A pilot impact study to test the effectiveness of FirstPass

Evaluation report

Dr Wayne Harrison and Dr John Brown

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.

BOLTON COLLEGE

FirstPass

Bolton College is one of the North West's largest providers of vocational training and further education. The college enables individuals, businesses and communities within Bolton and surrounding areas to reach their potential. They provide courses for a wide range of people of all different ages and are proud to be able to offer hundreds of courses to suit everyone.

What Worked Education

WhatWorked Education specialise in providing support to evaluate the effectiveness of interventions in the education sector using robust research designs.

For more information about NCFE or this report please contact:

Assessment Innovation Team

aif@ncfe.org.uk

www.ncfe.org.uk

Contents

Executive Summary

Project Summary

FirstPass is an innovative online platform designed to enhance the learning experience by allowing teachers to present open-ended questions and providing students with real-time feedback. Students can respond using free-form text, with answers ranging from a single sentence to multiple pages. As students compose their responses, FirstPass offers textual and graphical feedback through tick boxes, indicating whether the expected learned items are included. This immediate feedback enables students to reflect, edit, and improve their answers before submission.

One of the features of FirstPass is its use of Natural Language Classification (NLC). NLC technology allows the platform to analyse free-form text and accurately label sentences based on learned items. The system's accuracy improves with increased training data, enabling the creation of multiple natural language classifiers for different subjects. By providing real-time feedback, FirstPass reduces the time students typically wait for teacher feedback, fostering a more efficient and responsive learning environment.

The purpose of this pilot is to evaluate the effectiveness of FirstPass in enhancing learner attainment across different educational contexts. The pilot aims to assess the impact of FirstPass on learner attainment for the CA1 Child Development topic in the VCERT qualification, comparing its effectiveness as a supplement to traditional teaching methods. Additionally, the evaluation seeks to determine if the effect of FirstPass on attainment is greater for students eligible for pupil premium, whether there is a correlation between the impact on attainment and the amount of time students spend on the platform, and how the impact of FirstPass varies between different schools, teachers/classes, and courses.

The pilot also aims to evaluate the impact of FirstPass on learner attainment for the EYPS 5 Understanding How to Support Children's Development topic in the Level 2

Diploma qualification, comparing its effectiveness as a supplement to traditional teaching methods. The evaluation will seek to determine if there is a correlation between the impact on attainment and the amount of time students spend on the platform, and how the impact of FirstPass varies between different training providers.

The research utilised a randomised controlled trial (RCT) design, with learners randomly allocated at the class level for VCERT school delivery and at the individual level for diploma learners in training providers. Two distinct research protocols were developed for each context, but a common assessment was used for both evaluations due to the similarity in course content for the child development topic.

The primary outcome measures selected for the study were independent exam style questions for the NCFE CACHE Level 1/2 Technical Award in Child Development and Care in the Early Years. As the qualification is relatively new, with limited sample questions available to providers, NCFE commissioned the creation of an independent pre- and post-assessment for the CA1 Child Development content. The assessment consisted of one three-mark question and three six mark extended writing questions.

The sample for the VCERT evaluation included three schools and four classes of learners, with a sample size of fifty-nine KS4 learners. The sample for the diploma involved one training provider and twenty-one learners. The evaluation is designed to allow for further replications in the next academic year to increase the sample size and strengthen the evidence base.

The FirstPass platform was tested over a six-week period, with all learners completing the pre-assessment. Intervention learners were set nine tasks to complete on the platform linked to the topic. The post-assessment was then completed with both the control and intervention groups, with access to FirstPass enabled for the control learners after the completion of the post assessment

The pilot combined a light touch implementation and process evaluation, using online teacher surveys for both the control and intervention groups. A short online

questionnaire was sent to learners in the intervention group to receive user feedback and usage data was collected from the FirstPass platform on the completion of tasks.

Key conclusions

When deploying FirstPass with the VCERT cohort, students receiving First pass improved significantly more than those who did not, even after taking into account differences between the groups at baseline *ANCOVA (2,48) f=19.36 P<0.05*. On average, students receiving FirstPass improved 2.65 points more than controls.
The difference in post-test scores between those receiving first pass and controls indicates a small effect (Cohens' d 0.018, SE 0.29). The large standard error suggests that there is a large variation between individual scores.

A large, positive effect was observed in school B which reported a high fidelity of usage (ES 1.13, SE 0.41). However, due to the small sample size, caution must be used when interpreting these results.

There was no significant difference in improvement observed for students eligible for free school meals when compared to students who were not eligible for free school meals.

FirstPass shows promise as an effective feedback system, however the process evaluation highlights a few areas of consideration when implementing the platform. Firstly, ensuring variability in set tasks is key to retaining the interest and engagement of learners. In the six-week intervention period, teachers reported that the tasks had become monotonous, and this impacted learner engagement.

Secondly, the design of the platform could be improved to make this more user friendly for both learners and teachers. For example, teachers found marking difficult when moving from one student to another, increasing the time it takes to mark work.

When deploying FirstPass with the diploma cohort, the evaluation has demonstrated that it is difficult to engage teachers and learners in using the system. As the learners are working full time on an apprenticeship and the final assessment is not examination focused, the addition of the FirstPass component was overwhelming for students.

Feedback from teachers confirmed that this did not align with coursework and created additional work for learners. Consequently, before deploying FirstPass with learners in this context, identifying a clear need for how the system can be used should be considered.

Overall, the FirstPass platform has shown promise in how technology can be used to effectively support teachers in providing feedback to learners using an AI based system. Further replications of the evaluation will help to strengthen the claims that can be made regarding the effectiveness of the product.

Introduction

For a century the use of closed questioning techniques has monopolised the field of computer mediated assessment. Computers have excelled in supporting teachers to assess closed questions such as multiple choice, yes/no or drag and drop activities. Historically, computers have not been able to mediate semantic assessments of open-ended questions where students respond using free-form text answers. FirstPass has been designed to extend computer assessment support into the large domain of the meaning of written responses to open-ended questions.

FirstPass is an online platform that gives teachers the opportunity to present open-ended questions to their learners. Learners can respond using free-form text. Answers can be a single sentence in length or over multiple pages. As students compose their answers FirstPass offers real-time textual and graphical feedback. This feedback presents students with tick boxes, one for each learned item the teacher expects a good answer should include. When an answer includes one of the expected learned items, the box is ticked (coloured). This provides students with feedback on whether text includes expected learned items and the extent to which they have covered the items teachers expect. This feedback offers students the opportunity to reflect, edit and improve upon their answers before submitting them to their teachers. On receipt of the student work, teachers can add final commentary and feedback. The ability to offer learners effective real-time feedback from FirstPass reduces the time that learners would typically have to wait before they receive feedback from their teachers.

One of the key features of FirstPass is its use of Natural Language Classification (NLC). NLC enables a computer to analyse free-form text and associate with labels to the text that it has seen. FirstPass' ability to correctly label sentences improves as the volume of example sentences or training data rises. These sentences are stored in virtual files that are called natural language classifiers.

Multiple natural language classifiers can be created by teachers for different subject topics.

FirstPass is designed to enable teachers in different schools or colleges across the UK and further afield to train classifiers. In addition, as students respond to open-ended questions their responses can also be added to subject topic classifiers if they are fit for purpose. This means that FirstPass can operate with a high degree of accuracy as the volume of example sentences or training data rises as more teachers and students make use of the platform.

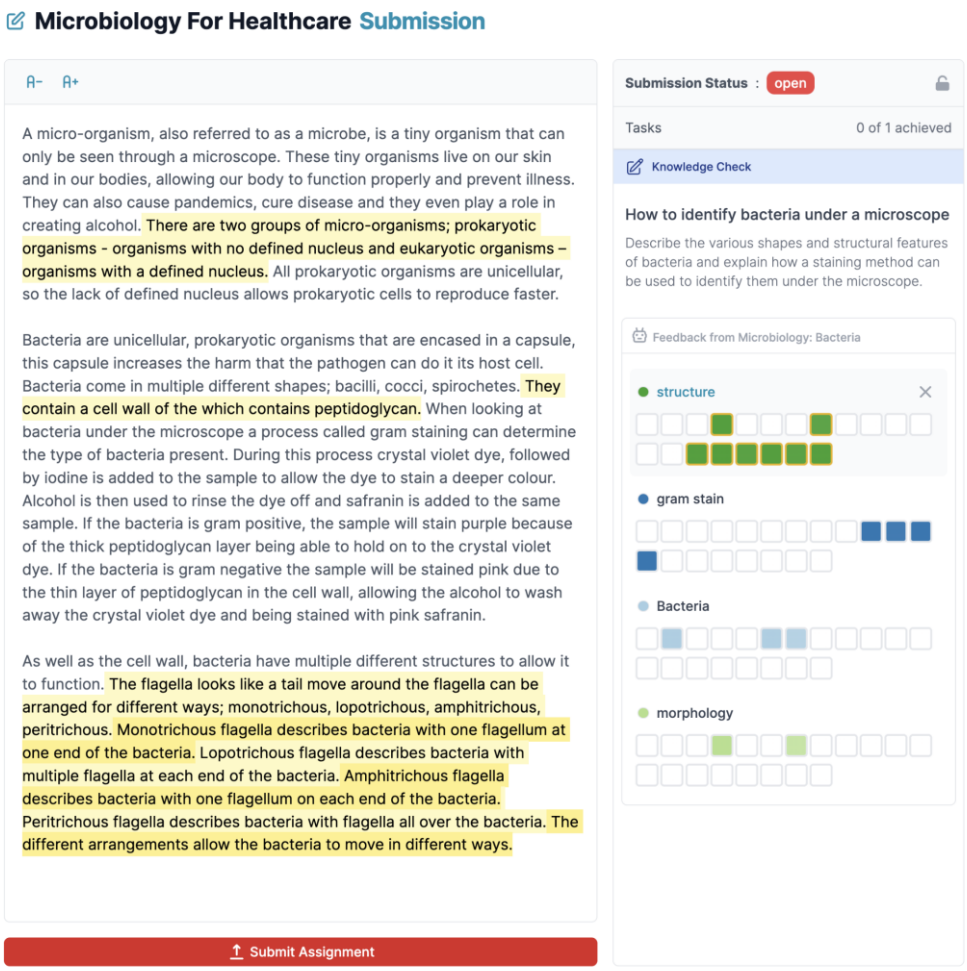


Figure 1. This screenshot shows how learners receive feedback in the FirstPass platform.

FirstPass uses NLC to automatically label text written by the learner. It can do this following the training of a subject topic classifier, which is a collection of pre-written and labelled text for one topic. When a teacher launches an assignment, he or she picks a set of subject topic classifiers from the library which tells FirstPass what elements to look for and label within learner work. Using NLC, FirstPass is then able to recognise when a learner provides accurate information on a topic using free form language, and instantly confirms this to the learner as they create their response. Consequently, this allows students to receive immediate feedback on whether they have included items expected by the teacher and the extent they have included all or most of the expected items.

The system does not provide students with guidance about the content of items expected by the teacher that have not yet been included in answers, thus FirstPass is a system for confirming when students have included expected learning. Consequently, FirstPass requires students to infer which learned items are expected, drawing upon their experience in the taught course, e.g., lecture notes, reading. Thus, the system seems likely to encourage students' attention to and retention of specific items of learning content.

One possible limitation of the system is the lack of scale of the educational value of specific expected learned items, where all items in FirstPass have equal weight or value in courses where certain items have more value than others. This issue seems likely to be addressed in the development of the project, for example by instructing students to minimally ensure the first classifier (the most important learning item) is always recognised as included by FirstPass before submission.

Subject topic classifiers can be pre-trained and saved in the library or created as needed by the teacher. New subject topic classifiers can be used by everyone to ensure that a library is quickly and efficiently built. The FirstPass development roadmap includes a rating system and quality assurance process for subject topic classifiers, making them even more usable for teachers.

For the current pilot, external experts recruited by NCFE will create the topic classifiers for the NCFE Level 1/2 Technical Award in Child Development and Care in the Early

Years (603/7012/9) Unit CA1 Child Development. This ensures FirstPass can operate with a high degree of accuracy as the classifiers underpinning the model are quality assured and not open for teachers to add their own classifiers.

Theory of change

The research evidence is consistent with the positive impacts of formative assessment and feedback on the academic attainment of students (Kingston and Nash, 2011; Black and William, 2009; EEF 2023). If high quality feedback is frequently provided by teachers to students, evidence shows that this improves attainment (Hattie, 2009; Hattie and Timperley, 2007). The FirstPass system is designed to provide timely and constructive feedback, confirming the inclusion of expected items of learning to students when they answer extended writing questions.

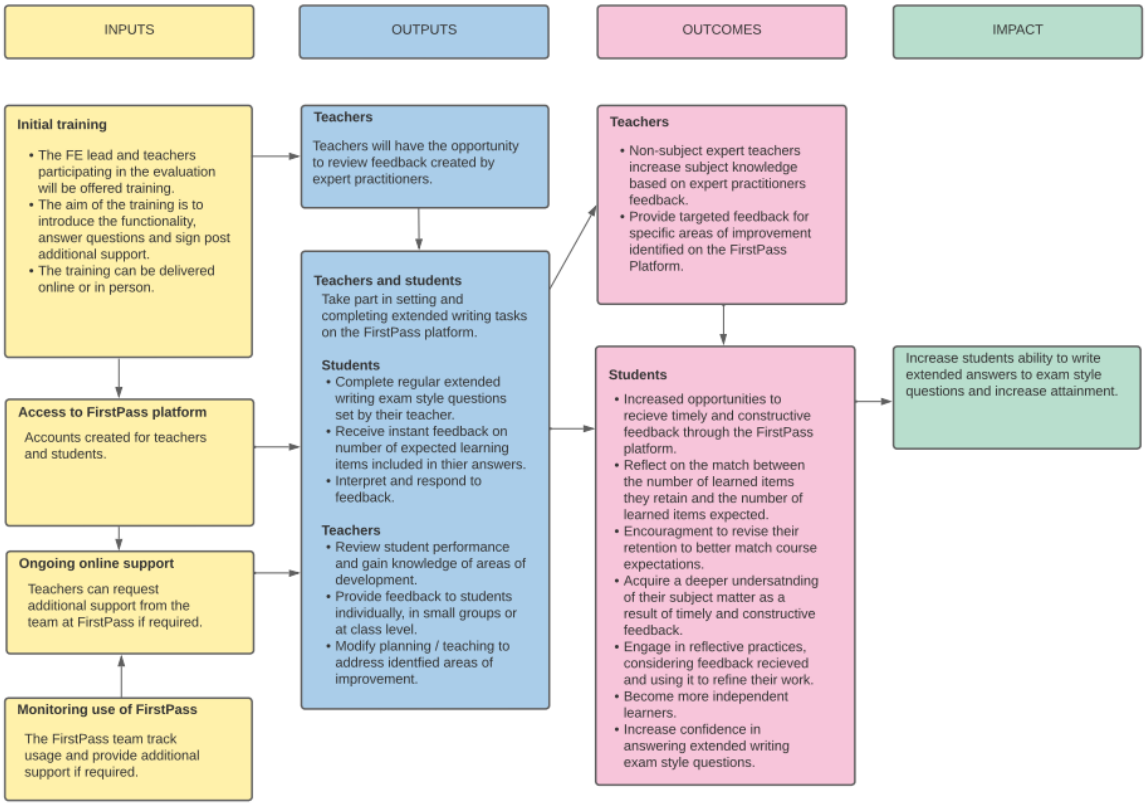
In this specific pilot, the FirstPass system uses classifiers created by subject matter experts from the childcare sector and is aligned to two NCFE childcare qualifications. Consequently, the platform is expected to provide real-time high-quality feedback interactions for learners with prompts to help learners include all or most of the learning items expected to exam style extended writing questions.

FirstPass aims to provide teachers and students with these benefits:

- Saving time for teachers by providing constructive feedback on written assessment tasks.
- Increasing opportunities for students to receive timely and constructive feedback, improving their learning experience.
- Enabling students to acquire a deeper understanding of their subject matter because of timely and constructive feedback on the extent to which their answers include expected learning.
- Supporting students to engage in reflective practices, encouraging them to self-identify the extent their retention of learned items matches the expectations of a course, using this to refine their work.

These elements of the FirstPass platform are anticipated to translate into higher quality feedback interactions and lead to improved attainment in extended writing exam style questions. The theory of change in figure 2 represents the inputs, outputs, outcomes, and impact explaining the rationale for the intervention.

Figure 2. Logic model for the proposed theory of change



Significance

To our knowledge FirstPass is the first and only automatic assessment and feedback system to offer educators the ability to guide students' free written open-ended responses to learning courses. As free writing is the main form of assessment for most higher education worldwide, the potential significance of FirstPass to add value by improving educational efficiency seems large.

Through the Assessment Innovation Fund, NCFE agreed to provide a grant to Bolton College to validate the benefits of using the FirstPass platform for NCFE learners and their teachers. This project began in August 2023 and finished in July 2024. The project had three objectives:

1. To complete the development of the FirstPass MVP.
2. To evaluate the ability of the FirstPass MVP to add value to NCFE customers.
3. To complete the development of the road mapped items.

This report assesses the 2nd objective, to evaluate the value to learners and educators of using the platform to deliver formative feedback across two NCFE qualifications to support the VCERT Child Development and Level 2 Diploma for Early Years Practitioner qualifications.

Research Methodology

The main report will outline the research methodology for the VCERT pilot evaluation, Appendix A includes the research methodology for the Diploma evaluation.

Research questions

The impact evaluation is designed to answer the following questions:

What effect does FirstPass have on the attainment of learners for the CA1 Child development topic in the VCERT qualification as a supplement to teaching compared to teaching as usual?

In addition to the primary research question, as the data aggregates for the evidence base for the intervention, secondary research questions include:

- Is the effect of FirstPass on attainment higher for students eligible for pupil premium?
- Is there a correlation between impact on attainment and the amount of time students spend on the platform?
- To what extent does the impact of FirstPass vary between schools, teachers/classes and courses?

The process evaluation component will address the following questions:

- How is the intervention implemented? What are the enablers and barriers to implementation of the programme? Is implementing the programme feasible?
- What constitutes ‘usual practice’ in the intervention and control schools, and does this change over the duration of the trial? Are control classes using similar interventions to FirstPass that might be considered close substitutes?

- How and why does the implementation of the programme vary? To what extent does any variability affect the achievement of expected outcomes?

Design

This study was designed as a randomised controlled trial with classes assigned at random to either intervention (exposure to FirstPass) or control conditions (non-exposure to FirstPass) on a 1:1 basis.

Randomisation

Randomisation used a tool created by WhatWorked Education to independently randomly allocate classes and record the process. This was anonymised to comply with data sharing agreements.

Participants

Schools were identified using NCFE's own database of schools who were implementing the NCFE Level 1/2 Technical Award in Child Development and Care in the Early Years (603/7012/9) qualification and approached to determine their interest in participating in the study.

The course is delivered at Key Stage 4 to students aged 14 to 16 years in UK schools. It was expected each school would only provide one class of students to participate in the evaluation, based on estimated class sizes from the uptake of the qualification.

Sample

The pilot aimed to recruit six schools with a class of students (15 learners per class), giving an initial sample size of 90 learners. For context, the ESSA standards in the USA require a sample size of 350 for robust studies. The design of the evaluation

allows for future replications that will increase the sample size as more schools participate in the evaluation.

Initially six schools expressed an interest to participate, with nine classes of students planned to participate in the evaluation. However, two schools withdrew, one prior to starting and one in the first few weeks of the project. The final sample involved three schools and fifty-nine learners.

Table 1: School level characteristics

Anonymous School ID	Progress 8 score	% Free school meal (FSM)	% English not first language	Type of school
A	0.18	11.1	5.3	State-funded secondary
B	-0.35	21	0	State-funded secondary
C	n/a	47	0	State-funded secondary

Outcome measures

The primary outcome measure selected for the study were independent exam style questions for the NCFE CACHE Level 1/2 Technical Award in Child Development and Care in the Early Years CA1 Child Development.

As the qualification was relatively new, with limited sample questions available to providers, NCFE commissioned the creation of an independent pre- and post-assessment for the CA1 Child Development content. The assessment consisted of one three-mark question and three six mark extended writing questions. The maximum assessment score was 21 marks. Learners completed the assessments under exam style conditions with a duration of twenty-one minutes for the assessment.

A mark scheme was provided for teachers to assess the pre- and post-assessment, with scores recorded in a pre-specified CSV file. The anonymised data was then shared with the evaluation team.

Analysis plan

Statistical analysis used Analysis of Covariance (ANCOVA) to examine whether there are probabilistic and reliable differences in scores between the treatment and control group, adjusting for any differences in pre-test scores between the treatment and control groups. This adjustment was carried out to test the assumption that the groups might perform differently because they contain students with different levels of aptitude, despite our attempts to ensure a fair comparison through random assignment.

Analysis used the same techniques to examine whether some schools gain more from using FirstPass than others and relating this to the process evaluation to see if schools' reports of how they use FirstPass can shed light on how FirstPass can be used for greatest effect.

The relationship between the usage of FirstPass and assessment scores will also be examined to inform understanding of optimal use.

Implementation and process evaluation

The evaluation team conducted a light touch process evaluation to assess the implementation and feasibility of the evaluation. The process evaluation provided an opportunity to understand any variation in the intervention across the trial schools and how any variation may have affected the primary outcome measure.

The process evaluation will attempt to assess:

- Fidelity – has the teacher delivered the programme as intended by the developer?
- Dosage – How much of the intervention has been delivered?
- Reach – What is the reach and scope of participation?

- Monitoring of control groups – What other interventions are currently used in these classes?

Methodology

The process evaluation will adopt the following methods for data collection:

- Online teacher surveys
- Analysis of data from the FirstPass data system.

To ensure the anonymity of schools, students, and teachers as outlined by NCFE in the MoU agreement, the evaluators provided the online survey questions, and these were sent to participating schools by the NCFE team. The responses were saved into a shared data folder between NCFE and the evaluation team with identifiable information anonymised.

As the study was a pilot and with limited timescales involved in recruitment, a pragmatic approach to the process evaluation was adopted. To reduce administrative pressures for teachers involved in the pilot, only one survey was delivered towards the end of the programme.

Process evaluation timeline

The following timeline was planned for the main activities that formed the process evaluation:

Table 2 Process evaluation timeline.

Date	**Activity**
October 2023	Online survey designed.

January 2024	Schools recruited start the pilot.
February – April 2024	Online survey completed by control class teachers.
February – April 2024	Online survey completed by the intervention class teachers.
April 2024	Analysis of data from the FirstPass system.

Ethics and data protection

Any risk of harm was mitigated by the design (e.g., wait list) and learners were undertaking teaching and learning activities that are typical in participating schools. As all data was anonymised, head teacher school consent for the evaluation is sufficient. Data was collected on the attainment scores for the pre and post-assessment and students' use of the FirstPass platform. Additional data included some socio-demographic information (for example pupil premium eligibility, which the school provide to NCFE). Student data was not linked to the National Pupil Database.

Legal basis for processing

The impact evaluation complied with the GDPR legal basis for processing personal data as this research project was in the public interest. Data sharing was necessary for the parties to undertake a research project into the effectiveness of the FirstPass platform aimed at students in Key Stage 4. This project was in the public's interest as the results will help to assess the impact of the FirstPass platform on student achievement. The FirstPass platform is designed to deliver formative feedback for extended writing. The collection and sharing of data from students participating in the research project was necessary for the parties to evaluate the effectiveness and impact of FirstPass on students' attainment and achievement in NCFE CACHE Level 1/2 Technical Award in Child Development and Care. In addition, the collection and sharing of student data and surveys from teachers participating in the project is

necessary to assess the process evaluation. All data was anonymised for pupils and schools before this was shared with the WhatWorked evaluation team.

Appendix B includes the personnel and project timelines.

Impact Analysis

VCERT

Data analysis

Students without post-test results, due to absence from the test, were excluded from analysis. Seven (12%) of 59 original students were excluded on this basis. Two participants had no pretest score but did provide post-test scores. To maximise the amount of data to support interpretations these students were included in analysis by interpolating their pretest scores as the average of all pretest scores, as is normal with replacement of missing data.

Results

Overall, students receiving First pass improved significantly more than those that did not, even after taking into account differences between the groups at baseline *ANCOVA (2,48) f=19.36 P<0.05*. On average, students receiving FirstPass improved 2.65 points more than controls (see table 3).

Table 3. The average pretest and post-test scores for all schools

			Pretest		Post-test		Gain
Condition	Free School Meals	Usage %	Average	sd	Average	sd	Average
Control	0.48	0.00	6.24	4.05	7.29	3.99	1.05
Intervention	0.43	39.99	4.43	3.23	8.13	4.85	3.70

Examining individual schools, School B improved most on average (4.47 points). However, this figure may be misleading as all the students used FirstPass and so improvements may be due to other practises in this school. Importantly, students in School C who received first pass improved significantly more than students in the

same school who did not, even after accounting for differences at baseline *ANCOVA (2,21) f=85.83 p<0.05*.

Table 4. The average pretest and post-test scores by schools

				Pretest		Post-test		Gain
Condition	School	Free School Meals	Usage %	Average	sd	Average	sd	Average
	A	0.42	0.00	7.42	4.93	8.50	4.46	1.08
Control	C	0.56	0.00	4.67	1.66	5.67	2.69	1.00
	B	0.33	72.58	3.67	2.61	8.13	4.93	4.47
Intervention	C	0.53	7.40	5.20	3.69	8.13	4.94	2.93

School B had a high level of fidelity, with 72.58% completion rates for learners submitting tasks on the FirstPass platform. The effect size for School B comparing pretest to post-test is large as we would expect based on the averages (ES 1.13 SE 0.41).

Pupil Premium/ Free school meals

No significant difference in improvement was observed between students eligible for free school meals and students who were not eligible for free school meals.

Effect Size

The size of the differences in post-test between those receiving first pass and controls indicate a small effect (Cohens' d 0.018 SE 0.29). The large standard error indicates that there is a large difference between individual scores.

Attrition

In the sample of 59 learners, seven learners did not complete the pre and post-assessment. This attrition rate is relatively low at 11.8%.

Diploma

Data analysis

Twenty-one students initially took part in the study. Only twelve completed post-tests, of these nine received first pass and three were in the control group. As only three individuals were included in the comparison group, it is highly likely individual differences between these students would have a very large effect on the average for the group as a whole. Averages of this group are likely to reflect the individuals involved and unlikely to represent averages in the population the study aims to draw conclusions about, therefore analysis is likely to be misleading. Consequently, inferential analysis was not carried out.

Table 5. Average pretest and post-test for those receive first pass and comparison students

		Pretest		Post-test		Gain
Condition	usage	Average	sd	Average	sd	Average
Control	~	10.33	2.52	14.00	3.46	3.67
Intervention	29.3	10.33	3.35	15.11	1.83	4.78

Attrition

In the initial sample of 21 students in the training provider, 12 completed the pre and post assessments. The high rate of attrition in this group (43%) highlights the challenge in implementing the FirstPass platform with these learners.

Process evaluation

The process and implementation evaluation for this research was designed to be minimally intrusive, employing online surveys as the primary means of data collection. Intervention and control teachers were requested to complete a brief survey upon completion of the programme. In addition, a short online survey was sent to learners in the intervention group to capture feedback on the FirstPass platform.

The process evaluation includes feedback for both the VCERT and diploma responses to provide insight into platform implementation in both contexts.

The first process evaluation questions focused on the implementation of the FirstPass Platform.

How is the intervention implemented? What are the enablers and barriers to implementation of the programme? Is implementing the programme feasible?

How is the Intervention Implemented?

The implementation of the FirstPass platform involved a training session focused on how FirstPass works, how to set up the platform and an overview of the research for the evaluation. The implementation of FirstPass can be split into the initial training session, set up and integration.

Training session

All VCERT intervention teachers attended a training session and found this to be helpful; agreeing or strongly agreeing with the statement "Overall, the training session was helpful". All teachers also agreed or strongly agreed with the statement “The training session was pitched at the right level for me”. For diploma intervention teachers, the responses were either strongly disagreed or disagreed with the statement.

Setup and Integration

The setup experience for VCERT showed variation between teachers and learners. The statement "I encountered minimal IT issues when implementing FirstPass with my class" received positive responses from teachers who either strongly agreed or agreed with this statement. However, in response to the statement “The students in my class had minimal issues accessing the FirstPass platform”, teachers either strongly disagreed or disagreed . Furthermore, in response to the statement “The FirstPass platform is easy to use to monitor student performance”, teachers strongly disagreed or disagreed. A teacher comment highlights the challenge learners encountered:

“I found the platform easy to access but difficult to find and mark their answers. The students also found it difficult to find the question as these were hidden by the question mark.”

The diploma teachers either strongly disagreed or disagreed with the following statements: “The FirstPass platform was easy to set up for my learners”, “I encountered minimal IT issues when implementing FirstPass with my learners” and “The learners had minimal issues accessing the FirstPass platform”. This highlights potential implementation issues in the diploma setting for the deployment and use of the FirstPass platform. One example of an issue encountered included:

“I also found that when exploring it with him writing full sentences wouldn't register down the side but if I just wrote the exact words from the side bar it would register it”.

Enablers and Barriers to Implementation

When implementing FirstPass in schools, the main enablers to implementation appeared to be the effectiveness of training and platform demonstration. The training sessions were positively received, indicating that they provided a solid foundation for

using FirstPass, with all responders agreeing or strongly agreeing that "The training session was pitched at the right level for me" .

Process evaluation responses from teachers involved in implementing the FirstPass platform highlight the following main barriers:

Engagement

The VCERT teachers reported that learners initially enjoyed FirstPass but felt that the tasks were too similar to retain their interest over the six-week period. For example, “Whilst the similarity in the format was good, the monotonous tasks did not keep their interest”.

The diploma teachers reported that learners have several platforms to check and use for current coursework and “found it confusing and overwhelming”. This reduced engagement as learners stopped using the FirstPass platform. A further valid point raised by the diploma process evaluation was the time commitment for learners. Noted difficulties in “getting the learner to engage when they work fulltime as well their apprenticeship” highlights how the two courses for VCERTs and the diploma are structured differently. If learners are working full time, completing coursework and using the FirstPass platform, this additional workload may disengage learners.

In the recruitment stage of the project, engagement with training providers was a challenge. Only one training provider was able to be recruited to the project, highlighting the difficulties in engaging teachers with FirstPass.

Accessibility

One school reported that FirstPass had to be used in school time due to pupil premium students finding it hard to access outside of school. Another school commented that “the students being able to see the questions and specifically what they needed to do”

was not clear for learners. When asked how the platform could be improved, comments such as “I think there needs to be a few minor redesigns to make it more user friendly” highlight that the user design of the product may need minor improvements.

In terms of accessibility for teachers, a few comments on the user interface were highlighted. For example, one teacher commented that “marking was difficult, as you have to go out of tasks to go to the next student. It would be easier and less time consuming to flick from one task to another”.

The second process evaluation question thread focused on the control group.

What constitutes ‘usual practice’, and does this change over the duration of the trial? Are control students using similar interventions to FirstPass that might be considered close substitutes for it?

As the FirstPass evaluation focused on the feedback the system provided, familiarity with the style of questions set on the platform could be a plausible alternative explanation for improvements in attainment. Therefore, all control students were provided access to the same weekly tasks set on the platform. In control schools, these were set as class activities and completed weekly over the duration of the trial. One school reported using an alternative homework platform for generic homework tasks, but these do not appear to be close substitutes for the FirstPass platform.

In the diploma, no control teachers completed the process evaluation survey so we are unable to verify whether they were using similar software for feedback.

The third process evaluation questions focused on the variability of the implementation.

How and why does the implementation of the programme vary? To what extent does any variability affect the achievement of expected outcomes?

It is important to note that the sample size is small for the VCERT pilot and caution needs to be used when interpreting the variability of the results. In School B, 67% of learners completed over 85% of the tasks set on the FirstPass platform demonstrating good fidelity to treatment. The effect size for School B comparing pretest to post-test is large (ES 1.13, SE 0.41), demonstrating a positive impact. In School C, fidelity to treatment was relatively low, with no students completing more than 30% of the tasks on the platform. This shows that engagement with the platform in this school was much lower than in School B.

The overall effect size of the differences in the post-test scores between those receiving FirstPass and controls indicates a small positive effect (Cohens' D 0.018, SE 0.29). The large standard error suggests that there are large differences between individual scores. Therefore, until further replications of the trial can be completed with additional schools, it is not possible to explore the extent that any variability will affect the achievement of expected outcomes.

Due to the limited data returned in the diploma evaluation of the FirstPass platform, it is not possible to examine the variability of the implementation on expected outcomes.

Lessons Learned

The following lessons have been learnt during the FirstPass project:

1. The method used to train the subject topic classifiers on FirstPass proved to be time consuming and costly. The advent of large language models addresses this shortfall, enabling teachers to set up open-ended questions to their students in minutes. The advent of generative AI and the use of large language models negates the use of pre-trained classifiers.
2. Signing up schools to be part of pilot studies is always a challenging task, and the project team experienced the same difficulty during the project. If we were

to repeat this process, we would engage with schools much earlier in the process.

3. Timely information, advice, guidance and support was welcomed by schools. The schools who tapped into this support, particularly noticeable during onboarding, were more engaged with the project, which subsequently led to better quality research outputs.

Conclusion

When deploying FirstPass with the VCERT cohort, test scores of students receiving FirstPass improved significantly more than those who did not receive FirstPass, even after taking into account differences between the groups at baseline (*ANCOVA (2,48) f=19.36 P<0.05)*. On average, students receiving FirstPass improved 2.65 points more than controls.

The size of the differences in post-test scores between those receiving FirstPass and those who did not indicates a small positive effect (Cohens' d 0.018, SE 0.29). The large standard error suggests that there are large differences between individual scores.

In School B with a high fidelity of usage, the effect size comparing pretest to post-test is large (ES 1.13, SE 0.41), demonstrating a positive impact. However, due to the small sample size caution must be used when interpreting these results.

No significant differences in improvement were observed between students eligible for free school meals and students who were not eligible for free school meals.

FirstPass shows promise as an effective feedback system, however the process evaluation highlights a few areas of consideration when implementing the platform. Firstly, ensuring variability in set tasks is key to retaining the interest and participation of learners. In the six-week intervention period, teachers reported that the tasks had become monotonous, and this impacted learner engagement.

Secondly, the design of the platform could be improved to make this more user friendly for both learners and teachers. For example, teachers found marking difficult when moving from one student to another, increasing the time it takes to mark work.

When deploying FirstPass with the diploma cohort, the evaluation has demonstrated that it is difficult to engage teachers and learners in using the system. As the learners are working full time on an apprenticeship, and the final assessment is not examination focused, the addition of the FirstPass component was overwhelming for students. Feedback from teachers confirmed that this did not align with coursework and created additional work for learners. Consequently, before deploying FirstPass with learners in this context, identifying a clear need for how the system can be used should be considered.

Post-pilot platform enhancements

Since the completion of this project, Bolton College have undertaken substantial development of its FirstPass platform. FirstPass v2 will leverage the use of agentic AI and knowledge graph retrieval techniques. These will further the platform's ability to support students and teachers as they seek feedback on complex free-form text responses to open-ended questions. In addition, FirstPass will provide students and teachers with inline commentary and feedback. The College will also develop the platform so that it can review images, charts and graphs which are included in student work. Finally, students will be able to post their own files, notes and website links to the chatbot interface within FirstPass so that they have additional resources to support the development of their work before submission to teachers.

A generative AI feedback tool will be introduced to students enabling them to receive feedback through an on-demand GenAI feedback button or through a GenAI chatbot. Figure 3 shows how GenAI feedback is presented to students. This shows the conversational interface between a student and an assignment GenAI chatbot. Students converse with the GenAI chatbot to garner feedback to inform their work.

Teachers remain in control of these interactions, a cornerstone of FirstPass' design, though a prompt creation interface that details how the assignment GenAI chatbot will deliver feedback to students as they respond to an open-ended question (Figure 4).

Overall, the FirstPass platform has shown promise in how technology can be used to effectively support teachers in providing feedback to learners using an AI based system. Further replications of the evaluation will help to strengthen the claims that can be made regarding the effectiveness of the product. During 2024-25, 1200 GCSE English Language students and teachers will make use of FirstPass v2 and the additional capabilities that it offers.

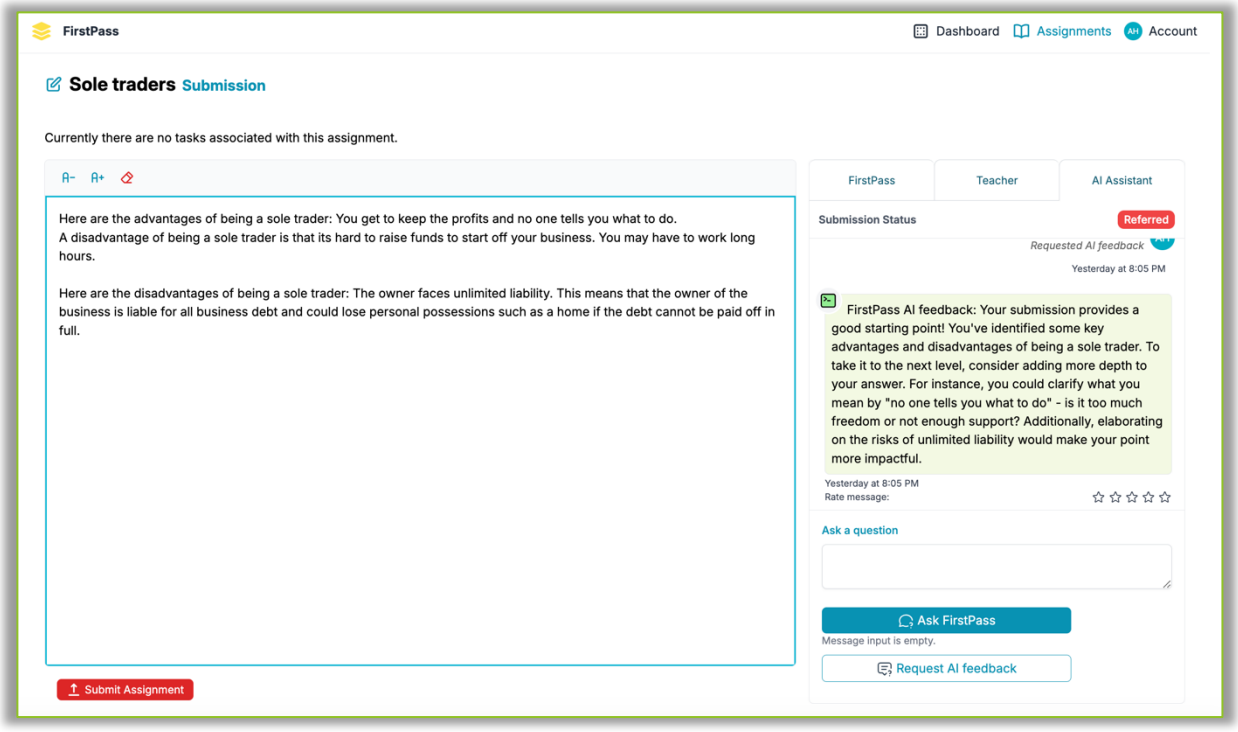


Figure 3: Student Assignment Conversational Interface

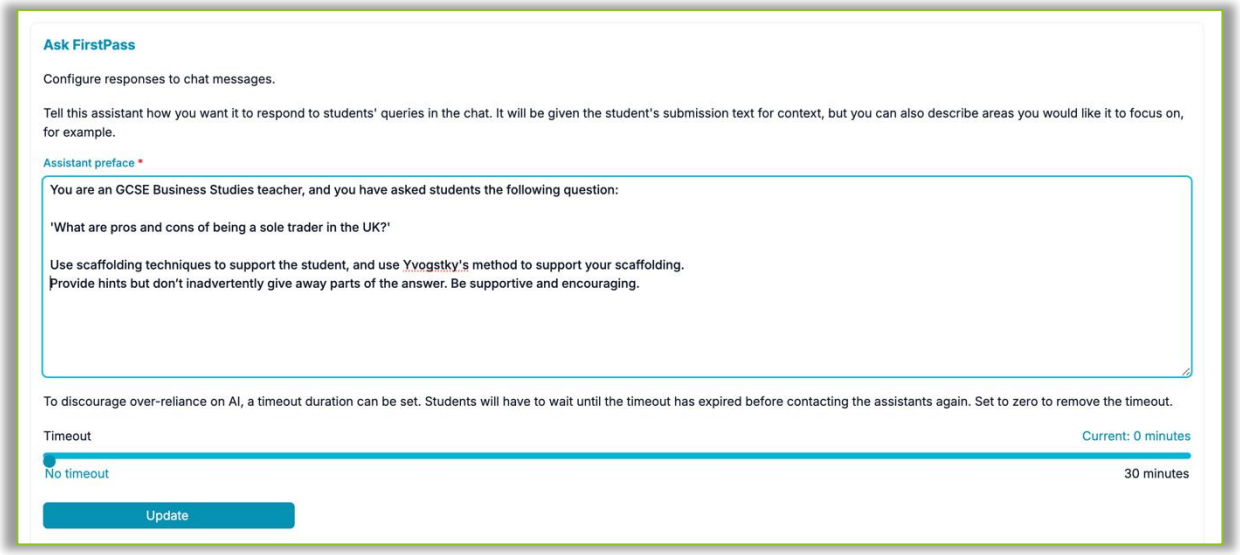


Figure 4: Prompt creation interface for teachers.

References

Black, P. and Wiliam, D. (2009) 'Developing the Theory of Formative Assessment', Educational Assessment Evaluation and Accountability, 21, pp. 5–31. 10.1007/s11092-008-9068-5

Education Endowment Foundation (2023). Teaching and Learning Toolkit: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/feedback

Hattie, J. (2009) Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement, London: Routledge.

Hattie, J and Timperley, H, (2007), The Power of Feedback, Review of Educational Research, 77 (1),81–112, p.82.

Kingston, N. and Nash, B. (2011) 'Formative Assessment: A Meta-Analysis and a Call for
Research', Educational Measurement Issues and Practice, 30 (4), pp. 28–37.

Appendices

Appendix A – Full research methodology section for the Diploma evaluation.

Research questions

The impact evaluation is designed to answer the following questions:

What effect does FirstPass have on the attainment of learners for EYPS 5 understanding how to support children's development topic in the Level 2 Diploma qualification as a supplement to teaching compared to teaching as usual?

In addition to the primary research question, as the data aggregates for the evidence base for the intervention, secondary research questions include:

- Is there a correlation between impact on attainment and the amount of time students spend on the platform?
- To what extent does the impact of FirstPass vary between training providers?

The process evaluation component will address the following questions:

- How is the intervention implemented? What are the enablers and barriers to implementation of the programme? Is implementing the programme feasible?
- What constitutes ‘usual practice’ for the control students, and does this change over the duration of the trial? Are control students using similar interventions to FirstPass that might be considered close substitutes for it?
- How and why does the implementation of the programme vary? To what extent does any variability affect the achievement of expected outcomes?

Design

This study was designed as a randomised controlled trial where individual students are assigned at random to either intervention (exposure to FirstPass) or control conditions (non-exposure to FirstPass) on a 1:1 basis.

Randomisation

Randomisation was be carried out by a randomisation tool created by WhatWorked to randomly allocate learners and record the process. This will be anonymised to comply with the data sharing agreements.

Participants

Training providers were identified using NCFE's own database of training providers who are implementing the NCFE Level NCFE CACHE Level 2 Diploma for the Early Years Practitioner (QN: 603/3723/0) qualification and approached to determine their interest in participating in the study.

Apprentices studying towards the Early Years Practitioner Standard must achieve a Level 2 Early Years Practitioner qualification as part of their apprenticeship. The apprentices that will undertake the Child Development Pilot are studying for their apprenticeship whilst employed in early years settings, under the guidance and support of Training Providers. There is no age restriction applied to studying as an apprentice. On the completion of their mandated qualification they enter Gateway, which is when apprentices will sit End Point Assessment, an objective and independent assessment that includes a Knowledge Test.

It was expected each training provider will only provide one cohort of students to participate in the evaluation.

Sample

The pilot aimed to recruit two training providers with a cohort of students (25 to 30 learners per course). However, only one training provider could be recruited with a sample of twenty-one students. For context, the ESSA standards in the USA require a sample size of 350 for robust studies. The design of the evaluation allows for future replications that will increase the sample size as more schools participate in the evaluation.

Outcome measures

The primary outcome measure selected for the study were independent exam style questions for the NCFE CACHE Level 1/2 Technical Award in Child Development and Care in the Early Years CA1 Child Development. The reason for using the extended writing questions from this qualification is that the content is identical to the diploma unit of work for ‘understanding how to support children’s development.’

As the qualification is new with limited sample questions available to providers, NCFE commissioned the creation of an independent pre- and post-assessment for the CA1 Child Development content. The assessment consisted of one three-mark question and three six mark extended writing questions. The assessment maximum score was 21 marks and learners completed the assessments under exam style conditions with a duration of twenty-one minutes for the assessment.

A mark scheme was provided for teachers to assess the pre- and post-assessment with scores recorded in a pre-specified CSV file. The anonymised data was then shared with the evaluation team.

Analysis plan

Statistical analysis used Analysis of Covariance (ANCOVA) to examine whether there are probabilistic and reliable differences in scores between the treatment and control group, adjusting for any differences in pre-test scores between the treatment and control groups. This adjustment is carried out to test the assumption that the groups might perform differently because they contain students with different levels of aptitude, despite our attempts to ensure a fair comparison through random assignment.

Analysis also used the same techniques to examine whether some schools gain more from using FirstPass than others and relating this to the process evaluation to see if training providers' reports of how they use FirstPass can shed light on how FirstPass can be used for greatest effect.

The relationship between duration of FirstPass use and assessment scores will be examined, to inform understanding of optimal use.

Appendix B

Personnel

WhatWorked Personnel

- Dr Wayne Harrison – Evaluation lead and project manager.
- Dr John Brown – Evaluation deputy and leads on the data analysis.
- Prof Steve Higgins – Research methodology Quality Assurance and ethics.

FirstPass Personnel

- Aftab Hussain – Bolton College Project Manager
- Jonathan Hart – Technical Manager

NCFE Personal

- Gray Mytton – NCFE Project Manager
- Janet King – Early Years
- James Lane – Digital Specialist

Timeline

The table below shows the dates for activities to be undertaken during the project. A detailed project plan will be created to support the planning and implementation of the pilot evaluation.

Please note that the start and end dates are only indicative, and these will be revised into specific dates upon the commencement of the main contract for the evaluation.

Work package	Start date	End date
Discovery stage		
Initial project planning meeting	Sept 2023	Sept 2023
Project planning and risk assessment	Sept 2023	Sept 2023
Assessment discovery stage	Sept 2023	Sept 2023
Planning Stage		
Project protocol agreed	October 2023	October 2023
Assessments created by NCFE	October 2023	November 2023
School recruitment	October 2023	November 2023
MoU created	October 2023	October 2023
Project planning and roles assigned for NCFE / WhatWorked Education	October 2023	October 2023
Process evaluation surveys created	October 2023	November 2023
Implementation Stage		
Schools sign MoU	November 2023	December 2023
School onboarding - Training scheduled and learner accounts created by FirstPass	December 2023	January 2024
Pre-test assessment window opens	January 2024	March 2024
Post-assessment window opens.	February 2023	April 2024
Delivery Stage		

FirstPass launches in each school for the intervention classes following the pre-test.	January 2024	March 2024
Implementation and usage weekly monitoring for the intervention students.	January 2024	April 2024
Process evaluation – Completed by the FE lecturers involved in implementing the programme.	February 2024	April 2024
Evaluation		
Usage data anonymised and shared with WhatWorked.	June 2024	June 2024
Data analysis	June 2024	June 2024
Draft evaluation report	July 2024	July 2024
Report published	July 2024	July 2024