Education Endowment Foundation

Voice 21

Pilot report and executive summary

June 2018

Independent evaluators:

Jenny Smith, Dr Anna Grant, Naomi Horrocks, Dr Kathy Seymour, Andrew Boyle, Lawrence Bardwell, Matthew Turner

AlphaPlus

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

Danielle Mason
Head of Research
Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

p: 020 7802 1679
e: danielle.mason@eefoundation.org.uk
w: **www.educationendowmentfoundation.org.uk**

About the evaluator

AlphaPlus Consultancy Ltd. is one of the UK's leading educational consultancies drawing on the professional experience of a team whose work spans the public and private sectors, covering education, assessment and evaluation. We can draw upon a team of over 400 associates, each of whom brings many years' experience in education, covering both schools and post compulsory education, as well as roles in many of the supporting agencies and government departments. We are actively engaged with the latest educational developments, on both a strategic level, through our research and evaluation projects for UK government, and at a practical level, through the experience of our wider team in all aspects of educational development and change.

We have an international reputation for robust and meaningful evaluations, which support our clients to understand the impact on their stakeholders of policy change and/or interventions. We work in collaboration with our clients to ensure that evaluation findings are evidence based and useful. We ensure that our data collection, analysis and interpretation is undertaken with a solid understanding of the stakeholder context and within ethical guidelines.

Our team has expertise in assessment, evaluation and statistics, fieldwork, and evidence gathering – but we are also excellent communicators, speaking and writing clearly and accurately and giving messages that are to the point and free of jargon. We provide realistic and evidence-based reports for our research and evaluation clients.

Contact details

Andrew Boyle
Director of research
AlphaPlus Consultancy Ltd.
e. **Andrew.boyle@alphaplus.co.uk**

Contents

Executive summary

The project

The Voice 21 Oracy Improvement Programme supports schools to develop pupils' use of speech to express their thoughts and communicate effectively. The aim of the programme is to improve these oracy skills with the expectation that this will improve wider academic outcomes.

The one year pilot programme was based on School 21's Oracy Skills Framework and consisted of:

- one hour per week of lesson time dedicated to developing four key areas of spoken language: physical, linguistic, cognitive, and social and emotional;
- materials for an oracy curriculum, including a mandatory unit that prepared pupils to do a five-minute individual talk;
- activities to promote an 'oracy culture' in the school, including building oracy into assemblies and cascading the principles of oracy to teachers and staff; and
- use of an oracy assessment measure developed by School 21 in collaboration with the University of Cambridge.

The aim of the pilot evaluation was to test the feasibility of the programme, its evidence of promise, and the reliability of the oracy assessment measure.

The Voice 21 team provided two days of training to a designated oracy lead at each school at the beginning of the programme. Additional training and support was available on request throughout the project. The oracy lead was responsible for cascading the training to other teachers involved in delivery.

Twelve schools were recruited to the pilot, but one dropped out before delivery started. The programme was designed for Year 7 pupils, but one school delivered it to Year 8. Initial training took place at School 21 in July 2016; schools delivered the pilot from September 2016 to July 2017.

The programme was developed as part of a collaboration between School 21 in East London, and the University of Cambridge with funding provided by the Education Endowment Foundation. This pilot was delivered by Voice 21, School 21's charitable outreach arm.

Key conclusions	
1.	Teachers reported that pupils' oracy skills improved as a result of the pilot; assessment results also showed that pupils' oracy skills improved. However, as there was no comparison group, it is not possible to say whether these changes would have happened anyway.
2.	Many schools were beginning to develop a whole-school oracy culture by the end of the programme, but felt that only limited change was achievable in one year and when focusing on only one year group.
3.	Most teachers were positive about the programme and agreed that it would work in most schools with minimal adjustments.
4.	The Voice 21 oracy assessment measure used in the pilot did not provide sufficiently reliable data. A revised or alternative impact measure would be needed for a trial.
5.	Delivery was not uniform across schools, or within schools, in part due to an initial lack of clarity about which elements of the programme were mandatory and which optional. The core components of the programme would need to be clearly articulated at the outset of a trial while maintaining the flexibility in delivery that was popular with teachers.

What are the findings?

Teachers generally reported improvements in pupils' oracy over the course of the pilot. However, they were not confident that the improved oracy skills they observed could have an immediate impact on academic attainment, although some felt this could be a longer-term outcome. The pilot evaluation found that the oracy assessment measure had limited reliability. Pupil's oracy, as measured using this assessment tool, improved during the pilot, however, given the limited reliability of the measure and the lack of a comparison group, we cannot conclude from these results that the programme improved oracy.

Participating schools were positive about the pilot. While most reported some difficulties with developing a whole-school oracy culture—and delivering oracy assemblies in particular—on the whole, there was widespread agreement that the programme would work in most schools.

Teachers thought the costs of the programme were acceptable. The most significant cost was the time required by the oracy lead for training, planning, and delivery of the programme.

Delivery across schools, and within schools, varied significantly. The flexibility in how to deliver the programme was seen as a key strength by participating schools, but will also present a challenge if the programme is taken to trial. Another challenge that was highlighted was maintaining the quality of training and ongoing support to schools, which was seen as critical to successful implementation if the programme was scaled up.

How was the pilot conducted?

Schools were selected to participate in the pilot based on their geographical location (to ensure that it would be feasible to attend on-site training at School 21) and their proportion of pupils eligible for free school meals (those with higher proportions of FSM pupils were prioritised, although no threshold was set).

Interviews and online surveys were carried out with oracy leads, participating teachers, and members of the senior leadership team at various points during the year to explore changing attitudes towards the programme and gather information about its implementation.

Schools were asked to use School 21's oracy assessment measure to test the oracy skills of approximately 60 pupils at the beginning and end of the programme. Each school used its own criteria to select this group. The tests were used to measure improvements in pupils' oracy skills over the course of the programme and analysis of the reliability and validity of the assessment was carried out.

Summary of pilot findings

Question	Finding	Comment
Is there evidence to support the theory of change?	Yes, but limited	All school staff reported some improvement to pupils' oracy. Oracy was also measured using the School 21 assessment and was found to have improved. However, given the limited reliability of the assessment, and the lack of a comparison group, we cannot conclude from these results that the programme improved oracy. The pilot did not measure impact on academic attainment.
Was the approach feasible?	Yes	The programme was well received across the pilot. Teachers felt it could be implemented in most school contexts, given the necessary support from senior leadership.
Is the approach ready to be evaluated in a trial?	Yes, but with some caveats	A clear definition of the programme's core components is needed before it can go to trial. More work should also be done to improve the oracy assessment measure so that it produces reliable data. Alternatively, another suitable attainment measure could be selected.

Introduction

Intervention

This pilot of the Voice 21 Oracy Improvement Programme trialled a year 7 oracy curriculum and assessment toolkit in 12 schools, although one school left the pilot shortly before the start of term (the reasons for not taking part are not known), leaving 11 schools implementing the programme.

The curriculum and assessment toolkit had been produced as part of an initial development phase between January 2013 and July 2014. This development phase was a collaborative project between School 21 and the Faculty of Education, University of Cambridge, and was funded by the Education Endowment Foundation (EEF) (Maxwell *et al.*, 2015).

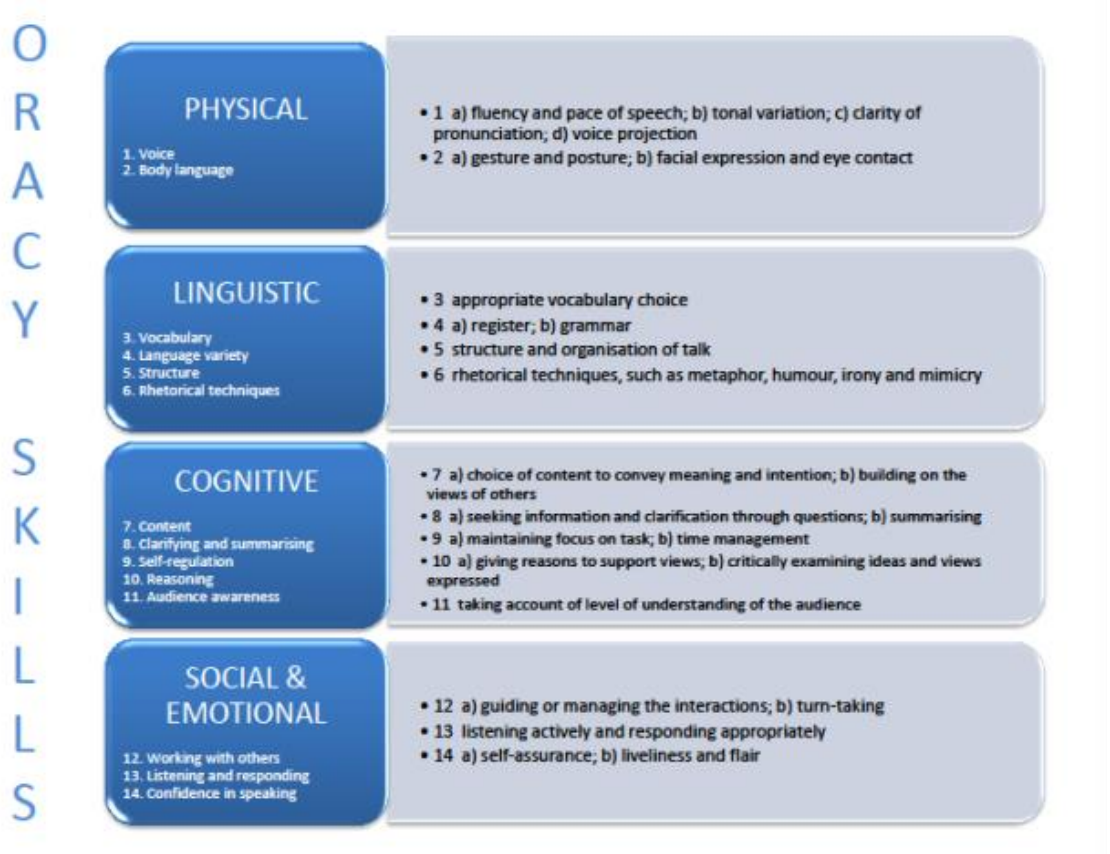
The delivery organisation for the current pilot was Voice 21, the charitable arm of School 21, a free school in East London with a strong commitment to oracy across the curriculum and the development of an oracy culture within the school. Initial training for schools took place in July 2016, with a one-year pilot phase from September 2016 until July 2017.

The pilot intervention

Oracy Skills Framework

The intervention was based on an Oracy Skills Framework that identified four areas of spoken language skills: physical, linguistic, cognitive, and social and emotional. The framework identified the key generic components for each of the areas—see Figure 1 below.

Figure 1: Oracy Skills Framework



Curriculum

The programme comprised an outline 'curriculum' for Year 7 stand-alone oracy lessons consisting of four units originally developed by School 21 for their own use as part of a designated one-hour oracy lesson held once a week. For the current pilot, the final unit, 'Ignite', was the only mandatory element of the curriculum, although schools were expected to implement the oracy framework through stand-alone lessons and by embedding oracy in other lessons for the entire 2016/2017 academic year. The Ignite unit did not have a prescribed number of lessons, but schools were recommended to allow a minimum of one half-term's worth of lessons—ideally as many as 9 to 12 lessons—for this unit. The curriculum was not intended to be prescriptive. The expectation was that schools would adapt the materials and activities to meet their own curriculum needs. Lesson plans were only shared with pilot schools on request. A brief overview of each unit is given below:

Unit 1: Finding our voice

The focus of this first unit is to familiarise pupils with the Oracy Skills Framework and to introduce them to a range of strategies and protocols to support the development of their talking and listening skills, especially in group work situations.

Unit 2: Performance poetry

This unit focused on the physical and social- and emotional areas of oracy. Pupils use performance poetry to support the development of effective talking and listening skills in small groups.

Unit 3: Persuasive techniques

Pupils consider a range of formal and informal talk scenarios which use persuasive techniques. Pupils are encouraged to reflect on the effectiveness of a range of these techniques. This unit focuses on the linguistic and social and emotional areas of the oracy skills framework.

Unit 4: Ignite

The final Ignite unit was the only mandatory part of the curriculum for the pilot phase. It prepared pupils for a five-minute individual talk on a subject of their own choice. The final unit was intended to encourage pupils to draw on skills and techniques developed from all strands of the Oracy Skills Framework.

Teacher training and support

A two-day initial training course was developed and delivered at School 21 with additional training and support when requested, including specific training for the Ignite unit and its assessment. The two-day initial training course modelled ways of talking and listening that could be translated into classroom activities, such as Harkness discussions, talking points, and debating.

Assessment activities

The programme included assessment activities that could be used as diagnostic tools and as indicators of progression for use at the start and end of Year 7. These covered the four areas identified in the Oracy Skills Framework. The three assessment activities were:

- a short presentation task;
- an instructional activity where one pupil instructs another to complete a specified task; and
- a 'talking point' activity where three pupils discuss a specific topic and are asked to reach a conclusion.

There was also a stand-alone Ignite assessment—a five-minute individual talk.

Materials

Materials to support the curriculum and the assessments were provided, including, for example, guidance on how to manage and deliver the talking and listening activities. Video clips and marking guidance to support standardisation of assessment outcomes were made available to pilot schools.

While the programme was intended to be delivered to Year 7 pupils, at one school, both Year 7 and Year 8 were included (although only Year 7 data was included in the assessments); at another school, Year 8 only received the programme.

Background evidence

School 21 was founded in 2012 based on an inclusive ethos, which included putting 'oracy' at the heart of the curriculum. The Voice 21 Oracy Improvement Programme uses the term 'oracy' to include talk and listening skills as a tool for learning in the socio-cultural context of the classroom, as well as a wider purpose to support pupil agency, wellbeing, and presentational skills for life and work. The term 'oracy' is explained as follows:

> *Oracy is to speech what literacy is to reading and writing, and what numeracy is to maths. An 'unappealing neologism', the word is nonetheless useful for suggesting that spoken language is an acquired, teachable skill, one that enables pupils as much as literacy does.*
>
> [Voice 21]

The overall ethos of the programme is to support young people to 'find their voice' by teaching oracy language skills explicitly for use across a range of contexts, formal and informal.

> *Oral language skills should continue to be instructed … probably well beyond the conventional 'speaking and listening' goals commonly adhered to within the English National Curriculum.*
>
> Law *et al.* (2011)

Research has evidenced the importance of talk for learning across curriculum areas, especially for the teaching and learning of higher-order concepts which require explicit, conscious effort and direct intervention (Alexander, 2008; Mercer, 2013). Studies that focus on the relationship between dialogue and cognitive development, such as those of Mercer and Littleton (2007: 29), have investigated and evidenced how 'ways of thinking are embedded in ways of using language'. Utterances are seen as 'thinking devices' when treated dialogically (Lotman, 1988). The emphasis is on the use of language to develop reasoning and problem-solving skills as learning tools—as pupils collaboratively engage in learning and understanding together. It has been claimed that pupils require a broad repertoire of talk to support their learning and wider development:

> *Pupils need, for both learning and life, not only to be able to provide relevant and focused answers but also to learn how to pose their own questions and how to use talk to narrate, explain, speculate, imagine, hypothesise, explore, evaluate, discuss, argue, reason and justify.*
>
> Alexander (2012:4)

In 2013, the EEF funded an initial oracy development phase and pilot within School 21 that enabled School 21 to work with a team from the Faculty of Education, University of Cambridge to create a framework for oracy and a range of assessment tools. Ofsted judged School 21 to be outstanding in all areas in 2014 with oracy highlighted as a key factor in the school's success.

Findings from the evaluation of the initial development and School 21 pilot phase (Maxwell *et al.*, 2015) concluded:

- the Oracy Skills Framework provided a useful tool for schools to review and develop their approach to oracy;

- the assessment toolkit provided a diagnostic approach for tracking pupils' progress in developing oracy skills;
- the curriculum and assessment toolkit as implemented in School 21 offered a sound foundation for the development of oracy skills, and in particular supporting persuasive talk and talk for presentational purposes, and talk in formal contexts; and that
- further refinement of the curriculum and resources was required for the development of exploratory talk and to ensure a wider range of opportunities for oracy, both formal and informal, were provided.

In addition, the evaluation report concluded that for some schools, the School 21 programme may require a fundamental shift in approach if adopting the full package of potential changes, including, for example, the cultural change required and dedicated curriculum time.

Research questions

The current pilot and its evaluation had three key aims. These focused on establishing evidence to support the theory of change and assess the feasibility and readiness for trial of the School 21 oracy model in a range of different schools. The questions the pilot evaluation was designed to answer were:

Evidence to support theory of change

1. To what extent is it plausible that the School 21 model would result in (positive) changes to teaching and learning oracy across a school?
2. To what extent is it plausible that any changes in teaching oracy translate into improvements at the pupil level (in oracy, reasoning skills, attainment or other)?
3. To what extent do we see changes in pupils' oracy on pre- and post-measures of oracy?

Feasibility

4. To what extent are schools able to deliver the Voice 21 curriculum, assessment, and training 'package'? (Cf. What does 'school ready' look like?)
5. Are the quality assurance / fidelity markers appropriate?
6. Is the process of identifying gaps in quality assurance appropriate?
7. How appropriate is the use of hubs as a means of rolling out the programme?

Readiness for trial

8. Is there a School 21 curriculum, assessment, and training 'package' that could be rolled out to schools (with minimal modifications)?
9. Is the School 21 oracy measurement a valid and reliable tool for use in future trials?

Ethical review

All schools that were involved in this pilot project were asked to sign a memorandum of understanding (MOU) which was designed by Voice 21, Alpha*Plus*, and the EEF. It stated the roles, responsibilities, and obligations of being a pilot school and detailed the ethical conduct of the evaluation activities and the data-handling protocol. The Alpha*Plus* ethics advisor (Professor Roger Murphy) reviewed the research strategy and the MOU as part of the Alpha*Plus* internal process of ethical clearance. A copy of the MOU is included in Appendix 1.

Project team

Table 1 details the team members who worked on this project and summarises their roles and responsibilities.

Table 1: Project team members

Name	Roles	Organisation	Responsibilities
Jenny Smith	Project manager / lead evaluator	Alpha*Plus*	Responsible for all aspects of the evaluation; researcher in process evaluation strand.
Dr Anna Grant	Senior researcher / evaluator	Alpha*Plus*	Contributing to all aspects of the evaluation 'life cycle'; researcher in process evaluation strand.
Naomi Horrocks	Advisor: teacher training qualitative researcher	Alpha*Plus*	'Critical friend' teacher training; contributing to process evaluation.
Dr Kathy Seymour	Researcher	Alpha*Plus*	Researcher in process evaluation strand, data analysis and report writing.
Andrew Boyle	Assessment lead/ quantitative researcher	Alpha*Plus*	Leading work on assessments and suitable measurements for the pilot.
Lawrence Bardwell	Statistician	Alpha*Plus*	Assessment data analysis and reporting.
Matthew Turner	Statistician	Alpha*Plus*	Assessment data analysis and reporting.
Prof Neil Mercer	Advisor: oracy	University of Cambridge	Advisor: oracy.
Dr Ayesha Ahmed	Advisor: oracy assessment	University of Cambridge	Advisor: oracy assessment.
Dr Dougal Hutchison	Senior statistician	Alpha*Plus*	Advising on statistical analysis of quantitative outcome data.
Prof Roger Murphy	Chair Ethics and Quality Assurance Board	Alpha*Plus*	Sign-off ethics protocol and research instruments.
Beccy Earnshaw	Director	Voice 21	Voice 21 Oracy Improvement Programme lead.

Methods

Recruitment

Twelve schools were selected to take part in this pilot, these were selected in regional clusters representing the North East, North West, South East and South West of England. Schools were initially selected by Voice 21 from those who had expressed an interest in the oracy programme and were subsequently asked to answer a range of questions relating to their school's characteristics and their attitudes towards, and abilities to deliver, the programme as part of the pilot. The criteria for selection were as follows:

- Schools should be varied but should preferably have a high proportion of pupils eligible for free school meals as is consistent with EEF's aims (no specific numeric threshold was set for the percentage FSM eligibility).
- Most schools should be enthusiastic to take part, that is, have a genuine commitment to introducing/ integrating oracy, but must not already have an established approach to oracy or teach oracy in dedicated lessons. This criterion was specified to ensure a commitment to the programme and minimise attrition.
- Schools should represent different regions in England but be fairly accessible from London for ease of delivery, and schools should form a set of regional 'hubs'.

One school left the project shortly before the programme delivery started, and another school took part in the initial delivery and evaluation activities but no longer participated in any evaluation activities after Easter 2017; it is not known whether it continued to deliver the programme after this point.

Data collection

There were three main strands to this research: the process evaluation, pre- and post-intervention assessments, and a review of the oracy assessment tools.

Process evaluation

The process evaluation began with a workshop to define the intervention and the overarching theory of change. During the pilot year, fieldwork which contributed to the process evaluation was conducted via a series of face-to-face and telephone interviews with oracy leads, teachers delivering the programme, and SLT members with oversight of the programme. In addition, there was an online survey of oracy leads and teacher deliverers.

The theory of change workshop was held with the Voice 21 Oracy Improvement Programme lead in Spring 2016. The aim of the workshop was to:

- clarify understanding of the underpinning theory for the Oracy Improvement Programme approach and intervention (for pupils and teachers);
- define the intervention and the extent to which the intervention would be the same for everyone;
- clarify the desired outcomes and how to obtain evidence on them; and
- agree research questions and methods.

An initial activity used concept mapping to identify the role of different theoretical debates in the development of the programme and how these had influenced the programme's inputs and expected outputs and outcomes. It was identified that there was no consensus about the key purpose of the intervention across the programme team: some were more focused on the wellbeing and pupil agency aspect of the programme rather than outcomes that were easier to measure, such as academic

attainment or improvement across specific aspects of the oracy framework. As a result of this discussion, a logic model was created to articulate the theory of change as applied to the pilot project.

The interviews which formed part of the process evaluation were undertaken at the following times:

- beginning-of-year, face-to-face interviews to establish baselines and expectations, as well as any initial perceptions of the programme and its implementation—October to November 2016, with oracy leads, teachers, and SLT members (one oracy lead and one SLT member was interviewed in each school, and schools were asked to put forward a minimum of two teachers for interview; in some instances more than two teachers were interviewed, but in others only one or no teachers were available for interview; Tabl below provides further detail);
- mid-year telephone interviews to explore perceptions of changing practice—March 2017, with oracy leads only; and
- end-of-year, face-to-face interviews to gauge the perspectives of all stakeholders involved in delivering the programme, its implementation, impact and feasibility—June to July 2016, with oracy leads, teachers, and SLT members (again, one oracy lead and one SLT member in each school and attempts were made to interview at least two teachers per school, see Tabl).

Stakeholders in different roles were invited to participate in the interviews (and the online surveys discussed below) in order to gain a range of perspectives on the programme and its implementation. All participants were assured that their views and responses would be treated in confidence and that they would not be identified in any outputs resulting from the pilot. To help maintain anonymity, the 11 schools that implemented the programme were given a letter identifier from A–K. Throughout this report schools are referred to by this identifier rather than their school name.

Table 2 below shows how many people were interviewed each time. Note that in all cases except schools A and K, an oracy lead and an SLT member were interviewed in the start- and end-of-year interviews. In School A, only the oracy lead was interviewed in the beginning of year and mid-year interviews (and this was by telephone in both cases due to the school being unable to accommodate in-person interviews); in School K, only the oracy lead and SLT member were interviewed in the first round of interviews.

Table 2: Number of people interviewed on each occasion by school

School	Number of interviewees— at beginning of year	at mid-year (oracy leads only)	at end of year	Total across all thee interviews
A	1	1	school withdrew	2
B	5	1	3	9
C	6	1	5	12
D	4	1	3	8
E	3	1	3	7
F	2	1	4	7
G	4	1	3	8
H	5	1	5	11
I	5	1	3	9
J	4	1	4	9
K	2	1	3	6
Total	**41**	**11**	**36**	**88**

Short online surveys were administered to oracy leads and teachers. The oracy leads' survey was administered in March 2017 and gathered factual data about the method of delivery of the programme. Oracy leads at the 11 pilot schools completed the online survey. The survey of teachers delivering the programme was administered in March 2017 and again in July 2017 and sought their opinions on aspects of the programme and the impact it had had via a series of statements to which they indicated their levels of agreement. Table 3 shows how many responded to the teachers' survey on each occasion. Note that the survey was distributed among relevant colleagues by the oracy lead at each school; it is not therefore known exactly how many teachers were asked to complete the survey, therefore an assessment of the response rate is not possible. School A had dropped out of the evaluation activities at the point when the surveys were administered.

Table 3: Number of responses to the teachers' survey per school

School	No. of responses to the March survey	No. of responses to the July survey
B	0	2
C	2	3
D	0	5
E	2	2
F	2	3
G	0	1
H	3	1
J	0	2
K	2	1
Total	11	20

Baseline and post-intervention testing

Reason for the choice of the particular tool

There was considerable debate at the start of the project about which instrument (and more widely, which approach to assessment) was most appropriate to assess oracy in this context. In the previous EEF pilot of School 21's oracy approach, Maxwell and her colleagues had used The Raven's Progressive Matrices Test. This is a standardised test of pupils' non-verbal reasoning ability. Another option would have been to use a published standardised test of a mainstream school subject (English, maths, or science). Any such option would have had the advantage that the test had been standardised and therefore we could have been reassured that it would give robust results.

However, the development team at Voice 21 (as oracy advocates) took the view that none of the proposed standardised assessments captured the 'essence' of oracy as they would wish. Therefore, they designed a bespoke and direct assessment of this construct for use in this pilot.

Other options for an outcome measure were also discussed but were discounted for various reasons. For example, if these had been Year 6 pupils, it might have been fruitful to access their national curriculum test scores from the national pupil database (NPD). But since these were Year 7s, this was not an option.

We came to the view that there was probably no 'perfect' approach to assessing oracy in the context of this pilot. We set out the drawbacks of possible alternative approaches in the table below.

Table 4: Drawbacks, limitations, and risks of various approaches to assessing oracy4

Course of action	Drawback/limitation/risk
Work on the assessment system (e.g. firm up standardisation, constrain markers more, etc.) in the hope that this will result in more reliable measurement.	• If you constrain markers a lot, you might lose the essence of oracy (i.e. prioritise reliability over validity). • It might be a lot of work. • It might not actually deliver results (i.e. you could spend a lot of effort firming up the assessment procedures and improve reliability by only a small amount).
Use some other proxy as a measure of oracy (e.g. a reasoning test).	The proxy would never represent the construct of oracy fully (e.g. reasoning isn't the whole of oracy).
Accept that oracy measurement is just inherently error prone (legitimate differences of opinion between professional judges, etc.).	It will be harder to show progress in oracy using their assessment approach (due to the impact of measurement error). Therefore less likely to show 'success' within the EEF paradigm.

Given that any assessment would have had limitations, we considered it reasonable to go with the development team's suggestion of using a version of their oracy toolkit.

Nature of the tool

An oracy assessment tool was developed with the Faculty of Education, University of Cambridge as part of an earlier EEF-funded development phase. The assessments used for the baseline (pre-intervention) and post-intervention comprised two tasks—a talking point task and a presentation task. Within each task, the four main oracy skills (physical, linguistic, cognitive, and social and emotional) and several sub-skills were assessed. Initially, a third task had been developed which involved instructional dialogue but this was not used for the purpose of the evaluation as it was felt that the presentation and talking points tasks adequately covered the curriculum. Figure 2 shows the structure of the assessment tasks.

Figure 2: Components of the assessment tool

	Oracy skill	Sub-skill
Talking Points Task	Physical	Voice
		Body (expression/eye contact)
	Linguistic	Register & Grammar
		Range of vocabulary
	Cognitive	Content & reasoning
		Building on views of others, summarising & critically examining
	Social & Emotional	Turn taking, guiding & managing interactions
		Active listening
Presentation Task	Physical	Voice
		Body language
	Linguistic	Vocabulary & Grammar
		Register & Rhetoric
	Cognitive	Content & reasoning
		Structure & self - regulation
	Social & Emotional	Confidence & flair
		Audience awareness

According to a School 21 guidance document (School 21, 2016), there is a maximum of 48 marks (combined across the two tasks), where each skill was marked on a scale from 0 to 3 as described in Table 5. Table 6 shows how these combined marks place pupils in one of eight ordered categories and the verbal description/interpretation associated with each.

Table 5: Marking categories for baseline assessments

Mark	Category	Explanation
0	Foundation	Pupil demonstrates no evidence of skill.
1	Beginner	Pupil demonstrates limited signs of skill but it may not be intentional, effective or support the purpose.
2	Developer	Pupil demonstrates some signs of purposeful and effective use of this skill.
3	Confident	Pupil clearly and competently uses this skill purposefully, naturally and effectively.

Table 6: Ordered categories, their interpretations and associated mark ranges

Ordered category	Interpretation	Marks
1	Foundation—emerging	1–6
2	Foundation—secure	7–12
3	Beginner—emerging	13–18
4	Beginner—secure	19–24
5	Developer—emerging	25–30
6	Developer—secure	31–36
7	Confident—emerging	37–42
8	Confident—secure	43–48

Markers and standardisation

Teachers within each school marked pupils' oracy. At the start of each school's assessment activity, a standardisation session was led by the school oracy lead.

Voice 21 gave us access to a Dropbox, which contained:

- a set of supporting documents for participating schools, including marking criteria and a baseline assessment guide; on standardisation, this stated:
    - [performance on the two tasks] can be marked live or recorded on video;
    - [Voice 21 requests that] three filmed samples from each school [are returned to Voice 21]; and
    - markers should conduct a standardisation process to compare marks of three students;
- a set of video recordings of example performances on the two tasks; and
- completed marksheets with detailed explanations of why Voice 21 would award the pupils in the video exemplified tasks particular marks.

We reviewed this material and commented on its suitability and any issues arising from it. This was particularly in the light of literature on standardisation and assessment of group work.

Selection of pupils to participate in pilot

Each pilot school was asked to provide the oracy assessment results for a minimum of 60 pupils at the beginning and end of the year. Schools were not asked to apply any specific criteria when

selecting which pupils to assess; instead, schools were asked to provide background data on the entire cohort the programme was delivered to (in other words, not just the 60 pupils who underwent assessments) and an assessment of representativeness was applied at the analysis stage (see the Participants section of this report, specifically Table 10.

Schools approached the assessments with some trepidation (this was apparent at the training day held at School 21) and it was felt that asking them to undertake a selection process on top of the other demands of applying these unfamiliar assessments to 60 pupils would prove to be too burdensome and might deter participation in the assessment, or the pilot overall. Assessment data for both the baseline and follow-up assessments was received from 10 of the 11 participating schools (the remaining school administered and sent data for the baseline but not the follow-up assessments).[1] The two datasets were analysed using t-tests to assess the progress made between the two points.

Table 7 summarises the achieved sample within the assessment data. There were certain children who had been tested initially that did not get tested again and vice versa. Only nine of the schools followed up had supplied data on both tasks. School B only had follow-up data available for the talking points task.

Table 7: Achieved sample of school and pupils returning complete data sets

	Number of pupils completing tasks—			
School ID	at baseline only	at follow-up only	at both time periods	Total
C	0	0	64	64
D	0	0	55	55
E	1	2	60	63
F	1	0	67	68
G	0	6	54	60
H	0	4	64	68
I	1	3	60	64
J	0	0	50	50
K	25	3	27	55

Each of the schools had a similar number of pupils that had taken both tasks at baseline, at follow-up, or during both time periods. Only three of the schools had all pupils that completed both tasks during both time periods. Additional pupils undertook the task during the follow-up with the additions ranging from two (School E) to six (School G). There were four schools where some pupils completed the task at the baseline but were not followed-up. However, three of these schools only had one pupil who did not take the task at the follow-up and in two of these schools, additional pupils had been added in the follow-up. School K had 25 pupils that had completed the task at the baseline but were not followed up, however this school also had additional pupils at follow-up.

Any pupils that completed only one part of the assessment were not included in the t-tests to analyse the difference in scoring of the baseline and follow-up tasks. All pupils that completed the baseline task were included in the regression analysis to analyse scoring and potential relationships to other background predictor variables.

Review of the oracy assessment tool

The review of the oracy assessment tool involved the analysis of the assessment content and procedures and addressed issues of content validity and procedural best practice. This involved the

[1] Cases were excluded from the data set on a 'listwise deletion' basis; that is, if a pupil's data set was incomplete it was not analysed. This aided an already somewhat complex analysis.

study of a map of the Oracy Skills Framework and the assessment outcomes covered by the assessment tasks (with a view to examining the extent to which the assessment outcomes cover the whole framework) and a review of videos and accompanying materials provided by Voice 21 to participating pilot schools in the light of the literature and best practice in assessment standardisation.

The *validity* of an assessment tool refers to the extent to which it measures the construct it is intended to measure (in this case, oracy), while the *reliability* of an assessment tool refers to the extent to which it provides stable and consistent results. Both of these aspects of the oracy assessment tool were explored through a series of analyses. Note that these analyses were based on the data returned by ten of the pilot schools from their baseline assessments conducted in autumn 2016. (Although 11 schools were participating in the pilot at this time, one school had not returned any baseline assessment data at the point at which these analyses were undertaken.) An initial technical report on these analyses was reviewed by the advisory group prior to the analysis of progress (i.e. the t-tests) using baseline and post-intervention assessment data.[2]

Voice 21 gave the Alpha*Plus* evaluation team access to a range of materials and resources associated with the oracy assessment tool. This included:

- a set of supporting documents for participating schools, including marking criteria and a baseline assessment guide;
- a set of video recordings of example performances on the two tasks; and
- completed marksheets with detailed explanations of why Voice 21 would award the pupils in the video-exemplified tasks particular marks.

Assessment instruments and procedures

Two analyses of the assessment content and procedures were conducted. Broadly, these addressed issues of content validity and procedural best practice. The following analyses were undertaken:

Analysis of curriculum coverage

This involved studying a map of the oracy curriculum and the assessment objectives covered by the two assessment tasks used in the initial oracy assessments. The extent to which assessment objectives in the two tasks covered the whole curriculum was reviewed.

Evaluation of best practice in assessment standardisation

Videos and accompanying material provided by Voice 21 to the other schools participating in the oracy pilot were reviewed and commented upon.

There is a vast literature surrounding standardisation of assessments (especially productive skills such as speaking and writing), and we cannot hope to do more than dip into such a vast field. In an earlier review of the standardisation literature, we had pinpointed a consensual approach to standardisation, which we described as follows (Alpha*Plus*, 2013, p. 56*)*:

> *In contrast to the hierarchical approaches to standardisation that appear to predominate in recent literature, Brown (1999) describes a consensual approach to standardisation under New Zealand's National Education Monitoring Project (NEMP). Children undertook tasks – either individually with a teacher, or in groups. These tasks were video recorded. Brown (1999) describes a process she calls 'cross marking'. In this process teacher-markers viewed a succession of video performances and discussed proper scoring in a group of up to 20. This process was repeated until consensus was felt to have been reached on features of performances that were associated with particular scoring levels.*

[2] The Advisory Group comprised Prof Neil Mercer and Dr Ayesha Ahmed from the University of Cambridge and Dr Dougal Hutchison from Alpha*Plus*.

Brown (2009, p. 10) argues that cross marking enhances validity as follows:

> *Cross-marking allows markers to apply their professional judgement to these issues and then receive feedback from others. In doing so, markers develop a robust understanding of the task construct and the qualities associated with each grade, which can then be applied to the range of responses that are generated in the NEMP data. Cross-marking therefore facilitates the development of a sense of 'ownership' amongst markers which is used to aid consistency when making judgements on student performance. Discussions which occur during cross-marking also allow markers to share their experience of a range of student responses, and in so doing they may collectively identify the need for additional categories which are not covered by the existing marking criteria. Cross marking therefore enhances the validity of the marking process by allowing a more accurate and representative picture of student achievement to emerge.*

Given that schools participating in this pilot were attempting to develop a community of teacher expertise, we considered that the New Zealand example (and its justification by one of its originators) could provide valuable insight for the oracy pilot participants.

There is substantial literature on the assessment of group work in Higher Education, and we referred Voice 21 to sources such as Nordberg (2006), O'Neill, (2013), and Carnegie Mellon University (2015). This literature tends to suggest that assessing group work is a valuable thing to do, but it does not provide a single 'silver bullet' or a simple template that can be followed without thought to effect high quality assessment of group work.

Rather, this literature points thoughtful practitioners towards some important issues that they need to bear in mind. These include:

- Assessment of group work needs to be supported by clear rating scales, and mark schemes.
- Those assessing group work could consider either scoring individuals separately for their work within groups, or giving all members of the group the same score (because they collaborated well or badly, and the collective performance was the thing being assessed).
- If individuals are given separate scores for group work, assessors need to have a position regarding how to score certain types of interactions; for example, they need to develop a position in circumstances where one individual dominates the interaction and thus inhibits others from participating. Conversely, some individuals could 'slipstream'—they might be in a very articulate, well-functioning group, but not contribute much themselves. Although there is no simple or catch-all approach to scoring such performances in group work, it is important for assessors to be aware of the phenomena.

Evaluation of the data set generated by the initial assessments

A teacher assessment of a productive skill such as oracy might not generate a robust and internally consistent data set. This is not intrinsically problematic in many circumstances (for example, for formative assessment), but if large quantities of measurement error variance (residual variance) are apparently present in the data, and/or other features abound (such as apparently inconsistent application of the standards), it will constitute an important inhibitor to move this project to a subsequent trial.

Assessment research, or the study of validity and reliability, is the mainstream approach to evaluating the quality of educational assessments and their outcomes. In the research protocol that governs this project, the grantors and grantees agreed that:

> *[A]s assessment researchers, our natural approach is to scrutinise the validity and reliability of any assessments used.*[3] *Validity and reliability are generic properties of 'good assessment', but researchers can choose to focus on particular facets or aspects of each. What amounts to sufficient and suitable evidence to consider an assessment (and its use) sufficiently reliable and/or valid depends upon context. For example, if an assessment is used low stakes to help teachers' judgements of students' progress, then it probably doesn't matter so much if the instrument does not provide consistent measurement between judges and/or centres. However, if the use is to provide scores for a highly quantitative comparison such as a Randomised Controlled Trial (RCT), then it is essential that an improvement of X score units in school 1 means the same thing as a rise of X score units in school 2.*

There is a massive corpus of literature on assessment research. One publication that has been described as 'canonical'[4] is Educational Measurement (Brennan, 2006), but many other publications also set out tenets of assessment research.

The above evaluations of assessment content and standardisation procedures address facets of validity; but there are also specific considerations when one is dealing with a data set generated by a relatively sophisticated (or complex) assessment technique that has many 'moving parts'. In particular, we want to know, for instance, that differences in scoring represent real differences in pupils' oracy ability, and not (for instance) differences in the leniency of markers, or the difficulty of assessment tasks.

To investigate the data sets generated by complex assessments, two approaches are well accepted in educational research. Generalisability theory (g-theory) uses techniques derived from analysis of variance (ANOVA) to show how much of the variance in scoring can be attributed to particular sources (for example, pupils, markers, schools, questions—which we refer to as 'items', or tasks). Furthermore, g-theory lets us think about how variables are related to each other; for instance, pupils might be 'crossed with' items (all the pupils answer all the items). Alternatively, items might be 'nested within' tasks—that is, that particular item only exists within the context of that particular task. To understand the contribution that an item makes to the measurement, it is necessary to model it in the context of the task within which it sits.

G-theory is also a generalisation from 'conventional' reliability analyses, such as internal consistency studies using indices such as Cronbach's alpha, or inter-rater studies. Such studies would be limited in so far as they would not account for various sources of scoring variance (as listed above). Thus, a g-study is preferable in this instance.

Rasch FACETs analysis derives from a different theoretical tradition to g-theory, but can used profitably on complex assessments. FACETs' insights derive back to Georg Rasch's measurement model that seeks to conceptualise test-takers' abilities and questions difficulties in a simple relationship (Rasch, 1960, 1980). The Rasch model has been extended in various ways; of particular relevance for the current work is Michael Linacre's FACETs software (Linacre, undated). This application allows analysts to extend the simple test-taker–question relationship to include other facets, such as items nested within tasks, pupils nested within schools, and so on. To some extent, the outputs of FACETs bear comparability with those of g-theory. However, the Rasch model's particular contribution and emphasis is on scaling facets relative to each other. In this work, a variable map is provided that shows the relative locations of various facets (schools, pupils, tasks, and items within them) in respect of a single scale or 'ruler'. More details on the 'FACETs ruler' is given when we report results below.

[3] We would argue, in fact, that researchers who use tests in their projects can tend to pay insufficient heed to the reliability and validity of assessments they use. The disciplined way in which the validation of educational assessment is carried out (for example, by exam board research departments) can bring substantial insight to generic educational research programmes that employ tests.
[4] By Newton *et al*. (2008).

Many publications elaborate the bases of g-theory and Rasch FACETs. On g-theory, Brennan (2001) is authoritative, and the reliability chapter in *Educational Measurement* also has useful exposition of the approach (Brennan, 2006). Eckes (2011) is a readable introduction to Rasch FACETs. In a U.K. schools context, Johnson and Johnson (2012a) give a comprehensive run-through of the bases of g-theory, and Baird *et al.* (2013) show appropriate uses of g-theory and Rasch FACETs in evaluating complex U.K. school assessments.

As with any statistical technique, g-theory and Rasch FACETs modelling come with some underlying assumptions. G-theory has less strict assumptions, being only limited by the assumptions underlying ANOVA theory (for example, normality of data). Rasch modelling, in contrast, has been said to be subject to stricter assumptions—specifically, the assumption that score variance can be understood in respect of a single dimension of ability. However, because multi-faceted Rasch analysis is a substantial extension of core or primitive Rasch models, it is debatable as to how significant the 'unidimensionality assumption' truly is. Indeed, one may argue that such an assumption will apply in any circumstance in which one uses a single test score to summarise a learner's ability (rather than a skills profile or descriptive grid of some sort).

The range of assessment research

The range of assessment research that was carried out is described below.

Generalisability theory analyses

Generalisability theory (g-theory) is a suite of analytical approaches that derive from the analysis of variance (ANOVA), and give several different outputs of interest.

In this project, multivariate g-theory analysis using Brennan's URGenova programme was run to enable the analysis of samples of pupils that differed in size for each school (that is, each school returned datasets containing slightly differing numbers of cases).[5]

A complex, nested measurement design was specified to reflect the sources of variance, or measurement facets[6] that influenced scoring in the Voice 21 oracy assessments,[7] and the following outputs were derived from the URGenova programme:

- the relative generalisability coefficient, which reflects the amount of variance attributable to differences between assessment participants;
- the absolute generalisability coefficient (labelled 'phi') which takes into account variance between test-takers, but also counts, as a differentiation facet, factors such as assessors and the particular items and tasks presented; and
- the portions of variance that may be explained by different elements of the assessment procedure; quantifying portions in this way allows us to see how much of the variance can be attributed to oracy ability, and how much can be attributed to elements of the assessment design.

A contextualised explanation of 'nesting', 'crossed facets', 'differentiation', 'instrumentation', and 'stratification facets' is given at Table 27 and surrounding text (p. 56).

Rasch FACETs analysis

As noted above, the main output from multi-faceted Rasch analysis was a 'variable map' or 'Facets ruler' (Eckes, 2011, p. 40). This map, or ruler, allowed us to position several variables pertaining to the

[5] Brennan's URGenova programme: https://tinyurl.com/yd2lzekv
[6] Or, more colloquially, these elements of the assessment design that influenced scoring on the assessment.
[7] This includes both individual sources of variance/Facets (e.g. teachers, schools, tasks), and the way that they are related to each other (e.g. Facet A may be nested within Facet B, but crossed with Facet C).

assessment procedure relative to each other. It allows us to compare different entities within the assessment procedure—for example, allowing us to see whether certain schools are more lenient than each other, or whether certain tasks appear easy or hard.

Multiple regression analysis

In conducting the Rasch FACETs analysis to derive the variable maps, one important point became very clear: the measurement design contained an important confound. If we tried to compare schools' oracy standing relative to each other, we could not tell whether school X scored very highly because their pupils had genuinely high oracy skills, or whether that school simply had more lenient oracy markers.

To disentangle this confound, a multiple regression model was set up. This aimed to predict a pupil's oracy score from some background variables about each pupil—such as the school they attended, prior attainment, EAL (English as an additional language) status, SEN status, and an indicator for Pupil Premium.

The model selected can then be used to assess what combination of variables is correlated with a higher score for a pupil on the oracy assessment. The regression model shows the contribution to a pupil's oracy score for each background variable including the school that a pupil attended as a fixed effect. This identifies which, if any, of the schools confers an advantage onto the pupil regardless of other background variables that are controlled for; for example, one possible contributing factor would be whether they appear to contain unduly lenient markers. Schools could have coached pupils more, or there may be other reasons, but the analysis controlled for the most important pupil-level covariates and there were still significant school effects.

By including covariates for ability in the form of KS2 results, it was attempted to control for the most important confounding factors enabling the generation of estimates of the effect schools had on the pupils. The combination of the covariates included were an individual pupil's EAL status, their prior attainment, SEN, and FSM. These are some of the most powerful predictive covariates at a pupil level.

Theoretical issues concerning progress scores

This project is at the pilot stage of the EEF's process. Therefore, there is no formal requirement to produce effect size calculation to show the impact of oracy. However, it is useful to consider how such an effect size should be calculated. This is particularly so given the phenomenon of 'regression to the mean' and the likelihood that the measures of oracy that we derive contain substantial amounts of residual or error variance.

To analyse the progress of learners in the oracy pilot, repeated measures of pupils were used. Each pupil in the trial is assessed for oracy ability before any teaching occurs, and at the end of the trial they were assessed again, and then change scores were computed. This is the difference between the follow-up and baseline scores.

Let us denote the observed baseline and follow-up scores as S_b and S_f respectively. Note that both scores are observed with error. It is useful to be able to quantify the amount of error present in these measurements. In the g-theory analyses, the proportion of variance attributed to instrumentation was calculated and, when added to the residual variance, this resulted in a substantial amount of 'error' variance present in the scoring.

We can also directly calculate the variance of the change score statistic, $S_f - S_b$, by using the variance sum law:

$$Var(S_f - S_b) = Var(S_f) + Var(S_b) - 2Cov(S_f, S_b)$$

$$= \sigma_{S_f}^2 + \sigma_{S_b}^2 - 2\rho\sigma_{S_f}\sigma_{S_b}$$

Where ρ is the correlation between baseline and follow-up scores.

If there is strong correlation between the two scores and ρ is large then the amount of variance of the change scores is reduced and may even in some circumstances be quite small. However, due to the nature of oracy assessment having a larger error associated with it than more traditional assessments—and the g-theory analyses above—we expect variances to be quite large.

Some problems with change scores are described in Allison (1990). In addition to the lack of a comparison group, the two most notable problems described in that work were the unreliability of change scores compared to their component baseline and follow-up scores as well as the phenomenon of regression to the mean. Regression to the mean is a common problem when estimating the effect size in trials of this type as baseline values can be negatively correlated with change because pupils with low scores at baseline generally improve more than those with high scores (Vickers and Altman, 2001).[8] These effects can both lead to spurious conclusions being drawn.

To properly assess the effect of the oracy trial on pupils, a different study design than that used here would be needed. A design that randomly assigns pupils into an experimental group (those who receive instruction on oracy) or a control group (who receive no instruction) would ideally be used.

This trial did not have a control group so any conclusions that can be drawn from it are limited. For example, if the oracy ability of every pupil in the pilot improved it would not be possible to say for certain that this pilot was the cause. Also, it would not be possible to quantify by how much the trial improved oracy ability over and above any normal progression.

Representativeness of sample of pupils

In any pilot that aims to investigate the effectiveness of some intervention on a sample and then to generalise this to the wider population, it is important to ensure that participants (in this case Year 7 and 8 pupils) are representative of their cohort. This involved establishing (for example) whether schools participating in the pilot had selected a representative sample of pupils to take part in the pilot (although as discussed previously, in order to minimise the burden on schools, they were free to select the 60 pupils on any basis—there were no specific criteria specified for selection), or whether the pupils were unrepresentative of all the school years' pupils in terms of some background variables. The background variables considered were SEN (special educational needs) status, Pupil Premium, and EAL.

These variables were chosen as a relatively lightweight suite of proxy measures for learners' disadvantage and specific issues that might be relevant given the nature of oracy as a construct. We needed to have relatively well-known and straightforward variables as we were gathering the data from schools and did not wish to impose onerous burdens on them. Other variables could have been chosen—such as free school meals rather than Pupil Premium—but in other work we have found FSM to be a somewhat unreliable indicator of disadvantage (see, for example, Coughlan, 2017).

The representativeness of the sample of pupils is discussed further in the Participants section of this report.

Timeline

The project timeline is summarised in Table 8 below.

[8] The 'regression to the mean' problem could be avoided if one did not use change scores (for example, only used a post-intervention score) to evaluate an intervention in an RCT. This can be legitimate, and many RCTs do this. However, if one only used an average score at the end of the intervention, there would be no sense of how much pupils progressed during the trial. High scoring pupils at the end of the pilot might have been good before experiencing the intervention.

Table 8: Timeline

Date	Activity
Spring 2016	**Recruitment of schools to the pilot, development of MOU, theory of change workshop.**
July 2016	**Pilot School training at School 21.**
September 2016	**Schools start delivering the programme.**
October–November 2016	**Schools undertake baseline assessments (pre-intervention).**
November 2016	**Face-to-face interviews with SLT members, oracy leads, and teachers at pilot schools.**
December–February 2017	**Analysis of baseline assessment and review of assessment tool.**
February–March 2017	**Telephone interviews with oracy leads.**
March 2017	**Online surveys of oracy leads and teachers.**
June 2017	**Second round of oracy assessments (post-intervention).**
June–July 2017	**Face-to-face interviews with SLT members, oracy leads, and teachers at pilot schools.**
July 2017	**Online survey of teachers.**
July–September 2017	**Analysis of assessment data, analysis of interview and survey data.**
October–November 2017	**Report writing.**

Findings

Participants

An initial 12 schools were selected for inclusion in this pilot. Three were located in the North East of England, three in the North West, four in the South East, and two in the South West.

The characteristics of these schools are presented in Table 9 below. All schools are situated in areas classified as 'urban' and all have a comprehensive admissions policy. School L dropped out of the pilot at the start of the first term, and School A began the pilot and took part in evaluation activities up to spring 2017, but then was non-responsive to both the evaluation team and the Voice 21 team and took no further part in any evaluation activities.

Table 9: Characteristics of pilot schools

ID	School size (capacity)	% FSM	Ofsted	Type	Age range
A	2,144	30.1	Requires improvement	Academy	11–18
B	1,350	13.6	Good	Community school	11–16
C	1,484	12.6	Good	Academy	11–18
D	1,500	18.2	Good	Academy	11–16
E	1,175	10.3	Requires improvement	Voluntary aided school	11–19
F	1,000	27.2	Outstanding	Voluntary aided school	11–18
G	900	19.7	Good	Community school	11–18
H	764	24.5	Good	Voluntary aided school	11–18
I	950	7.4	Good	Academy	11–19
J	1,152	32.1	Inadequate	Academy	11–19
K	1,349	5.2	Good	Academy	11–18
L	900	28.5	Good	Academy	11–16

At the pupil level, all pilot schools were asked to provide background characteristics on the entire Year 7 cohort (or Year 8 where this was the year group included) along with the oracy assessment results (at the start and end of the year) for a minimum of 60 pupils. Assessment data for both the baseline and follow-up assessments were received from 10 of the 11 participating schools (the remaining school administered and sent data for the baseline but not the follow-up assessments). Table 10 summarises the complete baseline assessment datasets received and the number of pupils both baseline and end-of-year assessments were returned for (only the baseline assessments were used in the reliability and validity analyses). Note that the reliability and validity analyses were conducted on the baseline assessment data returned by nine schools rather than ten because School K returned only 10 assessment results for both tasks, not a sufficient number to warrant inclusion on the g-theory and FACETs analysis.

Table 10 below summarises the number of pupils at each school for whom both sets of assessment data were provided.

Table 10: Achieved sample of schools and pupils returning complete baseline and follow-up assessment data sets

School ID	No. of pupils completing the baseline assessment tasks	No. of pupils completing tasks at both time periods
B	66	
C	64	64
D	55	55
E	62	60
F	68	67
G	60	54
H	68	63
I	63	60
J	60	50
K	30	27

Pupils' background characteristics were also requested to help establish whether schools participating in the pilot had selected a representative sample of pupils to take part in the pilot. The background variables considered were SEN status, Pupil Premium, and EAL status.

These variables all have a binary classification ('yes' or 'no') enabling analysis via a series of Chi-squared tests. Several of these tests were performed to see whether the trial participants differed to non-trial participants from each school on the three background variables mentioned above. Results are shown in Table 11.

Table 11: Chi-squared test statistics and p-values

School ID	SEN status		Pupil Premium		EAL	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
B	2.43	0.119	**11.1**	**4.40e-4**	1.09e-30	1
C	1.07	0.302	0.642	0.423	0.220	0.639
D	**5.63**	**0.00883**	**4.74**	**0.0147**	**9.67**	**9.36e-4**
E	0.0819	0.775	**10.0**	**7.77e-4**	0.0871	0.768
F	1.50e-30	1	1.63	0.201	0.0871	0.768
G	0.727	0.394	0.430	0.512	2.25	0.133
H	5.47e-31	1	0.0355	0.851	2.27	0.132
I	**19.0**	**1.29e-5**	0.782	0.377	0.0275	0.868
J	0.318	0.573	3.28e-30	1	1.96	0.161
K	**14.2**	**8.24e--5**	**5.05**	**0.0123**	3.02e-30	1

Values that were significant at the 5% level are highlighted in bold text.

From Table 11, it is evident that most schools provided samples of trial participants that were representative of their wider cohort at the 5% significance level. Some schools, however, such as D and K, provided samples of trial participants that differed substantially on at least two out of the three background variables considered.

The power of these tests is dependent on the effect size, sample size, and significance level. We looked at the power of a two sample comparison of proportions test.

To give a concrete example of the power of this test, we assume, and found, that the alternative hypothesis of the test is true and that the two proportions that we are testing, p_1, p_2, differ. Assume that $p_1 = 0.6$ and $p_2 = 0.8$ with when assuming a size of n = 70 observations=80 in each sample (close to the average size in this study) and with true proportions of 0.6 and 0.8, so an effect size of 0.2 and a significance level of 0.05, then the power of this test is 0.738794 which we would class as good. Of course, when the true effect size is smaller, the power does decrease rapidly, however this is to be expected. For the sample sizes we have in this problem, and the effect sizes we wish to detect, we believe the power of this test is adequate.

Evidence to support the theory of change

Two logic models were created to articulate the theory of change underpinning this intervention, one was School 21-centred and based on how the programme was originally conceived and is currently implemented at School 21, and the other was pilot school-centred and reflected implementation for this pilot phase. The logic models are presented in Figures 3 and 4. The **inputs** describe the core elements of the intervention, that is, who is doing what to or with whom. The **outputs** might here be conceived as necessary, intermediary outcomes of the intervention in pupils and involve assumptions about the preconditions for the intervention to work as expected. The **outcomes** articulate the short and medium term positive changes the Voice 21 oracy model is seeking to achieve in pupils, whereas the **impact** is about longer term as well as socially important intended changes. Taken together, the outputs and impact text boxes explicate the rationale for the intervention.

Based on the theory of change workshop and the team's resulting understanding of the project, three specific research questions were developed to support the theory of change, as follows:

1. To what extent is it plausible that the Voice 21 model would result in (positive) changes to teaching and learning oracy across a school?
2. To what extent is it plausible that any changes in teaching oracy translate into improvements at the pupil level (in oracy, reasoning skills, attainment, or other)?
3. To what extent do we see changes in pupils' oracy on pre and post measures of oracy?

These research questions were linked to the key outcomes and impacts of progression in oracy, perceptions of changes in the academic and socio-emotional areas such as attainment, wellbeing, and impacts on different groups of pupils.

Figure 3: School 21-centred logic model (based on the programme as originally conceived)

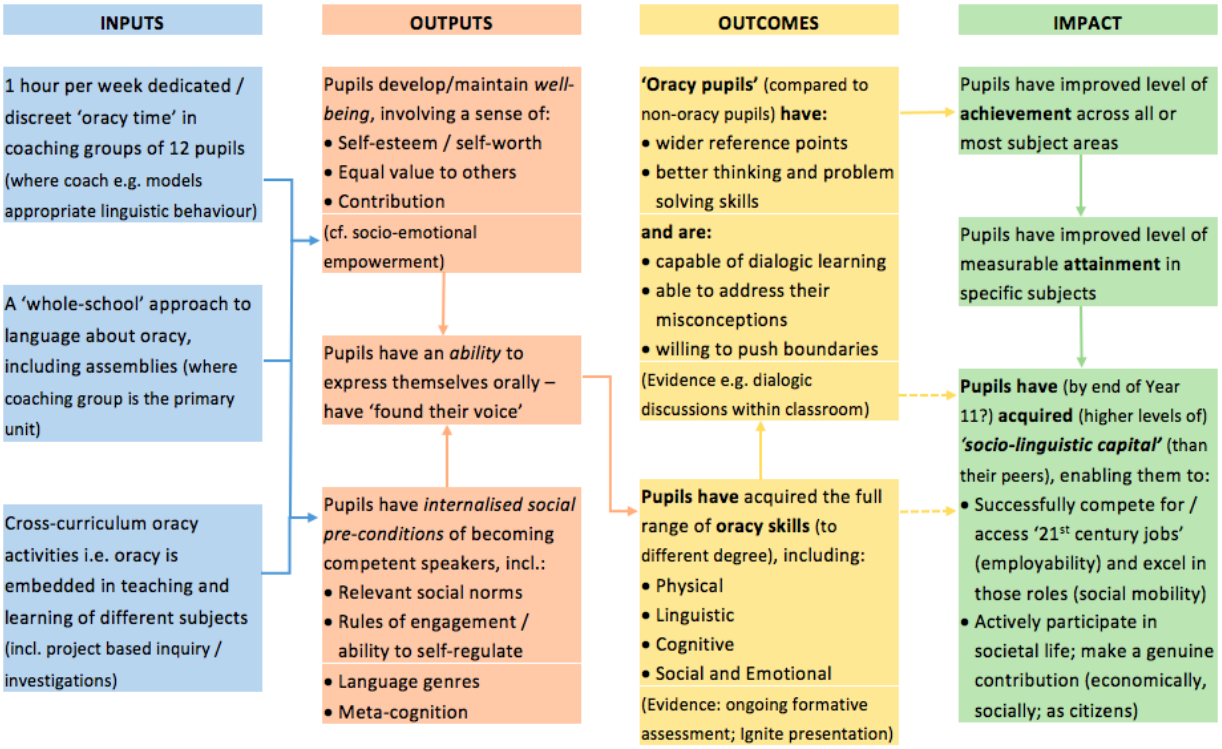
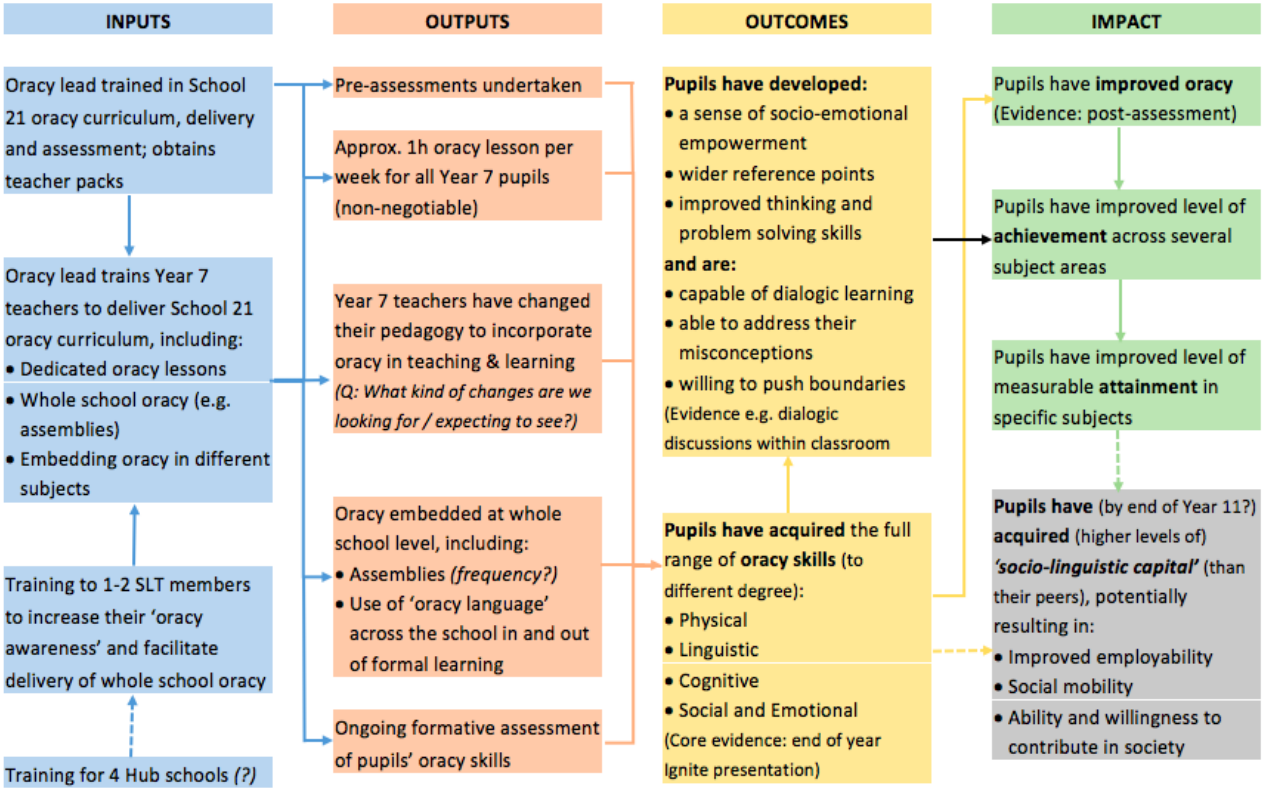


Figure 4: Pilot school-centred logic model (based on implementation in this pilot)



Research question 1: To what extent is it plausible that the School 21 model would result in (positive) changes to teaching and learning oracy across a school?

To set the context for this research question, we first outline the findings from (a) the interviews relating to the expectations of deliverers and (b) the provision at the school prior to the programme.

Expectations of the programme

Table 12 shows interviewees' expectations of the programme when first interviewed during November. Given that the oracy programme was intended to be delivered to Year 7 pupils only, the expectations of oracy leads, teachers, and SLT members in terms of the potential impact the programme could have on the school as a whole were understandably modest and realistic. However, there were several mentions of the oracy skills development and general ethos or culture of oracy spreading beyond the cohorts receiving the oracy lessons. (The category 'other' in Table 13 includes expectations that were mentioned by only one person each.)

During the end-of-year interviews, six interviewees (two oracy leads, three SLT members, and one teacher) said that their prior expectations of the programme—as expressed in earlier interviews—had been met or exceeded. One SLT member, however, described how their expectations had changed during the pilot year from initially wanting the programme to enhance pupils' oracy skills in order to equip them for life after school, to the more immediate benefit of enhancing their general wellbeing and providing the skills to express themselves and explore ideas and issues. Likewise, an SLT member at another school felt that the programme had not had the impact that they had expected, in this instance they felt that there had been improvements in pupils' confidence but that this had not been as far-reaching as anticipated.

Table 12: Summary of deliverers' expectations of the programme as at the first round of interviews in November 2016

Expectations of the pilot	Number of individuals mentioning each expectation	Number of different schools giving each expectation
Improve pupils' oracy skills.	8	5
Enhance oracy skills in all subjects, both in and out of the classroom.	7	5
Increase pupils' wellbeing, making them confident and socially competent.	7	5
Increase pupils' confidence.	4	3
Giving teachers confidence and skills to develop oracy.	3	2
Improve behaviour / behaving in the 'right way', etc.	3	2
Realistic in the context that this is a pilot.	3	3
Change the culture of teaching and learning to focus more on oracy.	2	2
Improve collaborative teaching and learning culture.	2	2
Other.	3	3

Provision prior to the pilot programme

To establish the 'baseline' of current practice in pilot schools, initial interviews with teachers, oracy leads, and SLT members asked whether oracy was already being taught as a distinct skill, and in what form (some schools were already using other programmes or initiatives).

Only two interviewees (from different schools) said that no existing provision was in place; however, most people's understanding of what constitutes 'oracy skills development' was somewhat looser than that intended by the Voice 21 programme.

Three interviewees (all from different schools) said that oracy was already being taught to some extent, but in a very unstructured way, and a further four (representing three schools) said that their existing teaching tended to focus on speaking, not on listening.

Two interviewees (in different schools) felt, prior to the pilot, that oracy was taught mainly through drama and the oracy lead at one school explained that they had already been working around the different departments giving oracy coaching to staff. Where specific initiatives had been in place before getting involved in the oracy pilot, these were as follows (note that Pixel and Articulacy UK were both offered at the same school):

- Listeners' Project (focusing on active listening skills at the end of each class: two minutes to reflect on what they had learned, two minutes to talk to a partner, and one minute of feedback from each pair);
- Pixel programme (public speaking coaching for sixth form pupils); and
- Articulacy UK (working with Year 8 pupils on their oracy skills with a view to achieving the English Speaking Board level one qualification).

These first interviews also asked about interviewees' knowledge and awareness of oracy as a skill. Eight respondents felt that their awareness was good, although this was usually because they were drama or English subject specialists and felt that it was an important element of their subjects. A further seven described their knowledge and awareness of oracy as fairly limited (again, based mainly on the extent to which it formed an inherent part of their subject specialism), and in two cases, interviewees said that the term 'oracy' was new to them although they could now see that they were teaching aspects of it but under different terminology.

During the second set of interviews—with oracy leads only at the mid-point in the pilot year—eight of the eleven interviewees indicated that their experiences of teaching oracy via the Voice 21 programme had changed their perceptions of what oracy is. These changes generally reflected key elements of the programme, for example, two people described how the cognitive elements of oracy had been emphasised by the programme (a layered process of thinking, drafting, redrafting, and then speaking), while another had come to realise the importance of listening skills.

Perceived changes to teaching and learning

The extent and type of changes stakeholders reported making to their teaching and learning evolved throughout the pilot year and were gauged via the interviews conducted at the beginning, middle, and end of the academic year. During the first set of interviews, most interviewees indicated that at this early stage in the programme, the main changes to their teaching practice were in using the techniques and skills both in their oracy lessons and in other lessons (in many cases they stated that the extension of these techniques into other lessons was not always a conscious decision, they seemed to 'creep in' once they had used them in the oracy lessons).

Some had noticed broader changes to their practice. For example, three interviewees felt that they had become better at setting up and facilitating discussions in classes (outside of the oracy lessons); two teachers said that they had a better awareness of, and hence focus on, the non-speaking elements of

oracy such as listening and gestures during lessons. Two interviewees said that they now tended to emphasise speaking before writing and reading in their classes, for example, encouraging discussion in pairs before starting any written work, and a further two said that they found themselves focusing on developing cognitive skills, in particular, encouraging pupils to think before speaking.

At the point of the second interviews, two oracy leads indicated that they had not made any major changes to their teaching practice. In one instance, the teacher felt that it simply involved minor changes associated with teaching any new curriculum and—because a drama teacher—felt that many of the techniques were an integral part of their normal teaching anyway. In the other case, the oracy lead felt it had not affected their own teaching practice because the subjects they teach (e. g. maths) were not felt to be conducive to the skills and techniques taught in oracy sessions, this person felt it was a more natural step to incorporate oracy teaching techniques in subjects such as English. Note that this school was the one that did not participate further in the evaluation activities after the second round of interviews, therefore it is not known whether changes to teaching practice did take place later in the year.

Three oracy leads commented that the changes mostly affected their teaching in the oracy sessions rather than their wider teaching, and that these changes involved implementing the skills and techniques required to deliver the programme. Eight said that they had used the skills and techniques in other lessons outside of the dedicated oracy sessions, and with other year groups, and some were also aware of colleagues doing so. The main areas of specific change mentioned included becoming better at facilitating discussions in class, and being better at establishing roles and structures in spoken communications.

In the third set of interviews towards the end of the pilot year, some of these changes were mentioned again (in most cases by the same teachers). However, the most frequently mentioned change to teaching practice described in this final set of interviews was the use of the techniques and terminology in other lessons (outside of the oracy lessons) and with other year groups. Twenty-two of the 36 interviewees (representing all ten schools that took part in these interviews) stated that this had been the main change in their teaching practice. Some interviewees described how well other colleagues and departments had embraced and implemented oracy techniques and skills, and in some cases, this was evident in subjects that they had not originally expected to adopt oracy teaching and learning in the way that they had. Several mentioned that teachers from humanities and modern foreign languages subjects had adopted techniques from the oracy programme, and in some instances, interviewees expressed surprise at how well colleagues teaching subjects that did not initially seem to lend themselves to the oracy techniques (such as maths) had embraced the skills and techniques. The most frequently mentioned ways in which the oracy techniques were being adopted outside of the oracy lessons was in facilitating discussion by using the talk protocols and assigning roles to pupils such as instigators, builders, and challengers.

School ethos and culture

During the first set of interviews, two interviewees said that one of their expectations was that the programme might help to embed an ethos of oracy across the school, and a further teacher indicated that they expected oracy skills and techniques to extend beyond the Year 7 cohort they were delivering the programme to. At these first interviews it was too early for anyone to say whether any progress had been made towards these goals, however, during the second interviews oracy leads were specifically asked whether they had noticed any changes in the overall school ethos in relation to oracy since beginning the programme. Four said that they had started to see a shift in the culture around oracy with staff and pupils taking on board the skills and techniques and that there was a more general awareness around oracy. Three felt that the shift in the oracy culture would happen but that it had not yet made any tangible impact. In a couple of cases this was dependent on outside factors: one felt this would only happen if there were no further initiatives implemented that might detract from or compete with oracy, and the other felt it was dependent on GCSE success, stating that if GCSEs did not go as well as hoped

then there was less chance for oracy to become more embedded in the school ethos. There was anecdotal evidence of additional work that some oracy leads were undertaking to help embed oracy in the school culture, for example, having school-wide 'no pens' days, and placing posters in classrooms relating to the oracy techniques and principles.

In the final interviews in the summer term, no one felt that they had yet achieved a whole-school oracy culture. However, 13 interviewees (from eight different schools) said that they did feel that the ethos had shifted during the pilot year and that good progress was being made towards this, while a further seven interviewees (from six different schools) felt that progress had been very limited on this front.

Research question 2: To what extent is it plausible that any changes in teaching oracy translate into improvements at the pupil level (in oracy, reasoning skills, attainment, or other)?

While the pilot project involved the collection of specific oracy assessment data (discussed further below), oracy leads, teachers, and SLT members were also asked to describe their perceptions of the effects of the oracy programme on their pupils. There were four main themes in their responses: general enhancements to oracy skills, social-emotional changes, academic performance or attainment improvements, and differential benefits across pupil groups; these are discussed below.

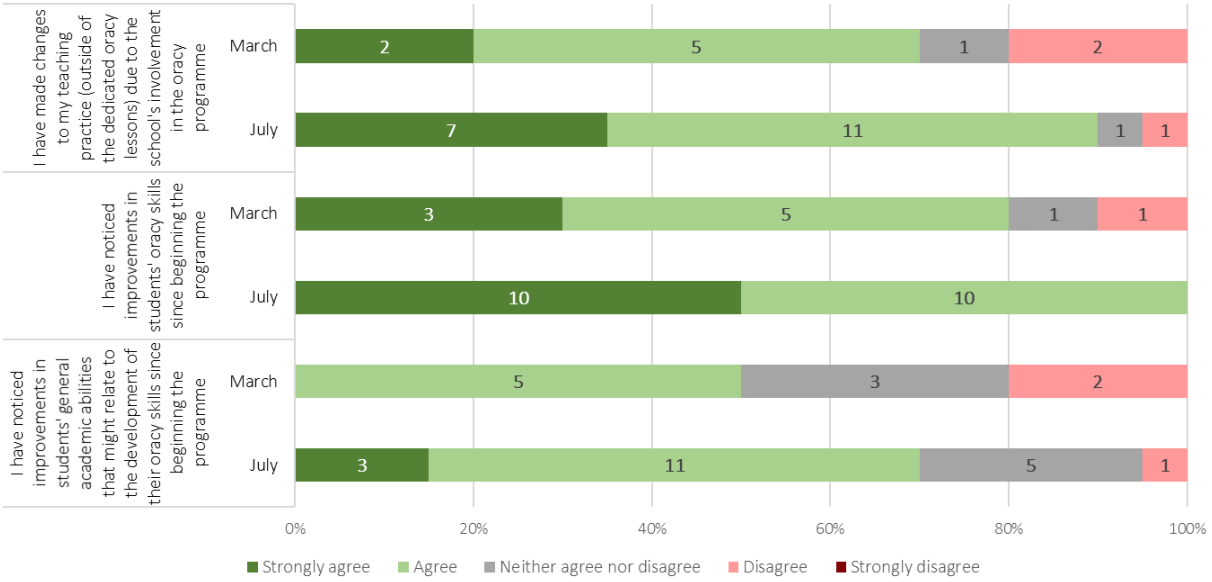
General enhancements to oracy skills

Throughout all of the interviews there were several mentions of how pupils' oracy skills were seen to be improving generally, and some expressed surprise at the effect it had had on pupils' listening skills (both in and out of oracy lessons, and in terms of listening to each other as well as to teachers). During the first set of interviews, interviewees were asked about the differences the programme was making more generally, and again, while some said it was too early to say, several interviewees described differences they had noticed even after just a few weeks of teaching. The most frequently mentioned difference noticed was that the programme had provided the required structure or scaffolding for pupils to enhance their oracy; eight teachers felt that this had been a particularly notable benefit of being involved in the programme. The benefits of having a structure or scaffolding for oracy was also mentioned by two teachers in the second round of interviews (with oracy leads only), and by three interviewees in the final set of interviews. Towards the end of the year, seven interviewees from five schools said that they had seen a general improvement in pupils' oracy skills as well as the more specific changes discussed below.

The online survey of teachers conducted in March and again in July 2017 also asked about perceptions of change across three areas. Teachers were presented with three statements which referred to potential changes that might be observed as a result of taking part in the pilot. There were high levels of agreement with all three statements and in each case the proportion of positive responses increased between the March and July surveys (Figure 5). Note that the bars on the chart represent the proportion of responses, but the labels on each bar show the actual number of respondents giving each answer. There was a particularly notable increase in the proportion of positive responses to the statement 'I have noticed improvements in pupils' oracy skills since beginning the programme'; all of the 20 respondents to the July survey agreed or strongly agreed with this statement. While half of respondents to the March survey agreed that they had 'noticed improvements in pupils' general academic abilities that might relate to the development of their oracy skills since beginning the programme', in the July survey, around two-thirds of respondents agreed, or strongly agreed, with this statement suggesting that the wider impact of the programme might be becoming increasingly evident towards the end of the pilot year. Note that this is slightly contradictory to the findings discussed under the heading 'academic performance' above whereby interviewees were generally less confident in the benefits of the programme to academic attainment. This might be because the people who completed the survey were not always the same as those interviewed. It could also be a by-product of the different methods of data collection (survey as opposed to face-to-face interviews) whereby at the interview it was clear that the interviewer would prompt for evidence of exactly where these benefits had been noted, and how they

were being measured. Whereas in response to a survey, coming as it did among a list of three statements about the programme, respondents might have been keen to give a positive view of the programme without the need to refer to direct evidence for any improvements in attainment and were therefore quite likely to agree or strongly agree with this statement.

Figure 5: Number of respondents agreeing or disagreeing with each statement on any changes noticed as a result of the pilot (March and July)



All but one of the 'disagree' responses from the March survey came from teachers at School K (one person answered 'disagree' to all three statements, while the other gave a 'disagree' response to the statement about making changes to their teaching practice outside of oracy lessons); the fifth 'disagree' response in the March survey was from a different school and was in response to the statement on improvements to general academic abilities. The five 'strongly agree' responses were from four respondents all representing different schools. In the July survey, the two 'disagree' responses were from one respondent, again representing School K, while the 'strongly agree' responses were fairly well spread across respondents and schools.

Social and emotional changes

During the first interviews, eight interviewees said that over the first few weeks of teaching oracy they had noticed pupils' confidence improve, for example, being more willing to speak in front of others. This theme recurred in the later interviews with five oracy leads stating that pupils' confidence had grown during the mid-point interviews and 18 interviewees described enhanced confidence during the final interviews. There were some descriptions of pupils 'finding their voices' and by the final set of interviews, some teachers recounted instances of pupils who had initially struggled with the demands of oracy taking on board the skills and techniques and approaching the end of the year with newfound confidence and skills. The quotes below illustrate this:

> *'They're developing their confidence. I mean, they've got loads of things to say. They always do, but now it's in more of a structured way and they are more aware of projecting their voice […], more aware of their choice of words they can use, and more aware of the scaffolding they were given, the introductory sentences […] that they could use to build their answers' (from the first interviews with a teacher).*

> *'I would say that first and foremost it's about building their confidence up and that's what I've seen the oracy programme do. Once they are confident, we can then accelerate their abilities. I feel like that's what oracy is, about communication' (from the final interviews with a teacher).*

In the first interviews, three interviewees said that the programme had helped to improve relationships and bonding between pupils and between pupils and teachers; this was felt to be due to the interaction required in the lessons and the removal of 'standard' classroom rules such as working quietly. This was not mentioned again in the mid-year and end-of-year interviews, but this is likely to be because pupils were new to the school shortly before the first interviews, therefore establishing relationships was likely to have been a prevalent concern when these initial interviews were conducted than at subsequent points in the year.

Nine interviewees (from four different schools) mentioned at their end-of-year interview the extent to which pupils seemed to have become more respectful towards each other (at least, in classroom situations), for example, by listening to one another, waiting their turn to speak, displaying empathy, and arguing less.

> *'I think that it has been good in that it gives students the opportunity to think about what their voice is. The Year 7s, I believe, are a little bit more mindful of the impact of their words socially and calling each other out' (from the final interview with an SLT member).*

> *'It's purposeful, ethical and productive. I think in the sense of […] empathy in the classroom as well, I noticed that earlier on and throughout. The quality of feedback that the students were giving one another was always really considerate so they're aware that it's a bit nerve wracking speaking in front of so many people so they're more sensitive with what they say. It's still constructive but they're sensitive about it, it's helped build those skills of empathy' (from the final interview with a teacher).*

Three interviewees (all from different schools) discussed elements of pupil wellbeing that had been enhanced by the programme, for example, by providing a channel for them to express themselves and come to terms with events in the news which might affect them (such as terrorist attacks), to assist them in articulating their needs, and in terms of sharing their opinions and feelings more openly and building more of a 'collective community' through the activities undertaken as part of the programme.

Another theme that was mentioned by two teachers in each set of interviews (not the same people each time) was the impact the programme has had on teachers' confidence. These interviewees pointed out that oracy is a major part of the role of teacher and they found themselves using the skills and techniques themselves as well as teaching them to their pupils. In the end-of-year interviews, four teachers mentioned the way in which the programme had helped them to develop as a teacher by providing oracy strategies and techniques that can be used in any classroom and when dealing with colleagues.

Academic performance

Very few interviewees felt that the oracy programme had had any noticeable impact on academic performance overall, and some stressed that a year was not enough time to realistically start seeing such improvements, however, two interviewees (from different schools) discussed how progress in English had been enhanced by the skills and techniques covered in the oracy sessions. In one instance this had been because it has allowed them to go 'back to basics' with English skills during the oracy sessions, for example through exploring the use of adjectives and emotive language, and looking at punctuation by delivering a speech and punctuating it in the air with the fingers. The other person who had noticed improvements in English said that this was in terms of writing.

One person said that the improvements in listening skills were the key to enhanced academic performance because pupils 'learn to listen and then listen to learn'. This person was optimistic that the skills acquired would stay with them through to GCSE and beyond and help to give them the best possible chance to do well.

Two interviewees made more general comments about the effect the programme has had on pupils who would normally struggle with writing since it allows them to express their opinions and demonstrate their knowledge orally rather than in writing, this is turn was felt to have boosted their confidence. A further two teachers described how pupils were using an expanded range of vocabulary as a result of the coaching the programme had offered (for example, through considering different sentence starters).

Differential benefits across pupil groups

Interviewees were asked whether they felt any specific groups of pupils were benefitting more from the oracy lessons than others. Table 13 below shows the different pupil groups mentioned in this context and how many individuals mentioned each as well as the number of different schools represented in these comments. It should be noted that in many cases interviewees stressed that it is very early days to notice differential benefits for different pupil types (even towards the end of the pilot year) and that where differences had been noted these were purely on an anecdotal basis and not based on formal assessments of skills or ability. Although the numbers of interviewees mentioning each pupil group is low, it is notable how variable the responses were across the year, particularly looking at the difference between the first and third set of interviews. This variation in opinion over the year, together with the fact that in the final interviews (when compared with the first round of interviews) more interviewees felt that ‘everyone benefits differently’, suggests that no specific pupil group benefits from the programme more than others. Rather than generalising the benefits of the programme, it seems, from the evidence gained during the interviews, that based on the experiences of this pilot, there are no specific groups or pupil types that are more likely than others to benefit from the programme.

Table 13: Pupil groups that appear to have benefitted the most from the oracy programme (based on teachers' perceptions and anecdotal evidence discussed in the interviews)

Pupil groups that seem to be particularly benefitting from the oracy lessons	Interview 1 Number of individuals mentioning each group	Interview 2 Number of oracy leads mentioning each group	Interview 3 Number of individuals mentioning each group
Everyone benefits differently	1	2	6
All pupils are taking on board and progressing well - expect it to be 'leveller' rather than benefitting some groups more than others	3	2	5
Lower ability pupils	5	1	4
Higher ability pupils	3	1	3
Pupils who struggle with writing	1		2
More talkative pupils are benefitting (e. g. being more considered in their speaking)	2	1	4
Quieter pupils	5		4
Pupils who haven't been brought up in a culture of talking	4		1
Pupils with additional needs	2		4
EAL pupils	1		3
Boys	2	2	2
Girls		1	1
Less 'well behaved' children	1		2
Pupils from 'difficult backgrounds' / lower socioeconomic groups	5		1

Research question 3: To what extent do we see changes in pupils' oracy on pre and post measures of oracy?

The findings from the analyses of the baseline and follow-up oracy assessments are presented below. The Methods section discusses in more detail the nature of the assessments and the achieved number of assessment results.

Baseline and follow-up assessment results

To determine if there had been an increase from the overall baseline scores to the overall follow-up scores, a one-sided t-test was conducted with the null and alternative hypothesises being:

H_0: Total mean score at baseline = Total mean score at follow-up

H_1: Total mean score at baseline < Total mean score at follow-up

For the t-tests, only pupils that had both a baseline and follow-up score were included. The results of this analysis can be seen in Table 14.

Figure 6 below shows the two raw score histograms plotted over each other for the baseline and post task scores. The distribution of the post intervention scores is clearly seen to be shifted towards the right with a higher mean than the distribution of the baseline scores.

Pre-intervention scores appear to be symmetrically distributed, and medium peaked (mesokurtic). In contrast, post-intervention scores seem left-skewed.

Figure 6: Histograms overlaid to show the distribution of scores for the baseline task and post task

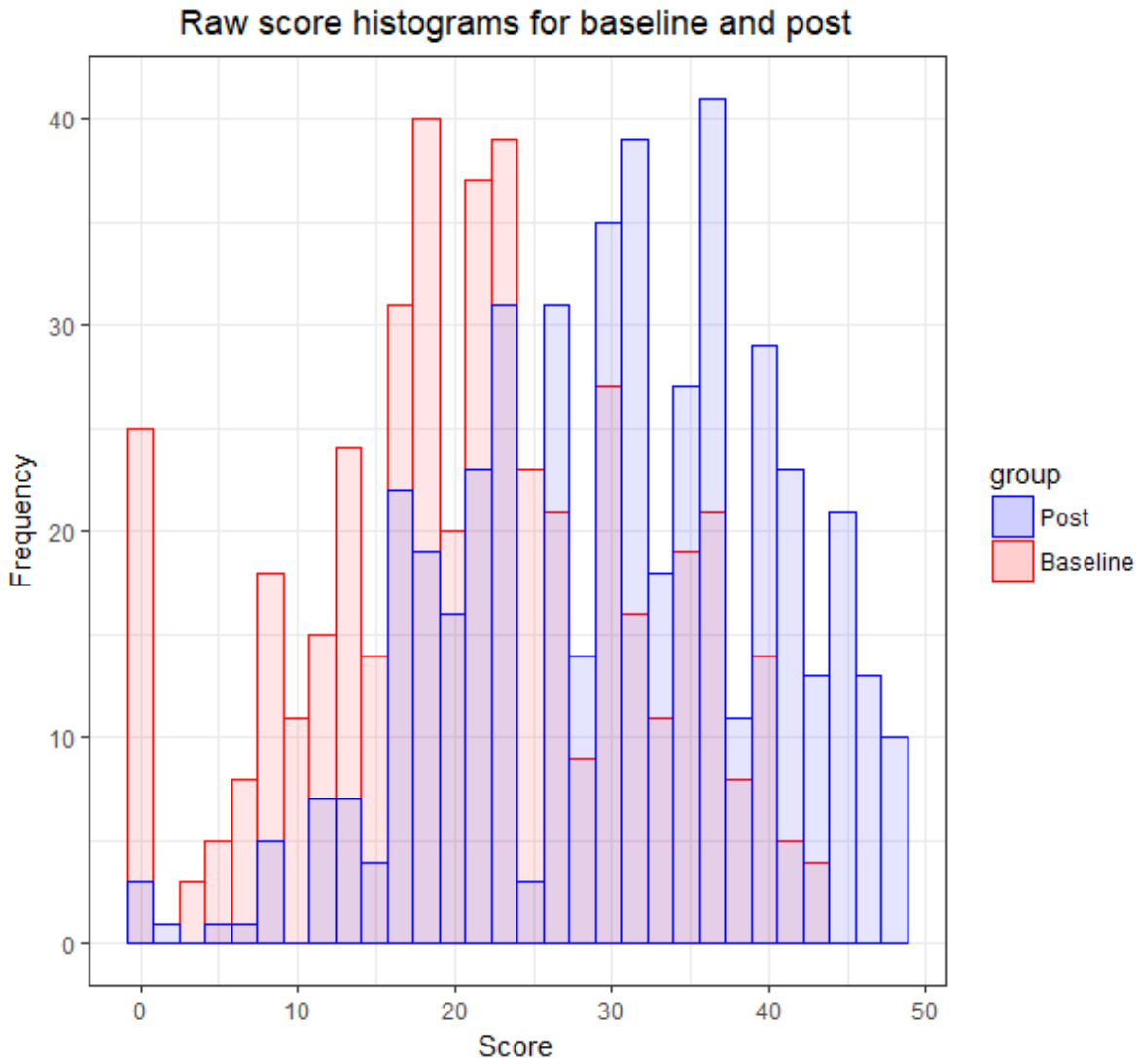


Table 14: T-test results with mean (sd) for pupils with baseline and follow-up scores

School ID	Number of pupils	Mean Before (sd)	Mean After (sd)	Test Statistic	P-value	Significance
C	64	19.92 (6.27)	29.13 (8.35)	-14.91	1.23E-22[9]	***
D	55	35.11 (6.51)	39.84 (5.82)	-4.88	4.83E-06	***
E	60	23.95 (8.80)	33.60 (6.47)	-7.13	8.13E-10	***
F	67	24.46 (9.07)	24.54 (8.41)	-0.09	4.64E-01	
G	54	20.59 (6.77)	34.59 (8.65)	-17.58	4.58E-24	***
H	63	22.67 (8.02)	30.33 (11.76)	-6.27	1.97E-08	***
I	60	18.80 (6.93)	30.47 (6.07)	-15.51	8.82E-23	***
J	50	17.78 (6.92)	21.88 (5.65)	-7.80	1.92E-10	***
K	27	14.33 (5.88)	32.56 (8.33)	-12.62	6.86E-13	***

*** ≤ 0.001, ** ≤ 0.01 , *≤ 0.05.

As indicated in the table, all of the schools did have an increase in the overall mean score obtained. However, there was not enough evidence for School F for the null hypothesis to be rejected. Whereas each of the other schools had a noticeable increase (with the increase ranging from 4.10 to 18.22), School F's overall score only increased by 0.07. School F's interview data was examined in the light of this finding in order to explore whether there might be any clues as to the lower average progress scores in this school. A couple of factors might have played a role: the oracy lead explained that the year group had experienced some staffing problems creating a lack of consistency in the year group teaching team, this might mean that different people undertook the baseline and follow-up assessments and therefore might not have standardised in the same way. Additionally, School F delivered the programme to Year 8 rather than Year 7, and in their final interview, the oracy lead acknowledged that a Year 7 cohort might have resulted in more substantial and noticeable changes in oracy skills and abilities:

> *'Had we had the same staff, same amount of time, same amount of enthusiasm with Year 7, I think this could have been double the impact. We could have had much more significance, because the Year 7 group are more malleable, they come in to secondary school with no prior expectations, they are ready to be moulded, that could have had a much bigger impact ' (School F, oracy lead).*

The fact that all the schools' scores increased may suggest that those who undertook the tasks at baseline had learned from their experience and were able to improve their overall scores as a result. However, if the pupil did not completely understand the tasks at baseline, this could have affected the scores they obtained (resulting in results close to zero). Equally, teachers could have marked strictly at baseline knowing there was going to be a follow-up exercise. Another possibility is 'natural progression'; pupils at the start of the Year 7 are probably quite shy and tongue-tied, but after a few months in school they may feel more confident and thus speak more fluently and interact more effectively. This would be true whether or not they experienced oracy tuition.

[9] Ex represents 10 to the power x (for example, 1E-2 = 1 x 10^{-2} = 0.01).

T-tests were performed on each of the tasks individually using the same hypotheses as before. The results for the talking points and presenting tasks can be seen in Table 15 and Table 16 below.

Table 15: T-test results with mean (sd) for talking points task only and pupils with baseline and follow-up scores

School ID	Number of pupils	Mean before (sd)	Mean after (sd)	Test statistic	P-value	Significance
B	72	7.14 (3.09)	10.26 (5.37)	-4.79	4.50E-06	***
C	64	9.78 (3.67)	14.61 (4.16)	-13.98	2.83E-21	***
D	55	17.51 (3.72)	20.35 (3.10)	-3.86	1.54E-04	***
E	60	12.43 (4.89)	17.20 (3.56)	-5.65	2.48E-07	***
F	67	12.10 (4.79)	12.21 (4.83)	-0.22	4.13E-01	
G	54	11.13 (3.45)	17.50 (5.85)	-9.28	5.36E-13	***
H	63	11.83 (3.99)	15.30 (7.88)	-4.18	4.58E-05	**
I	60	9.83 (4.81)	15.87 (3.50)	-10.15	7.24E-15	***
J	50	8.28 (4.65)	10.92 (3.17)	-7.40	7.88E-10	***
K	27	2.67 (4.53)	15.33 (4.82)	-10.72	2.44E-11	***

*** ≤ 0.001, ** ≤ 0.01 , *≤ 0.05.

Table 16: T-test results with mean (sd) for presenting task only and pupils with baseline and follow-up scores

School ID	Number of pupils	Mean before (sd)	Mean after (sd)	Test statistic	P-value	Significance
C	64	10.14 (3.12)	14.52 (4.70)	-10.06	5.01E-15	***
D	55	17.60 (3.81)	19.49 (4.10)	-3.89	1.38E-04	***
E	60	11.52 (4.60)	16.40 (3.94)	-7.28	4.59E-10	***
F	67	12.36 (4.70)	12.33 (4.17)	0.06	4.76E-01	
G	54	9.46 (4.13)	17.09 (3.79)	-17.49	5.70E-24	***
H	63	10.84 (5.33)	15.03 (6.52)	-4.49	1.60E-05	***
I	60	8.97 (3.51)	14.60 (3.28)	-12.84	5.02E-19	***
J	50	9.50 (3.42)	10.96 (3.21)	-4.10	7.71E-05	***
K	27	11.67 (3.95)	17.22 (5.14)	-5.67	2.92E-06	***

*** ≤ 0.001, ** ≤ 0.01 , *≤ 0.05.

Both above tables indicate the same results gained from when both test scores were combined. Table 15 added the scores from School B since the talking points task results were available for this school. Like the other schools (apart from F), there was an increase in the talking points score from baseline to the follow-up.

The talking points task had a wider range for the increase in the mean scores compared to the presentation task (2.64–12.67 and 1.46–7.63 respectively). In both tasks, School K had one of the largest increases in the mean from baseline to follow-up. The data collected from School K in the interviews and surveys that contributed to the process valuation were re-examined in the light of this finding and there was nothing to suggest an implementation-based reason for the large increase in scores between the baseline and follow-up assessments.

Table 17 summarises the progress made in each school and sets this against some contextual factors such as the representativeness of assessment-takers and the mode of delivery of the programme.

Table 17: Summary of progress outcomes and contextual factors

School ID	School FSM %	Representativeness of pupils assessed against the school cohort*			Assessment outcome**	Number of pupils completing baseline and post-intervention assessment	Delivery method (type of delivery, frequency of oracy lessons, duration of oracy lessons)
		SEN status	Pupil premium	EAL			
A	30.1%	NA	NA	NA	NA	NA	Standalone, weekly, 40 minutes
B	13.6%	2.43 (0.119)	11.1 (4.40e-4)	1.09e-30 (1)	NA***	NA	Standalone, fortnightly, 1 hour
C	12.6%	1.07 (0.302)	0.642 (0.423)	0.220 (0.639)	9.21	64	Standalone, weekly, 1 hour
D	18.2%	5.63 (0.0083)	4.74 (0.0147)	9.67 (9.36e-4)	4.73	55	Standalone, weekly, 1 hour
E	10.3%	0.0819 (0.775)	10.0 (7.77e-4)	0.0871 (0.768)	9.65	60	Standalone, weekly, 45 minutes
F	27.2%	1.50e-30 (1)	1.63 (0.201)	0.0871 (0.768)	0.08	67	Standalone, weekly, 50 minutes
G	19.7%	0.727 (0.394)	0.430 (0.512)	2.25 (0.133)	14	54	Standalone, weekly, 50 minutes
H	24.5%	5.47e-31 (1)	0.0355 (0.851)	2.27 (0.132)	7.66	63	Standalone, weekly, 1 hour
I	7.4%	19.0 (1.29e-5)	0.782 (0.377)	0.0275 (0.868)	11.67	60	Integrated, weekly, 55 minutes
J	32.1%	0.318 (0.573)	3.28e-30 (1)	1.96 (0.161)	4.1	50	Standalone, weekly, 1 hour
K	5.2%	14.2 (8.24e-5)	5.05 (0.0123)	3.02e-30 (1)	18.23	27	Integrated, weekly, 1 hour

* Test statistic and associated p-value in brackets for the two sample proportion test. Those significant tests at the 5% level are shown in bold text. The null hypothesis of the test is that the proportions of trial participants that had a characteristic were the same as the proportion of non-trial participants that had the characteristic given that they are from the same school.
** This is the difference between the mean scores on the baseline assessments and the follow-up assessments (based on both assessment tasks combined). All except for School F were significant at the ≤ 0.001 level.
*** School B only returned assessment results for one of the two assessment tasks.

Feasibility

To what extent are schools able to deliver the School 21 curriculum, assessment and training 'package'?

Delivery models

A recurring theme in both the interviews and online surveys was minor differences in the programme delivery methods reported across the pilot schools. In addition, some schools found it difficult to articulate exactly how they were delivering the programme, and whether this was consistent with the intended delivery model. However, where there were deviations from the intended delivery, these were generally minor differences within the programme model and several staff members commented during their interview that the flexibility that the programme offers in terms of delivery mode was one of its strengths.

The online survey of oracy leads offered three descriptions of different ways in which the oracy programme might be delivered at the schools: nine oracy leads said that oracy was being delivered as a 'standalone' programme with classes focusing entirely on oracy, although content might be brought in from other topics; the remaining two oracy leads described their delivery model as 'integrated' whereby oracy teaching is contextualised into a subject (in both cases this was English) and the learning outcomes can relate to either oracy or the particular subject. It is not entirely clear, in the case of these two schools, whether their delivery is truly 'integrated', but both oracy leads defined it in this way. One of the oracy leads explained at their end-of-year interview, 'We're implementing it through the English department and the English department runs bespoke oracy lessons.' However, this school had particular concerns about the lack of resources and content for oracy lessons and had therefore used English content which might have been what led them to feel that the 'integrated' model was the closest match to theirs. It was a similar situation at the second school that defined their delivery as 'integrated', although in this instance, the oracy lesson formed an additional timetabled session (where English previously had four sessions per week for Year 7, for the pilot year they were given five lessons with one being delivered as a dedicated oracy lesson), although again, the content was drawn from English. It is most likely that the learning outcomes relating to either oracy or the subject (English) that identified these two schools' delivery models as 'integrated' rather than 'standalone'. Table 18 shows the distribution of responses to this question.

Table 18: Oracy leads' descriptions of the delivery model for the oracy programme (based on survey responses from oracy leads)

Delivery model	n
Standalone: Classes focus completely on oracy. Content may be brought in from other topics or via project-based methods, but the primary learning outcomes relate to oracy.	9
Integrated: Oracy teaching is contextualised into a subject; learning outcomes can relate to either oracy or the subject.	2
Embedded: Oracy teaching is completely embedded into another subject. Oracy learning outcomes are secondary to those of the subject it is embedded within.	0
Total	**11**

One of the core elements of the programme is the delivery of one dedicated oracy lesson a week to Year 7 pupils. When asked how frequently the oracy lessons were delivered, all but one of the schools said that they had weekly timetabled oracy lessons; the remaining school was delivering the programme via fortnightly sessions. In the June interview with the SLT member at this latter school, it was explained that the reason for fortnightly sessions was because at the point at which they discovered they were going to be a pilot school, the timetable for the current year had already been agreed and it was only possible to fit in the oracy lessons once a fortnight. From following year, the programme would be delivered at this school via weekly hour-long sessions. In five of the 11 participating schools, the timetable structure meant that an hour-long lesson was not possible because teaching periods are timed at less than an hour (ranging from 40 to 55 minutes). Table 19 below shows the length of the weekly (or in one case, fortnightly) timetabled oracy sessions delivered for all 11 schools based on the responses to the online survey for oracy leads.

Table 19: Duration of timetabled oracy lessons (based on oracy leads' survey responses)

Length of oracy lessons	n
40 minutes	1
45 minutes	1
50 minutes	2
55 minutes	1
1 hour	6
Total	11

Enablers and barriers to delivery

At each set of interviews, interviewees were asked to describe any barriers or enablers to the successful delivery of the programme. These are reported below.

Barriers and challenges

The opinions of SLT members, oracy leads, and delivering teachers on the barriers to delivering the programme shifted somewhat over the one-year pilot period. At the initial interviews during November, interviewees described a range of specific challenges or barriers that they had encountered during the first few weeks of delivering the programme, and many of these related to the practicalities of delivering it and the 'newness' or unfamiliarity of the content and pedagogy. Table 20 shows the barriers and challenges mentioned during these initial interviews and how frequently they were mentioned. At each school, more than one person was interviewed, hence this table shows the number of individuals giving each response and the number of different schools represented by these individuals. The category 'other' includes challenges mentioned by only one or two interviewees, such as the number of other initiatives or programmes running at the school at the same time and the extent to which newly qualified teachers could, or should, be part of the delivery team.

Table 20: Summary of challenges and barriers to delivery of the programme as at the first round of interviews in November 2016

Challenges faced—based on initial interviews with oracy leads, teachers, and SLT members in November 2016	Number of individuals mentioning each challenge	Number of different schools giving each challenge
Timetabling issues	14	8
Practical / logistical issues, e.g. suitability of classrooms	12	8
Maintaining momentum / achieving high profile / changing the school culture permanently	9	6
The 'newness' of it all, having to become familiar with it as they teach it	9	6
Initial set-up work, e.g. creating schemes of work, training colleagues	8	6
Lack of information or late information provision	8	4
Pupils' behaviour or attitude (initially)	7	6
Pupils' reluctance or lack of confidence	6	5
Some staff members have been more resistant than others	5	3
Lack of structure or clear overarching goals or aims	3	2
Lack of time set aside to work on developing and implementing the programme	3	2
Lack of training / not all staff able to go to School 21	3	3
Other	10	6

The two most frequently mentioned challenges in these initial interviews both relate to practical concerns: timetabling issues were mentioned by 14 interviewees representing eight different schools, while general practical and logistical issues were mentioned by 12 interviewees (again representing eight schools). In terms of timetabling, most who commented recounted difficulties in fitting in the required one-hour oracy lesson per week. In most cases, interviewees pointed out that this meant something else had to 'give way' for the oracy lesson and it was often those who were attempting to integrate oracy with another timetabled session (such as guidance or tutor time) that experienced greater issues than those who were delivering oracy as one of the weekly lessons for another subject, such as English.

Twelve interviewees from eight different schools mentioned other practical and logistical issues; these tended to refer to the suitability of classrooms for the types of activities required during oracy sessions. In instances where the hall or studios were available this seemed to work better than 'standard' classrooms, although in some cases interviewees said that this was only an initial teething problem, and that once both pupils and the teacher had got used to adapting the classroom at the beginning of each session (for example, by moving desks to the side of the room), these practicalities were no longer an issue. These types of adaptations seemed to generally meet needs; there were no indications that the activities themselves were having to be adapted to meet the needs of the students or the environment. Assemblies were the greatest cause of comment on practical and logistical issues. Most who mentioned this pointed out that their pupil numbers were much higher than School 21 and that the practicalities of administering the assemblies in the format suggested were, in some cases, seemingly insurmountable.

Nine interviewees from six different schools described the challenges around maintaining momentum and achieving a permanent culture change in the school. Some who mentioned this described how the

initial responses to the pilot have been positive but that they were concerned that such enthusiasm will only last until the next new initiative comes along. Others were concerned that since oracy was currently only delivered to Year 7, it would take a long time for colleagues to see the all-round benefits of the programme, and hence, interest and enthusiasm might diminish if the returns are not immediately visible. Several of those who discussed this detailed how they were attempting to overcome this by offering regular training and updates and using posters and assemblies as a way of involving the whole school.

Nine respondents described how challenging it had been to familiarise themselves with the oracy programme and to be confident in delivering it to pupils. The new terminology was a specific issue for some who felt it took them a while to become familiar with it, particularly for those whose subject specialism was less closely related to oracy.

Another relatively prevalent issue (mentioned by eight interviewees from six different schools) which is also related to the newness of the programme was that some interviewees felt they had a significant amount of 'set-up' work to do to establish the programme, for example, training colleagues, and in some cases oracy leads were taking responsibility for developing week-by-week lesson plans for colleagues.

A further eight interviewees said that the lack of information provision or late provision of information had been a particular challenge. Some felt that the implementation of the programme had felt rushed given that the initial training was provided in the summer term with teaching due to start at the beginning of the autumn term; in addition, some interviewees noted that the information on the assessments had arrived later than anticipated.

The pupils themselves were mentioned among the challenges faced on three fronts: seven interviewees from six different schools said that some pupils struggled initially with the necessary changes to their behaviours and attitude in the oracy classes (for example, adapting to no desks); six interviewees (representing four different schools) mentioned issues with a minority of pupils being initially reluctant to join in with activities (particularly the quieter pupils); and one person said that it was a major challenge ltrying to overcome the ingrained 'bad habits' of pupils in their speech, for example, those associated with regional dialects.

Three interviewees (all from the same school) commented that there was a perceived lack of clarity as to what the overall aims of the programme were, or its 'end product'. These teachers were keen to ensure that everyone involved in delivering the programme at the school was working towards the same goals and has the bigger picture as to what they are trying to achieve.

There were also three complaints about a lack of dedicated time to look at the resources and work on plans for delivering the programme—perhaps something that needs to be addressed at local management level rather than an issue for the programme, but still a concern for the overall feasibility and success of the programme.

Three interviewees mentioned the challenges they had faced due to a perceived lack of training on the programme; this was particularly the case for staff who had not been able to attend the training at School 21.

The telephone interviews with oracy leads conducted at the mid-point of the pilot year also gathered information on any challenges or barriers oracy leads had encountered. While some of the challenges discussed were the same as those mentioned during the first interview (such as practical or logistical issues, problems caused by there being other initiatives and programmes running at the schools at the same time, and problems where some staff members were more resistant to embracing oracy than others), there were also some new challenges identified. These are summarised in Table 21 below.

Table 21: Summary of challenges or barriers to delivery of the programme as at the second round of interviews with oracy leads in February 2016

New challenges faced at the mid-point interviews with oracy leads	Number of oracy leads mentioning each challenge
'Ignite' speeches	3
Keeping staff on board and motivated	2
Other staff struggling with the curriculum or style of teaching	2
Staff changes during the academic year	2
Lack of reassurance that what they are doing is 'right'	1
Staff not seeing themselves as oracy deliverers	1
The school's focus on external exam results (e.g. GCSEs)	1
Wider introduction of oracy (e.g. into other subjects)	1

The most frequently mentioned new challenge in the second round of interviews was the Ignite speeches. Two of the three oracy leads who referred to the Ignite speeches as a challenge or a barrier did so on the basis of the scale of organising the speeches, particularly the practical issues of enabling pupils to give their speeches and ensure that they have an appropriate venue and audience for them. The third oracy lead expressed concerns about the resistance from colleagues in terms of supporting the preparation and execution of the speeches.

Many of the other new challenges identified related to other staff members, for example, two oracy leads said that a key challenge for them had been keeping their colleagues on board and motivated to deliver the oracy programme, and a further two commented that some colleagues were struggling to deliver the curriculum and adapt to the style of teaching required in the programme.

Table 22 below summarises the key challenges mentioned during this final set of interviews and it is notable that the number of individuals mentioning a challenge has reduced markedly since the first interviews. During the first set of interviews, 39 of the 41 people interviewed mentioned at least one challenge or barrier; during the final interviews, just 12 of the 36 interviewees mentioned at least one challenge.

Table 22: Summary of challenges or barriers to delivery of the programme as at the third round of interviews in June 2017

Challenges faced—based on final interviews with oracy leads, teachers, and SLT members in June 2017	Number of individuals mentioning each challenge	Number of different schools giving each challenge
Some staff members more resistant than others	3	3
Staff changes during the academic year	3	3
Maintaining momentum / achieving a high profile / changing the school culture permanently	2	2
School focus on external exam results (e.g. GCSE)	2	2
Timetabling issues	2	2
Practical or logistical issues, e.g. suitability of classrooms	2	2
Issues with subject matter or content rather than techniques and skills	1	1
Keeping staff on board and motivated	1	1
Lack of information or late information provision	1	1
Lack of structure or clear overarching goals or aims	1	1
Lots of other initiatives and programmes running at the same time	1	1
The 'newness' of it all, having to become familiar with it as they teach it	1	1
Other staff struggling with the curriculum or style of teaching	1	1

The perceived challenges or barriers discussed by interviewees at the third and final set of interviews during June 2017 represented a further shift in terms of the nature of the challenges. Rather than relating to primarily practical and logistical matters and issues linked to the unfamiliarity of the programme, as had been the case in the first two sets of interviews, challenges discussed during the final interviews tended to focus more on the issues around maintaining momentum, getting other staff members on board and motivated, and striving to achieve or lay the foundations for an oracy culture at the school. This suggests that any challenges or barriers experienced by the end of the pilot year tended to relate to matters external to the programme design and implementation; they were more issues for the individual schools to address and resolve internally rather than issues for the programme itself.

While deliverers' perceptions of the challenges or barriers encountered in the delivery of the programme represent something of a journey across the pilot year—from the initial struggles with the practicalities and the unfamiliarity of the programme through to the wider concerns about maintaining momentum and working towards an oracy culture in the school—perceptions of the enablers to the successful delivery of the programme were far more consistent.

Enablers

The quality of training and support from Voice 21 (both the scheduled training events and the ad hoc support received throughout the year), the support and 'buy-in' from colleagues, and the attitudes and willingness of pupils were all mentioned at all three interviews as enablers to the successful delivery of the programme. Additionally, in the final set of interviews, two interviewees (each from different schools) mentioned the flexibility of the programme as a major enabler to its successful delivery, for example, the freedom to adapt or design the resources, content, and delivery to suit the context of the school. Table 23 shows the enablers mentioned during each set of interviews and the number of interviewees who mentioned each.

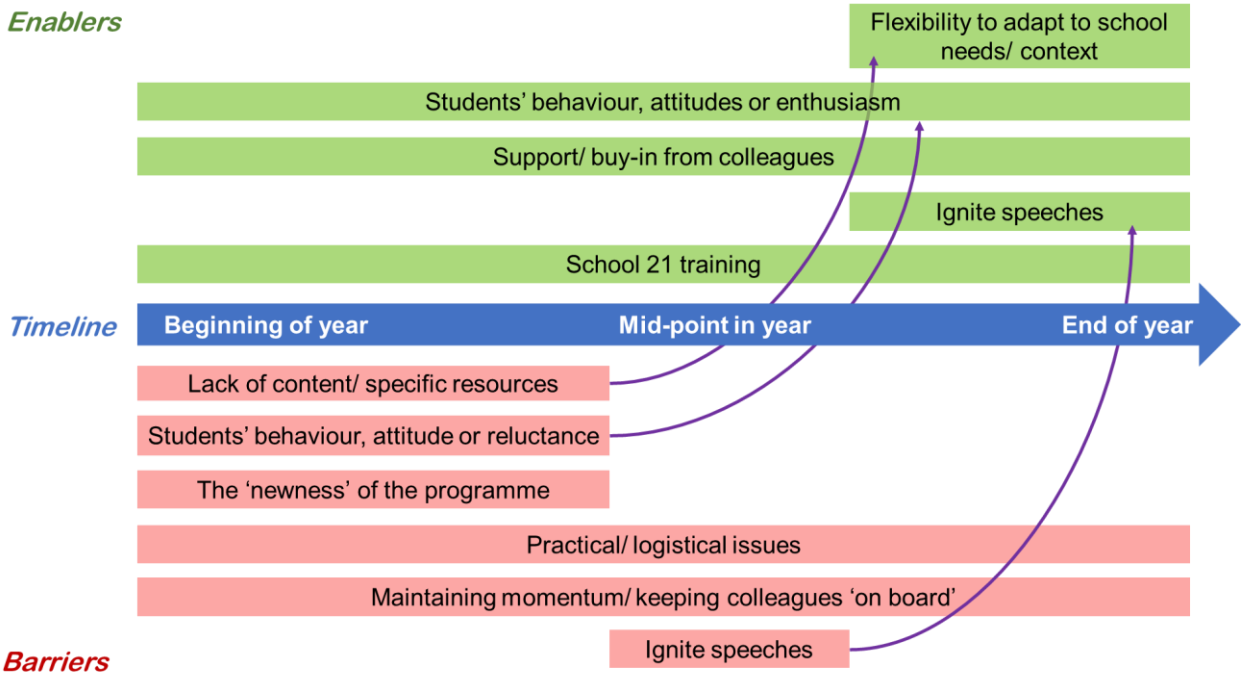
Table 23: Summary of the enablers mentioned at the three sets of interviews and the number mentioning each

Enablers mentioned across all three interview occasions	Interview 1 (n)	Interview 2 (n)	Interview 3 (n)
Students' behaviour, attitudes, or enthusiasm for the programme	3	2	3
Support and 'buy-in' from other colleagues	2	2	1
Headteacher or SLT are supportive / school culture is conducive to oracy	0	2	2
Lesson plans provided by the oracy lead	2	1	1
Quality of information, resources, and training provided by Voice 21	1	2	1
The relative freedom / flexibility in terms of content and context	0	1	2
'Ignite' speeches	0	0	2
Launch activities and training have gone well and encouraged 'buy-in' from colleagues	1	1	0
Quality of ad hoc support from Voice 21	1	0	1
Voice 21 willing to listen to feedback and make changes	1	0	0

Changing perceptions of barriers and enablers across the pilot year

The descriptions of the barriers or challenges and the enablers to the delivery of the programme shifted somewhat during the pilot year. Figure 7 below presents a visual summary of the most frequently mentioned barriers and enablers and shows at which points in the year they were most likely to be mentioned. In some cases, the barriers or enablers were consistently present throughout the year (for example, the 'practical and logistical issues' barrier and the 'support/buy-in from colleagues' enabler), while in other cases they were more transient (for example, some of the barriers that could be attributed to 'teething problems' were only prevalent issues during the first set of interviews). There were three areas which were described as barriers in the first and/or second set of interviews but which later transformed into enablers (denoted by the purple arrows in the diagram). The initial complaints about an apparent lack of content and specific resources could be said to have transformed into an enabler in the form of the flexibility to adapt the content—mentioned by some interviews at the end of the year. Similarly, the Ignite speeches were a cause for concern for some during the mid-year interviews with oracy leads, but this seems to have been caused by the unfamiliarity of the format and process for these speeches; at the end of the year these were seen as a positive element of the programme and had helped to promote oracy among staff, other pupils, and parents as well as providing a platform for pupils to demonstrate their skills. The behaviour and attitudes of pupils was an enabler that was mentioned at all three sets of interviews, but in the first set of interviews some interviewees were concerned that the more reluctant pupils might cause barriers to the successful delivery of the programme. In the event, these fears were largely unfounded and the topic did not re-emerge at either the mid-point or end-of-year interviews.

Figure 7: Visualisation of the occurrence of the main perceived barriers and enablers against the pilot year timeline

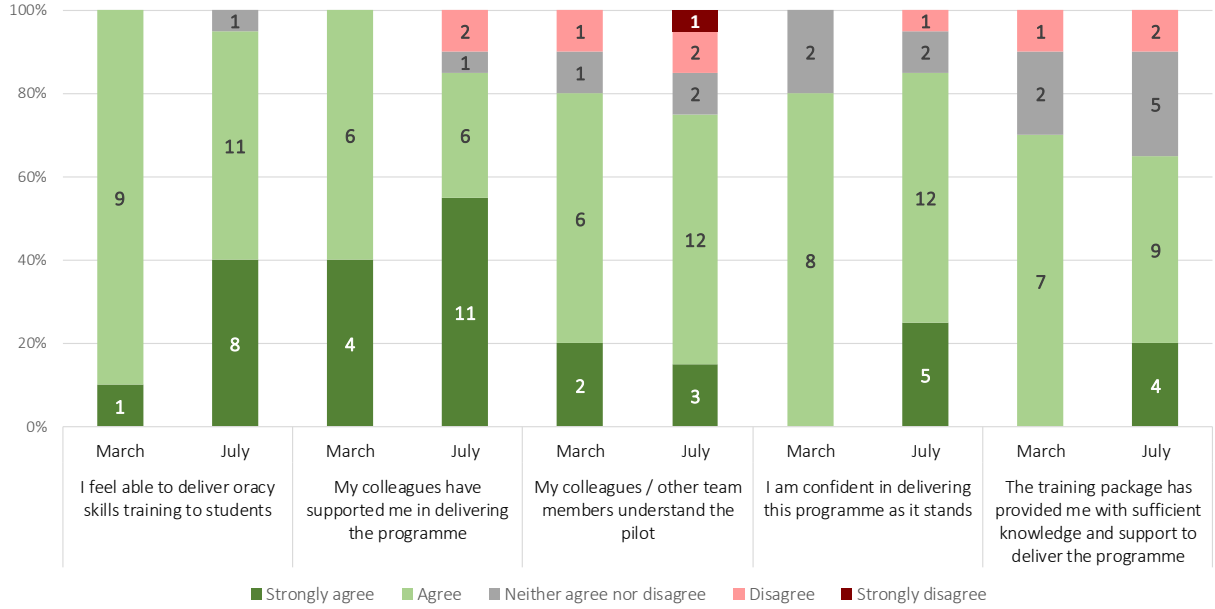


Overall perceptions of the programme

Interviewees were, overall, very positive about the programme and most (30 of the 36 people interviewed in the end-of-year interviews) expressed the view that their school's participation in the pilot had been worthwhile. In the first round of interviews, there were slightly more negative views than in the final interviews; for example, three interviewees indicated that the exact means by which oracy is incorporated into the curriculum needs careful consideration as in their view it is better suited to some subjects than others. However, by the end of the pilot year, this was not mentioned as a concern, indeed, as discussed above, some interviewees expressed surprise at how well other departments or subjects were incorporating oracy skills development, and in some instances these were subjects where they would have expected it to be more challenging, such as in maths. Sixteen interviewees made general positive statements about programme and a further 11 specifically stated that the programme is feasible and could be run in any school with minimal adjustments, however, during the final interviews one SLT member cautioned that it will take a long time before substantial change is achieved. Seventeen interviewees (from nine different schools) indicated that oracy should have a place in the secondary curriculum and 18 people (from eight schools) said that oracy should be taught at primary phase, with 11 people (from six schools) indicating that oracy should be incorporated into initial teacher training.

The online survey of teachers administered in March and again in July 2017 also gauged overall perceptions of the programme. Figure 8 shows teachers' responses to the statements about delivering the programme. Note that the bars on the chart represent the proportion of responses, but the labels on each bar show the actual number of respondents giving each answer. There were generally high levels of agreement with all statements. Across both surveys, all but one respondent indicated that they 'feel able to deliver oracy skills training for pupils' and the number of respondents who strongly agreed with this statement increased from one to eight between the two surveys. The two statements that elicited the fewest 'agree' or 'strongly agree' responses were 'the training package has provided me with sufficient knowledge and support to deliver the programme' and 'my colleagues/other team members understand the pilot', which suggests that where there are perceived weaknesses in the programme, these might relate to the training and preparation teachers received.

Figure 8: Number of respondents (teachers) agreeing or disagreeing with each statement on the oracy programme (March and July)



There were just two 'disagree' responses given in response to these statements in the March survey and these were both from the same respondent at School K. In the July survey, the 'disagree' and 'strongly disagree' responses were made by four respondents out of a total of 20, all from different schools. In two instances, each respondent gave only one negative response each, but three of the 'disagree' responses were made by one respondent (from School J), and another respondent (from School K) gave two 'disagree' and one 'strongly disagree' responses, suggesting that these two individuals had felt generally unsupported and ill-informed about the programme. The 'strongly agree' responses were fairly evenly spread across respondents and schools.

Conditions or prerequisites for the successful running of the programme

Most interviewees agreed that the programme would work in almost any school and during the final set of interviews they were asked whether they felt there were any specific conditions or prerequisites to the successful running of the programme. While many felt that there were none, Table 24 presents those that were discussed and indicates that SLT support and buy-in was the most frequently mentioned factor in the successful implementation (mentioned by eight interviewees at five different schools), followed by adequate initial training before starting to deliver the programme (mentioned by six interviewees at four schools, some of whom stated that it was important that this is undertaken at School 21). Four interviewees suggested that buy-in and support from other colleagues (that is, not SLT members) was an important condition and a further four indicated that a cohesive team to deliver the programme was important, for example, colleagues from one department who already have a close working relationship, rather than building a team based on more ad hoc factors such as who has enough spare capacity to deliver.

Table 24: Summary of conditions or prerequisites to the successful delivery of the programme as at the third round of interviews in June 2017

Conditions or prerequisites to successful programme delivery	Number of individuals mentioning each challenge	Number of different schools giving each challenge
Buy-in and support from SLT	8	5
Upfront training (preferably at School 21)	6	4
Buy-in and support from other colleagues	4	3
A cohesive team to deliver (e.g. all from one department)	4	4
Adequate time to prepare before delivering the programme	4	3
Enthusiastic and motivated oracy lead	3	3
The 'right' motivation for implementing it, e.g. whole-pupil ethos, or oracy being central to the way the school works	2	1
Time for a dedicated lesson in the timetable	2	2
Appropriate teaching rooms/spaces	1	1
Understanding of how oracy links to your own subject's curriculum	1	1
Gradual introduction and accept that there are few 'quick wins'	1	1
Ability and motivation to deliver assemblies	1	1
Small school or year group sizes	1	1

Consistency of delivery and fidelity to the intended programme

While (aside from assemblies) most oracy leads felt that they had made no, or only minor, adjustments to the School 21 model, a small number expressed concern over consistency of delivery, both in terms of fidelity to the intended programme and consistent delivery within their own school across the different deliverers. Four oracy leads said that they doubted all teachers at their school were delivering the programme in the same way, and this was said to be due to a number of factors, most usually the levels of buy-in and commitment they have to the programme and the amount of training they have received. In terms of fidelity to the intended programme, there appears to be a tension between the freedom to adapt the content to the context and the need to deliver a core curriculum.

Training and support

A key element of the programme as offered was the training and support from Voice 21. The training was intended to support the implementation and delivery of the programme and to ensure that, as far as possible, the programme was delivered in the intended manner. All oracy leads and the SLT member with responsibility for oracy were invited to attend training at School 21 before commencing delivery of the programme. It was expected that the oracy lead would then cascade training to colleagues involved in the delivery of the programme. Voice 21 also offered on-demand support via telephone and email. Some schools also invited a representative of Voice 21 to deliver a training session or presentation at their school, and there was also an online bank of resources and support materials provided by Voice 21 to help support programme delivery.

During all three waves of interviews, almost everyone interviewed was positive about the training and support they had received up to that point. There was a general perception that information and resources were widely available and where existing resources were unable to answer a query, the individual support from Voice 21 was usually very helpful. Those who attended it directed specific praise at the training held at School 21 in the summer and felt that seeing what is done at the school was extremely informative and motivating. In instances where individuals had been unable to attend the training at School 21, they usually felt that they had 'missed out' and suggested that this is a key factor in successfully implementing and delivering the programme.

In the first set of interviews, there were some criticisms and suggestions on the topic of training and support, the most prevalent of which was the suggestion that Voice 21 should have provided more week-by-week teaching guidance or resources such as lesson plans with clear objectives and suggested lesson content. Nine interviewees in six different schools talked about the difficulties of planning lessons and the perceived need for detailed schemes of work or lesson plans. However, by the time the final set of interviews was conducted, just four interviewees felt that more detailed information or better resources were needed. While detailed lesson plans or schemes of work would have ensured greater consistency of delivery both within and across schools, the freedom to adapt the content to suit the school context and curriculum needs was seen as a strength of the programme.

How appropriate is the use of hubs as a means of rolling out the programme?

The regional hubs were not a key factor in the delivery and success of the programme. Across all three sets of interviews, deliverers were asked whether they had made contact with other schools within their regional hub and, aside from meeting staff from the schools at the initial training at School 21, there had been no significant further contact between schools in the hubs. Most oracy leads and some of the teachers delivering the programme agreed that it would be a 'nice to have' element of the programme but that time constraints had prevented them from taking steps to liaise with the other schools in their hub.

In the first set of interviews, six interviewees from six different schools indicated that they would have liked some form of support or contact with other schools in their regional hub, and by the final set of interviews this had increased to nine such requests (representing five schools). While some acknowledged that in-person meetings with other schools in the hub were potentially the most helpful form of contact, these might be difficult to arrange. This led some to suggest communications between schools could have been facilitated through the use of an online forum.[10] It was felt that this would represent less of a demand on the time of teachers while still conferring many of the benefits of face-to-face meetings. However, even this online forum method of communicating was demanded to a greater extent in the early interviews when schools were struggling a little more with the initial implementation and seeming to need some reassurance that what they are doing is right, or simply another party to share ideas, resources and experiences with. Towards the end of the pilot year, appetite for such an online forum was slightly less evident in the interviews, perhaps due to school having 'found their own way' without the need for interaction with other schools.

Readiness for trial

Is there a School 21 curriculum, assessment, and training 'package' that could be rolled out to schools (with minimal modifications)?

Is there a defined programme that could be rolled out in schools?

The views of interviewees on the extent to which the Voice 21 Oracy Programme is 'school ready' seemed to shift slightly across the pilot year. During the initial interviews in November 2016, there was evidence of many oracy leads and teachers feeling overwhelmed by the task of delivering the programme, which for many was quite different to anything they had delivered before. There were several requests for some form of reassurance that what they were doing was right, and some were still grappling with the practicalities and implications of 'borrowing' a lesson a week from another subject. There was also a slight split in opinion in the first round of interviews as to whether oracy is best delivered as a distinct skill, embedded in other subjects, or a combination of both approaches. Perhaps

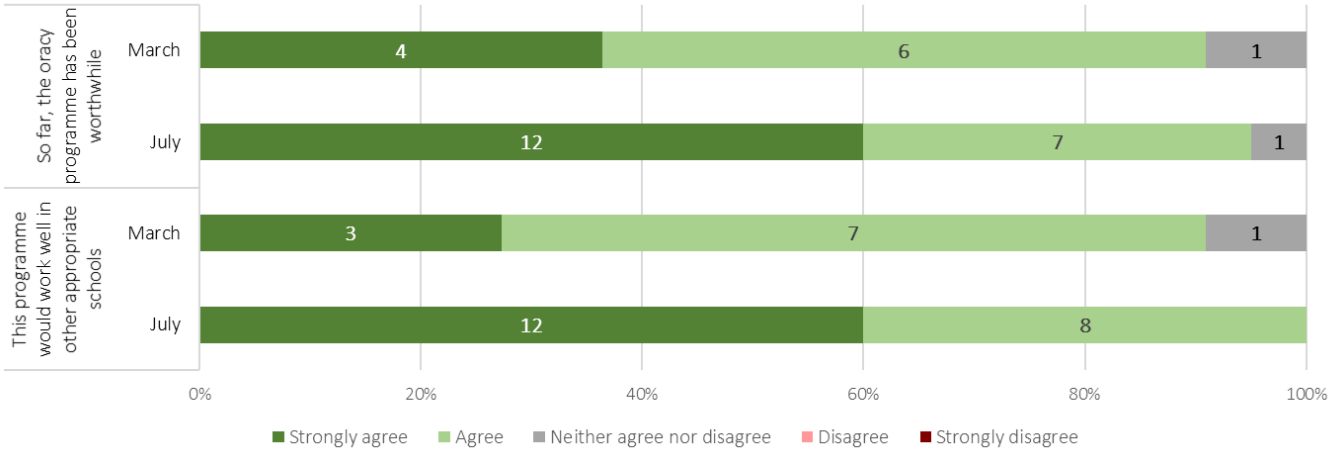
[10] Although this was the feedback from schools, we also note that Voice 21 did provide an online forum but that it wasn't widely used.

the greatest cause for concern expressed by interviewees during the first interviews was the apparent lack of content or prescribed subject-matter for the oracy sessions. However, it was particularly notable that these complaints reduced over the year and were non-existent in the final set of interviews, in fact, rather than complain about a lack of content, several interviewees felt that the way in which the content of teaching sessions can be adapted to meet the school context, the needs of the subject it is being delivered within, or to reflect current affairs and issues of importance to the pupils was one of the key strengths of the programme.

By the time the final interviews were conducted, many of the issues associated with the introduction and initial delivery of the programme had been resolved and almost everyone interviewed expressed the view that overall the programme is feasible and can be delivered in almost any context with minimal adjustments. The one area which did give rise to concerns in terms of feasibility throughout the pilot year was the delivery of oracy assemblies. Throughout all three sets of interviews, assemblies were repeatedly described as problematic in terms of the logistics of delivering these in the spaces available and with the pupil numbers involved (larger numbers than School 21). This may have implications for the types of school in which the programme as it currently stands can feasibly be delivered. There were also a small number of concerns expressed over the apparent 'untouchable' nature of assemblies which meant that there was an unwillingness on the part of leaders at the schools concerned to allow assemblies to be used for anything other than their existing purposes (this affected two schools in the pilot).

The online survey of teachers included two statements about teachers' overall perceptions of the programme and its feasibility. As Figure 9 shows, the vast majority of responses to both statements in both surveys were positive, and across the two surveys the proportion of 'strongly agree' responses increased and these represented a range of different schools. In the March survey, just one respondent (the same person in both cases) answered 'neither agree nor disagree' to both statements, and there was just one such response in the July survey to the statement 'so far, the oracy programme has been worthwhile'.

Figure 9: Number of respondents agreeing or disagreeing with each statement on their overall perceptions of the Programme



Costs of running the programme

At the end-of-year interviews, the interviewees were asked to indicate what they felt the cost of running the programme had been at their school, not just in pure financial terms but in terms of other resources such as time, materials, or equipment. In most cases, the main cost was the time taken to prepare for delivery of the programme, and this usually placed the greatest demand on the oracy lead. In the first set of interviews, nine interviewees in six different schools talked about the difficulties of planning lessons and the perceived need for detailed schemes of work or lesson plans, and in some instances

it was apparent that the oracy lead had taken on the role of lesson planning, and in at least three cases they were clearly struggling to keep on top of the demands of this role. This was felt to have a non-monetary cost since it was diverting them from other roles and responsibilities. While this issue was less prevalent in the final interviews, it was still considered to represent a 'cost' of delivering the programme and oracy leads were keen to be allocated sufficient non-teaching time to be able to fulfil the lesson planning requirements. Training and CPD were also classed as a cost. As well as the initial training at School 21, in all schools there was some degree of cascading the skills and techniques to colleagues and interviewees from five different schools described the CPD element as a cost of running the programme, again in terms of time and effort above financial costs (although there were some financial costs in terms of associated resources such as printed hand-outs). In some instances, the costs of running the programme had been absorbed by the 'lead' department, for example, one school was delivering the programme through their English department and said that the small financial costs (such as printing) came from the English budget. One person commented that the programme assumes access to technology such as tablets or recording equipment, and while most schools have these, this person found that their equipment was somewhat outdated and struggled with the demands.

No one indicated that the programme had particularly high or unacceptable 'costs' (financial or otherwise) associated with it. However, two interviewees (at different schools) acknowledged that from the next academic year there would be greater financial costs if they wished to continue with the training offered by Voice 21 due to the new arrangement whereby training is run as a collaboration between Voice 21 and the University of Cambridge and is charged for.

Is the School 21 oracy measurement a valid and reliable tool for use in future trials?

Assessing a productive skill such as oracy can be difficult to do objectively and robustly. Expert judges can take differing views of the same performance, and assessment results can show low levels of reliability and consistency. If such is the case in the current situation, it does not necessarily mean that the oracy approach is not a valid educational intervention. It may well remain so, even if it cannot be assessed very accurately. Further, one can envisage a perfectly valid formative assessment tool, which provides credible, useful information for teachers, while not necessarily exhibiting high levels of reliability using a conventional index such as Cronbach's alpha. It is just that in such a situation, it would be challenging to show the robust evidence that the EEF's approaches typically require.

A range of analyses were undertaken to attempt to assess aspects of validity and reliability in the oracy assessment tool; these were based on the baseline assessment (pre-intervention) data provided by ten pilot schools in autumn 2016.

Table 25 below provides a summary of the main analyses conducted, the key findings, and areas for further development and consideration emerging from these analyses.

Table 25: Summary of research findings divided between positive findings and points for further work

Researched area	Generally positive findings	Areas for further development or consideration
Curriculum coverage	The two assessed tasks covered the oracy framework to a substantial extent.	• The oracy skill that had the sparsest coverage across the two tasks was 'social and emotional'. Although we did not believe this was a catastrophic deficit, we did take the view that this was the type of 'softer skill' that can often be underrepresented in assessments. • Some of the oracy sub-skills in the tasks were re-worded versions of the same entity in the oracy framework. We suggested that sub-skills in the assessment should use the same wording as the framework unless there was a good reason to change the wording.
Assessor standardisation	Standardisation documents and videos were comprehensive and ought to give assessing teachers a good understanding of how the standard was meant to be applied.	• Some notes instructing standardisation facilitators in how to run sessions would be useful. • Oracy specialists could consider what features of EAL learners' developing English language skills should be considered to have an impact on their oracy attainment. • Oracy specialists could consider how to assess individuals' performance within groups so as to avoid inappropriate ceiling or floor effects.
Trial pupils' representativeness	Most schools entered pupils into the trial who broadly represented their cohorts in the year groups that were assessed.	• Trial participants from School D were significantly less likely to have statuses of SEN, Pupil Premium, or EAL compared to their year group. • Trial participants from School K were significantly more likely to have statuses of SEN or Pupil Premium compared to their year group.
Generalisability theory	The g-theory relative coefficient showing the measurement's quality for distinguishing pupils' scores from each other was reasonable for a teacher-assessed assessment for research purposes (0.741).	• The absolute g-theory coefficient—which shows whether a measurement can generalise across a wide range of contexts such as different curriculum areas, tasks, and so on—was low (0.468). This suggests that this measure could not easily generalise to such different contexts. • Differences between pupils' oracy abilities accounted for only around 27% of the variance in scoring. This was considered to be a low proportion. • Cofounded residual variance accounted for around 34% of the variance. • In so far as the EEF's approaches require robust measurement of traits, the existence of these two proportions may mean that this form of assessment is problematic.
Rasch Facets analysis	Many of the measured Facets were of similar levels (such as tasks, schools, and so on).	• School D appeared to stand apart from the other schools by giving higher scores on the oracy tasks. • Because our measurement design was insufficiently anchored, we were not able to disentangle whether School D had pupils who were genuinely very good at oracy, or whether that school's markers were particularly lenient.
Multiple regression analysis	We were able to fit two multiple regression models, which allowed us to disentangle the confound noted about concerning leniency vs. truly higher ability.	• School D's higher performance was not predicted by background variables in the regression model. We therefore suggested that that school had lenient markers. • School H also appeared lenient according to the regression model, although this school had not stood out in g-theory and Rasch Facets analyses.

Assessment instruments and procedures

Analysis of curriculum coverage

We took the oracy curriculum as described in Figure 1 and mapped it to the assessment tasks, oracy skills, and sub-skills in Figure 2. The resultant mapping is shown in Table 26. We started from the oracy skill and sub-skill in the left-most columns in the table, and mapped the terms used in the talking points and presentation tasks in the two rightmost columns.

We observed both where oracy skills or sub-skills did not appear to be represented in the tasks, and where the sub-skills had been reworded in the task (using an analogous term, rather than the particular term used in the framework).

Table 26: Mapping from oracy framework to assessment tasks

Terms from oracy framework		Actual term or analogue in task description	
Oracy skill	**Sub-skill**	**Talking points**	**Presentation**
PHYSICAL	Voice	Voice (actual term)	Voice (actual term)
	Body language	Body (expression/eye contact) (analogue)	Body language (actual term)
LINGUISTIC	Vocabulary	Range of vocabulary (analogue)	Vocabulary & grammar (analogue)
	Language variety	Register & grammar (analogue)	Register & rhetoric (analogue)
	Structure	Register & grammar (analogue)	Vocabulary & grammar (analogue)
	Rhetorical techniques	Register & grammar (analogue)	Register & rhetoric (analogue)
COGNITIVE	Content	Content & reasoning (analogue)	Content & reasoning (analogue)
	Clarifying and summarising	Building on views of others, summarising & critically examining (analogue)	
	Self-regulation	Building on views of others, summarising & critically examining (analogue)	Structure & self-regulation (analogue)
	Reasoning	Building on views of others, summarising & critically examining (analogue)	Content & reasoning (analogue)
	Audience awareness		Audience awareness (actual term)
SOCIAL & EMOTIONAL	Working with others	Turn-taking, guiding & managing interactions (analogue)	
	Listening and responding	Active listening (analogue)	
	Confidence in speaking		Confidence & flair (analogue)

We make the following observations, based on this table:

- A large majority of the oracy skills and sub-skills are well represented across both tasks.
- A few sub-skills are missing from one or other tasks; for example, ‘clarifying and summarising’ is not assessed in the presentation task, and ‘audience awareness’ is not assessed in the talking points task.
- The oracy skill that has the sparsest coverage across the two tasks is ‘social and emotional’. Two of its sub-skills are not assessed in the talking points task, and one is not assessed in

the presentation task. This degree of lack of curriculum coverage is not, in our opinion, serious, but it is worth noting that 'social and emotional' is perhaps the kind of ephemeral or hard to assess skill that often is missing in assessments.

- While most of the cells in the two rightmost columns are filled, we can also see that many of the terms used in the task descriptions are analogues, rather than the actual term from the framework. Again, this is not a critical failing, but it may give rise to confusion; if there is no special reason for using a different term, the best advice is probably to adopt the term that originates from the framework.

Evaluation of best practice in assessment standardisation

We watched standardisation videos and compared them to example marksheets. In general, we found the marksheets to be clear and fully descriptive; we considered that a teacher watching the videos with the marksheets would have a good chance of understanding how the assessment originators intended the standard to be implemented.

It may be helpful if Voice 21 provided a set of notes for facilitators leading assessor standardisation sessions in pilot schools. For example, particularly salient points could be pointed out in session leaders' notes, and advice about how to make judgments concerning oracy could be cascaded down via standardisation session facilitators' notes or a guidance pack.

On specific videos and marksheets, we observed the following:

- On the example, 'Presentation task, pupil 5', the marksheet noted the following issue with the pupil's intonation (in the 'voice' oracy skill):

  Speaks audibly although flow of presentation is sometimes difficult to follow because of intonation.

When we watched this video, we noted that the girl concerned was an EAL learner, probably from an Eastern European heritage. We wondered whether her 'flat intonation' was, in fact, a language learning matter. Was it that she was a little 'inexpressive' because her English speaking skills had not yet developed sufficiently?

When considering this issue, we surmised that oracy advocates could clarify how such issues related to non-native speakers' language learning (such as non-standard intonation) fitted within the definition of oracy; this would ensure no bias against non-native speakers in the definition of the construct—particularly important in schools with high numbers of EAL learners, such as School 21.

- In several of the exemplified talking points tasks, we asked ourselves about the extent to which it was possible to assess individuals' performance in a group. That is, could one avoid either 'ceiling' or 'floor' effects? In the former case, if one's fellow group members were very dominant, one might not get a chance to speak and so there would be a ceiling put on one's potential oracy scoring. In the converse case, if you were a weak oracy performer, but your fellows were very facilitative, your oracy mark might be 'lifted up' in comparison to an oracy performer of a similar skill whose fellow group members were less helpful.

Evaluation of the dataset generated by the baseline assessments

Generalisability theory analyses

G-theory is an approach to evaluating the quality of assessments by quantifying the amount of variance in data that can be attributed to different variance components. It can help to establish how much of the observed variance can be attributed to 'differentiation facets' (that is, the thing the assessment purports to measure), and how much is attributed to 'instrumentation facets'. If the proportion of variance attributable to instrumentation is high, then it may be that the data contains a lot of 'error' or 'background noise' variance.

Table 28 shows an example data file to illustrate the measurement design in this g-study. The differentiation facets are listed in the first two columns of Table 28. Pupils are 'nested within' schools; that is to say, each pupil is assessed within their own school, and there are no assessors who assess across more than one school. In g-theory jargon, schools are referred to as a 'stratification facet'.

Conversely, the first three rows of Table 28 show differentiation facets. Sub-skills are nested within oracy skills, which are nested within tasks. In fact, because of a limitation of the g-theory programme being used to only one level of nesting in a differentiation facet, sub-skills have been nested (referred to as 'items').[11]

Thus, the g-theory design is as follows:

P:S / I:T

This can be read as: the differentiation facet is pupils (P) nested within schools (S), and this is crossed with the instrumentation facet, in which items (I) are nested within tasks (T).

A useful way to explain this design is to consider a short extract of a data file produced to illustrate the nature of the design—in particular, the concepts of 'nested' and 'crossed facets'.

This illustration is provided in Table 27.

Table 27: Simplified illustration of a data file for the School 21 assessment design

Schools	Pupils	Task 1				Task 2			
		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
School 1	Pupil 1								
	Pupil 2								
	Pupil 3								
	Pupil 4								
	Pupil 5								
School 2	Pupil 6								
	Pupil 7								
	Pupil 8								
	Pupil 9								
	Pupil 10								

The data file is in a grid or matrix format; all the blank cells in the main body would be filled with pupils' scores on each item. But if we look at how the row and column headings are arranged we can understand something about what we mean by nesting and crossed facets, and by differentiation (and stratification) facets and about instrumentation facets.

First, we explain the difference between crossed and nested facets. Items are 'nested within' tasks; items 1, 2, 3 and 4 only occur within the context of task 1 (item 3, say, cannot be within task 2). This is what we mean by 'nesting'. There is a similar issue concerning pupils and schools. Pupils are 'nested within' schools. Pupils 1 to 5 are only assessed within school 1, never within school 2.

However, all pupils answer all items, which is known as a 'crossed design'. But, the design is slightly more complex; we do not say simply 'pupils are crossed with items', rather, we say 'pupils nested within schools are crossed with items nested within tasks'.

[11] In g-theory programmes, we have to refer to facets by initial letters, and since 'schools', 'pupils', and 'sub-skills' all start with 's', we have named them something different.

We are also concerned about the difference between differentiation and instrumentation facets. Essentially, in a measurement design, our aim is to maximise the amount of variance (scoring effectively) that can be attributed to the differentiation facet(s). In other words, the differentiation facet is the 'thing we are trying to measure'. In this case, it is the difference between pupils' ability in oracy. However, in any measurement procedure we also have instruments. In this case, the instruments are items and tasks. In Table 27 instrumentation facets are the column headings and differentiation facets are the row headings.

This corresponds to our equation above:

P:S / I:T

—in which differentiation facets are to the left of the forward slash (/) and instrumentation facets are to the right.

There remains one final complication; it will be recalled that pupils only sit within their own school. This is a form of 'nesting', but because this is on the 'left of the slash', we refer to schools as a 'stratification facet'.

A further extract of the actual matrix is shown in Table 28, and the codes for item names are listed in Table 29.

Table 28: Example data file showing measurement design for g-study

School ID	Pupil ID	Talking Points Task								Presentation Task							
		Physical		Linguistic		Cognitive		Social & Emotional		Physical		Linguistic		Cognitive		Social & Emotional	
		TP_V	TP_B	TP_RG	TP_Rvoc	TP_CR	TP_VOSCE	TP_TTGMI	TP_AL	P_V	P_B	P_VG	P_RR	P_CR	P_SSR	P_CF	P_AA
D	C1	2	2	2	2	2	2	2	3	3	2	2	2	2	1	3	3
	C2	2	3	2	2	2	2	3	3	3	2	3	2	3	3	3	1
	C3	2	1	2	2	2	2	1	3	2	2	2	2	3	2	2	1
J	A1	3	3	3	2	2	3	3	2	2	1	2	1	2	3	1	1
	A2	3	2	3	3	2	2	2	2	3	1	3	3	2	2	2	2
	A3	2	2	2	2	2	2	2	2	2	1	2	2	3	2	2	2

* This table illustrates that 'oracy sub-skills' (which we refer to as 'items' in the g-theory analysis) were nested both within 'oracy skills' (physical, linguistic, …), as well as within tasks. Most g-theory approaches do not model more than one level of nesting, and so we did not model the nesting of sub-skills within oracy skills.

Table 29: Key to sub-skills labels

Item ID	Item in words
TP_V	Voice
TP_B	Body (expression/eye contact)
TP_RG	Register and grammar
TP_Rvoc	Range of vocabulary
TP_CR	Content and reasoning
TP_VOSCE	Building on views of others, summarising, and critically examining
TP_TTGMI	Turn-taking, guiding, and managing interactions
TP_AL	Active listening
P_V	Voice
P_B	Body language
P_VG	Vocabulary and grammar
P_RR	Register and rhetoric
P_CR	Content and reasoning
P_SSR	Structure and self-regulation
P_CF	Confidence and flair
P_AA	Audience awareness

Mean scores for levels within particular facets

URGenova returns mean scores for each level within facets. In the following tables, we report two sets of means. In the first column of means for each variable, the grand mean (or the mean over all observations in the input data set)[12] has been subtracted from each mean reported. This column in each table lets us see the relative standing of each level within each variable.[13] Then, we have added 'raw score means'; these are 'out of' the following totals:

- total score, 48;
- each task, 24;
- each item, 3.

Table 30: Mean scores for schools

School ID	N	Means for S	Mean total score
B	66	-0. 376	15.794
C	64	-0. 118	19.922
D	55	0. 831	35.106
E	62	0. 119	23.714
F	68	0. 143	24.098
G	60	-0. 107	20.098
H	60	0. 123	23.778
I	63	-0. 194	18.706
J	60	-0. 319	16.706

Table 31: Mean scores for tasks

Task name	Means for T	Mean total task score
Talking points task	0. 001	10.913
Presentation task	-0. 001	10.897

[12] In this context, we calculate the grand mean as follows:

$$Grand\ mean = \frac{Sum\ of\ scores}{No.of\ items\ X\ no.of\ pupils}$$

That is:

$$\frac{12{,}170}{(16\ x\ 558)} = 1.363$$

[13] And, by the by, it is these figures that go forward to calculate the subsequent variance components and indices.

Table 32: Mean scores for items (i.e. sub-skills) nested within tasks

Task	'Item' (i.e. sub-skill)	Means for I:T	Raw score mean for I:T
Talking points	Voice	0. 158	1.521
	Body (expression/eye contact)	-0. 092	1.271
	Register and grammar	-0. 100	1.263
	Range of vocabulary	-0. 153	1.210
	Content and reasoning	0. 047	1.410
	Building on views of others, summarising, and critically examining	-0. 042	1.321
	Turn-taking, guiding, and managing interactions	0. 081	1.444
	Active listening	0. 112	1.475
Presentation	Voice	0. 219	1.582
	Body language	-0. 204	1.159
	Vocabulary and grammar	0. 022	1.385
	Register and rhetoric	-0. 051	1.312
	Content and reasoning	0. 094	1.457
	Structure and self-regulation	-0. 032	1.331
	Confidence and flair	-0. 026	1.337
	Audience awareness	-0. 033	1.330

The Facets analysis and multiple regression analysis below presents more information about the relative position of schools. However, at this point, tasks appear very close together in difficulty (and items within tasks also appear largely of similar difficulty).

Most of the schools appear to have similar scores on the oracy tasks, with School D having scored notably higher than others (Table 30). This is discussed in more detail below.

Variance components and generalisability coefficients

URGenova gives two types of output relating to the amounts of variance that can be attributed to the differentiation facet (oracy). These are a variance components table (Table 33) and some values of generalisability coefficients—similar to reliability coefficients (Table 34).

Table 33: Variance components for baseline assessments design

Effect	df	T	SS	MS	VC	%
S	8	969.7732	969.77323	121.22165	0.116	20.45%
P:S	549	2766.992	1797.21828	3.27362	0.15166	26.73%
T	1	0.01614	0.01613	0.01613	-0.00238	-0.42%
I:T	14	108.5623	108.54615	7.7533	0.01233	2.17%
ST	8	999.9769	30.18757	3.77345	0.00453	0.80%
SI:T	112	1206.258	97.73462	0.87263	0.01098	1.94%
PT:S	549	3262.242	465.04631	0.84708	0.08184	14.43%
PI:ST	7686	4946.742	1478.21924	0.19233	0.19233	33.90%

S = schools; P = pupils; I = items; T = tasks.

The meanings of the column headings are as follows:

df—degrees of freedom;

T—uncorrected sums of squares;

SS—sums of squares;

MS—mean squares; and

VC—g-study estimated random effects variance components.

Table 34: True score, absolute and relative error variances, and relative and absolute g coefficients

Coefficient	Value
s2(T)	0.152
s2(D)	0.173
s2(d)	0.053
Er2	0.741
Phi	0.468

Table 33 shows that around 47% of the variance is accounted for either by pupils or by the stratification facet 'schools' (S + P:S). To some extent this is comforting, but it could also be argued that only around one quarter of the observed score variance can unequivocally be attributed to differences between pupils.

Most of the instrumentation facets (T, I:T, ST, and SI:T) are relatively small in and of themselves (all less than 3% of the variance), but PT:S accounts for around 14% of the variance. The highest order interaction effect is also confounded with residual variance (of unknown cause). This term (PI:ST) accounts for approximately one third of the variance (differences due to scoring) in the dataset.

Table 34 reports the relative and absolute g coefficients. The former (Er2) is also the true score variance —s2(T) divided by the total variance (true score variance plus relative error variance), or:

$$Er2 = \frac{s2(T)}{(s2(T) + s2(d))}$$

In contrast, the absolute g coefficient (phi) has the same structure, except that the absolute error variance is on the denominator:

$$Phi = \frac{s2(T)}{(s2(T) + s2(D))}$$

It might be considered that a relative coefficient value of 0.741 is reasonable for a teacher-assessed procedure for research purposes. However, the difficulty comes when the absolute coefficient is considered: this low value (0.468) suggests that scores from this assessment procedure would not readily generalise to different contexts (for example, different curriculum areas, different tasks, different assessors, and so on).

Figure 10 illustrates the impact of measurement imprecision. This figure shows the mean scores (the dot for each school). The lower and upper bands are confidence intervals (CIs) generated by multiplying the square root of the relative error variance by 1.96 and subtracting or adding that quantity to the mean to create the upper and lower bounds (whiskers).

Figure 10: Mean scores for schools with upper and lower confidence intervals of 95%

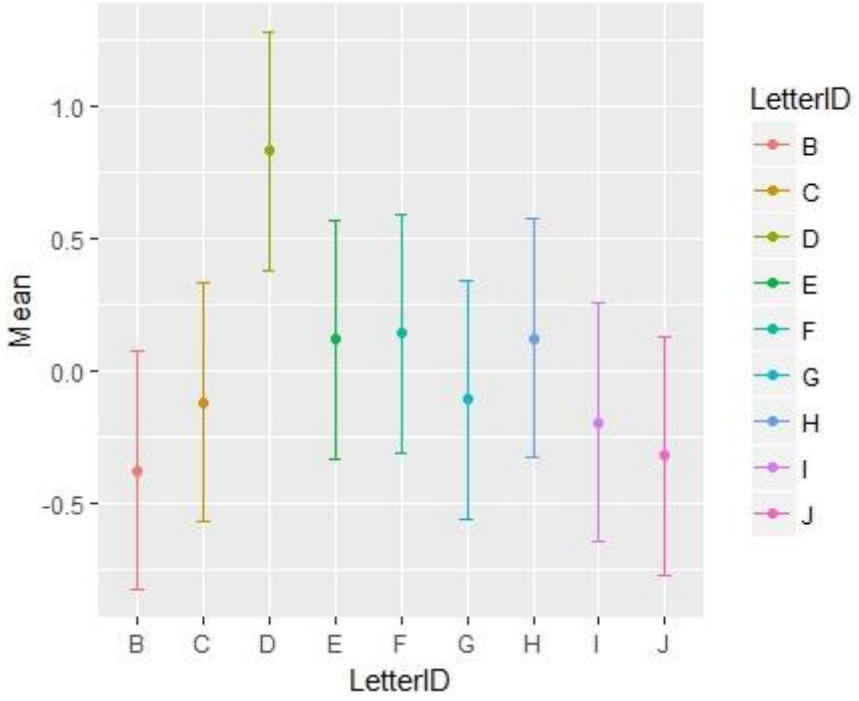
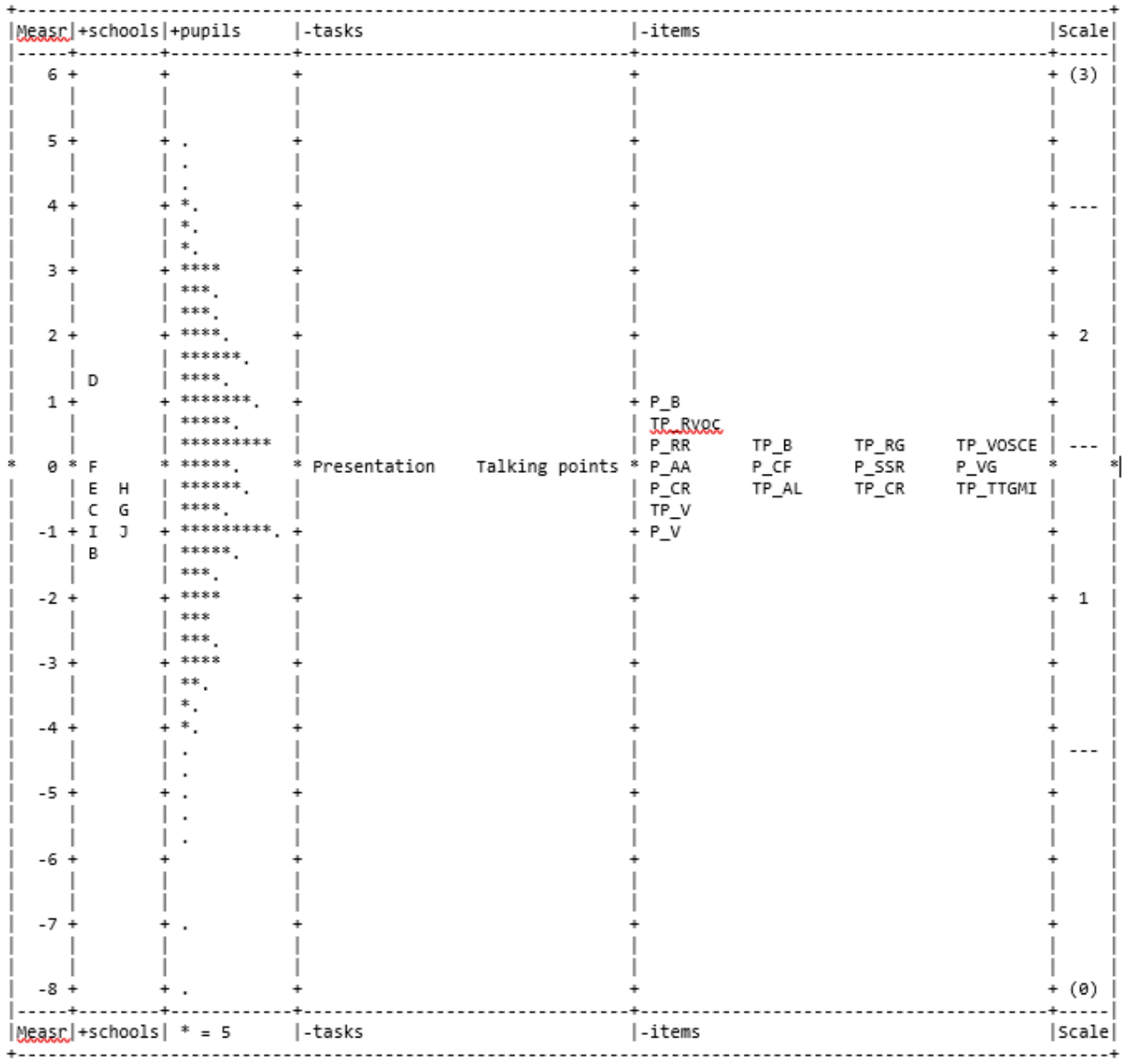


Figure 10 suggests that School D's means are abnormally large; indeed, the upper bounds for Schools B, C, G, I and J were lower that School D's lower bound.

Rasch FACETs analysis

Figure 11 shows a 'variable map' which places schools, pupils, tasks, and items (sub-skills) in respect of a single scale or 'ruler'. Sub-skill labels are set out in Table 29, above. The variable map shows the standing of different variables on a difficulty/ability trait relative to each other. It can be seen (for instance) that one school (denoted by its letter) was very high on the ruler. This would suggest that the school was high-attaining. That school can be compared both with other schools and with levels within other variables (such as tasks, items—pupils even). In the map, variables are signed either positive or negative: positive signs suggest a school or pupil has a lot of the trait (ability) if they are at the top of the map; a negative sign suggests a task or item is very difficult if it is towards the bottom of the map.

Figure 11: Variable map showing relative positions of schools, pupils, tasks, and items



The variable map largely confirms the mean scores derived in the g-theory process. School D remains somewhat apart from the other schools, and the two tasks appear almost identically difficult, for example.

As was the case in the g-theory analysis, the items do not have a huge range of difficulty. This is neither problematic nor consoling necessarily; given that this is a novel assessment method, we should be cautious in assuming that items would have particular characteristics—for example, it is quite possibly entirely reasonable that all items are roughly equally difficult.

As we saw in Figure 6, pupils' total scores on the pre-intervention assessment have characteristics consistent with the normal distribution. However, we must be aware of the issue of 'disconnected sub-sets'. In a design with disconnected sub-sets, it cannot be certain whether (for instance) School D has pupils who are (genuinely) gifted in oracy, or whether that school's assessors are somewhat more lenient than others. To resolve this confound, the 'anchoring design' could have been improved and assessors asked to visit other schools and assess some of their pupils. Such activity, however, was never envisaged in this project (it is not included in the research protocol, for example). Moving teachers

between schools would have been costly and onerous. There were a range of standardisation and QA steps in place (as discussed earlier in this report), which aimed to facilitate teachers' understanding and common interpretation of assessment standards.

In the absence of a linking or anchoring design, the next best approach is to see if School D's high performance on oracy appears likely, given background variables.

A multiple regression analysis was conducted to address such matters and is reported below.

Multiple regression analysis

In the initial stage of the trial each school provided assessment data for a sample of approximately 60 pupils, but also background data on the entire cohort (Year 7 in most cases).

This background data included:

- gender;
- SEN status;
- EAL indicator;
- Pupil Premium indicator; and
- KS2 performance (in writing, reading and mathematics).

One problem with the background data provided is that the KS2 writing performance of the pupils is teacher-assessed and although there are guidelines set out on how to report outcomes (Standards and Testing Agency, 2016) some schools reported numerical data instead of the categorical data that is recommended.

As KS2 writing performance is an important variable used later, the data is split into two parts for this regression analysis between the five schools that reported categorical data and the four that reported numerical data. This was done as two separate regression models needed to be undertaken because the categorical levels cannot be directly compared to the numerical values.

The different levels of KS2 writing and their meanings are as follows:

- BLW—below the standard of the interim pre-key stage standards;
- WTS—working towards the expected standard;
- EXS—working at the expected standard; and
- GDS—working at a greater depth within the expected standard.

Table 35 below summarises the percentage of pupils who achieved each of the four writing levels, along with their performance in KS2 reading, from the five schools that gave categorical writing levels. For the four schools that reported numerical writing scores, some summary statistics are reported in Table 36.

Table 35: Summary statistics for those schools that provided KS2 Writing categorical data

KS2 Writing Levels	Percentage of pupils
BLW	6.5%
WTS	34.5%
EXS	48.4%
GDS	10.7%

Summary Statistics	KS2 Reading
Min.	80
Mean	100.2
Max.	120
Missing	24
Standard Deviation	8.07

Table 36: Summary statistics for those schools that provided numerical KS2 reading and writing scores

Summaries	KS2 reading	KS2 writing
Min.	83	83
Mean	103.4	104.3
Max.	120	120
Missing	16	16
Standard Deviation	8.35	6.93

Additionally, in the regression analysis, School F was not included as it entered Year 8 pupils rather than Year 7 pupils.

First, the summary statistics for the five schools that provided categorical KS2 writing levels is considered. The boxplots in Figure 12 show the range of oracy scores for the four different KS2 levels and the different schools.

Figure 12: Boxplots showing oracy score divided into the different KS2 writing levels and schools

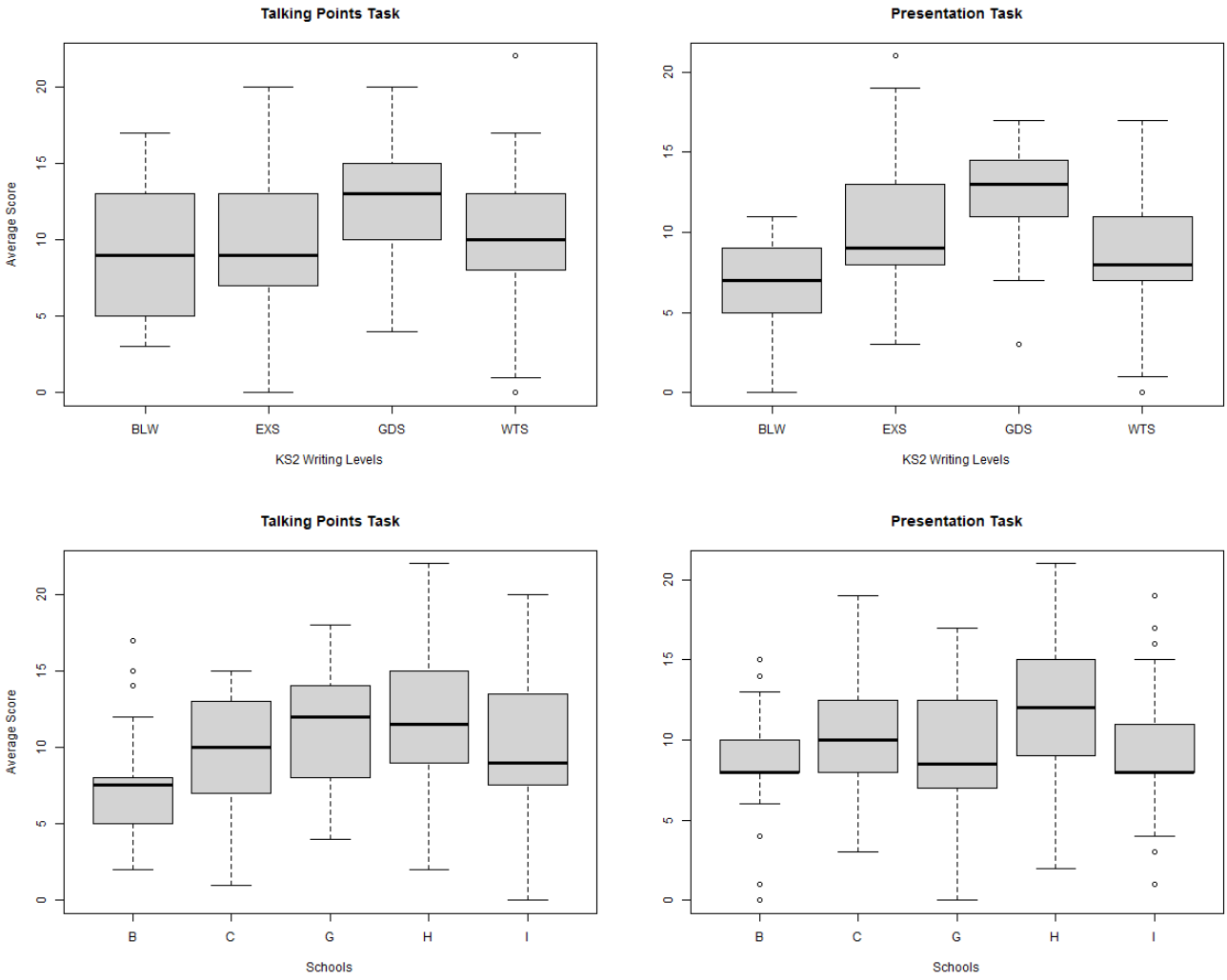


Figure 12 suggests that pupils with a KS2 writing level of GDS (the best possible) performed better on both oracy tasks than other pupils. The varying performance of the schools is a little subtler as some have large ranges in score but pupils from school B receive lower marks on average than their counterparts in other schools.

If the regression model is fitted to the data, the coefficient estimates and p-values in Table 37 are obtained.

Table 37: Coefficient estimates, standard errors, and p-values

Coefficients	Estimate	Standard Error	p-value	Significance
Intercept	8.428	0.941	<2e-16	***
School G	1.982	0.665	0.0031	**
School C	1.182	0.553	0.0332	*
School H	4.538	0.629	5.25e-12	***
School I	1.625	0.550	0.0034	**
KS2 writing: EXS	-0.203	0.874	0.8164	
KS2 writing: GDS	2.169	1.054	0.0406	*
KS2 writing: WTS	-0.523	0.860	0.5435	
KS2 reading	0.435	0.202	0.0323	*
EAL: yes	-1.481	0.520	0.0047	**

* significance at the 5% level; ** significance at the 1% level; *** significance at the 0.1% level.
N = 290 pupils.

Because there were several different levels, to fit a regression model, a baseline was required that, in this case, was pupils from School B with a KS2 writing level of BLW and an EAL status of 'no'. School B was a reasonable choice as baseline because this school had one of the lowest average oracy scores of the five schools in this sample.

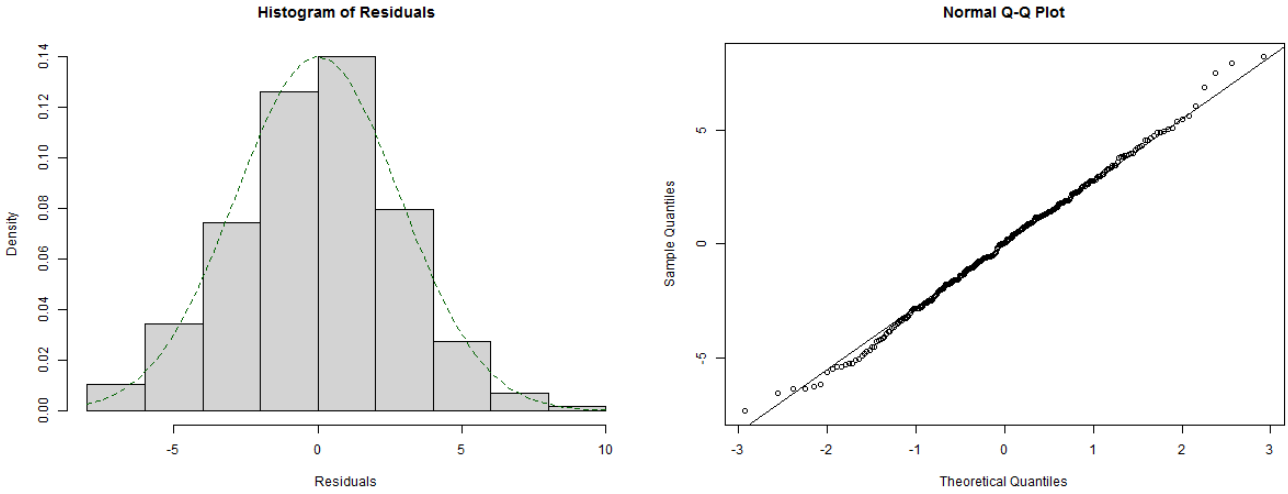
The fitted coefficients for the different schools in this sample show that they are all significant. However, School H stands out as potentially being unduly lenient due to its much larger coefficient value. Such leniency might result, for example, in pupils from School H being given much higher oracy scores than otherwise identical pupils—in terms of background variables—from other schools.

Among the writing levels, only the highest, GDS, was significant in that those pupils who achieved this were predicted to have a higher oracy score than others. Indeed, from the other three levels not being significant, it can also be said that having a KS2 writing level of EXS, WTS or BLW makes no difference to the pupils' predicted score. KS2 reading scores are also positively associated with oracy score (this being on a numerical scale), in addition, having English as an additional language was negatively associated with oracy performance.

To assess whether the regression model is valid, the assumption that the error terms for the fitted model are indeed normally distributed needs to be checked. Obtaining the error terms from the fitted model involves taking the differences between the actual oracy score for each pupil and the predicted score given by the model above.

There are two main diagnostic plots that are used to check that error terms are normally distributed—a histogram of the error terms, and a Q-Q plot which look at the distribution and spread of percentiles respectively compared to a normal distribution. For the model considered above, these plots are shown in Figure 13.

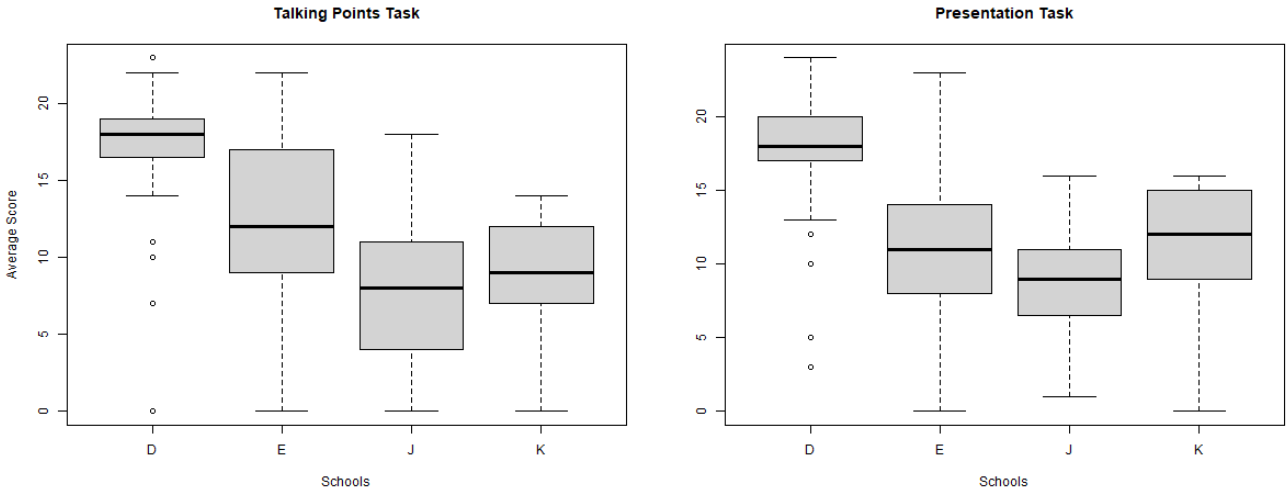
Figure 13: Assessing model fit via the error terms



Both plots show that the distribution of the error terms is approximately normal, so the predictive power of our model can be assumed with some confidence.

The boxplots in Figure 14 show the range of oracy scores for the four schools that provided numerical KS2 writing levels.

Figure 14: Boxplots showing oracy task scores for the four schools providing numeric KS2 assessment data



School D has a significantly higher average score for both oracy tasks and a much smaller range (apart from the four outliers) than the other schools.

The same regression model as before was fitted on these four schools and obtained the coefficient estimates and p-values reported in Table 38.

Table 38: Coefficient estimates, standard errors and p-values

Coefficients	Estimate	Standard Error	p-value	Significance
Intercept	8.371	0.736	<2e-16	***
School D	7.890	0.926	9.31e-15	***
School K	3.966	1.408	0.0054	**
School E	2.746	0.851	0.0015	**
KS2 writing	0.880	0.425	0.0399	*
KS2 reading	1.000	0.470	0.0350	*
EAL: yes	0.915	0.741	0.2184	

* indicates significance at the 5% level; ** significance at the 1% level; *** significance at the 0.1% level.
N = 173 pupils.

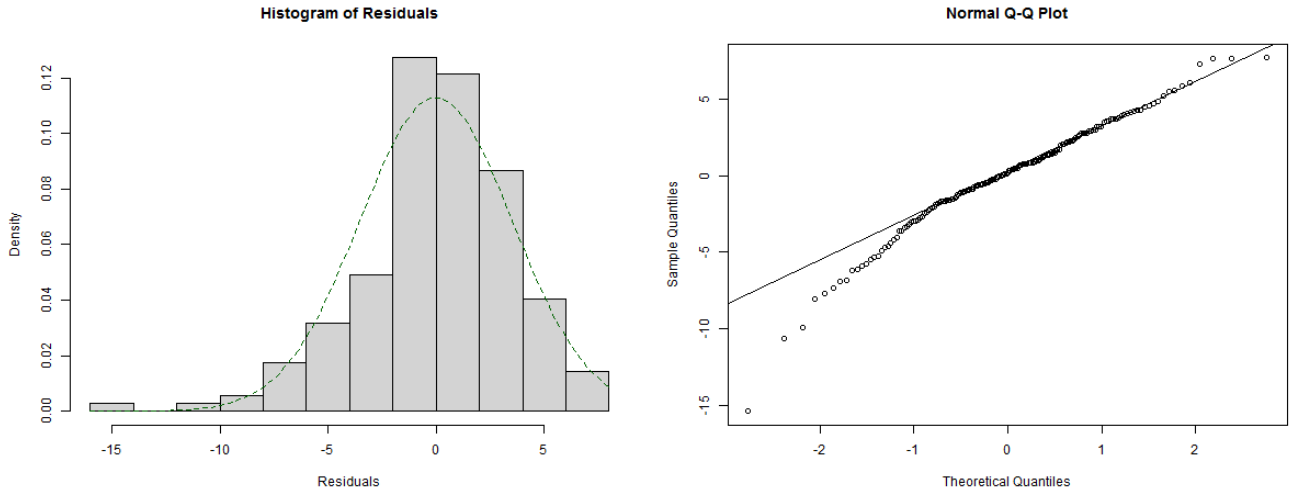
As in the previous model, there was a baseline scenario due to the different levels in the model; in this instance, the school was School J with an EAL status of 'no' and a score of zero in both pre-tests.

Looking at the fitted coefficients for the different schools, it can be seen that they are all significant. School D has an especially large positive impact on oracy score compared to the other schools. As in the previous model, this is a school effect and so it is concluded that the marker(s) at School D were much more lenient than at other schools and being a pupil at School D confers a significant advantage.

When looking at the KS2 reading and writing scores on a numerical scale instead of using levels, it was found that both factors have approximately the same positive effect on oracy score. This is contrary to KS2 writing being dominant in the previous model.

Contrary to the previous model, English as an additional language was not significant in this model. Figure 15 shows the diagnostic plots of the error terms.

Figure 15: Assessing model fit via the error terms



In Figure 15, the histogram and the Q-Q plot show that the distribution of the residuals is quite heavily skewed to the left compared to a normal distribution. This indicates that there were a significant number of pupils who underperformed their predicted oracy score.

The data was examined in more detail and it was found that School K was under represented in the sample.[14] However, a common feature of all the underperforming candidates was their poor performance on at least one half of the oracy assessment with some candidates obtaining a mark of zero in one of the tasks. This could potentially be due to outside factors such as absence as neither

[14] Note that School K was not included in g-theory and Facets analyses for this same reason.

one of the tasks was overrepresented for the pupils who performed poorly. There is insufficient data to investigate this thoroughly, but it is worthy of note as this effect does not occur in the first regression model.

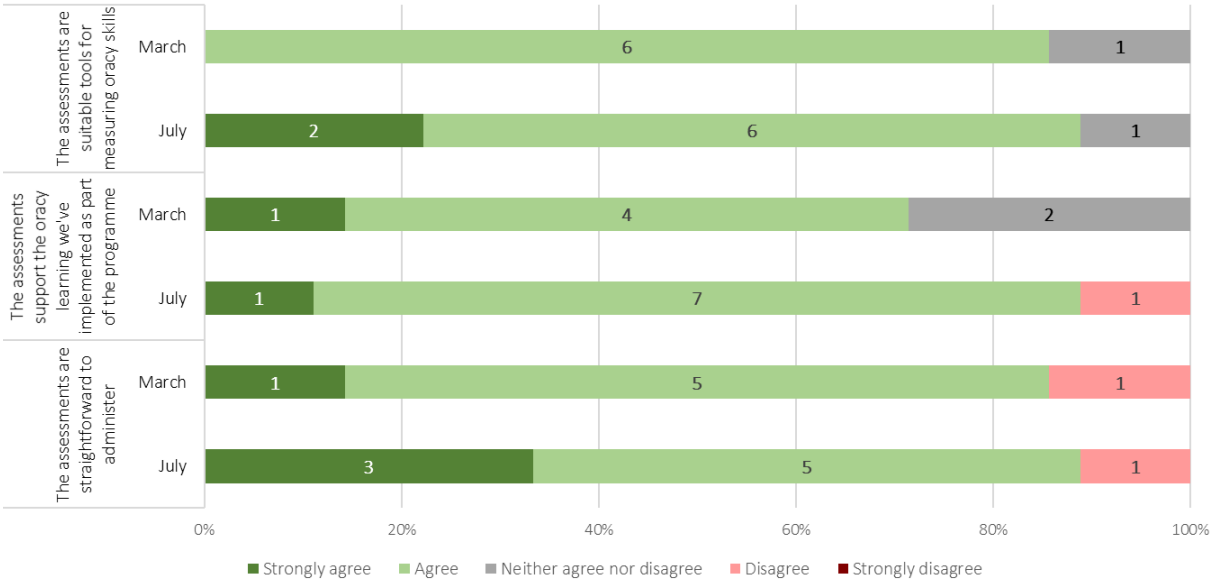
The main objective of this multiple regression analysis was to discover the relationships between a set of background variables and a pupil's oracy score. The analyses suggested that the most significant variables were the school attended, prior attainment, and EAL status. The fitted coefficients for the school effects from Schools H and D stand out in this analysis as high oracy scores cannot be fully explained by background data (such as high prior attainment). Our interpretation is that these results are due to lenient markers.

There was nothing in the process evaluation element of this research to suggest that there might be reasons other than marker leniency as to why these pupils performed better (for example, previous oracy initiatives at the school). While collection of other variables might give rise to contrary findings, our conclusion on the evidence we have is that these schools appear to be assessing oracy more leniently compared to other institutions.

Views on the assessments from the online survey of teachers

The online survey of teachers conducted in March and again in July 2017 also sought opinions of the oracy assessments. Seven respondents to the March survey and nine July survey respondents said that they had been involved in undertaking the baseline assessment with pupils during the autumn term. These respondents were asked to indicate how far they agreed or disagreed with three statements about the assessments. As Figure 16 shows, these teachers were largely positive about all three aspects of the assessments, and there was a slight increase in the proportion of positive responses to each statement from March to July. In both surveys, there was one respondent who disagreed with the statement 'the assessments are straightforward to administer' and there was also, in the July survey, one 'disagree' response to the statement 'the assessments support the oracy learning we've implemented as part of the programme'. (Note that in the July survey, it was the same respondent who disagreed with these two statements, and all three 'disagree' responses came from School E, suggesting it may have been the same person giving all negative responses.) There were six 'strongly agree' responses across this set of statements in the July survey and four of these were made by respondents from School F (one individual gave three and another gave one 'strongly agree' response), while the remaining two were given by one respondent from School J. One respondent in School F also gave both 'strongly agree' responses in the March survey.

Figure 16: Number of respondents agreeing or disagreeing with each statement on the oracy assessments (March and July)



Conclusion

Formative findings

The Oracy Improvement Programme was perceived positively by stakeholders—oracy leads, SLT members and teachers involved in programme delivery—in all schools involved in the pilot. The programme was generally considered to have been worthwhile and beneficial to pupils and teachers. Indeed, all ten schools that were involved in the final interviews indicated at that stage that the oracy programme would run again for the new Year 7 intake in September 2017 (although with minor adjustments in some cases, for example, some schools were planning to change the order of presentation of topics and skills, while others planned to move the lessons and responsibility from one department or subject to another). All but one of the ten schools also said that they would continue oracy teaching in some form or another for Year 8 (the cohort that had undergone the programme as Year 7 pupils). In many cases, the exact form of this continuation was still to be discussed, but embedding it across a range of subjects was the most frequently suggested delivery method. Two schools were also planning to extend some form of oracy coaching into Year 9, but exact details were yet to be finalised, although one school was considering a careers focus and another mentioned incorporating it into exam preparation or study skills sessions. There was also a plan at one school to present a condensed version of the programme to new sixth formers via an induction oracy day.

The areas in which improvements might be needed, based on this pilot project, lie largely in the assessments. The analysis of the assessment tool undertaken as part of this evaluation suggests that reliability is a potential issue, for example, it was difficult to explain whether progress between the two test events represented a genuine improvement in skills and abilities, or whether factors such as the leniency of markers were also having an effect.

The analyses conducted during this pilot indicate that the assessments used to measure oracy generate substantial amounts of instrumentation, residual, or error variance. To some extent, this may be a teething problem; a new approach to assessment could be ameliorated as assessors become more familiar with it. But, there are reasons to believe this is not the main explanation. A relative g-coefficient of around 0.75 for a teacher assessment of a performance skill is probably about right, representing the 'true' or most likely extent of reliability in such an enterprise.[15] This will mean, however, that any measure made using such a procedure would be likely to contain substantial amounts of error variance, as discussed above. This would inhibit this assessment approach as a tool for generating the kind of robust measures that the EEF's efficacy and effectiveness trials require. The suggestions in Table 39 are put forward as alternative options for oracy assessment, based on the experiences of this pilot.

[15] There are many treatments of assessment reliability in the literature. Harth and Hemker (2012) and Johnson and Johnson (2012b), however, suggest that the reliability of complex assessments (similar in nature to those trialled here) in U.K. school qualifications are in the general area to that found here.

Table 39: Pros and cons of alternative options for oracy assessments

Description of option	Pros	Cons
Amend the existing oracy assessments so as to make them generate more reliable data (e.g. make marks more objective). Assessors could use insights from medical assessments—such as objective structured clinical examinations (OSCEs), which combine checklist based assessments with impressionistic scores.	• Such an innovation would probably produce more reliable scoring. • Future trials could then continue to be based on a direct measure of oracy.	• Redesigning an assessment to increase standardisation as described may 'remove the essence' from oracy assessment. • As we understand it, substantial time has already been given to designing and redesigning oracy assessments. It may seem wasteful to revisit this.
For subsequent trials, do not use oracy scores as an outcome measure—use some other measure (such as a standardised test of English writing, or a reasoning test score, or KS2 / KS4 data).	• Such a measure would probably produce more reliable data. • Voice 21 could use such a measure without having to 'remove the essence from' its teacher assessments of oracy.	• Oracy advocates could argue that any measure of something other than oracy does not represent the trait; and thus, doubt would be raised as to how much improvement in the other measure could be attributed to improvement in oracy. • As we understand it, this approach was tried before and abandoned in favour of the current approach of having a direct assessment of oracy.
Consider the possibility that, while oracy is generally a good thing educationally, it may not be possible to measure a significant impact on attainment using the EEF's preferred approach to gathering and analysing experimental evidence.	• This could release oracy advocates from having to resolve the knotty and possibly irresoluble issues addressed in this report. • An educational innovation can still be perfectly reasonable, even if it cannot demonstrate efficacy within current conceptions of experimental evidence.	• If proof of impact is not attainable, then an important source of funding for the educational good of oracy may be cut off. • The EEF's central contention that educational innovations should be able to show a material effect is surely worth holding onto.

Some of the core elements of the programme worked better than others when taken outside of the School 21 context. Most notably, the oracy assemblies proved very difficult to organise and deliver in most of the pilot schools due to the practical and logistical challenges of delivering them in the spaces and with the pupil numbers at the school, but also, in a small number of schools, there was resistance from senior leaders to change the format and purpose of assemblies to accommodate the oracy assemblies. This, in turn, made it more difficult to start to adopt a whole-school oracy culture in the way that School 21 has. The assemblies were intended to help contribute to spreading the oracy ethos but this was one area in which all schools admitted that progress had been minimal (although this was expected to be the case as it was felt it would take more than a year to achieve any significant 'culture shift'). To a lesser degree, the requirement for one dedicated oracy lesson a week proved problematic in a small number of schools. However, most found a way to deliver this, and the biggest compromise in evidence on this front was that in one school pupils received one lesson per fortnight.

Interpretation

Evidence to support the theory of change

Referring back to the initial logic model used to articulate the overarching theory of change for the pilot schools, the outcomes and impacts identified are summarised in Table 40 below.

Table 40: Summary of the outcomes and impacts identified in the theory of change logic model

Outcomes	Impacts
Pupils have developed: • a sense of socio-emotional empowerment; • wider reference points; and • improved thinking and problem solving skills, a**nd are:** • capable of dialogic learning; • able to address their misconceptions; and • willing to push boundaries. Pupils have acquired the full range of oracy skills (to different degree): • physical; • linguistic; • cognitive; and • social and emotional.	• Pupils have improved oracy. • Pupils have improved level of achievement across several subject areas. • Pupils have improved level of measurable attainment in specific subject areas.

The oracy leads and teachers that were interviewed tended to view any improvements in pupils' oracy skills in a more generic way rather than relating them to the four skills areas or to the more specific sub-skills. The interviews suggested that the main shift in pupils' development was the increased confidence they exhibited, particularly in terms of presentations and speaking in front of groups. To a slightly lesser extent, pupils were seen to be using an expanded vocabulary in their classroom discussions and were said to be less likely to resort to slang and other conversational 'bad habits'. Another key enhancement noted by interviewees quite early on in the pilot year was the improved listening skills pupils were demonstrating both in listening to teachers and to fellow pupils. This was felt to be a particular benefit to learning which, in turn, could ultimately contribute to the intended impact—'pupils have improved level of achievement across several subject areas'—although this was not considered to have resulted from the pilot year alone. Our interviewees suggested that there was little evidence from the pilot project of any tangible improvements in pupils' achievement or attainment in specific subject areas. In their view, this was perhaps the result of the cognitive aspect of the programme being the most under-developed. As a counter, we observe that the pilot was not designed to collect data to show impact on attainment, and although teachers questioned any immediate impact on attainment, a few thought this might be a longer term outcome.

The extent to which pupils were able to draw on wider reference points and achieve a sense of socio-emotional empowerment was evidenced in some of the interviews where teachers described using current affairs topics (such as terrorist incidents) as the basis of oracy lessons. This had helped pupils in articulating and coming to terms with upsetting events that were close to them geographically and had directly or indirectly affected them.

Overall, the interviews, the teachers' survey responses, and the assessment data indicated that pupils' oracy skills had improved over the year, and for many, the Ignite speech at the end of term demonstrated the extent of such improvement.

The role of training was an important element of the successful delivery of the programme. The training provided at School 21 for the oracy leads and relevant SLT member was perceived by many as a crucial aspect in successfully delivering the programme. A small number of interviewees who were unable to attend the Voice 21 training were disappointed at missing out on this and felt that it had slightly hindered their introduction of the programme and its initial implementation at their school. Those who had a member of the Voice 21 team visit their school to deliver training or give a presentation were appreciative of the input, but in some instances felt that they had missed an opportunity by not providing Voice 21 with a clear specification on what they wanted to cover during the visit.

Towards the end of the pilot year, the role of cascading training to colleagues became more apparent, with many schools reporting that CPD sessions had been allocated as oracy training sessions to help bring all colleagues up to speed with what the programme involves and how they might be able to incorporate the skills and techniques into their own lessons. The extent to which this approach to training is sustainable must be considered. If the programme went to a larger scale trial, could Voice 21 still provide the amount and type of training and support enjoyed by the pilot schools? What would the implications for programme delivery be if they could not do so? Could schools continue to use the limited time they have for CPD on oracy training?

At the final interviews, some interviewees expressed concerns over maintaining the momentum of the programme in order to achieve the wider aims of embedding oracy across all subjects and shifting the culture at the school. Training could play a major role in keeping the programme at the forefront of people's minds and ensuring the necessary support and buy-in from colleagues. There was some evidence of moves being made towards embedding a wider oracy culture, for example, several oracy leads and teachers described how they had found themselves adopting oracy techniques in other lessons, for example when instigating discussions, and in getting children to talk about concepts before starting to write.

Feasibility of the approach

Overall the programme was viewed positively. In the interviews and in response to the online surveys there was widespread agreement that the programme was generally feasible and scalable and would work in most schools and contexts.

Practical and logistical matters were the main concerns that might affect feasibility, for example, finding one lesson a week in the timetable to dedicate to oracy, or having appropriate teaching spaces in which to accommodate the style of teaching. The timetabling issue was most frequently resolved by the pilot schools by assigning oracy to a specified department, for example, the English department, and using one of the timetabled sessions for that subject as the dedicated oracy lesson. In one school, responsibility for oracy delivery was being rotated across different departments on a termly basis to overcome any potential issues with just one department 'losing' a lesson a week on a permanent basis. In other schools, oracy was being delivered in PSHE sessions, which were generally taught by teachers from a range of departments. The problem with the suitability of teaching spaces tended to be a concern predominantly at the beginning of the year; towards the end of the year, many had discovered that almost any teaching space could be adapted to suit the needs of delivering an oracy lesson. More frequently, the delivery of assemblies was constrained or completely impossible due to logistics and the attitudes of leaders as to the purpose and format of assemblies.

The most frequently mentioned conditions or prerequisites for the successful implementation and delivery of the programme described by interviewees were the need for full buy-in and support from the SLT, high quality initial training (preferably run by, and held at, School 21), buy-in and support from other colleagues, a cohesive team to deliver the programme, adequate time to prepare before delivering the programme, and a motivated and enthusiastic oracy lead. There were just two suggestions that the school needed to be of a particular type or context in order for the programme to work: one person suggested that it would work better in smaller schools and another felt that it would work best where

the ethos of the school is 'whole-pupil centred' rather than having a primary focus on external examination results.

The regional hubs element of the programme was not seen to be essential to its successful delivery. Most schools had no contact at all with other schools in their hub and felt that this was unlikely to happen unless facilitated and encouraged by Voice 21.

No one indicated that the programme had substantial financial cost implications to the school. The greatest costs were hard to quantify and related, for example, to the time required—particularly from the oracy lead—to train colleagues and produce lesson plans or schemes of work—tasks often said to have been 'absorbed' by the oracy leads in addition to their existing responsibilities. The only minor reservations in terms of direct financial costs were the requirement for certain equipment (for example, iPads for recording pupils' work) and the cost of continuing beyond the pilot year when there would be more of a financial outlay for staff training.

Readiness for trial

This programme would be suitable for future efficacy or effectiveness trials with some adjustments to meet the needs of a larger scale trial. Three of the key components of the EEF efficacy and effectiveness trials are:[16]

- delivery of the intervention in a larger number of schools;
- a process evaluation; and
- a quantitative impact evaluation to assess the impact on attainment.

In terms of delivering the intervention in a larger number of schools, as discussed above, the programme was generally felt to be scalable and feasible, although the sustainability of the training and support offering was a potential issue. However, this could be built into any future trial by designing a training and support package that is feasible to deliver to all schools involved in the intervention.

A process evaluation requires an exploration of issues around fidelity (the delivery of the intervention as intended) and dosage (the level of exposure for participants). The oracy programme is, by design, a non-prescriptive 'outline' curriculum with content adjusted to suit the context of the school and the subject through which it is being delivered. This lack of clear definition of exactly what delivery of the programme should look like might make assessing fidelity, and (to a lesser extent) dosage, problematic. It is important to note that the oracy leads and other teacher deliverers interviewed during the pilot year stressed that the flexibility and non-prescriptive approach of the programme was one if the features they liked about it and was what made it a feasible intervention in almost any school regardless of context. There were also indications that beyond Year 7, several of the pilot schools were looking to embed oracy skills development within other subjects. If this programme was adopted as it currently stands in a wider trial, its flexibility and non-prescriptive nature could make it even more problematic to define the intervention and what constitutes delivery as intended, and what the desirable and actual 'dosage' of the programme is. To compound these issues, four interviewees indicated that they were fairly convinced that there was little consistency of delivery even within their own team of teachers delivering the programme, let alone between different schools. Although these issues might make it difficult to envisage a situation whereby fidelity to the intended programme is stable enough both within and across treatment schools to make a larger scale trial feasible, they are not insurmountable given careful planning and a clear definition of the programme at the start of any future trial.

The lack of firm evidence confirming that the assessment tool is reliable and valid creates some issues for any larger scale trial in terms of the need to conduct a quantitative impact evaluation to assess the impact on attainment. The discussion concerning the best assessment procedure between oracy

[16] Source: EEF evaluation, 'A cumulative approach', available at: https://v1. educationendowmentfoundation. org. uk/uploads/pdf/EEF_evaluation_approach_for_website.pdf [accessed 22 Nov 2017].

advocates at Voice 21, the project evaluators, and various advisors is summarised in Table 4, above, and the options on completion of the project are set out at Table 40. We could imagine other alternatives (for example collecting GCSE results to see whether pupils who had experienced oracy tuition had higher grades on average), however, it would be difficult to draw any meaningful conclusions when the oracy intervention and collected data point are so far apart in time, that is, an intervention that took place at the start of a pupils' secondary school career compared to their results at the end of compulsory schooling. Also, the outcome measure of GCSE grade does not map directly to oracy skill so any quantitative conclusion drawn from this may not be valid.

Limitations of the evaluation

As a small-scale pilot, attrition at any level is bound to have an effect on the evaluation. Although 12 schools initially signed-up to the pilot, only ten participated in the evaluation activities to the end of the pilot year. Although one dropped out before the programme had been implemented and therefore had not taken part in any evaluation activities, another school took part (albeit in a somewhat limited way) in all evaluation activities up until Easter 2017. Despite efforts to undertake one final interview with the school, it was not possible to do so, which was potentially detrimental to the evaluation since it would have been helpful to have been able to develop an understanding of the circumstances which led to the non-participation and whether this meant the programme had also been discontinued (and if so, why). Similarly, for the assessment data, there were some schools that did not return the full set of data, or for which the complete datasets from the two assessment occasions could not be matched.

The way in which schools were selected to take part in the pilot may have had some influence on the outcome of the pilot. Although with only 12 schools able to take part it is impossible to ensure good representativeness of different school types, one element that united all of those selected was that they had expressed an interest in the oracy programme; this in itself might slightly skew the sample, but seems a necessary concession to minimise attrition on the pilot.

Many schools had difficulty in articulating exactly how the programme was being implemented and exactly how they would define the oracy curriculum. Although outside of the scope and budget of this project, lesson observations might have been a helpful addition to address these matters.

Future research and publications

Our findings suggest that a key area for future research emerging from the experience of the pilot of this programme is in the suitability of oracy skills assessments. For example, there may be value in further exploring whether any existing outcomes such as standardised English writing tests or reasoning tests could adequately measure attainment and progress in oracy, or whether some other form of assessment might be better placed to address issues of reliability and validity in assessing oracy skills.

References

Allison, P. D. (1990) 'Change scores as dependent variables in regression analysis', *Sociological Methodology*, 20, pp. 93–114.

Alpha*Plus* Consultancy Ltd. (2013) *Standardisation methods, mark schemes, and their impact on marking reliability*, Coventry, U.K.: Office of Qualifications and Examinations Regulation.

Baird, J-A., Hayes, M., Johnson, R., Johnson, S. and Lamprianou, I. (2013) 'Marker effects and examination reliability: a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling', Office of Qualifications and Examinations Regulation, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378059/2013-01-21-marker-effects-and-examination-reliability.pdf

Brennan, R. L. (2001) *Generalizability theory*, New York: Springer-Verlag.

Brennan, R. L. (ed.) (2006) *Educational Measurement* (4th edn), Westport, CT: American Council on Education/Praeger.

Carnegie Mellon University (2015) *How can I assess group work?*: http://tinyurl.com/pqn4xma

Coughlan, S. (2017) 'Free school meals is "unreliable poverty measure"': http://www.bbc.co.uk/news/education-39479028.

Eckes, T. (2011) Introduction to many-Facet Rasch analysis: analysing and evaluating rater-mediated assessments, Frankfurt am Main: Springer Verlag.

Harth, H. and Hemker, B. (2012) 'On the reliability of results in vocational assessment: the case of work-based certifications', in Q. He and D. Opposs (eds), *Ofqual's reliability compendium*, Coventry, U.K.: Office of Qualifications and Examinations Regulation (pp. 321–364).

Johnson, S. and Johnson, R. (2012a) 'Conceptualising and interpreting reliability', in Q. He and D. Opposs (eds), *Ofqual's reliability compendium*, Coventry, U.K.: Office of Qualifications and Examinations Regulation (pp. 459–522).

Johnson, S. and Johnson, R. (2012b) 'Component reliability in GCSE and GCE', in Q. He and D. Opposs (eds), *Ofqual's reliability compendium*, Coventry, U.K.: Office of Qualifications and Examinations Regulation (pp. 67–90).

Law, J., McBean, K. and Rush, R. (2011) 'Communication skills in a population of primary school-aged children raised in an area of pronounced social disadvantage', *International Journal of Language and Communication Disorders*, 46 (6), pp. 657–664.

Linacre, J. M. (undated) *Facets. Many-facet Rasch analysis* [software]: http://www.winsteps.com/facets.htm

Maxwell, B., Burnett, C., Reid, J., Willis, B. and Demack, S. (2015) 'Oracy curriculum, culture and assessment toolkit: evaluation report and executive summary', London: Education Endowment Foundation.

Newton, P. E., Stobart, G., Goldstein, H., Harlen, W., Baird, J-A. and Winter, J. (2008) 'Return of the bible', *Assessment in Education: Principles, Policy and Practice*, 15 (3), pp. 307–320.

Nordberg, D. (2006) 'Fairness in assessing group projects: a conceptual framework for Higher Education': http://dx.doi.org/10.2139/ssrn.873605

O'Neill, G. (2013) 'Assessing group work (including online)': https://www.ucd.ie/t4cms/UCDTLE0065.pdf.

Rasch, G. (1960/1980) *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago, IL: University of Chicago Press.

School 21 (2016) 'Baseline assessment guide' (unpublished document).

Sibbald, B. and Roland, M. (1998) 'Understanding controlled trials: Why are randomised controlled trials important?', *British Medical Journal*: http://www.bmj.com/content/316/7126/201

Standards and Testing Agency (2016) *Key Stage 2: submitting teacher assessment data*: https://www.gov.uk/government/publications/key-stage-2-submitting-teacher-assessment-data.

Vickers, A. J. and Altman, D. G. (2001) 'Statistics notes: analysing controlled trials with baseline and follow up measurements', *British Medical Journal (Clinical research edn)*, 323 (7321), pp. 1123–1124.

Appendix: Memorandum of Understanding

A copy of the Memorandum of Understanding which all pilot schools signed up to is presented below.

Thank you for participating in the Voice 21 EEF Pilot. The project is led by Voice 21, part of the 21 Trust and funded by the Education Endowment Foundation.

This agreement outlines the responsibilities of Voice 21 and the Pilot Schools and the scope, nature and requirements of the evaluation (to be conducted by Alpha*Plus* Consultancy).

Purpose of the Pilot

In this pilot the 21 Trust will develop an oracy training package for other schools and the feasibility of the Voice 21 oracy approach will be tested in 12 other schools, which may have different challenges when implementing the approach. The pilot will also look for evidence that the intervention is likely to impact on academic attainment, including observing changes to teaching practice and measuring oracy improvements, as well as assessing the interventions readiness to be trailed as part of a large-scale randomised controlled trial.

Responsibilities

Voice 21 will:

- Conduct pre-delivery meeting at each school to agree expectations
- Deliver a two-day training programme for oracy leads and a member of SLT at School 21
- Provide bank of resources and materials to support oracy teaching
- Provide guidance for the delivery of core components
- Deliver 'in-school training in each partner school
- Provide assessment materials and guidance for in-school training and standardisation
- Provide ongoing training and support to each school throughout the duration of the project
- Provide regular dissemination updates for the partner schools throughout the project
- Be the first point of contact for any questions about the project
- Pay accommodation costs for schools travelling from outside of London

Pilot Schools will:

- Release relevant members of staff to attend the September training day and other additional training days as needed
- Schedule school-based CPD on oracy to be delivered by Voice 21
- Deliver weekly oracy lessons to Y7 cohort
- Encourage oracy-based teaching across all subjects
- Trial regular oracy-based assemblies
- Deliver the Ignite talks programme culminating in the performance of a speech without notes by year 7 pupils
- Conduct Baseline Assessments and Post Intervention Assessment of sample (to be agreed by evaluation team) of Y7 pupils
- Support the collection of data at the beginning and end of the evaluation

Data and Evaluation

The evaluation is being conducted by the Alpha*Plus* on behalf of the Education Endowment Foundation.

- The pilot school agrees to share relevant data with the external evaluators (Alpha*Plus* Consultancy) for evaluating the oracy pilot, to the extent that such information sharing is permitted by the Data Protection Act 1998. The external evaluators shall fully comply with all applicable laws and regulations relating to the processing or protection of any personal data including, but not limited to, the seventh data protection principle set out in the Data Protection Act 1998.
- The types of data the external evaluator are expecting to request from the pilot school include pupil oracy assessment data matched with specific pupil background/demographic data. It is assumed that the pilot school has appropriate permissions in place to share such pupil data with the evaluators, i. e. either has gained explicit parental consent or the school makes the choice to share such data *in loco parentis*.
- School leaders, the oracy lead and teachers at the pilot school will be expected to participate in essential evaluation activities, including semi-structured interviews with the external evaluators. The oracy lead will also be asked to complete a short online questionnaire during the training/ planning period and again at the end of the school year. Teachers may be requested to provide schemes of work/ lesson plans for analysis by the external evaluators. The evaluators will also require access to the oracy assessment pre- and post-intervention attainment outcomes.
- Neither the school nor any individuals will be identified in any reports or other publications arising from the oracy pilot evaluation. The information collected will be used for research and evaluation purposes only and no information that can identify individuals will be used for any other purpose without the explicit permission of the individual/s concerned. Any personal data collected will be destroyed by the external evaluators in accordance with the Data Protection Act 1998 when it is no longer required.

We commit to participating in Voice 21 Oracy project as detailed above:

School name: _____

Signature:_____ Date:_____

Email address:_____

Thank you for agreeing to take part in this project. Please return this form as soon as possible by email to: **Lizzie Lynch: Programme Officer, Voice 21.lizzie@voice21.org**

Or by post: Voice 21, School 21, Pitchford St, London E15 4RZ

This document is available for download at **www.educationendowmentfoundation.org.uk**

Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk