).

The Progressay AI tool was trained to replicate the teacher/examiner in the feedback and marking process, producing marks and comments that lined up with what the human examiners had given the training data student papers. The AI relies on thousands of data measurements that fall under two main categories; reading metrics and writing metrics. Reading metrics refers to the substance of what students write about and as such, includes, but is not limited to:

- Readability
- Keywords used

- Specific knowledge reference
- Ideas and arguments

Writing metrics refers to how the students write I.e surface level technical accuracy. This includes, but is not limited to:

- Spelling and Vocabulary
- Organisation
- Punctuation
- Grammar
- Sentence structure

The students ended their time on the platform by completing a post trial test to measure any change in performance. After cleaning the data produced from all these stages - discounting those students who did not complete both the pre and post trial tests- the baseline marks were compared to the summative test marks using mean scores to analyse the overall group performance.

Following the post trial test students participated in a focus group and qualitative data gathering question session. In these both quantitative and qualitative data was collected using surveys designed within a pedagogical-rooted framework for encouraging metacognition. This focused upon:

- time spent on the platform
- student opinions regarding the process
- feedback in relation to their own sense of efficacy

The resultant qualitative data was analysed thematically to identify interesting features relating to student efficacy, resilience, motivation, engagement, platform design, technical features, and efficiency/ speed. Student opinion was also sought more broadly regarding aspects of Progressay platform user experience and future

of AI marking in schools. By integrating the quantitative and qualitative analyses it was possible to correlate changes in attainment score with thematic insights.

Timescale

It should be noted that the pilot stage was only reached following the unanticipated marketing and recruitment of schools to the pilot, as well as unplanned data cleaning, additional data training (80+ rubrics), and unforeseen online course design/ building).

The planned Pilot ran from **January 2021** to **January 2022**, however as stated above a large proportion of time was spent before it could begin properly, this was due to having to address this series of unanticipated aspects of the project. It meant the project as a whole ran from **February 2022** until **June 2024**.

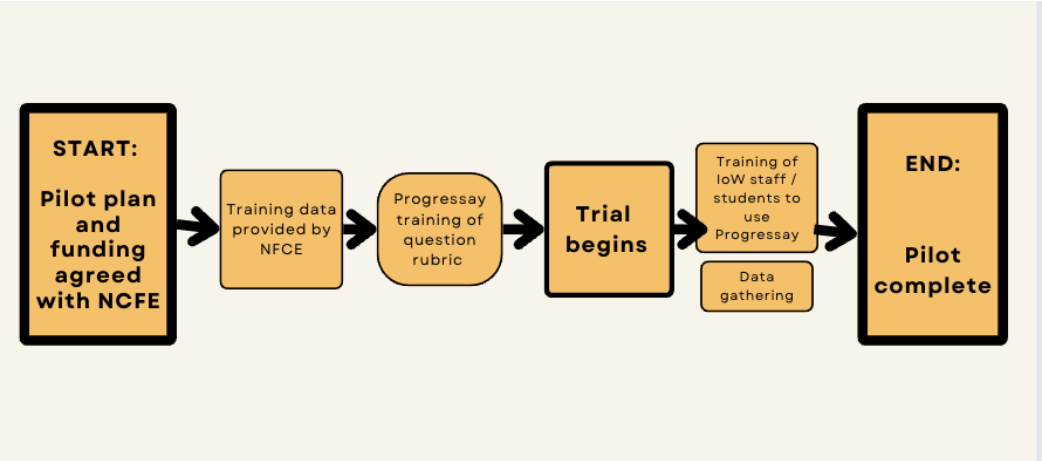
Before the pilot could begin the following (additional) work needed to be undertaken:

1) Recruitment of schools: A marketing campaign was run in order to recruit schools and participants to the trial. This involved a social media recruitment drive, as well as articles in the press. One of these articles prompted the inclusion of the work of NCFE and Progressay in the Oct 2023 POST note (Policy Office of Science and Technology) on AI in Education.
2) 30,000 training scripts were provided by NCFE, however these were largely unreadable due to the HTML code that was contained within the script. They were also a mix of Level 1 and Level 2 scripts (Progressay trial only focussed on Level 2). Progressay therefore built an automated data cleaning algorithm from scratch in order to make the scripts readable before they could begin training.
3) Progressay were required to train 80+ rubrics for a huge variety of questions, rather than the smaller sample size expressed in their trial plan.
4) As mentioned above, although we had initially intended on using NCFE English Functional Skills learning material for the online courses and to merely integrate this learning material with our Progressay Grading Assistant, due to such teaching and learning resources not being available, we instead had to develop our own interactive learning content to form the base of the

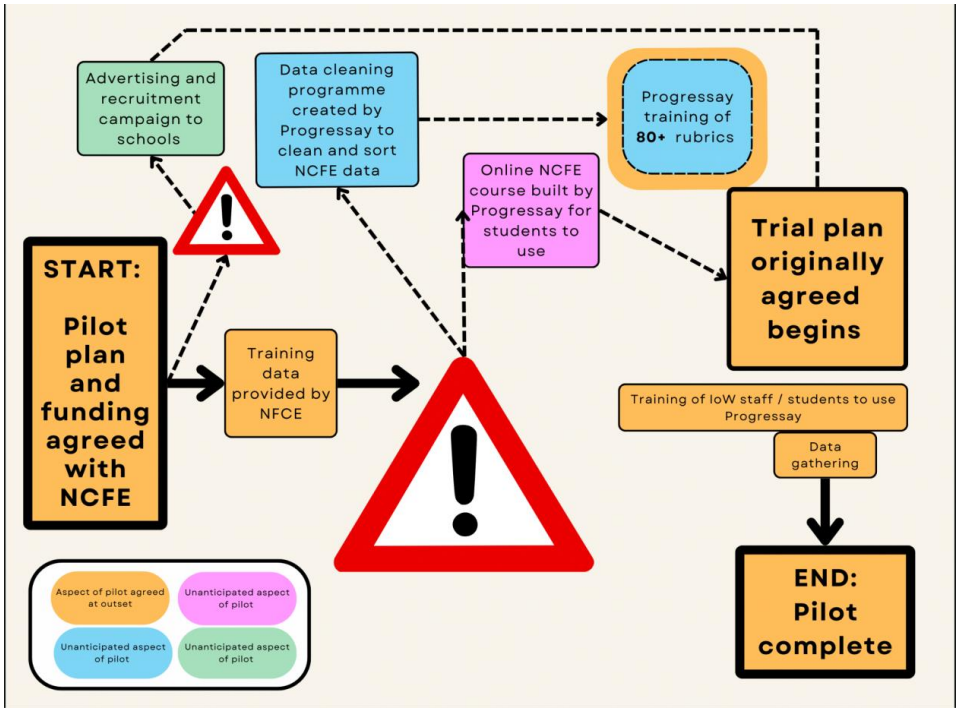
two online courses (i.e the NCFE Level 2 Reading and NCFE Level 2 Writing) as

5) Progressay also tested on an AQA GCSE paper in interim due to recruitment problems- efficacy remained the focus in this adjacent project.

Therefore although the pilot plan was originally intended to look like this:



The resultant process looked more like this:

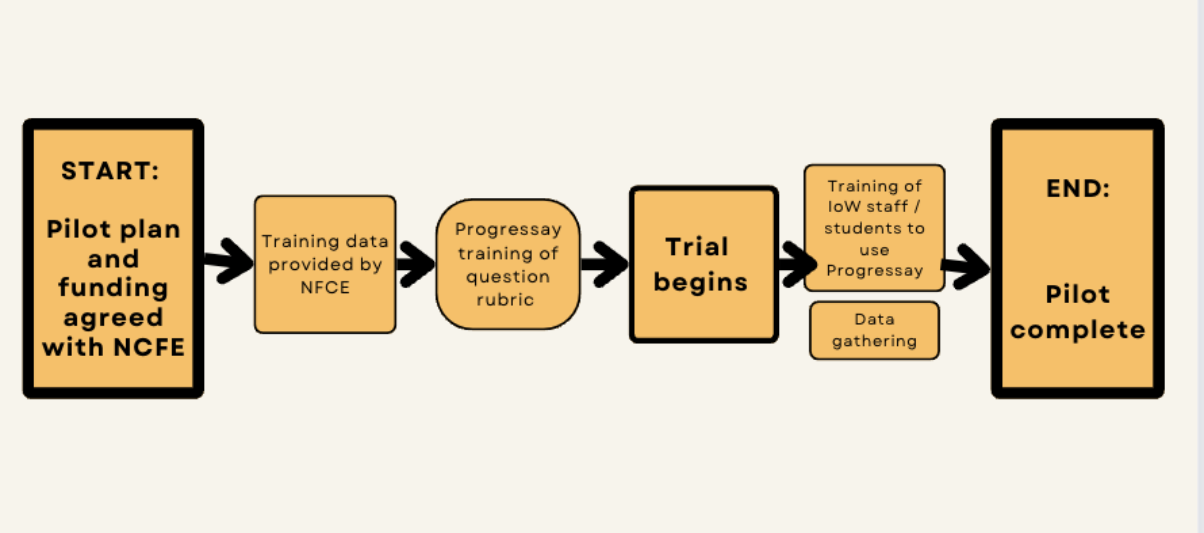


However, before focussing upon the additional aspects of this project it is important to foreground the pilot itself.

Introduction

An introduction to the pilot, background, project team and agreed research aims/objectives.

The Pilot:



Building upon previous work, this business-led project offered a highly innovative approach to essay marking, falling within the scope of the NCFE Innovation fund, the AI & Data Economy grand challenge and the Innovate plan-for-action. It significantly contributes to the ongoing development of best practice in the integration of AI technology into the educational environment. Through the provision of valuable insights into the effective use of AI augmented educational technology for the purposes of feedback and marking it clearly identifies one way in which such technology can be harnessed to improve student learning outcomes. While helping to develop greater student agency, and raise self efficacy the pilot shows increased

pupil achievement in the face of academic challenges. In demonstrating the impact of AI augmented, specific, timely, and constructive feedback on student essays, it offers the opportunity for schools and education providers to empower students to take greater ownership of their learning journeys.

Inspired by the research of Hattie and Timperley (2017), Deci and Ryan, Opitz et al. (2011) the aim was to evaluate whether the Progressay tool offering real-time AI-augmented feedback on low-stakes exam-style assessments positively impacts student attainment and learner efficacy.

Hattie and Timperley (2017) found that students who received swift feedback on low-stakes testing exhibited the most significant improvements in attainment. Opitz et al. (2011) similarly demonstrated that immediate feedback resulted in a substantially larger increase in performance compared to delayed feedback. Deci and Ryan (1985) highlighted the importance of ensuring learners understand that feedback is intended to help them compete against their personal bests rather than against one another. Pennebaker, as cited by Stenger (2014), emphasised that students must have access to information about their performance to determine whether they have mastered the material. He argued that providing students with insights into their studying, reading, searching for information, or answering questions can be invaluable. It is evident that providing students with prompt access to this information helps them develop a greater awareness of their learning, enabling them to more easily recognise mistakes and devise strategies to address these weak points.

The key features of the Progressay AI Automated real-time feedback tool were therefore:

- Fast (real time) feedback on student work on low stakes frequent assessment.
- Specific feedback comments - providing learners with information on what exactly they did well, and what may still need improvement.

- Involving learners in the process - making the comments available to students.
- Human- sounding - although offering a mix of numerical and prose style feedback it remained important that the student was not left feeling alienated from their specific learning experience.

The pilot had two research objectives:

1. Attainment:

- To what extent does real-time, computer-generated/ AI Augmented, diagnostic feedback impact upon student attainment?

2. Learner Efficacy:

- To what extent does real-time, computer-generated/ AI Augmented, diagnostic feedback impact upon fostering learner efficacy?

Research Method

Samples size, sampling technique, recruitment activities.

The pilot set out to explore whether real time AI Augmented feedback on low stakes exam style assessment impacts attainment and learner efficacy. Progressay designed a trial that sought to measure potential impact by comparing student attainment results from a pre-test and a post-test, as well as gathering qualitative data from focus groups and surveys designed within a pedagogical-rooted framework for encouraging metacognition.

The Pilot had two research objectives:

1. Attainment:

- To what extent does real-time, computer-generated, diagnostic feedback impact upon student attainment?

This was measured by comparing grades from a pre-trial and post-trial assessments of the user group. Considering their Baseline marks vs Summative marks to see if there has been an increase in attainment having received the AI Augmented real

time feedback on the low stakes assessment tasks that they submit via the Progressay online course.

2. Learner Efficacy:

- To what extent does real-time, computer-generated, diagnostic feedback impact upon fostering learner efficacy?

This was measured through surveys designed within a pedagogical-rooted framework for encouraging metacognition. Data collected concerning learner efficacy will relate to:

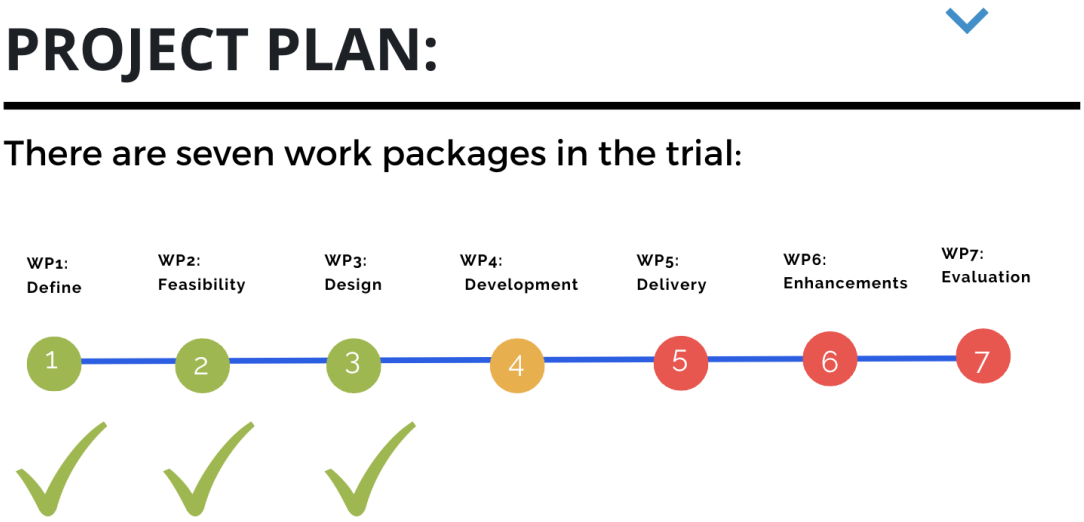
- resilience
- motivation
- engagement

The trial used Agile Research methods, drawing upon an iterative approach to research that emphasises flexibility, adaptability, and collaboration, thereby allowing researchers to quickly adjust their methods based on ongoing feedback and changing conditions to ensure the research process remains responsive and relevant.

Research Question	KPIS (Metrics to be collected)	Data collection model	Data collection means
1) Attainment	Change in attainment over time	Pre-test vs post-test	2 x NCFE Functional Skills English Level 2 assessments - a baseline testa and an End of Course test
	Change in English proficiency skills over time	Pre-test vs post-test	Automatically extracted by natural language processing AI/ ML
2) Learner Efficacy	Change in self-reported learner confidence over time	Pre-test vs post-test	2 x online surveys - - one issued at the start of the trial and one issued after the trial.
	Change in Learner engagement over time	Pre-test vs post-test	- Measuring lesson completion rate - Measuring time spent

Specific processes or cycles were implemented throughout the process of the project e.g. sprint planning, iterations, regular check-ins, and feedback loops. In addition

team members and stakeholders were engaged throughout the research process with their input shaping the direction of the research, contributing to problem-solving, and ensuring the research remained aligned with its goals. These are evidenced throughout the process of the work-packages outlined below:



Recruitment of pilot participants was difficult but the pilot ended up relying on a snowball technique. Once a keen teacher at the IoW school had been recruited the Progressay team relied on her input to ensure that students from the cohort were engaged within the tasks.

A total of 119 students, aged 14-16 and enrolled full-time at the Isle of Wight School, were given access to the Progressay online platform, which includes the Level 2 Functional Skills Reading and Writing course.

1. Progressay ran a ‘live’ online student onboarding process, with a subsequent mop-up/ follow up, to ensure that users were able to access the required information.

2. Students had asymmetric start times for their engagement so that they were able to complete the modules at their own pace.

3. Students first took a pre-trial test to establish their baseline performance in English Reading and Writing Level 2.

4. Over a month, students participated in 10 lessons on the online Progressay NCFE Functional skills Reading and Writing platform.

5. Embedded within the lessons were a series of low stakes, exam-style submissions for the students to complete as practice. The Progressay platform provided real-time AI-augmented feedback, including both formative and summative comments, presenting numerical data and descriptive narrative (e.g., “You made X spelling errors, aligning your spelling with the class average”). The AI tool aimed to replicate examiner feedback and was trained to produce marks consistent with human examiners, analysing factors such as readability, spelling errors, inference, and complex sentence structure.

6. Following their ‘practice’ the pilot concluded with students taking a post-trial test to measure their performance changes.

7. After data cleaning—excluding students who did not complete both pre and post-trial tests—the baseline marks were compared to the summative test marks using mean scores to analyse overall group performance.

8. Following the post-trial test, students participated in online focus groups and qualitative data gathering sessions. Both quantitative and qualitative data were collected through pedagogically-rooted surveys promoting metacognition, focusing on:

    - Time spent on the platform
    - Student opinions on the process
    - Feedback on their sense of efficacy

9. The qualitative data was thematically analysed to identify features related to student efficacy, resilience, motivation, engagement, platform design,

technical aspects, and efficiency/speed. Broader student opinions were also gathered on the user experience of the Progressay platform and the future of AI marking in schools. Integrating quantitative and qualitative analyses allowed for correlating changes in attainment scores with thematic insights

Results

Presentation of pilot data, outcomes, learner experience data.

The trial had two research questions. It produced both quantitative and qualitative results for both research questions.

1. Attainment:

- To what extent does real-time, computer-generated/ AI augmented, diagnostic feedback impact upon student attainment?

This was measured by comparing grades from a pre-trial and post-trial assessments of the user group. Considering their Baseline marks vs Summative marks to see if there has been an increase in attainment having received the AI augmented real time feedback on the low stakes assessment tasks that they submit via the Progressay online course.

2. Learner Efficacy:

- To what extent does real-time, computer-generated/ AI augmented, diagnostic feedback impact upon fostering learner efficacy?

This was measured through surveys designed within a pedagogical-rooted framework for encouraging metacognition. Data collected concerning learner efficacy will relate to:

- resilience
- motivation
- engagement

Outcomes for the pilot are detailed below:

Research Question	KPIS (Metrics to be collected)	Data collection model	Data collection means	**Outcomes**
1) Attainment	Change in attainment over time	Pre-test vs post-test	2 x NCFE Functional Skills English Level 2 assessments - a baseline testa and an End of Course test	**Overall, 57% of learners for the Reading course improved and 30% of students for the Writing Course improved.**
2) Learner Efficacy	Change in self-reported learner confidence over time	Pre-test vs post-test	2 x online surveys - - one issued at the start of the trial and one issued after the trial.	**+12% increase in students trusting the platform when the pre and post trials were compared - (33% rated an AI marking platform with 4 and 5 stars in the pre-trial survey which went up to 45% in the post-trial survey).**
	Change in Learner engagement over time	Pre-test vs post-test	- Measuring lesson completion rate - Measuring time spent	**+7% increase in students saying they found AI marking "Very useful" in the post-trial survey vs students who felt that teacher feedback is "Very useful" in the pre-trial survey.**

1. Attainment:

- **To what extent does real-time, computer-generated/ AI Augmented, diagnostic feedback impact upon student attainment?**

This section will initially consider the quantitative data collected regarding any changes in attainment overall for both Reading and Writing. It then looks at a breakdown of changes in attainment per question, the amount of time spent per question, time spent per paper on both Reading and Writing, and course completion rate.

Change in attainment overall for Reading and Writing papers

57% of learners for the Reading course improved and **30%** of students for the Writing Course improved.

Pilot Results summary:

The quantitative data that was collected regarding student attainment suggested that there was an increase in the total marks scored on average over the duration of the course. Whereas for the writing course, there was a marginal decrease in the total marks scored on average over the duration of the course.

All Assignments - Progress over time

	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
Total Marks available per student	29	26.0	-3.00	Decrease	42	42.0	0.00	No Change
Total Marks available for all students	1305	962	-343.00	Decrease	2856	1974	-882.00	Decrease
Total Marks scored by all students	233.68	214.58	-19.10	Increase	1082.10	675.17	-406.93	Decrease

For the reading course, of the 34 students that completed both the Baseline and Summative assessments, there was a marginal increase in the total marks scored on average over the duration of the course.

For the writing course, of the 44 students that completed both the Baseline and Summative assessments, there was a marginal decrease in the total marks scored on average over the duration of the course.

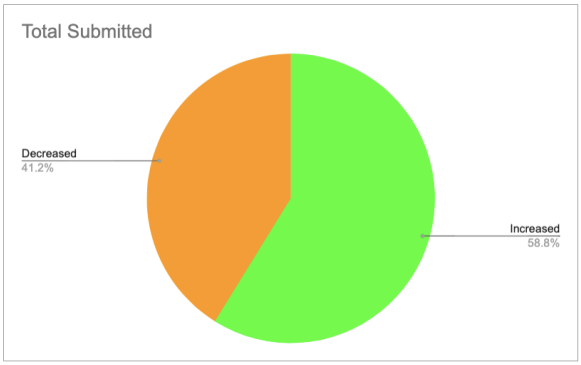
	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
No of Students completed both pre and post tests	34.00		/	/	44.00		/	/
Total Marks available per student	29	26.0	3.00	Decrease	42	42.0	0.00	No Change
Total Marks available for submitted students	986	884	-102.00	Decrease	1848	1848	0.00	No Change
Total Marks scored by submitted students	233.68	214.58	-19.10	Decrease	764.32	665.62	-98.70	Decrease
Total Marks scored as %	23.70%	24.27%	0.01	Increase	41.36%	36.02%	-4.76	Decrease
Avg mark for submitted	0.35	0.36	0.003	Increase	16.62	14.47	-2.15	Decrease

Course 1: READING

Overall, **59%** of students made progress over the duration of the reading course. As such, there appears to have been a positive trend in educational outcomes across these levels.

Progress over time

Progress over time	Total Submitted	%
Increased	20	58.82%
Decreased	14	41.18%
Total	34	100.00%



Overall, 59% of students that submitted both Baseline and Summative Assessments (20 of the 34 students) improved their marks over the trial compared to 43% (14 students) whose marks did not improve

Overall Progress in Marks

% of students that improved	58.82%

Overall, 59% of students made progress over the duration of the course i.e when the average mark for the baseline assessment in lesson 1 is compared to the average mark for the Summative Assessment in Lesson 12.

Progress by Classes

Classes	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Public Services	32	18	9	50.00%	9	50.00%
Sport	26	1	0	0.00%	1	100.00%
Travel	61	15	11	73.33%	4	26.67%
Total	119	34	20	58.82%	14	41.18%

There were three classes involved in the trial; Public Services, Sport and; Travel:

- Public Services students improved by 50% (9/18);
- For Sport, there was only onestudent that submitted and;
- For students studying Travel, 74% improved (11/ 15).

Progress by SEND Status

SEND Status	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Y	10	6	4	66.67%	2	33.33%
N	109	28	16	57.14%	12	42.86%
Total	119	34	20	58.82%	14	41.18%

There were 10 students with SEN overall. Of these 10 SEND students, 6 completed the course. Of this, 4 out of 6 improved (i.e 67%)

57% of students of the 28 students without SEN improved (i.e 16/28).

Progress by Level

Year	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
2	35	20	11	55.00%	9	45.00%
3	84	14	9	64.29%	0	0.00%
Total	119	34	20	58.82%	9	26.47%

Level 2 Students improved **more** than Level 3 students.

55% of the 20 Level 2 students that submitted (i.e 11 students) improved.

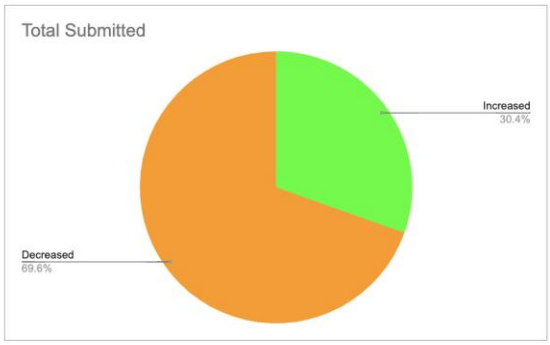
64.29% of the 14 Level 3 students (ie 9 students) improved.

Course 2: WRITING

Overall, **31%** of students made progress over the duration of the writing course.

Progress over time

Progress over time	Total Submitted	%
Increased	14	30.43%
Decreased	32	69.57%
Total	46	100.00%



Overall, for the writing course, 30.43% of students that submitted both Baseline and Summative Assessments (14 of the 46 students) improved their marks over the trial compared to 70% (32 students) whose marks did not improve

Overall Progress in Marks

% of students that improved	30.43%

Overall, 30% of students made progress over the duration of the course i.e when the average mark for the baseline assessment in lesson 1 is compared to the average mark for the Summative Assessment in Lesson 10.

Progress by Classes

Classes	Total Submitted	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Public Services	21	21	5	23.81%	16	76.19%
Sport	35	11	1	9.09%	10	90.91%
Travel	63	14	8	57.14%	6	42.86%
	119	46	14	30.43%	32	69.57%

There were three classes involved in the trial; Public Services, Sport and; Travel:

- Public Services students improved by 24% (5/21);
- For Sport, only 9% improved i.e (1/11);
- For students studying Travel, 56% improved (8/ 14).

Progress by SEND Status

SEND Status	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Y	10	8	2	25.00%	6	75.00%
N	109	38	12	31.58%	26	68.42%
Total	119	46	14	30.43%	32	69.57%

There were 10 students with SEN overall. Of these 8 SEND students that submitted, 2 improved (i.e 25%)

32% of the 38 students without SEN improved (i.e 12/38).

Progress by Level

Year	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
2	35	16	3	18.75%	13	81.25%
3	84	30	11	36.67%	19	63.33%
Total	119	46	14	30.43%	32	69.57%

Level 3 Students improved more than Level 2 students.

19% of the 16 Level 2 students that submitted (i.e 3 /16) improved.

37% of the 30 Level 3 students (ie11/30) improved.

Students appeared to improve more in the reading course compared to the writing course.

The principal difference between Reading and Writing was that there were more AI essays in the reading compared to the writing, and by extension, more opportunities to receive AI marks and AI feedback.

This data can be broken down into more detail:

The question types demonstrated a slight decrease in attainment (1 mark) from Baseline to Summative test for the reading paper and no change for the writing paper.

	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
No of Questions	16.00	15.00	-1.00	Decrease	2.00	2.00	0.00	No Change
No of Assignments	14.00	13.00	-1.00	Decrease	2.00	2.00	0.00	No Change
% of Assignments	88%	87%	-0.01	Decrease	100.00 %	100.00%	0.00	No Change
No of MCQs	2.00	2.00	0.00	No Change	0.00	0.00	0.00	No Change
% of MCQs	13%	13%	0.01	No Change	0.00%	0.00%	0.00	No Change

Reading - Progress by class:

There were three classes involved in the trial; Public Services, Sport and; Travel. Students studying Travel experienced the greatest improvement in attainment.- **73% of Travel students improved.** Public Services students also improved, however by a little less i.e 50% (9/18). For Sport, there was only one student that submitted, who did not improve.

Progress by Level:

For the reading course, Level 2 Students improved more than Level 3 students. 69% of the 16 Level 2 students that submitted (i.e 11 students) improved. 30% of the 30 Level 3 students (ie 9 students) improved.

Progress by SEND:

There were 10 Students that identified as having Special Educational Needs and Disabilities (SEND) overall. Of these 10 students, only 6 actually completed the course. 4 of the 6 saw an improvement, (i.e 67%). Although the data is limited here, there is some indication that SEND students benefit from access to the platform.

Time taken - Baseline vs Summative

More time was spent on average on the Baseline assessment for the reading course compared to the writing course.

	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
Average Time Spent	0:13:55	0:09:13	0:04:42	Decrease	0:08:08	0:03:54	0:04:14	Decrease

This decrease / retention of score should be taken in the context of the decrease in time spent (4 minutes less) on the summative test as compared to the baseline test. More time was spent on average on the Baseline assessment for the reading course compared to the writing course.

	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
Average Time Spent	0:13:55	0:09:13	0:04:42	Decrease	0:08:08	0:03:54	0:04:14	Decrease

Total time spent on the Reading Course:

Reading			
Lesson 1 Time Spent	Lesson 12 Time Spent	Difference in time take	Difference in %
9:58:22	5:53:56	4:04:26	59.15%

A comparison of the total time spent by all students shows that almost 10 hours was spent on the Baseline Assessment whereas only around 6 hours was spent on the Summative Assessment. Around 4 hours less was spent on the Summative assessment vs the Baseline.

Average time spent on the platform:

Reading			
Lesson 1 Time Spent	Lesson 12 Time Spent	Difference in time take	Difference in %
0:13:55	0:09:13	0:04:42	66.21%

This is also the case when we look at the average time spent per student. The Baseline, each learner spent around 14 minutes.

More students started the course than those that completed the course; 10 less students completed the End of Course lesson (35) compared to the Baseline Assessment (25).

This must be considered within the context of time spent on each course (more details are provided below). In a nutshell, more time was spent on the Reading Course:

	Reading	Writing
Difference in time spent for Baseline compared to Summative (minutes)	-00:04:42	-00:04:14
Difference in % of time spent for Baseline compared to Summative	-33.79% decrease	-52.05% decrease

Reading			
Lesson 1 Time Spent	Lesson 12 Time Spent	Difference in time take	Difference in %
0:13:55	0:09:13	0:04:42	66.21%

This is also the case when we look at the average time spent per student. The Baseline, each learner spent around 14 minutes.

More students started the course than those that completed the course; 10 less students completed the End of Course lesson (35) compared to the Baseline Assessment (25). Despite an attrition rate of around 30%, the total marks for all students increased by 4.40%; for the Baseline Assessment, 17.47% of the total available marks were awarded compared to 22.31% for the Summative Assessment.

Course 1: READING

Overall Course Completion Rates

The course completion rate was especially low. From 119 students signed up we ended up with 29% completing all elements of the pilot - pre & post trial for Reading, and 37% for the Writing course.

This table shows information on student course completion - baseline vs summative:

	Course 1: READING				Course 2: WRITING			
	Lesson 1	Lesson 12	Difference	Status	Lesson 1	Lesson 12	Difference	Status
Total Students	119.00	119.00	0.00	No Change	119.00	119.00	0.00	No Change
No of Absent Students	74.00	82.00	8.00	Increase	51.00	72.00	21.00	Increase
No of Present Students	45.00	37.00	-8.00	Decrease	68.00	47.00	-21.00	Decrease
No of Students completed pre and post	34.00				44.00			
Percentage of Students that	29%				37%			

completed both Baseline and Summative Test (%)				

	READING COURSE	WRITING COURSE
	Freq	Freq
Total Lessons	15	12
Total lessons set for all students	1785	1547
Total Lessons completed for all students	642	737
Lesson Completion Rate (%)	36%	48%
Total Students	119	119
Total No. of Students that completed both Baseline and Summative Test	34	44
Percentage of Students that completed both Baseline and Summative Test (%)	29%	37%
Average time spent	2:49	4:01

The Reading Course was longer than the Writing Course. This perhaps explains why there is a greater completion rate for the writing course compared to the reading course.

Course Completion Rates by class

For the Reading course, the majority of the students that completed the course were from the Public Services class.

Course 1: READING				
	Public Services	Sport	Travel	Total
Total Students	32	38	49	119
Total No. of students that Completed Both Baseline (Pre-test)	20	4	10	34

and Summative (Post-test)				
%	62.50%	10.53%	20.41%	

The same is true for the Writing course, with the majority of the students completing the course being from the Public Services class:

Course 2: WRITING				
	Public Services	Sport	Travel	Total
Total Students	32	38	49	119
Total No. of students that Completed Both Baseline (Pre-test) and Summative (Post-test)	21	11	14	46
%	65.63%	28.95%	28.57%	

Students were asked to complete the Writing Course first. More students completed the Baseline and Summative for the writing course compared to the Reading Course.

Progress over time:

READING

Overall Progress in Marks

% of students that improved	58.82%

Overall, 59% of students made progress over the duration of the course i.e when the average mark for the baseline assessment in lesson 1 is compared to the average mark for the Summative Assessment in Lesson 12.

Progress by Classes

Classes	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Public Services	32	18	9	50.00%	9	50.00%
Sport	26	1	0	0.00%	1	100.00%
Travel	61	15	11	73.33%	4	26.67%
Total	119	34	20	58.82%	14	41.18%

There were three classes involved in the trial; Public Services, Sport and; Travel:

- Public Services students improved by 50% (9/18);
- For Sport, there was only onestudent that submitted and;
- For students studying Travel, 74% improved (11/ 15).

Progress by SEND Status

SEND Status	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Y	10	6	4	66.67%	2	33.33%
N	109	28	16	57.14%	12	42.86%
Total	119	34	20	58.82%	14	41.18%

There were 10 students with SEN overall. Of these 10 SEND students, 6 completed the course. Of this, 4 out of 6 improved (i.e 67%)

57% of students of the 28 students without SEN improved (i.e 16/28).

Progress by Level

Year	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
2	35	20	11	55.00%	9	45.00%
3	84	14	9	64.29%	0	0.00%
Total	119	34	20	58.82%	9	26.47%

Level 2 Students improved **more** than Level 3 students.

55% of the 20 Level 2 students that submitted (i.e 11 students) improved.

64.29% of the 14 Level 3 students (ie 9 students) improved.

WRITING COURSE - PROGRESS OVER TIME

Overall Progress in Marks

% of students that improved	30.43%

Overall, 30% of students made progress over the duration of the course i.e when the average mark for the baseline assessment in lesson 1 is compared to the average mark for the Summative Assessment in Lesson 10.

Progress by Classes

Classes	Total Submitted	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Public Services	21	21	5	23.81%	16	76.19%
Sport	35	11	1	9.09%	10	90.91%
Travel	63	14	8	57.14%	6	42.86%
	119	46	14	30.43%	32	69.57%

There were three classes involved in the trial; Public Services, Sport and; Travel:

- Public Services students improved by 24% (5/21);
- For Sport, only 9% improved i.e (1/11);
- For students studying Travel, 56% improved (8/ 14).

Progress by SEND Status

SEND Status	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
Y	10	8	2	25.00%	6	75.00%
N	109	38	12	31.58%	26	68.42%
Total	119	46	14	30.43%	32	69.57%

There were 10 students with SEN overall. Of these 8 SEND students that submitted, 2 improved (i.e 25%)

32% of the 38 students without SEN improved (i.e 12/38).

Progress by Level

Year	Total Students	Total Submitted	Increased	Increased as %	Decreased	Decreased as %
2	35	16	3	18.75%	13	81.25%
3	84	30	11	36.67%	19	63.33%
Total	119	46	14	30.43%	32	69.57%

Level 3 Students improved more than Level 2 students.

19% of the 16 Level 2 students that submitted (i.e 3 /16) improved.

37% of the 30 Level 3 students (ie11/30) improved.

Average time spent on the platform:

Attainment over the courses must be considered within the context of the time spent on each course (more details are provided below). In a nutshell, more time was spent on the Reading Course:

	Reading	Writing
Difference in time spent for Baseline compared to Summative (minutes)	-00:04:42	-00:04:14
Difference in % of time spent for Baseline compared to Summative	-33.79% decrease	-52.05% decrease

This should also be evaluated within the context of the amount of data provided for each course. There was significantly more data available for reading compared to writing. This meant that most of the course for writing was based on MCQs compared to the reading course.

Writing			
Lesson 1 Time Spent	Lesson 12 Time Spent	Difference in time take	Difference in %
0:08:08	0:03:54	0:04:14	- 47.95%

For the Baseline test, each learner spent an average of around 8 minutes across the course. This went down to an average of 4 minutes per student for the Summative Test. This shows there was a drop of almost 50% in terms of how much time was spent on the actual course.

The fact that learners spent less time at the end of the course compared to at the start could be used to show that engagement dropped as the course progressed. However, it could also be the case that less time was allocated by the teacher.

The fact that 47.95% less time per student was spent on the summative assessment compared to the

baseline assessment, could explain why learners scored 4.76% less on average.

2. Learner Efficacy:

- **To what extent does real-time, computer-generated, diagnostic feedback impact upon fostering learner efficacy?**

Survey and focus group results

Below is a comparison of two questions raised in the Pre-trial and Post-trial surveys respectively:

Pre Trial			**Post Trial**			
Question	Score	Results	**Question**	Score	Results	Results
4. How useful do you find feedback when given by your teachers?	**16%**	16% reported finding teacher feedback "very useful"; 49% reported finding teacher feedback "useful"; 26% reported finding teacher feedback "kind of useful"; 9% reported finding teacher feedback "Not very useful";	**4. How useful did you find the feedback?**	**23%**	23% found the platform's feedback "Very useful". 37% found the platform's feedback "useful". 19% found the platform's feedback "Kind of useful". 21% of respondents reported finding the feedback "Not very useful". It seems many students did not access any feedback at all.	**+7% increase in students saying they found AI marking “"Very useful" in the post-trial survey vs students who felt that teacher feedback is "Very useful" in the pre-trial survey.**

10. Finally, to what extent would you say that you could trust a platform (i.e one that could read, understand and generate feedback for open-ended text) could help with learning?	33%	When asked the following question: “To what extent would you say that you could trust a platform (i.e one that could read, understand and generate feedback for open-ended text) could help with learning?” 12% selected 5/5 21% selected 4/5 42% selected 3/5 12% selected 2/5 14% selected 1/5	**10. Now that you have used the auto-grading and feedback system, to what extent would you say it could help with learning?**	45%	When asked *"Now that you have used the auto-grading and feedback system, to what extent would you say it could help with learning?"* 12% selected 5/5 33% selected 4/5 40% selected 3/5 7% selected 2/5 9% selected 1/5	**+12% increase in students trusting the platform when the pre and post trials were compared - 33% rated an AI marking platform with 4 and 5 stars in the pre-trial survey which went up to 45% in the post-trial survey.**

Below is a more detailed break down of the pre-trial and post-trial surveys:

Pre-Trial Survey Results:

Current Progress:

- 18% of students felt they were doing "great."
- 71% felt they were doing "good."
- 10% felt their progress was "not so great."

Struggles:

- 53% struggled with "Remembering knowledge taught."
- 14% struggled with "understanding concepts."
- 14% struggled with "remembering feedback."

Feedback Openness:

- 76% were open to a system providing more and quicker feedback, believing it could help them progress faster.
- 93% of learners used feedback to improve their learning.
- 6% did not use feedback to improve their learning.
- 34% wanted feedback “straight after doing it.”
- 42% wanted feedback a few days later..

Following completion of the online course and assessment on the Progressay platform pupils were then asked to complete a series of questions as part of the post-trial survey. The post-trial survey results indicate a largely positive experience with Progressay's AI-augmented real-time feedback system, though there are some mixed feelings among the students.

Post-Trial Survey Results:

- **Positive Feedback:** 67% of respondents rated the feedback as either "Great" or "Good," indicating a generally favourable reception.
- Uncertainty: 14% of respondents were unsure about how the system knew the answers, suggesting some confusion about the AI's functionality.
- Mixed Feedback: 65% were happy with the feedback, while 30% were not, pointing to a significant minority who were dissatisfied.

Usefulness of Feedback:

- High Usefulness: 23% found the feedback "Very useful," 37% found it "useful," and 19% found it "Kind of useful."
- Lower Usefulness: 21% reported the feedback as "Not very useful," indicating room for improvement in perceived feedback effectiveness.

Willingness to Use Feedback for Improvement:

- Positive Intent: 72% of respondents indicated they would use the feedback to improve, showing a strong inclination to leverage the feedback for better performance.
- Reluctance: 10% said they would not use the feedback to improve, highlighting a small but notable resistance.

Clarity of Next Steps:

- High Clarity: 77% of respondents felt clear about the next steps needed after reading the feedback.
- Moderate Uncertainty: 2% were "kind of" clear, while 9% were not clear about the next steps.

Perception of the Feedback System:

- Positive Perception: 21% described the platform as "great," 33% as "Good," and 30% as "Kind of good."
- Negative Perception: 16% described the platform as "Not good."

Impact on Learning Progress:

- High Expectations: 84% believed that a system providing quick marks and feedback would help them progress faster.
- Scepticism: 4% were uncertain, and 12% did not think it would help them progress faster.

Overall Rating of the Platform:

- High Ratings: 12% rated the platform 5 out of 5, 33% gave it 4 out of 5, and 40% gave it 3 out of 5.
- Lower Ratings: 7% rated it 2 out of 5, and 9% rated it 1 out of 5.

Insights on Student Efficacy:

Positive Indicators:

- A majority of students found the feedback useful to varying degrees, with 67% rating it positively.
- Most students (77%) felt clear about the steps they needed to take to improve, suggesting that the feedback provided actionable insights.
- A significant majority (72%) were willing to use the feedback to improve their learning, indicating a positive attitude towards leveraging feedback for self-improvement.
- However, there was a rise in the number of students who could not see themselves using the system more from the start of the trial to the end. When questioned further about this in the focus groups it transpired that many initially replied in the ‘hypothetical’- “if such a system were available then it would be nice to use it”. Following the concrete experience of using such a system they came to question the general availability of such a system in their educational setting.

The students were also asked to contribute their thoughts about the platform to a Post-it note board as well as in the discussion groups. Taking into account the feedback from both, the strengths of Progressay appeared to revolve around efficiency, user-friendliness, clarity, accessibility, and the objectivity of the feedback provided.

Themes	Description
Efficiency and Speed	All of the students highlighted the speed and quick response of Progressay's feedback. This indicates that pupils appreciated the platform's efficiency in delivering feedback promptly.
User-Friendly Design	The pupils found the design of the platform easy to navigate which suggests that Progressay prioritises user experience and has a well-designed interface. Pupils suggested that both the content and the platform itself are easy to understand and use. Pupils really appreciated the simplicity and clarity of Progressay's approach.
Objective Feedback	The discussion indicated that the pupils appreciate the impersonal, objective nature of the feedback provided. They would rather it was harsher feedback as a result. It seems that the feedback Progressay platform provided was perceived as objective by the students without personal biases, which they viewed as a strength, valued for its fairness and impartiality.

From the more negative perspective the prevalent themes seem to be as follows:

Themes	Description
Lack of Feedback Specificity	Pupils expressed a desire for more detailed and precise guidance. They commented that the feedback was "too nice" and that they wished it to be more direct. Additionally, they expressed a need for clearer steps on how to improve rather than just being told what was good.

Technical Skills	There was extensive discussion regarding the digital divide, with some pupils displaying more confidence in their IT skills than others. The importance of technical proficiency and computer literacy when using such a system was emphasised as being essential.
Spell Check and Technical Features	Some users noted that the spell check function was not disabled. This resulted in all users scoring well in spelling, which might not be indicative of their true ability. Users expressed the importance of having control over this feature, with the option to enable or disable it, as they sometimes wished to assess their score or ability independently of spelling and grammar (SPAG) considerations.

In summary, the identified themes revolve around the quality and specificity of feedback, user interface design and readability, the importance of technical skills, and the availability of technical features.

Analysis

Overall, there are modest improvements to student attainment evident from the trial. However, **what is very interesting is that despite spending almost 66.21% less time per student on the summative assessment compared to the baseline assessment, learners still scored 4.40% more on average.**

The fact that learners spent less time at the end of the course compared to at the start could be used to show that engagement (and possibly efficacy) dropped as the course progressed. However, it could also be the case that less time was allocated by the teacher. That there are (and will always be) a multitude of uncontrollable variables when running trials in educational settings ought to be taken into account.

Course completion

More students completed the Baseline and Summative for the writing course compared to the Reading Course. One reason for this is because the students were asked to complete the writing course first.

Attainment - Time relationship

Reading:
Overall, 59% of students made progress over the duration of the reading course. As such, there appears to have been a positive trend in educational outcomes across these levels, suggesting that interventions or teaching methods used may be effective in enhancing student performance.

This is interesting considering the students spent less time on the Summative compared to the Baseline. **This perhaps could suggest that access to AI could obtain the same attainment whilst investing less study time.**

SEND Impact

There were 10 Students that identified as having Special Educational Needs and Disabilities (SEND) overall. Of these 10 students, only 6 actually completed the course. 4 of the 6 saw an improvement, (i.e 67%). Although the data is limited here, **there is some indication that SEND students benefit from access to the platform.**

Writing
For the Writing Course, 30% of students improved attainment across both Level 1 (L1) and Level 2 (L2). There were 10 Students that identified as having Special Educational Needs and Disabilities (SEND) overall. Of these 10 students, 8 completed the course (2 more than the reading course), only 2 of which went on to see an improvement, (i.e 25%). However, any progress data for the writing course should be contextualised against the fact that students had less time for the summative assessment compared to the baseline assessment. It should also be noted that there was far less training data available for the writing course compared to the reading course, which meant students had far less access to automarking and auto-feedback in this course.

Overall it seems SEND students did better on the Reading course compared to the Writing course. This is interesting considering there were more AI essays in the Reading compared to the writing requiring more from the students. It is possible that these students were the ones that spent longer on the platform and engaged more

fully in the process, however due to data restrictions we are unable to match these aspects up.

Another reason may also be related to the fact that most of the training data provided by NCFE related to reading. It should be noted that there were only 8 assignments in the entire writing course due to limitations around training data provided. It was therefore possible to create more assignments in the reading course compared to the writing course and as a result students were able to practise a greater variety of questions as there were more variables trained for. This meant that students had the potential to engage in the iterative process of learning, identifying more areas of weakness via their multiple submissions. This suggests that **the more assignments available in a course, the better the students will improve - a finding that applies to SEND students as much as it does those who have more 'normal' learning profiles.**

Engagement with feedback

The interviews with teachers and pupils indicated that some pupils had not chosen to access the feedback at all. The teachers in the focus group speculated that this was indicative of a lack of confidence on the students part. One intimated that the students would not read the feedback as they were, in her experience, fatalistic about their outcomes and lacked any sense that they had real control over their outcomes. There was fruitful discussion about whether the feedback being presented without any action on the students' part would be a positive move - i.e. if the comments simply appeared in popup as opposed to relying on the student to click the relevant button. The teachers generally agreed that it would be better than the pupils were faced with their outcomes in some form. The students remained divided.

In addition to this there was also a wider discussion about **the pupils feeling happier about the feedback from computer being negative as it "wouldn't be personal" -** there was a greater sense that the AI generated feedback would be commenting in a non emotional way about the work produced whereas there was potential for personal relationships (or lack of) to get in the way when the feedback is produced by a human. **There was less scepticism and fear about the production of AI feedback from the pupils than from the teachers.**

Research by Hattie and Timperley (2011) indicated that quick feedback provides students with a sense of control over their learning, it also makes the comments more relevant and actionable, as students can easily connect it to their recent efforts. **The majority of the participants in the focus group discussions liked the speed of the system in providing them with a response.** However, it should be noted

that although a large proportion of students surveyed reporting they wanted their feedback within a few days, it seemed that some students found the provision of immediate feedback quite hard to manage. There was discussion around this feeling that a task was never completed as the feedback would always suggest improvements. For some students being able to hand work in and forget about it for a few days until the marking came back was important. Student response to AI augmented feedback.

It was interesting to see that **the perceived 'distance' of the AI offering them feedback meant the students were more open to criticism.** For example many of them found the comments were too kind. They wanted the AI to be more direct with what went wrong. The impersonal nature of technological feedback arguably empowers individuals to interpret and utilise the information based on their own judgement and context. This autonomy in handling feedback fosters a sense of personal responsibility and control over one's learning or development process. Rather than feeling like a comment on the student personally, the increased sense of objectivity allows individuals to critically evaluate the feedback, discern its relevance. **By offering a neutral perspective, technological feedback has the potential to encourage a more self-directed and reflective approach to personal and professional growth.**

In summary, the identified themes from the focus groups revolve around the quality and specificity of feedback, user interface design and readability, the importance of technical skills, and the availability of technical features. It will therefore be important that the digital divide in schools/ colleges is addressed to ensure that all pupils have the technological know-how to be able to utilise these platforms. Simple access is not equitable enough.

Limitations of pilot

Reliance on pre-test vs post-test model

One shortcoming with the trial is certainly the reliance on pre-test vs post-test data, as this model unfortunately limits the overall reliability of the data, due to the fact that we cannot be sure to what extent, if any, that access to the computer-assisted marking and feedback actually impacted on the learner attainment and learner efficacy. There could have been other factors, not least the teacher's teaching, that contributed to the increase or decrease in learner attainment and learner efficacy. It should however be noted that we had initially set out to adopt the Randomised

Control Trial (RCT) model. However, due to the significant delays in securing participant schools, this was later replaced with the more accessible pre-test vs post-test model.

Inability to control variables

A lack of a uniform and consistent onboarding experience for learners meant that not all learners accessed the platform in the same way. Unfortunately, many learners did not complete every lesson or every question, opting in many cases to skip questions. Having more control of the environment would also improve reliability. For instance, we cannot even say for sure if the assigned learners answered questions on their logins as theoretically, a learner could have shared their login with another student etc.

Inability to maintain clear lines of communication with IoW College

Not being able to communicate with the participating college in a clear way unfortunately meant that of the three classes that had initially signed up, only one had really completed the course. Even getting this one class over the line was only possible through the help of direct support from NCFE.

Conclusion

Summary of findings in the research and recommended next steps:

Research question 1:

Student attainment was increased in the Reading paper by 56.7% and in the Writing paper by 30.4%.

It seems that when there was greater access to the Progressay AI augmented real-time feedback, students performed better. Students were happy with the platform with 86% of them rating it 3* or above (even with limited technical support on platform use) indicating that it was intuitive and easy to use.

Research question 2:

Fostering self-efficacy through mastery experiences can enhance students' academic self-beliefs and, consequently, their motivation and performance, and while this proved harder to measure, it seems that the speed of feedback, the level of student access to the comments, and the process of iterative improvement through low stakes testing, proved important in the learning process.

Therefore, the Progressay AI Augmented real time marking pilot, focusing on NCFE Functional Skills English Language Level 2 Reading and Writing, offers significant insights and implications for education.

Recommended next steps:

- AI Augmented real time marking available in schools at scale

One of the key findings from this pilot programme is the value to be given through the production of timely and constructive feedback at scale. Quick feedback on student work is often considered more effective for student improvement and is something that a vast number of teachers would like to be in a position to consistently provide. However, large class sizes, the logistics of course delivery, and the increasing pressures on teacher time often prevent this from being feasible.

The Progressay platform however offers formative feedback within a few moments and, as the wider research has shown, when feedback is delivered shortly after the student completes a task or assignment, the context and specific details of their work are still fresh in their minds. This makes such feedback more relevant and actionable, as students can easily connect it to their recent efforts. Moreover, Progressay provides feedback in a way that is far quicker and on a much larger scale this in turn allows for students to immediately address any mistakes or misconceptions highlighted and then apply the feedback to their current work. This appears to have been particularly effective for those studying at the lower level (L2). Although there was some reluctance reported by some students with regards to wanting feedback quickly - reporting a need to feel that they could step away from work as it being ‘finished for now’ - there is no denying that the speed of Progressay’s feedback can be seen to support an iterative learning process as students can submit multiple drafts or attempts, with the expectation of receiving feedback and improving their work over time. This iterative approach, promoting deeper understanding and mastery of the subject matter, is something that would be considered “best practice” for teachers however, is not practical within the current

working environment for many. Progressay is providing the mechanism within which active engagement with pupil written work is more easily facilitated thereby allowing teachers to maintain high standards of teaching and learning. This in turn contributes to the student's overall academic development as well as having a positive impact on motivation and engagement. For example, if students receive timely feedback, they see the direct results of their efforts and get recognition for their successes. This immediate reinforcement keeps them motivated and encourages continued investment in their learning journey.

Therefore, while it is important to note that none of the knowledge regarding efficacy, speed of feedback, mastery learning etc is not new, the Progressay tool seeks to maintain the high standards of teaching and learning using strongly established norms within education. Progressay is simply embracing the power of AI to protect and enhance the core function of teachers.

Furthermore, the study emphasises the importance of feedback quality and specificity. In the post trial survey students expressed a desire for more detailed and precise guidance, indicating that feedback should not merely be "nice" but also direct and actionable. The objectivity of the feedback provided by AI technology was appreciated, highlighting its value in fostering fairness and impartiality in the educational process.

Despite the potential benefits of such technology, the study also highlights challenges. The digital divide, with some students displaying more confidence in their IT skills than others, is a significant concern. It seems that those students who reported that there "was no feedback" or the "system was slow" in the post trial survey had not accessed the system correctly or properly submitted their work and it was this that had resulted in a lack of feedback. Furthermore, of those who reported that they did not see themselves using a system like Progressay it seems that some of the reasoning behind this was due to their belief that it would be expensive and therefore unavailable to schools more widely. Addressing this divide is essential to ensure equitable access to educational technology and opportunities.

There are those who worry that technology and AI in particular will "take teachers' jobs", however, this research underscores the vital role of teachers in nurturing students' self-efficacy beliefs through supportive feedback and creating a positive learning environment. While AI augmented technology offers valuable feedback, teachers continue to influence students' perceptions and their responses to feedback and are in no danger of being replaced. Teaching and learning remain core to Progressay - this platform has been specifically designed to allow for the

teacher/student focus to remain upon the provision of high quality provision of this rather than the technology itself.

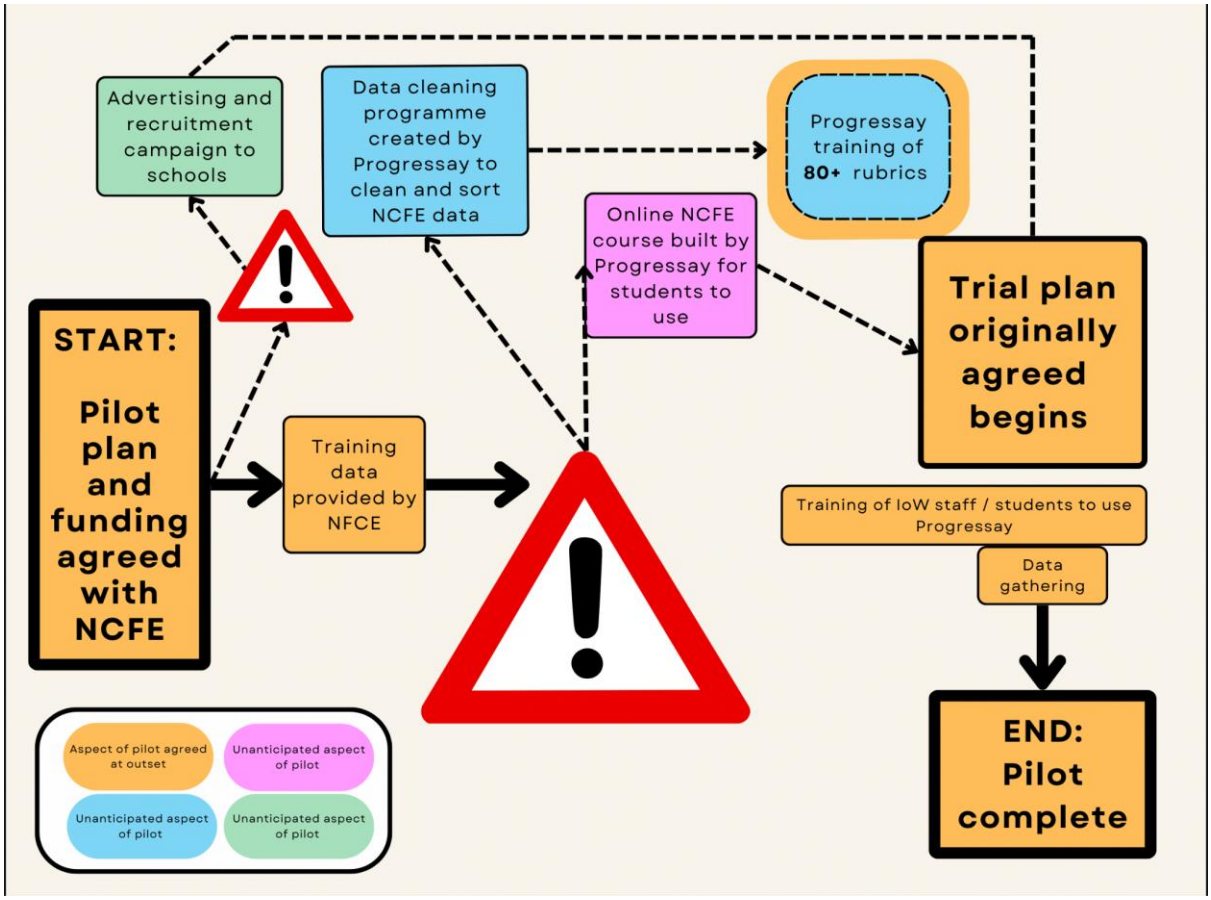
Following the pilot project taking place, it is clear from the lessons and findings from it that Progressay AI Augmented marking has the potential to positively impact education by enhancing self-efficacy, providing timely feedback, and promoting personalised learning experiences. The tool itself will continue to develop to ensure that teachers are provided with the most up to date AIEdtech developments from a feedback and assessment perspective. It is clear that teachers remain pivotal in supporting students' self-efficacy and it is vital that they are supported when ensuring effective use of AI technology in educational contexts. This research contributes to a growing body of knowledge on how AI can improve education and foster a culture of continuous improvement in teaching and learning.

Additional Results

Spill Over Effects

Before Progressay could start the Pilot a variety of additional, unplanned tasks needed to be completed. The outcomes of these are additional benefits for the NCFE.

It proved especially important to use an Agile methodology due to these issues.



While some can be described as ‘added bonuses’ the majority of these issues positioned themselves as obstacles that needed to be overcome before the pilot could begin. As such they have produced their own findings which feed into a much richer overall result for NCFE.

As a result, the trial produced several “Spillover Effects”

1) Marketing strategy for the recruitment of schools.
2) Building an automated data cleaning algorithm from scratch in order to make the 30,000 NCFE scripts readable for use in training.
3) Training of 80+ rubrics using the 30k+ training essays
4) Building two interactive online courses for the NCFE Level 2 Reading and for the NCFE Level 2 Writing
5) Testing for AQA in interim due to recruitment problems.

6) Testing at University on Level 4 student undertaking Education Studies degree (teachers of the future)
7) NCFE speaking to University Education undergrads about career pathways with NCFE and how the Innovation Fund works. This was a useful brand awareness exercise for NCFE.

Here are the findings from some of these aspects:

1) Marketing Strategy for the recruitment of schools

Having been awarded the funding the need to recruit NCFE schools / participants to the Progressay trial became clear. Therefore Progressay designed a series of marketing campaigns, utilising our educational networks and social media.

It proved incredibly difficult to recruit schools due to:

a) the school year timings (school holidays reduced the level of impetus as we had to wait for students to return in September)

b) Many schools remain sceptical / concerned about the use of AI in relation to academic work - a large proportion discourage their students from using it - therefore engaging in pilot trials was complicated in relation to school policies.
However, in response to this we co-developed an approach that drew on the NCFE existing relationships with some schools / providers. Eventually, through in-bound marketing we signed up IoW College.

Building an automated cleaning algorithm

While we were recruiting the schools ready for the live testing we needed to train the model on existing data. For this aspect over 30,000 training essays were provided by NCFE.

Initially however, the data provided was not viable due to the HTML format it was provided in. As such, we were first tasked with cleaning the data to remove the HTML code.

Example essay	Desired Output
<s id="1" ua="1" um="0"><c id="1" subtyp="0" typ="1"></c><c id="2" subtyp="0" typ="1"></c><c id="3" subtyp="0" typ="1"></c><c wei="1" id="4" ie="1" subtyp="0" typ="11" ua="1" um="0">Hello Sam, #!##!#I am writing you to thank you for planning our work trip. I think hosting staff day trip really helped to improve team spirit because people were able to talk to their friends and collegues without having to worry about deadlines or anything. #!##!#Given the positive feedback share amongst the teams, I think it would be beneficial to have these days often - lets say once every quater - as a way to reward people of their efforts and to encourage them to continue to work harder. #!##!#I think a few useful day trips would be: a team walk, a beach cleaning day or volunteering at a local charity. This was we build bonds within teams and the local community. #!##!#Thank you once again for hosting, and please let me know if you would like to discuss these ideas further. #!##!#Sincerely, #!##!#Margretsam.singh@zmail.bizComments on staff trip</c><c id="7" subtyp="0" typ="2"></c><c id="8" subtyp="0" typ="1"></c><c id="9" subtyp="0" typ="1"></c><c id="11" subtyp="0" typ="1"></c></s><s id="3" ua="0" um="0"><c id="1" subtyp="0" typ="1"></c></s><itemComment />	Hello Sam, I am writing you to thank you for planning our work trip. I think hosting staff day trip really helped to improve team spirit because people were able to talk to their friends and collegues without having to worry about deadlines or anything. Given the positive feedback share amongst the teams, I think it would be beneficial to have these days often - lets say once every quater - as a way to reward people of their efforts and to encourage them to continue to work harder. #!##!#I think a few useful day trips would be: a team walk, a beach cleaning day or volunteering at a local charity. This was we build bonds within teams and the local community. Thank you once again for hosting, and please let me know if you would like to discuss these ideas further. Sincerely, Margret

Steps for cleaning:

	Steps
step 1:	Find greeting at the start e.g "Hello Sam"
step 2:	If no greeting present, flag as no greeting
step 3:	If greeting present, delete everything before greeting
step 4:	Find sign-off phrase
step 5:	If no sign-off phrase- flag as "no sign-off phrase"
step 6:	If sign-off phrase present, cut off after that

step 7:	If sign-off present, after sign off extract name and email
step 8:	Remove buffer characters i.e

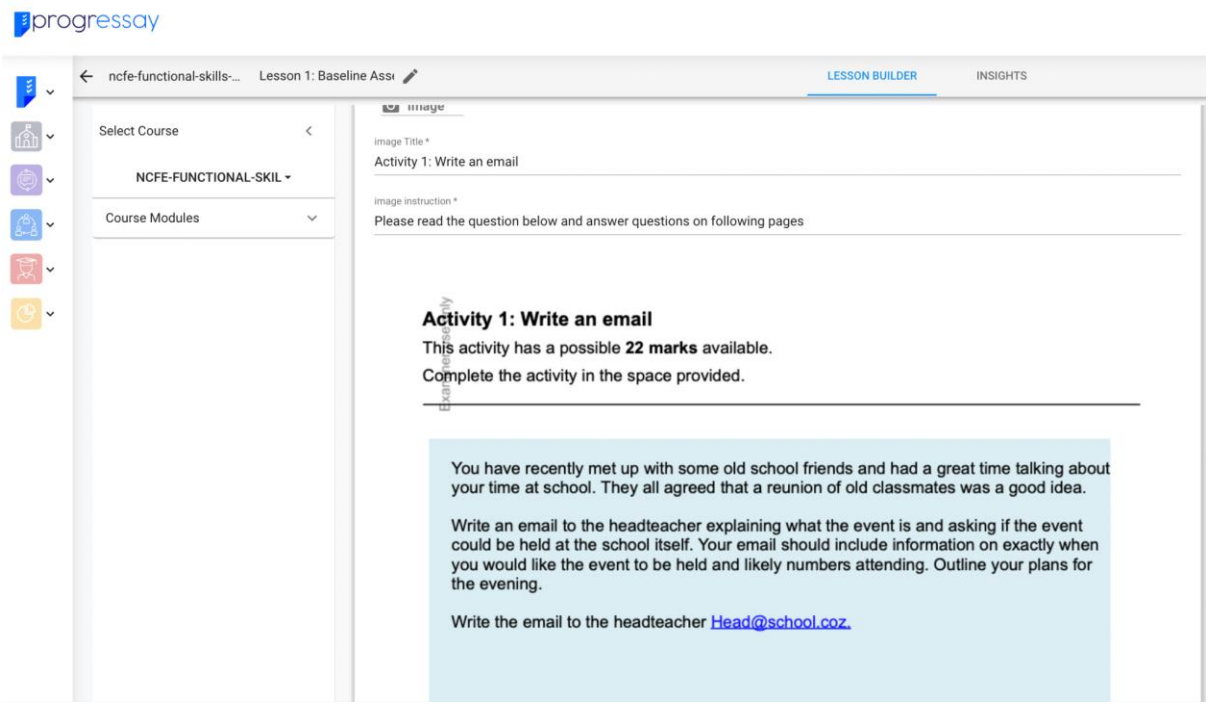
Essentially, Progressay built an automated data cleaning algorithm from scratch as a spin off from the agreed pilot. This was essential so as to make the 30,000 NCFE provided scripts readable for use in training.

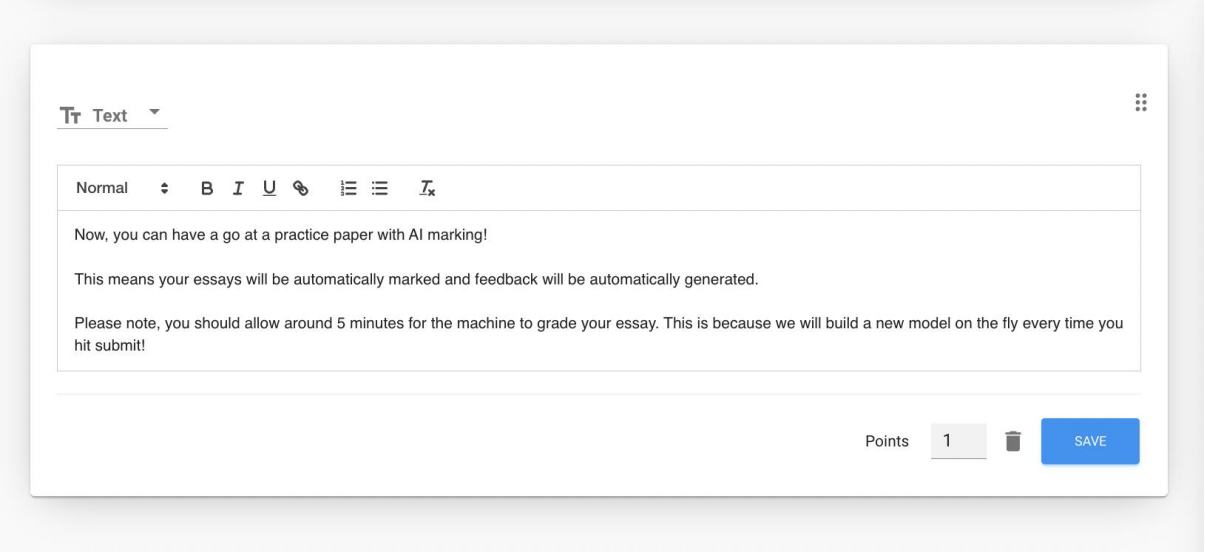
Design and build an online NCFE Level 2 Reading and Writing Course

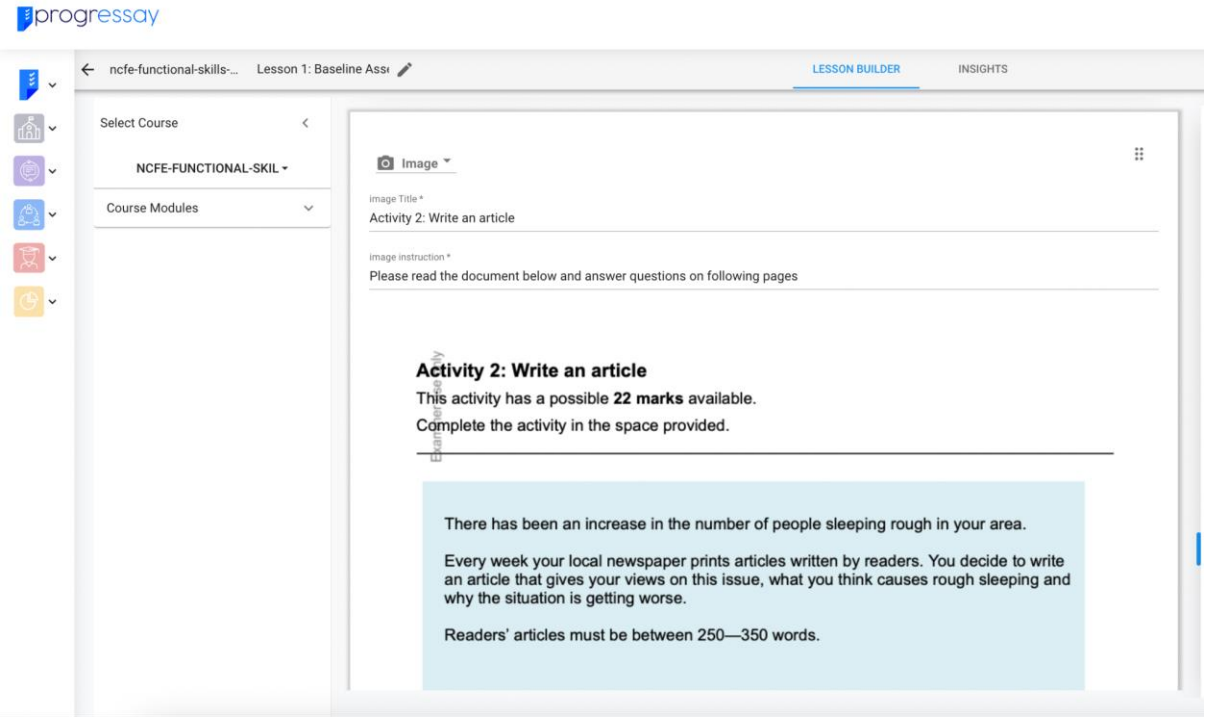
It quickly became apparent that there was no interface for the students to interact with the Progressay AI Augmented tool. Rather than 'plug in' to any existing NCFE online pages Progressay needed to build a online course. This was done using past paper structure and some activities. 10 Lessons were built containing a mixture of practice questions and activities. Here are some examples of the type of content they interacted with on the platform:

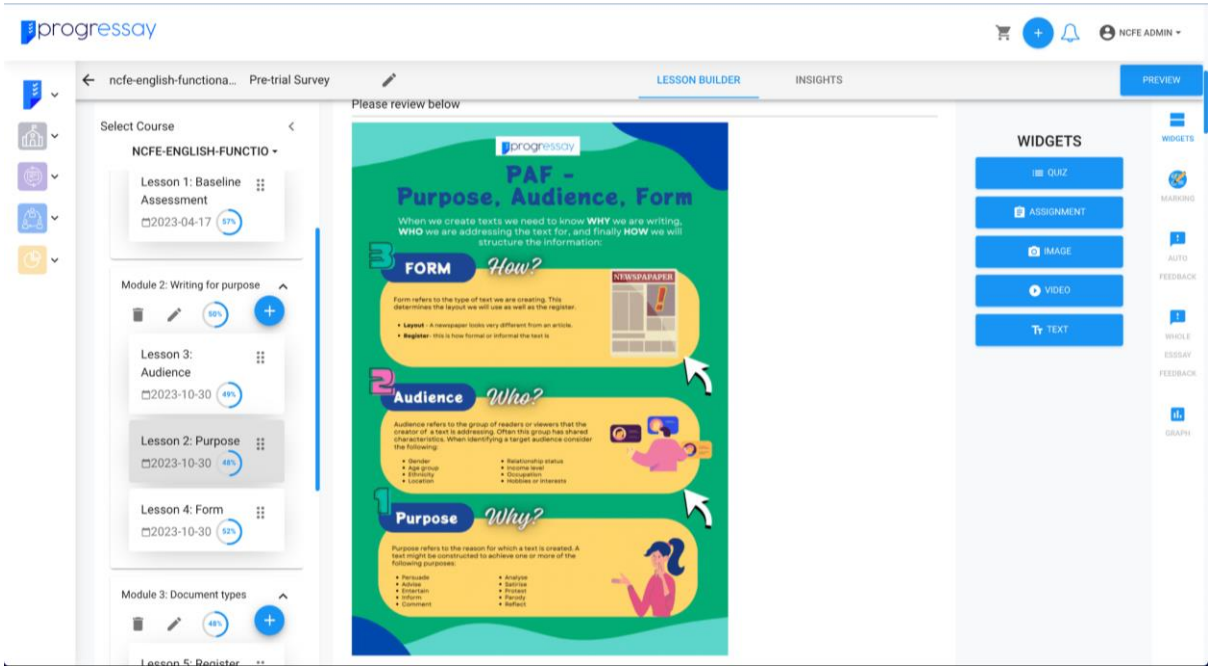
A lesson would consist of some content and then a formative assessment task or tasks.

Please review below




Text
Normal
Now, you can have a go at a practice paper with AI marking!
This means your essays will be automatically marked and feedback will be automatically generated.
Please note, you should allow around 5 minutes for the machine to grade your essay. This is because we will build a new model on the fly every time you hit submit!
Points
1
SAVE


progressay
ncfe-functional-skills-...
Lesson 1: Baseline Ass
LESSON BUILDER
INSIGHTS
Select Course
NCFE-FUNCTIONAL-SKIL
Course Modules
Image
Image Title *
Activity 2: Write an article
Image instruction *
Please read the document below and answer questions on following pages
Activity 2: Write an article
This activity has a possible 22 marks available.
Complete the activity in the space provided.
There has been an increase in the number of people sleeping rough in your area.
Every week your local newspaper prints articles written by readers. You decide to write an article that gives your views on this issue, what you think causes rough sleeping and why the situation is getting worse.
Readers' articles must be between 250—350 words.



Below is the structure of the online NCFE Level 2 Reading course that Progressay built to enable the pilot to go ahead:

Reading Course					
Module	**Lesson Number**	**Lesson Title**	**Paper Code**	**Level**	**Description**
1	0	Lesson 0			Pre-trial survey
	0	Lesson 0			Course Introduction
	1	Lesson 1: Baseline Assessment	P001274	L2	Lesson 1: Baseline Assessment (Level 2) -P001273
2	2	Lesson 2	P001272	L1	Lesson 2: What is the purpose of this text? (L2.2.11)
	3	Lesson 3			Lesson 3: Different types of texts (L2.2.11)
	4	Lesson 4			Lesson 4: Organisational features (L2.2.16)

3	5	Lesson 5			Lesson 5: Sources of Reference (L2.2.15)
	6	Lesson 6			Lesson 6: Relationship between features and meaning (L2.2.14)
	7	Lesson 7			Lesson 7: Fact vs Opinion
	8	Lesson 8			Lesson 8: Writer's style
4	9	Lesson 9			Lesson 9: Comparing texts and answering test questions
	10	Lesson 10			Lesson 10: Close reading for understanding
	11	Lesson 11			Lesson 11: Formal vs Informal
	12	Lesson 10 End of Unit Assessment	P001273	L2	Lesson 12: EoU Assessment (P001274)
	13				Post Trial survey

Here is the structure of the online Writing Course that was built by Progressay to enable to facilitate the pilot:

Writing Course					
Module	**Lesson Number**	**Lesson Title**	**Paper Code**	**Level**	**Description**
1	0	Lesson 0			Pre-trial survey
	0	Lesson 0			Course Introduction
	1	Baseline Assessment	P001277	L2	Activity 1
2	2	Lesson 2	P001277	L2	Activity 2
	3	Lesson 3			
	4	Lesson 4			
3	5	Lesson 5			
	6	Lesson 6	P001276	L1	Activity 2

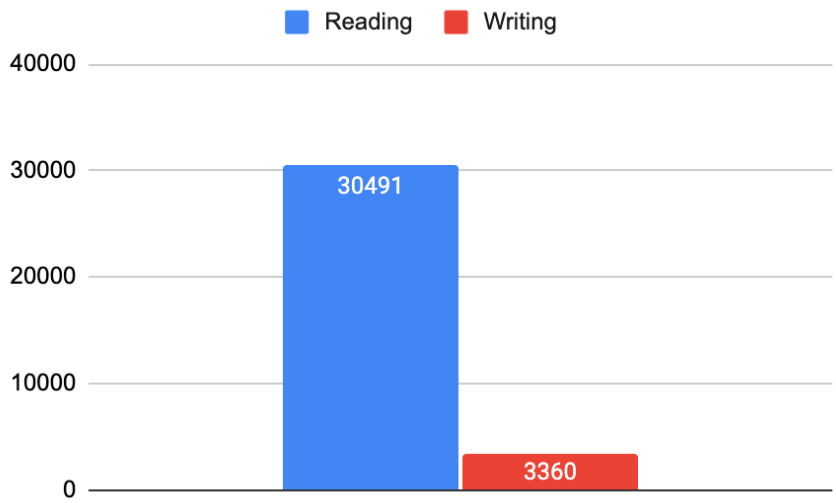
	7	Lesson 7	P001276	L1	Activity 1
	8	Lesson 8	P001275	L1	Activity 1
4	9	Lesson 9	P001275	L1	Activity 2
	10	End of Unit Assessment	P001278	L2	Activity 1
	11	End of Unit Assessment		L2	Activity 2
	12				Post Trial survey

Training data and building rubrics

As stated above in the section on data cleaning, over 30,000 training essays were provided by NCFE.These 30,000 training essays were comprised of reading and writing essays at both Level 1 and 2 as detailed below:

-

Training data provided



It is clear that there was far more training data for the Reading than for the Writing paper. We trained the following NCFE paper rubrics:

	Type	Level	Paper	Freq
1	Reading	L1 Functional Skills in English - Reading	P001271	5603
2	Reading	L1 Functional Skills in English - Reading	P001272	5512
5	Reading	L2 Functional Skills in English - Reading	P001273	11522
6	Reading	L2 Functional Skills in English - Reading	P001274	7854
			Total	30491

	Type	Level	Paper	Freq
3	Writing	L1 Functional Skills in English - Writing	P001275	550
4	Writing	L1 Functional Skills in English - Writing	P001276	696
7	Writing	L2 Functional Skills In English - Writing	P001277	932
8	Writing	L2 Functional Skills In English - Writing	P001278	1182
			Total	3360

Total	33851

Each paper contains a variety of questions - marks available range between 1 and 22. Each question type required discrete training to ensure that the grading was accurate and reflective.

Training Data by Question:

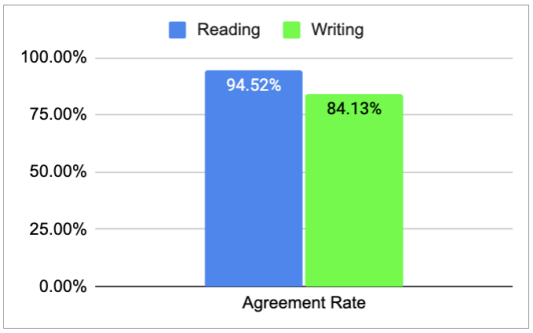
The amount of data provided for each course determined the type of questions we could build into each course.

As there was significantly more training data available for reading compared to writing, there are more AI essays in the reading course compared to the writing course:

	READING	WRITING
No of AI Assignments	39	7
% of Ai Assignments	75%	25%

Agreement rates

Overall there were high levels of computer-human agreement. There was higher agreement for the reading course compared to the writing course, likely due to the fact there was more training data related to reading than writing.



Below is a question by question breakdown of computer-human agreement:
(It should be noted that this pilot did not set out to prove agreement rate - it was focussed on efficacy and attainment. These exceptional results are a ‘spill over effect’).

Reading

Full Name	Level	Paper	QuestionNo	Type	Total Marks Available	Average Mark	Average Mark (%)	Rubric Link	Agreement Rate
Lesson 1	L2	P001273	Q1	Assignment	1	0.35	35.37%	https://progressay.com/rubric/629/?isPopup=false&step=1&mode=new	91%
	L2	P001273	Q2	Assignment	3	0.47	15.61%	https://progressay.com/rubric/630?isPopup=false&&step=1&mode=new	73.91%
	L2	P001273	Q3	Assignment	2	0.98	49.20%	https://progressay.com/rubric/1819/?isPopup=false&step=4&mode=new	100%
	L2	P001273	Q4	Assignment	1	0.46	45.62%	https://progressay.com/rubric/632/?isPopup=false&step=1&mode=new	100%
	L2	P001273	Q5	Assignment	1	0.32	31.65%	https://progressay.com/rubric/633/?mode=new	100%
	L2	P001273	Q6	Assignment	2	0.60	29.86%	https://progressay.com/rubric/927/?mode=new	100%
	L2	P001273	Q7	Assignment	2	0.19	9.69%	https://progressay.com/rubric/928/?mode=new	90.14%
	L2	P001273	Q8	Assignment	2	0.06	2.86%	https://progressay.com/rubric/929/?mode=new	98.77%
	L2	P001273	Q9	Assignment	1	0.12	12.27%	https://progressay.com/rubric/930/?mode=new	100%

	L2	P001273	Q10	Assignment	1	0.11	11.38%	https://progressay.com/rubric/931/?mode=new	100%
	L2	P001273	Q11	Quiz	1	0.44	43.51%	Quiz	Quiz
	L2	P001273	Q12	Assignment	2	0.43	21.72%	https://progressay.com/rubric/932/?mode=new	90.91%
	L2	P001273	Q13	Assignment	2	0.23	11.65%	https://progressay.com/rubric/933/?mode=new	91%
	L2	P001273	Q14	Assignment	2	0.31	15.58%	https://progressay.com/rubric/1073/?isPopup=true&mode=new	92%
	L2	P001273	Q15	Assignment	3	0.22	7.30%	https://progressay.com/rubric/934/?mode=new	85.57%
	L2	P001273	Q16	Assignment	3	0.11	3.72%	https://progressay.com/rubric/935/?mode=new	97%
Lesson 2	Quiz	Quiz		Quiz	1	0.59	59.46%	Quiz	91%
	L1	P001271	Q1	Assignment	1	0.54	54.32%	https://progressay.com/rubric/1054/?isPopup=false&step=4&mode=new	100%
Lesson 3	L1	P001271	Q19	Assignment	1	0.18	18.30%	https://progressay.com/rubric/1870/?isPopup=true&mode=new	100%
Lesson 4	L2	P001274	Q2	Assignment	1	0.08	7.89%	https://progressay.com/rubric/952/?isPopup=true&mode=new	100%
Lesson 5	L1	P001271	Q2	Assignment	1	0.54	54.32%	https://progressay.com/rubric/1055/?isPopup=true&mode=new	87.56%
Lesson 6	Quiz	Quiz		Quiz	1	0.00	0.14%	Quiz	Quiz
	L1	P001271	Q5	Assignment	2	0.25	12.51%	https://progressay.com/rubric/1058/?isPopup=true&mode=new	94.76%
Lesson 7	Quiz	Quiz	Quiz	Quiz	1	0.02	1.57%	Quiz	Quiz

	L1	P001271	Q4	Assignment	2	0.22	10.89%	https://progressay.com/rubric/1873/?isPopup=false&step=4&mode=new	100%
Lesson 8	L2	P001274	Q5	Assignment	2	0.51	25.46%	https://progressay.com/rubric/957/?isPopup=true&mode=new	92.58%
Lesson 9	L1	P001271	Q20	Assignment	3	0.30	10.11%	https://progressay.com/rubric/1079/?isPopup=true&mode=new	87.38%
Lesson 10	L1	P001271	Q3	Assignment	1	0.71	70.65%	https://progressay.com/rubric/1056/?isPopup=true&mode=new	86.53%
	L1	P001271	Q5	Assignment	2	0.16	7.92%	https://progressay.com/rubric/1058/?isPopup=true&mode=new	94.76%
Lesson 11	Quiz	Quiz	Quiz	Quiz	1	0.43	43.24%	Quiz	Quiz
	L1	P001271	Q19	Assignment	1	0.19	19.38%	https://progressay.com/rubric/1870/?isPopup=true&mode=new	100%
Lesson 12	L2	P001274	Q1	Assignment	1	0.28	28.23%	https://progressay.com/rubric/952/?isPopup=true&mode=new	100%
	L2	P001274	Q2						
	L2	P001274	Q3						
	L2	P001274	Q4	Quiz	1	0.31	30.83%	Quiz	Quiz
	L2	P001274	Q5	Assignment	2	0.56	28.22%	https://progressay.com/rubric/957/?isPopup=true&mode=new	92.58%
	L2	P001274	Q6 Part a	Assignment	1	0.17	16.84%	https://progressay.com/rubric/954/?isPopup=false&step=4&mode=new&processId=155	100%
	L2	P001274	Q6 Part b	Assignment	1	0.29	28.86%	https://progressay.com/rubric/955/?isPopup=false&step=4&mode=new	100%

	L2	P001274	Q7	Assignment	2	0.04	2.23%	https://progressay.com/rubric/1049/?mode=new	99%
	L2	P001274	Q8	Assignment	2	0.26	13.20%	https://progressay.com/rubric/1050/?isPopup=false&step=4&mode=new&processId=185	97%
	L2	P001274	Q9	Assignment	1	0.46	45.70%	https://progressay.com/rubric/1051/?isPopup=false&step=4&mode=new&processId=186	91%
	L2	P001274	Q10	Assignment	2	0.29	14.58%	https://progressay.com/rubric/956/?isPopup=false&step=4&mode=new&processId=157	97%
	L2	P001274	Q11	Assignment	2	0.96	47.88%	https://progressay.com/rubric/1053/?isPopup=false&step=4&mode=new	95%
	L2	P001274	Q12	Assignment	1	0.21	21.26%	https://progressay.com/rubric/1436/?isPopup=true&mode=new	91%
	L2	P001274	Q13	Assignment	4	0.27	6.87%	https://progressay.com/rubric/1872/?isPopup=true&mode=new	92.58%
	L2	P001274	Q14	Quiz	1	0.32	32.00%	Quiz	Quiz
	L2	P001274	Q15	Assignment	3	0.22	7.22%	https://progressay.com/rubric/1438/?mode=new	86%
	L2	P001274	Q16	Assignment	3	0.38	12.61%	https://progressay.com/rubric/1438/?isPopup=true&mode=new	95%

Rubrics Trained	Total Marks available	Total Marks scored	Total percentage of	Agreement Rate

			marks scored%		
46	77	15.17	23.47%		94.52%

Writing

Lesson Name	Level	Paper	QuestionNo	Type	Out of Mark	Avg Mark	Avg Mark (%)	Rubric Link	Agreement Rate
Lesson 1: Baseline Assessment	L2	P001277	Activity 1	Assignment	22	7.26	33.01%	https://progressay.com/rubric/1411/?isPopup=true&mode=new	79.80%
Lesson 1: Baseline Assessment	L2	P001277	Activity 2	Assignment	22	8.38	38.08%	https://progressay.com/rubric/1412/?isPopup=false&step=4&mode=new	83%
Lesson 6	L1	P001276	Activity 2	Assignment	20	4.93	24.65%	https://progressay.com/rubric/1403/?isPopup=true&mode=new	88.18%
Lesson 7	L1	P001275	Activity 1	Assignment	20	5.65	28.26%	https://progressay.com/rubric/1048/?isPopup=false&step=4&mode=new	88.71%
Lesson 8	L1	P001275	Activity 1	Assignment	20	5.02	25.08%	https://progressay.com/rubric/1048/?isPopup=true&mode=new	88.71%
Lesson 10: End of Unit Assessment	L2	P001278	Activity 1	Assignment	22	6.44	29.29%	https://progressay.com/rubric/1410/?isPopup=true&mode=new	84.85%

Lesson 10: End of Unit Assessment	L2	P001278	Activity 2	Assignment	22	5.03	22.87%	https://progressay.com/rubric/1406?isPopup=true&mode=new	75.26%

Rubrics Trained	Total Marks available	Total Marks scored	Total percentage of marks scored%	Average Agreement Score
8	148	42.71	28.75%	84.07%

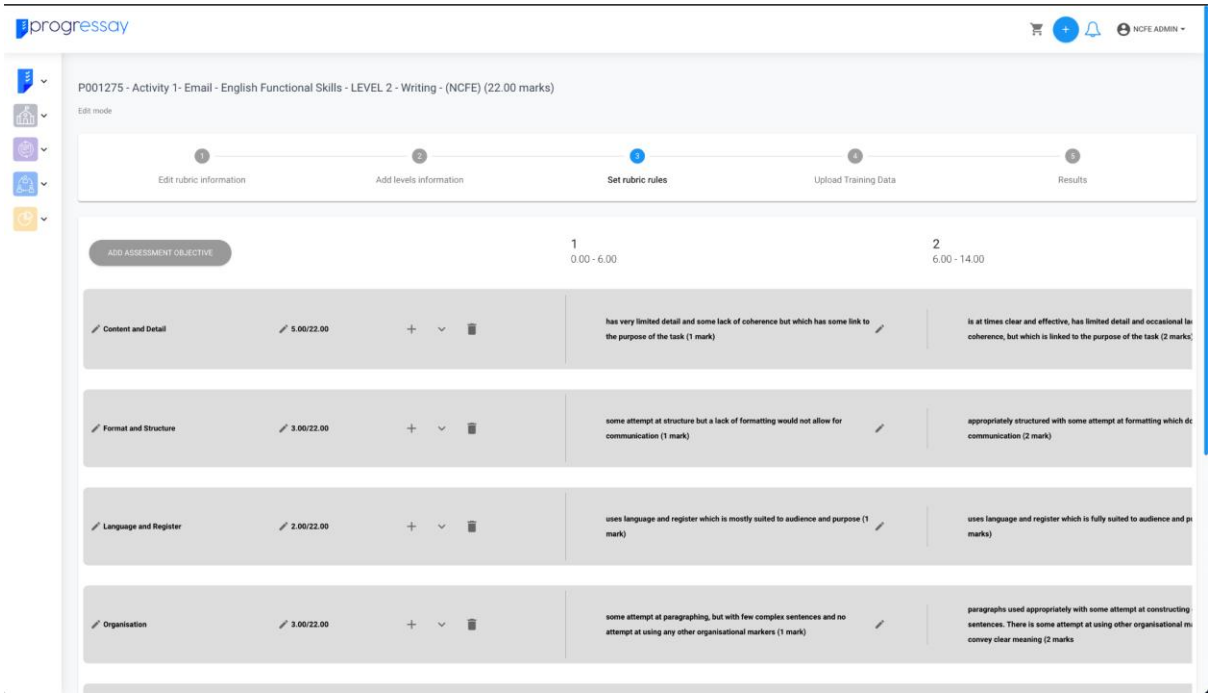
Rubric training - Writing - P001277 - Level 2

To illustrate how the AI rubrics were developed, below is information related to the following writing rubric in particular:

Writing - P001275 - Level 2 - Link	https://progressay.com/rubric/1942/?isPopup=false&step=4&mode=new&processId=447

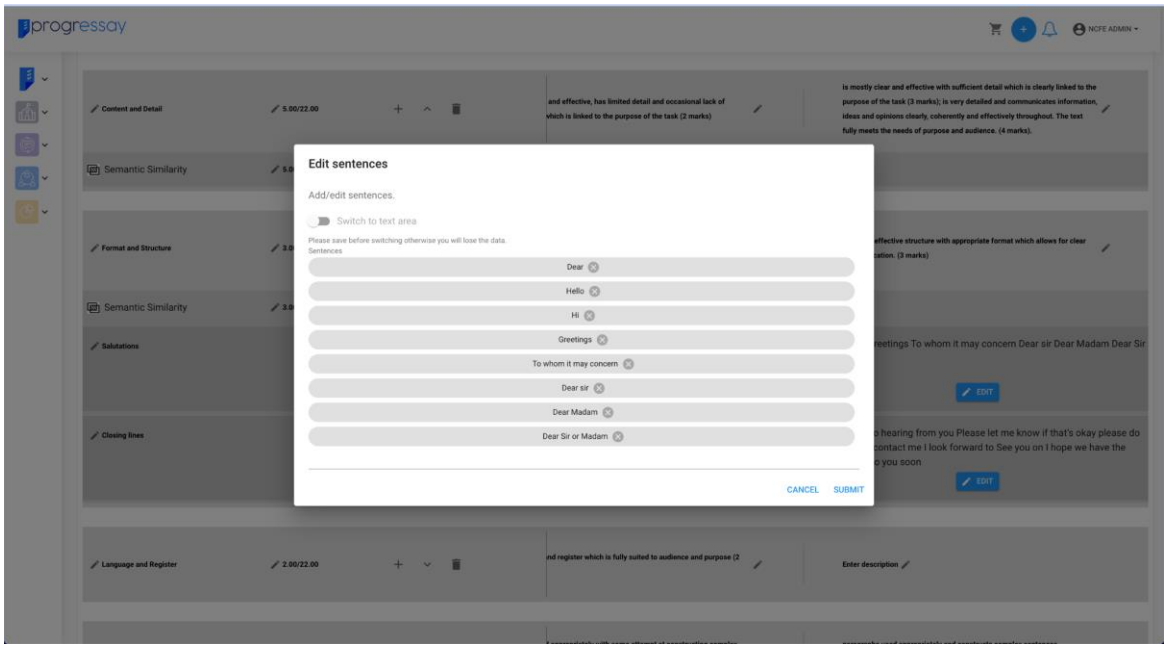
AI rubric - Customising rules

To allow for bespoke customisation of each rubric, our platform invites users to add the levels and descriptors for each rubric. A screenshot of this uI is shown below:



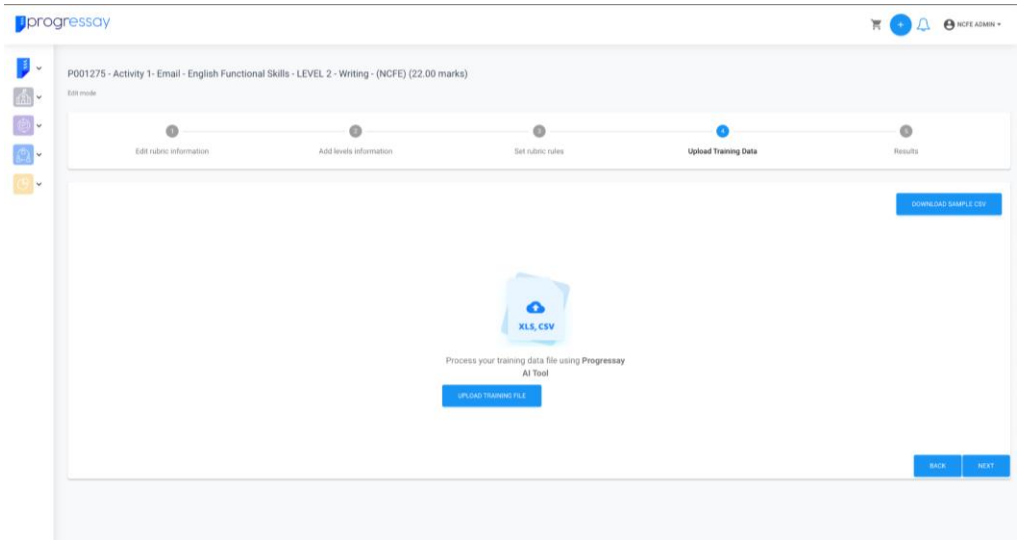
Rubric customisation - Indicative Content

Customisation can include inputting indicative content, bespoke to each question, which can include either keywords or phrases. This is seen below:



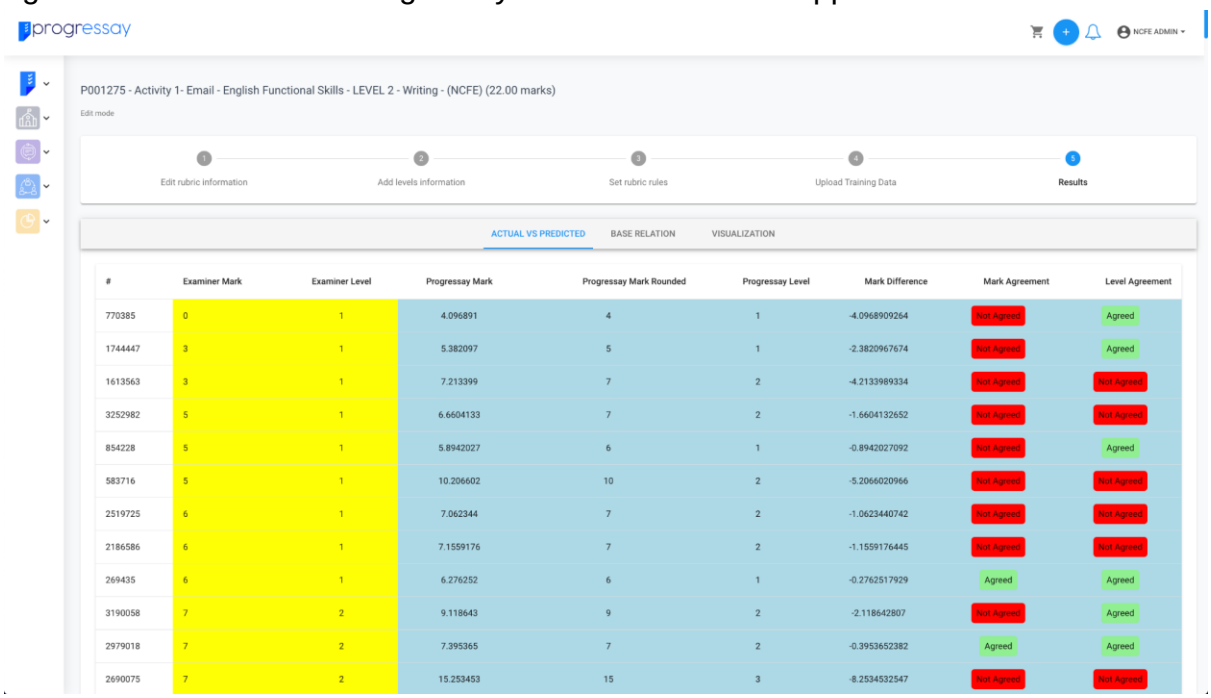
Uploading Training data

The main part of training an AI rubric, entails uploading pre-marked training essays, in the form of a CSV file. Here is a screenshot of the upload screen:



Training data Results

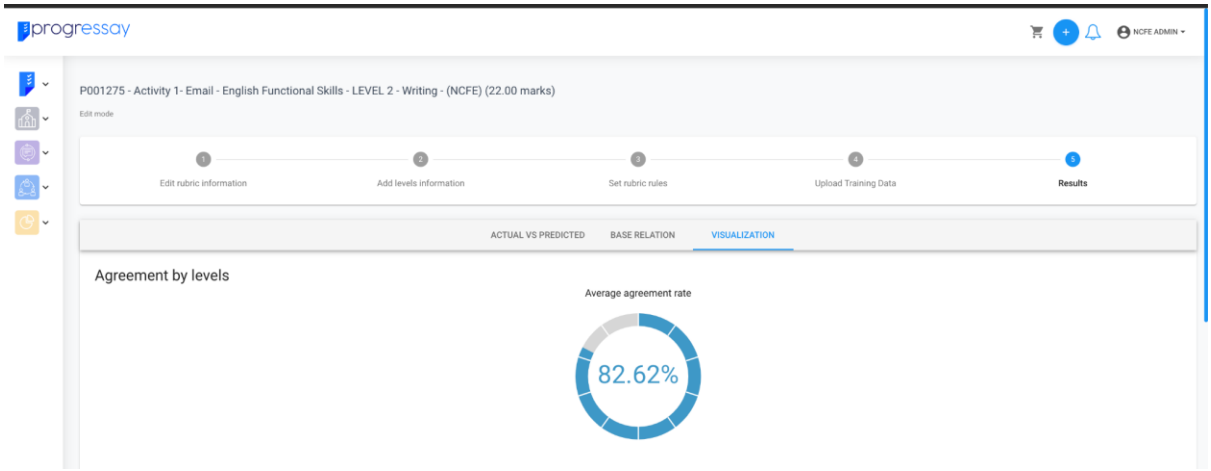
Once the training is complete (usually after 5 minutes or so), a table comparing the agreement between the Progressay and the human will appear:



#	Examiner Mark	Examiner Level	Progressay Mark	Progressay Mark Rounded	Progressay Level	Mark Difference	Mark Agreement	Level Agreement
770385	0	1	4.096891	4	1	-4.0968909264	Not Agreed	Agreed
1744447	3	1	5.382097	5	1	-2.3820967674	Not Agreed	Agreed
1613563	3	1	7.213399	7	2	-4.2133989334	Not Agreed	Not Agreed
3252982	5	1	6.6604133	7	2	-1.6604132652	Not Agreed	Not Agreed
854228	5	1	5.8942027	6	1	-0.8942027092	Not Agreed	Agreed
583716	5	1	10.206602	10	2	-5.2066020966	Not Agreed	Not Agreed
2519725	6	1	7.062344	7	2	-1.0623440742	Not Agreed	Not Agreed
2186586	6	1	7.1559176	7	2	-1.1559176445	Not Agreed	Not Agreed
269435	6	1	6.276252	6	1	-0.2762517929	Agreed	Agreed
3190058	7	2	9.118643	9	2	-2.118642807	Not Agreed	Agreed
2979018	7	2	7.395365	7	2	-0.3953652382	Agreed	Agreed
2690075	7	2	15.253453	15	3	-8.2534532547	Not Agreed	Not Agreed

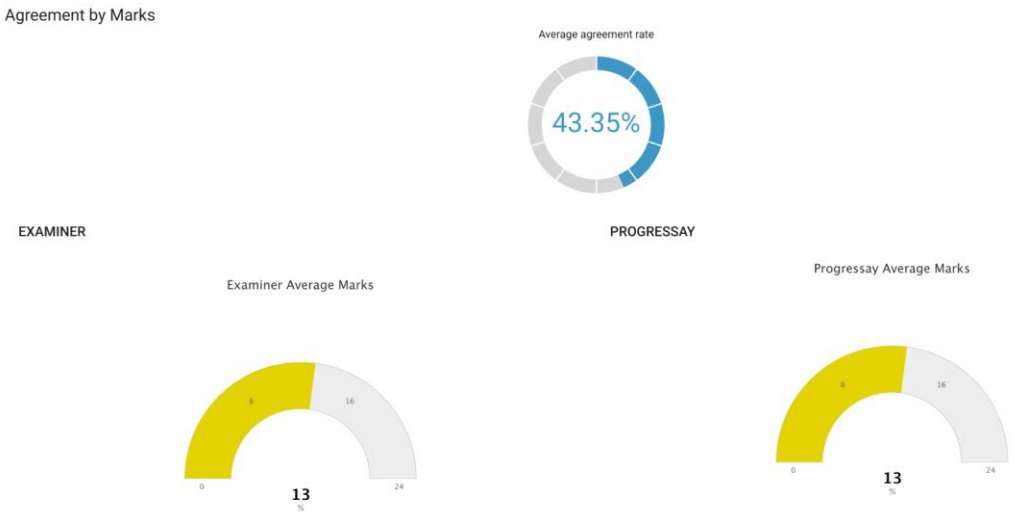
Computer-to-human Agreement rate- Levels

The table below underpins the “Computer-human Agreement rate”, presented under the "Visualisation” tab:

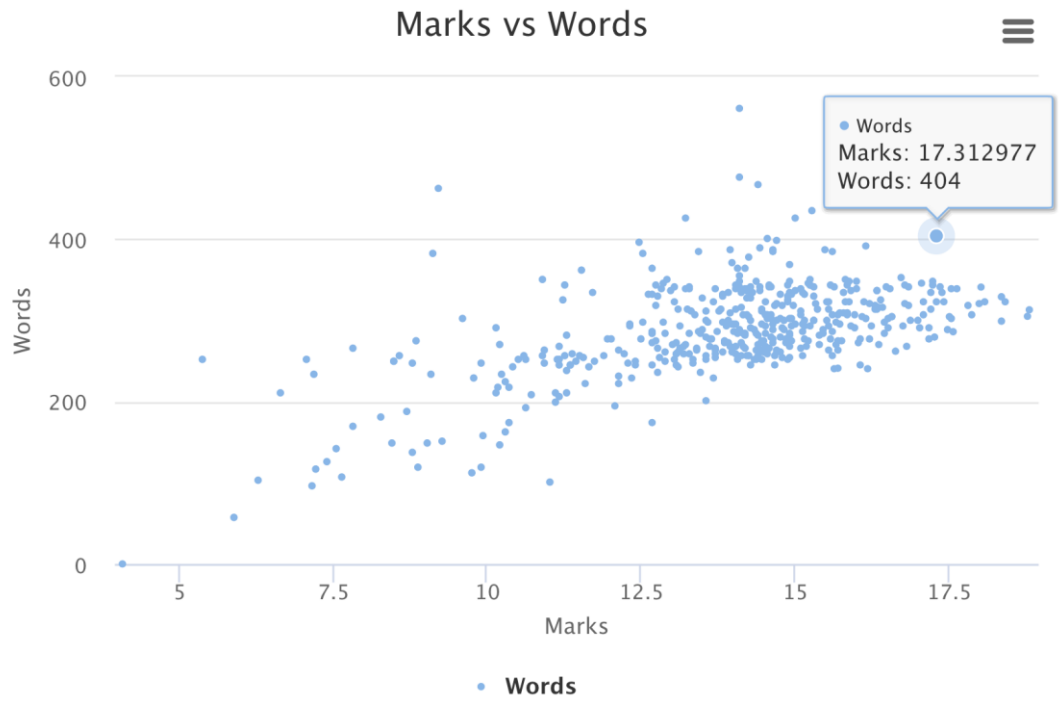


Computer-to-human Agreement rate- Levels

Agreement will also be presented in relation to marks, as seen below: In this case, there are 22 marks available so it makes sense that there is low agreement as this refers to matching by mark across the 22 mark scale. It is also worth noting, Progressay grades to 2 decimal places and as such, we round up if the mark different is equal to or above 0.5 and round down if the mark difference is less than 0.5 marks away.



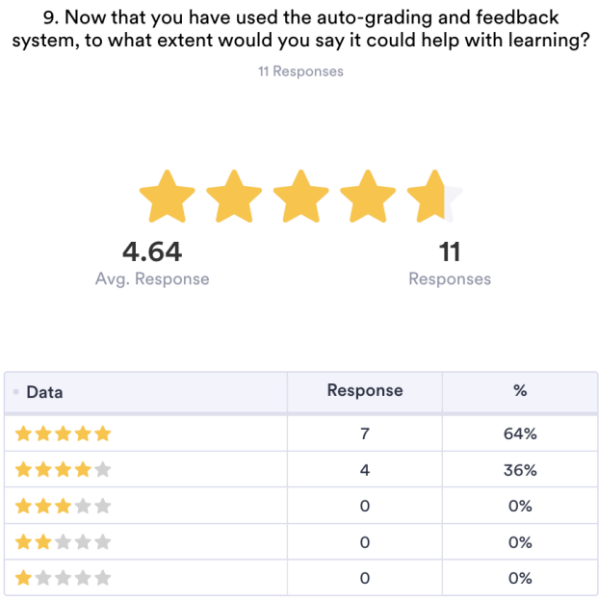
Text-Analytics for training data:



In addition, the platform presents text-analytics data to reveal patterns detected in the training data corpus. For example, this graph shows the correlation between marks scored out of 22 against words. There appears to be a positive correlation between the two, i.e typically, the more students write, the better they do in this question.

Tested for AQA GCSE English due to recruitment problems

As there were a number of issues with recruitment and we were struggling to find NCFE centres that were in a position to engage with the pilot we decided to use the time and run a pre-pilot tested on an AQA paper in the interim. Customer facing data drawn suggested high levels of satisfaction with the platform and the use of Progressay AI augmented real time marking.



Of the 11 in the class that responded 64% gave the Progressay's AI augmented real time marking 5 stars out of 5 with the remaining 4 who completed the survey gave it 4 stars. This gives an overall score of 4.64 stars for user satisfaction in the pre-pilot.

References

Edward L. Deci, Richard M. Ryan. (1985) *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum

Department for Education. (2023). *Generative artificial intelligence (AI) in education*. GOV.UK. https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Opitz B, Ferdinand NK, Mecklinger A. Timing matters: the impact of immediate and delayed feedback on artificial language learning. Front Hum Neurosci. 2011 Feb 1;5:8. doi: 10.3389/fnhum.2011.00008. PMID: 21344008; PMCID: PMC3034228.

Parliamentary Office of Science and Technology. (2023). *Artificial Intelligence in Education (POSTnote Number 712)*. https://post.parliament.uk/research-briefings/post-pn-0712/

Sharma, M. (2020). The role of self-efficacy in academic achievement. Journal of Educational Psychology, 112(3), 567-582. https://doi.org/10.1037/edu0000456[1]

Stenger, M. (2014, August 6). *5 research-based tips for providing students with meaningful feedback*. Edutopia. https://www.edutopia.org/blog/tips-providing-students-meaningful-feedback-marianne-stenger