The Brilliant Club

Research Report Series

Assessing Pupils' Academic Writing Skills Using a Standardised Mark Scheme

October 2017

About The Brilliant Club

The Brilliant Club exists to increase the number of pupils from under-represented backgrounds that progress to highly-selective universities. We do this by mobilising the PhD community to share its expertise with state schools. In pursuit of this mission, The Brilliant Club delivers two programmes:

The Scholars Programme recruits PhD researchers, trains them as university access professionals and places them as tutors in schools to deliver academically rigorous programmes to small groups of high-potential pupils.

Researchers in Schools recruits PhD graduates, places them as trainee teachers in schools and supports them to develop as excellent teachers and research leaders committed to closing the gap in attainment and university access.

The **Research Report Series** forms part of our Research and Impact Series, which provides three ways to engage with our work and that of our partners. Please click on the icons below to find out more:

Research Seminar Series

Impact Case Study Series

Research Report Series

About the Authors

This research project was designed and conducted by Dr Celeste Cheung and Dr Lauren Bellaera from The Brilliant Club's Research and Impact Department.

Dr Celeste Cheung is the Research and Evaluation Manager at The Brilliant Club. She manages the charity's evaluation and innovation projects to evaluate pupil performance, programme measures and innovations. She completed her PhD in Developmental Cognitive Neuroscience at King's College London, her previous research focused on attentional processes and self-regulation in autism and attention-deficit hyperactivity disorder (ADHD).

Dr Lauren Bellaera is the Research and Impact Director at The Brilliant Club. She is responsible for evaluating the impact of the charity's programmes on pupil outcomes. Prior to this, Lauren worked at the University of Cambridge where she evaluated the impact of a set of online educational resources on university students' conceptual understanding and critical thinking skills. Lauren is trained as a cognitive psychologist and has worked with a number of educational organisations, including IGGY and Macat.

Contact Details

Dr Celeste Cheung, Research and Evaluation Manager
Email: celeste.cheung@thebrilliantclub.org

Contents

Executive Summary

This report details the methodology and findings from one of our internal research projects that evaluated pupils' academic writing skills. We hope this report will be useful for others in widening participation, who are seeking a systematic approach to evaluate their own programmes.

The project

At The Brilliant Club, we want to understand if our programmes are helping pupils to develop three key academic skills that are important for progressing to and succeeding in higher education: written communication, critical thinking and subject knowledge (hereafter, academic writing skills). This research project aimed to evaluate the impact of one of our programmes – The Scholars Programme – on pupils' academic writing skills and to test the feasibility of applying a standardised mark scheme to assess programme performance. The project involved 83 Key Stage 3 and 4 pupils from nine schools. Pupils received six university-style tutorials delivered by five doctoral researchers (PhD tutors), and completed written assignments at the start and end of the programme. Assignments were marked by PhD tutors using a standardised mark scheme.

Key findings

- Overall, pupils academic writing scores improved by 18.6%. The increase in performance was uniform across all aspects of academic writing that we measured.
- The positive effect of The Scholars Programme did not vary by gender or Ever6FSM status. However, the improvement was more pronounced for pupils in the STEM subject stream compared to pupils in the Arts/Humanities subject stream.
- Qualitative feedback from PhD tutors suggested that it is feasible to apply a mark scheme with a standardised structure across Key Stages for a range of subject areas.

Recommendations for the wider sector

A standardised mark scheme with clearly defined outcomes that are explicitly communicated to pupils can be an effective way of measuring pupils' progress. Our results indicate that the same assessment tool can work well to measure academic writing skills across broad subject areas. Given the complex and variable nature of school and pupil characteristics, evaluation can be made more rigorous by being systematic (e.g. using pre-post designs and controlling for extraneous variables), and by being aware of the varying sources of influence (e.g. reliability of the marking process and unconscious bias).

Recommendations for The Scholars Programme

PhD tutors should be trained to use the mark scheme consistently and effectively. The average score from the mark scheme can be used for reporting on a group level (e.g. impact reports for schools). However, subscale scores (which break down academic writing into its constituent skills) should be used to guide individual feedback on which aspects of academic writing skills pupils can improve. Further research with a larger sample is necessary to draw robust conclusions about whether the impact of our programme varies by pupil characteristics or subject streams. Follow-up evaluation work should be conducted to further validate the standardised mark scheme on a larger scale.

1. Introduction

1.1 Writing skills in higher education and beyond

The ability to convey and exchange knowledge, ideas and information effectively is crucial for many aspects of our lives. Young children use oral language as the primary medium to communicate and learn. Writing skills become increasingly important as children enter the education system, as writing is a core medium through which knowledge of the curriculum is demonstrated and assessed.

Effective writing is key to success in higher education. In the UK, the Organisation for Economic Co-operation and Development (OECD) considered written communication as a generic skill that is expected of all students in higher education (OECD, 2012). Similarly in the US, written communication was a skill that academics from top US institutions most frequently reported as critical to both academic and career success (Educational Testing Service, 2013).

Beyond higher education, writing skills are also emphasised across the workplace. In a survey of 431 employers in various industries and workforces, 93% identified this skill as being "very important" (Casner-Lotto & Barrington, 2006).

1.2 Multiple components of academic writing skills

A detailed review by Sparks and colleagues examined a range of definitions for effective writing in a higher education context (Sparks, Song, Brantley, & Liu, 2014). Similar concepts emerged across different frameworks: 1) Analysis – the ability to analyse and act on understanding of audiences, context and purposes of the text; 2) Critical evaluation - the ability to make thoughtful decisions and evaluate the source's credibility, bias and reasoning; 3) Language and style – the ability to convey information appropriately, clearly using appropriate functional grammar, language and style; 4) Developing an argument – the ability to formulate one's argument in a convincing manner taking account of various viewpoints; 5) Research process – the ability to conduct research and use a range of reliably sourced evidence to support claims and ideas.

A deep level of subject knowledge has also been considered as key for effective writing (Hayes & Flower, 1980). It was argued that writing is a complex interaction between the task environment (e.g. topics, features of writing assignment and context or purpose), the writer's knowledge of the topic and the audience (e.g. general or specific audience), and the writing process (e.g. planning and reviewing). The former two components involve knowledge acquisition to understand the nature of the task and the topic of interest, while the writing process requires self-regulation and is goal-directed (Hayes & Flower, 1980).

Academic writing is a multifaceted construct that encompasses a full range of skills and requires the ability to demonstrate a deep understanding of a topic, thorough relevant research and critical analysis of the content, as well as stylistic techniques. This definition forms the basis of the framework that we use in our work to assess pupils' academic writing.

1.3 Assessing academic writing skills

The definition of academic writing skills and its underlying construct is particularly important for providing a consistent framework on which assessments can be based. In England, for example, writing skills are assessed against the expected standards in primary school pupils (Standards and Testing Agency, 2016).

In secondary education, written examinations are typically assessed using a standardised mark scheme. The assessment objectives for most subjects are structured using Bloom's original taxonomy of cognitive thinking (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), which includes six main categories: 1) knowledge; 2) comprehension; 3) application; 4) analysis, 5) synthesis; and 6) evaluation. For example, to obtain the highest level in GCSE English Writing, pupils are expected to "use an extensive vocabulary strategically with rare spelling errors" (Pearson, 2017), "communicate clearly, effectively and imaginatively" and "organise information and ideas using structure....and structural features to support cohesion and coherence" (AQA, 2017). A revised version of the Bloom's taxonomy has since been published, which includes elements of creativity – "putting elements together to form a coherent whole to make an original project" (Anderson, Krathwohl, & Bloom, 2001). Similarly, written assessments in higher education are typically graded against standardised criteria informed by Bloom's framework, with the highest grades awarded for evidence of critical analysis and the ability to synthesise issues or ideas in an original way.

Designing the assessment framework

In higher education, the assessment tools are typically used to either support institutional or individual goals, this may influence how the mark scheme is constructed and how the scores from the assessments are used (Sparks et al., 2014). For example, institutions may run group-level analyses to examine overall student performance, to make improvements to the curriculum and teaching instructions. Writing assessments may also be intended to support individual learners via student feedback. In this case, a more granular level of outcomes with subscale scores is useful for providing specific and structured feedback. The diagnostic information that subscale scores provide can have added value beyond overall scores, supporting teaching and learning for both institutions and individual learners (Haberman, 2008).

A close alignment between the teaching instructions and assessment, which are clear and explicitly communicated to the students, are important factors for effective writing interventions (Haswell, 2008). The assessment results should therefore inform institutions about which aspects of writing are the most challenging for pupils, these can then be addressed through further teaching instructions.

The most valid measures of writing ability are those that require pupils to write extended text, where pupils are given the opportunity to demonstrate the full range of skills (Fowles, 2012). Most writing assessments do not monitor the planning or revision processes of writing, as examinees respond to a single prompt in a limited amount of time. Designing assessment tools that capture the planning and reviewing process of pupils' writing has been recommended as a useful way to scaffold learning (Sparks et al., 2014).

1.4 How does effective academic writing relate to The Brilliant Club's programmes?

Our five-year strategy *The Path to Outcomes* outlines six competencies (Table 1) that have been shown to be important for progressing to highly-selective universities, and are used to demonstrate our shorter-term impact of our programmes at The Brilliant Club. This report focuses on three of these competencies that can be assessed through written assessments (written communication, critical thinking and subject knowledge). Throughout this report, we refer to these three competencies, collectively, as 'academic writing skills'. For both The Scholars Programme and Researchers in Schools, these competencies are core skills that we aim to help pupils to develop through a series of university-style tutorials.

Table 1. Six competencies for assessing our programme impact

Competencies	Definitions
Written and Verbal Communication	The ability to use written information or spoken language clearly and appropriately to convey ideas.
Subject Knowledge	Having a deep-level of understanding on the topics studied in The Scholars Programme and Researchers in Schools.
University Knowledge	New knowledge about university options and how to successfully apply to study at university in Year 13.
Motivation and Self-Efficacy	Motivation refers to what causes an individual to want to do one thing and not another (intrinsic motivation). Self-efficacy measures pupils' belief in their ability to achieve future goals or influence future situations.
Meta-Cognition	The ability to think explicitly about one's own learning.
Critical Thinking	The ability to analyse and evaluate a subject objectively to form a judgement.

In the original version of The Scholars Programme (delivered prior to September 2017), all pupils were asked to complete a baseline assignment following their first tutorial, which was used as a proxy to assess pupils' general performance at the beginning of the programme. They were also required to complete an extended written assignment at the end of the programme (hereafter, final assignment), which was marked, moderated and fed back to the pupil by their PhD tutors. The format and quality of the baseline assignment was not standardised in the original version of the programme, resulting in some variation in the type of baseline assignments across courses – ranging from short-answer questions to open-ended questions. The baseline assignments were set as homework but these were not systematically used to monitor and evaluate pupils' progress. The final assignments were marked by PhD tutors using a mark scheme that they designed, based on the skills and knowledge that they aimed to help their pupils to develop. PhD tutors were given some guidance on how to structure the mark scheme, but the content and structure of the mark schemes were not standardised between PhD tutors.

As part of a planned programme re-design launched in September 2017, we implemented a new assessment feature to The Scholars Programme - using a standardised mark scheme to systematically assess the baseline and final assignment. The standardised mark scheme approach was chosen as it is representative of how universities typically assess student performance. The design of the mark scheme was informed by academic literature (e.g. Sparks et al., 2014), which recommends the inclusion of subscale scores to enable PhD tutors to delineate pupils' performance on different aspects of academic writing. The content of the mark scheme

aligns with the competencies that we aim to help our pupils to develop (written communication, subject knowledge and critical thinking), and are generic academic skills relevant to all subject areas.

The efficacy and feasibility of this assessment approach was first evaluated, and the results were then fed directly into the implementation of the re-designed programme. The subsequent sections of this report outline the aims, methodology and findings from this evaluation project.

1.5 Project aims

The aims of the evaluation project were twofold: 1) Outcome evaluation: to evaluate the impact of The Scholars Programme on pupils' academic writing skills. 2) Process evaluation: to test the feasibility of using a standardised mark scheme in the context of The Scholars Programme.

2. Methodology

2.1 Participants

We selected a sample of Key Stage 3 (age 11-14) and Key Stage 4 (age 14-16) pupils from non-selective state schools in England, who were taking part in The Scholars Programme at The Brilliant Club in the Spring term of 2017. We selected a series of The Scholars Programme course handbooks where the format of the baseline assignments and final assignments were the same (i.e. open-ended essay format). Course handbooks contain key programme materials (e.g. assignment titles, marking criteria, relevant reading and homework assignments). The structure of the course handbooks is similar across subject areas, but the content of the course handbooks differs. After selecting the handbooks that fitted our criteria, we contacted the corresponding PhD tutors and schools to invite them to participate in this project. A total of 105 pupils and five PhD tutors in nine schools were invited to take part in the project (Table 2).

Table 2. Descriptive statistics of pupil demographics from each school

Schools	Tutor	Pupils recruited (n)	Pupils Ever6FSM (%)	Year 7 (%)	Year 8 (%)	Year 9 (%)	Year 10 (%)
A	AA	12	42	0	0	50	50
B	AA	10	90	0	0	0	100
C	BB	15	33	0	0	100	0
D	BB	12	33	25	33	8	33
E	CC	10	30	0	0	60	40
F	CC	11	36	0	0	0	100
G	DD	11	82	0	0	45	55
H	DD	12	50	0	0	0	100
I	EE	12	33	100	0	0	0
Total/average	5	105	47.67	14	4	29	53

Four PhD tutors (AA-DD) taught two placements at different schools and one tutor (EE) taught one placement at only one school. The percentage of pupils who had been eligible for free school meals during the previous six years (Ever6FSM), the proportion of pupils from different year groups and the submission rate of the baseline and final assignment from each school is presented in Table 2.

2.2 Design

Pre-post-test design without a control group

This evaluation project followed a pre-post-test design (Figure 1). Pupils participating in the original version of The Scholars Programme received six university-style tutorials, taught by a PhD or postdoctoral researcher on a topic of their expertise. At the end of their first tutorial, pupils were set a baseline homework assignment (pre-test), which consisted of an open-ended essay question between 300-500 words. The content of the tutorials aimed to help pupils develop their academic writing skills, which was then assessed through the final assignment (post-test) – a 1500 (Key Stage 3) or 2000 (Key Stage 4)-word essay. Examples of the baseline and final assignment question titles used in this project can be found in Appendix A.

Figure 1. A schematic of the research design

Marking the assignments

Pupils submitted the baseline assignment online through The Brilliant Club's Virtual Learning Environment (VLE). PhD tutors were given the standardised mark scheme (Appendix B) and were asked to use this to mark the submitted essays, and to record each pupils' marks using the mark sheet we provided (Appendix C). The same process was carried out for the final assignment. PhD tutors were also asked to give an overall grade for each final assignment script which was then moderated (see below for moderation process). PhD tutors were given approximately two weeks from the date of their second tutorial to mark all the essays using the mark scheme. Any assignments submitted a week after the deadline were deemed late and were not included in the following analysis.

Corresponding to a mark scheme commonly used in universities in the UK, a score below 40 was considered a fail, between 40-49 was equivalent to a third, between 50-59 was equivalent to a lower second-class mark (2:2), 60-69 was equivalent to a higher second-class mark (2:1) and 70+ was equivalent to a first-class mark.

Moderating the assignments

All PhD tutors were given training on how to mark and moderate the assignments before the first tutorial. For each placement, two out of 12 final assignments were second marked (blind) by

another PhD Tutor to give an 'impression grade' (moderated mark), this was then compared with the PhD tutor's initial mark. If there was a significant discrepancy between the marks, scripts were moderated for the third time by a staff member at The Brilliant Club (ex-teacher) before applying a rule if necessary. For example, if the PhD Tutor's initial mark was 52 and the moderated mark was a low 2:2, then no rule was applied and the final moderated mark would be 52. However, in the case that the moderated mark was a high 2:2, then a rule of +3 was applied to all the final assignments on the placement to give the final moderated mark. If the moderated mark and the PhD Tutor's initial mark was different by more than a grade (e.g. mid-high 2:1 vs 52), then a rule of +10 was applied. The same rule would then be applied to the baseline and final assignment subscale marks, on the basis that the PhD Tutor must have generally 'mis-calibrated' their marking. The marks were only adjusted if there was a large discrepancy between the PhD tutors' initial marks and the moderated marks (e.g. low 2:2 vs high 2:2 or high 2:2 vs low 2:1).

2.3 Measures

Academic writing skills were assessed through a baseline and final assignment using a standardised mark scheme. The constructs and content of the mark scheme were informed by academic literature and frameworks outlined by Spark et al (2014). The mark scheme included six skills (subscales) that are critical for academic writing at undergraduate levels (addressing the question, developing and argument, use of evidence, critical evaluation, structuring, and use of language), each graded between 0 and 100. As our programme aims to support the development of these specific skills across all age groups, we applied the same mark scheme to pupils in both Key Stages 3 and 4. PhD tutors were asked to provide a mark for each subscale and an overall average. The ranges of scores used to assign university-style class marks is explained in Section 2.2 above.

The outcome variables used to assess academic writing skills were the subscale and average scores from the baseline and final assignments and the difference between the average baseline and final assignment scores, hereafter, the change scores.

2.4 Analysis

Data cleaning process

Of the 105 pupils recruited, we received baseline data from 83 (79%) pupils and final assignment data from 61 (58%) pupils. Final assignment data from school G and school H were missing as one tutor was unable to complete the marking sheet due to time constraints. Of the 83 pupils who submitted the baseline assignment, data from three pupils were excluded due to late submission; all three from School B. No exclusions were made to final assignment submission.

We compared the PhD Tutor's initial marks with the moderated marks to check for anomalies and to flag up discrepancies between the two marks. Of the twelve scripts that were moderated, a moderation rule of -10 was applied to two scripts indicating that two PhD tutors (Tutors CC and EE) were overly generous with their marks (Appendix D). As such, we deducted 10 marks from each of the subscale scores from these pupils for the final analysis, for both the baseline and final assignments. We also re-ran the main analysis excluding these scores where a moderation rule was applied, to ensure that the results from the main analysis were not due to variability in the moderation process.

Analysis steps

The full sample was included in the analyses on the baseline (n=83) or final assignment (n=61) separately (Section 3.1.3). However, all analyses that assessed the change scores between baseline and final assignment included only pupils who had completed both the baseline and final assignment (n=56) (Sections 3.1, 3,1.1, 3.1.2).

The baseline and final assignment average scores and the change scores were all normally distributed. Paired sample *t*-tests were used to analyse the within-subject change scores from baseline to final assignment for each subscale score and the overall average score.

To examine whether the effect of The Scholars Programme on academic writing skills can be explained by other factors including gender, Ever6FSM and subject stream (Arts/Humanities or STEM), we ran a multiple regression using the average change score as the dependent variable and the extraneous variables as independent variables.

Finally, we analysed PhD tutors' qualitative feedback to confirm the feasibility of including the standardised mark scheme in the context of The Scholars Programme.

3. Results

3.1 The impact of The Scholars Programme on pupils' academic writing skills

Across the whole sample, pupils showed an 18.6% (11 percentage points) increase in average scores from baseline (mean score of 59) to final assignment (mean score of 70), which was statistically significant (t (55) = 8.11, $p<0.01$).

An increase in average academic writing score was observed uniformly across all six subscales of academic writing skills, with the biggest increase found in critical evaluation (23.3%) and the smallest increase in use of language (15.4%). Paired sample *t*-tests indicated statistically significant shifts from baseline to final assignment for all six subscales and for the overall average (Figure 2, Table 3).

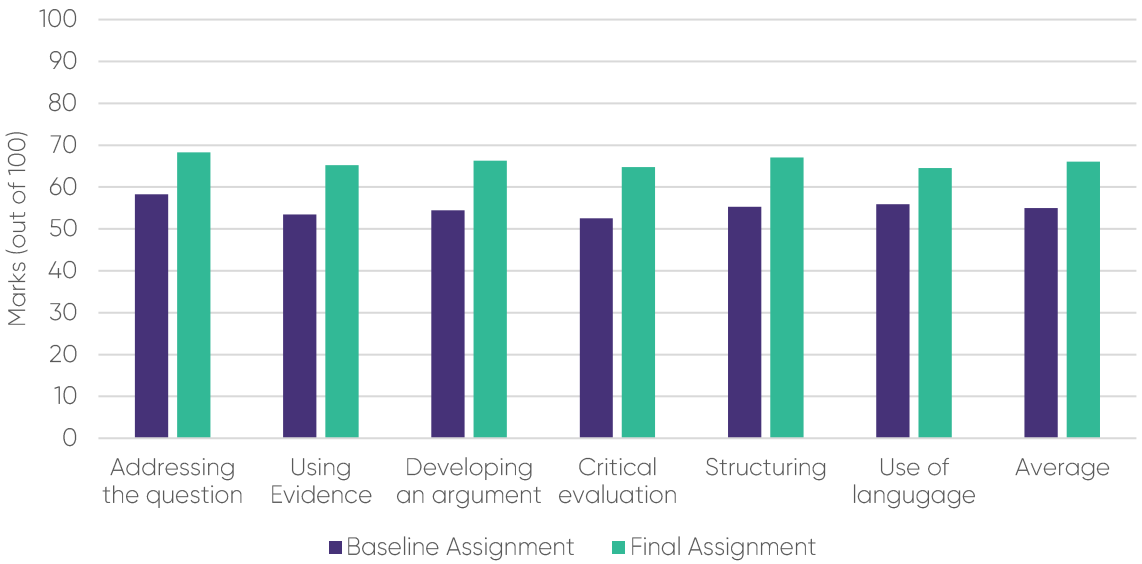

Figure 2. Scores on the six subscales of academic writing skills and the overall average from the baseline and final assignment (n=56)

Table 3. Average scores (and standard deviation) of the baseline and final assignment marks, and change scores (final assignment – baseline assignment marks)

Academic Writing Skills	Baseline Assignment Marks	Final Assignment Marks	Change Scores	*t*-values	*p*-values
Addressing the Question	58.26 (9.76)	68.29 (8.30)	10.02 (10.54)	7.12	<0.01
Using Evidence	53.48 (9.49)	65.21 (10.48)	11.73 (10.96)	8.01	<0.01
Developing an Argument	54.43 (10.00)	66.27 (11.18)	11.84 (12.06)	7.35	<0.01
Critical Evaluation	52.55 (10.39)	64.79 (11.31)	12.23 (11.93)	7.68	<0.01
Structuring	55.32 (9.39)	67.07 (9.32)	11.75 (12.33)	7.13	<0.01
Use of Language	55.94 (1.04)	64.57 (7.75)	8.64 (11.22)	5.76	<0.01
Average	54.99 (8.13)	66.03 (9.34)	11.04 (10.18)	8.11	<0.01

3.1.1 Can the positive effect of The Scholars Programme on academic writing skills be explained by the moderation process?

We applied the same moderation rule to the baseline assignment marks due to limited resources to second-mark the baseline assignment scripts, based on the assumption that even though a PhD Tutor may be overly generous or strict in their marking they do so consistently.

To ensure that the interpretation of our outcome evaluation still holds in the case that this assumption is violated, we performed an additional analysis step now excluding data from pupils taught by Tutors CC and EE. The results remained unchanged: data from the 35 pupils taught by Tutors AA, BB and DD indicated a significant increase in average scores from the baseline (mean=56.39, S.D.=7.43) to final assignment (mean=64.24, S.D.=8.20; t (34) = 6.54, p<.01).

3.1.2 Does pupils' progress in academic writing skills vary by gender, Ever6FSM status and subject stream?

The subject stream of the course (STEM or Arts/Humanities) had a significant effect on the change score (t (55) = 3.89, p<0.01), but this was not significant for pupils' gender or Ever6FSM status. Table 4 displays the average and change scores split by gender and Ever6FSM status. Although non- Ever6FSM pupils had higher scores in both baseline and final assignments, and showed greater change scores, this difference was not statistically significant (p=0.15).

Table 4. Means and standard deviations of the baseline and final assignments and the change scores for boys and girls, Ever6FSM and non- Ever6FSM pupils

Outcome Measures	Boys	Girls	Ever6FSM	Non- Ever6FSM
Baseline Assignment Marks	n=26 61.17 (8.43)	n =38 60.15 (6.80)	n=35 58.56 (8.07)	n=48 61.34 (5.93)
Final Assignment Marks	n =15 67.18 (11.35)	n =34 67.81 (10.06)	n=24 64.56 (12.92)	n=37 72.74 (9.28)
Change Scores	n =15 8.10 (8.75)	n =30 9.29 (8.96)	n=20 8.38 (10.67)	n=36 12.51 (9.73)

Detailed breakdown of scores split by subject streams, schools and PhD tutors is presented in Appendix D. Pupils who took part in the STEM programme had significantly higher levels of change in academic writing skills (n=16, mean=21.76, S.D. = 9.87) compared to pupils who completed the Arts/Humanities programme (n=40, mean = 6.75, S.D. = 6.52).

To confirm that the significant increase in academic writing score across the whole cohort was not driven only by pupils on the STEM course, we repeated the original analysis removing the STEM pupils. The results remained the same: a significant increase in academic writing skills from baseline to final assignment (t (39) = 6.54, $p<.01$).

3.1.3 What is the relationship between the subscale and average scores?

The relationship between the six subscale scores at both baseline and final assignments were explored to understand how different components of academic writing skills related to each other in this sample. Overall, the associations were stronger in the final assignment than in the baseline assignment, with correlations ranging between 0.85 and 0.95, and between 0.43 and 0.85, respectively (Table 5). The correlations between baseline and final assignment for each of the six subscales were similar in magnitude (ranged between 0.35 to 0.50). The baseline and final assignment average scores were correlated at 0.50, and this was similar for all four PhD tutors who had submitted data for both baseline and final assignments (Figure 3). For one PhD Tutor (Tutor BB) this association was stronger, as this was driven by an outlier with low scores for both baseline and final assignments.

Table 5. Pearson's r correlations of the subscale and average scores in the baseline and final assignment.

BA \ FA	Addressing the Question	Using Evidence	Developing an Argument	Critical Evaluation	Structuring	Use of Language	FA Average
Addressing the Question	.47	.83	.85	.87	.90	.87	.93
Using Evidence	.52	.47	.88	.88	.87	.82	.93
Developing an Argument	.68	.71	.46	.95	.92	.89	.96
Critical Evaluation	.64	.69	.85	.50	.93	.87	.97
Structuring	.60	.53	.79	.75	.35	.93	.97
Use of Language	.52	.43	.68	.71	.71	.42	.94
BA average	.79	.77	.93	.92	.86	.79	.50

Note: Along the diagonal in white are the correlations of each category in the baseline (BA) and final (FA) assignment. Below the diagonal in light grey are the correlations between subscales at baseline and above the diagonal in dark grey is the correlations between subscales in the final assignment.

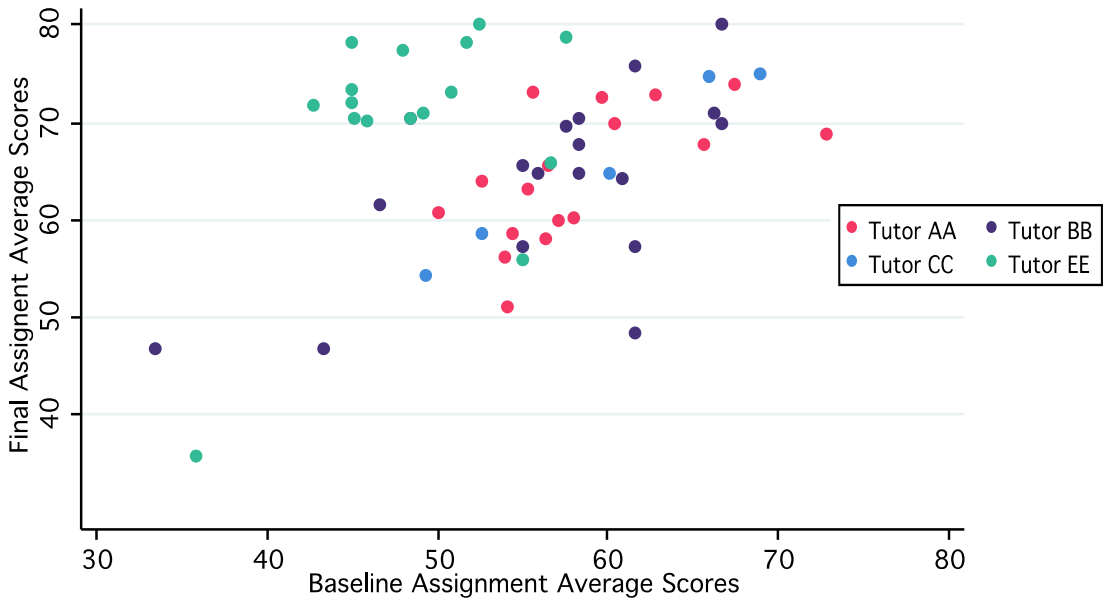

Figure 3. The relationship between baseline and final assignment split by PhD tutors

3.2 Feasibility of using a standardised mark scheme

Qualitative feedback from the PhD tutors about the mark schemes was generally positive. There was consensus on its use in articulating and describing the skills that pupils should aim to achieve. One PhD Tutor found that the mark scheme was *"interesting, and brought home how far some seemed to have developed their style."* This PhD Tutor also felt that the baseline assignment was a helpful and simple exercise to *"kick off the course and set everyone at their ease before we dived into more complex ideas later"*. However, the PhD Tutor found it challenging to apply the standardised mark scheme to the baseline assignment, which was *"very brief and introductory without much analysis involved in the process."* (Tutor AA).

Another PhD Tutor found the mark scheme helpful but also described some challenges with balancing the precision and validity of the marking process: *"the marking scheme was really helpful in marking and writing feedback for the students. However, I found it a bit tricky to put exact marks onto the marks scoring sheets. Indeed, it is always hard to give the exact percentage. Perhaps it could be easier if we could mark them high/low 2.1, 2.2. etc"*. (Tutor BB). There were some questions about the validity of the mark scheme, particularly for some aspects of pupils' abilities and performance that it was not able to capture. For example, a PhD Tutor indicated to have given a higher mark to pupils who have *"written more since these are usually the pupils who have put in more work, thought and care"*.

In the follow-up interview with the PhD tutors, they were asked if in future they would prefer to either use a standardised mark scheme that is not editable, or a mark scheme with standard structure but with the option to modify some elements of the specific skills. The PhD tutors expressed a consistent response, with a preference for the latter option to allow flexibility to assess skills specific to each subject area. Feedback from PhD tutors was used to inform the implementation of the re-designed programme launched in September 2017.

4. Discussion and Recommendations

4.1 Impact and feasibility

The project assessed the impact of The Scholars Programme on a range of academic writing skills and examined the feasibility of applying a standardised mark scheme to assess pupils' progress in academic writing. Our results demonstrate that The Scholars Programme has a positive impact on pupils' academic writing skills, and that it is feasible to use a standardised mark scheme in the context of The Scholars Programme. The PhD tutors found the standardised mark scheme helpful in identifying their pupils' strengths and weaknesses, and in guiding individualised feedback. However, they expressed a preference for a mark scheme with some degree of flexibility to allow them to assess skills specific to their subject area.

Recommendations
The mark scheme should aim to assess the same academic skills across all subject areas, but PhD tutors should be allowed some flexibility to edit the content to make it relevant for their course.

4.2 Subscale and average scores

The use of subscale scores is beneficial for three reasons: 1) to increase understanding of which aspect of pupils' writing skills the programmes have the biggest impact on; 2) to help develop and improve the course curriculum to maximise impact for different areas of academic writing; and 3) to inform PhD tutors, pupils and teachers which areas of academic writing their pupils require the most support. Our data indicate that on a group level, all six subscales of academic writing skills showed similar levels of positive increase between baseline and final assignment, no significant difference was observed between any of the subscales for either the baseline or final assignment. Therefore, the data suggest that at least for the selected groups of pupils, The Scholars Programme had a similar degree of positive impact on all areas of academic writing skills.

The correlations between subscales show that some subscales are more closely related than others. For example, in the baseline assignment, the correlation between 'developing an argument' and 'critical evaluation' was twice as high as the correlation between 'use of language' and 'using evidence'; while in the final assignment, all six subscales were strongly correlated to one another. This indicates two things: 1) developing an argument and critical evaluation may tap into similar cognitive constructs and are skills that are likely to develop together; 2) at the start of The Scholars Programme, pupils who are good at language and style in writing may not necessarily also be good at using a wide range of evidence.

The baseline and final assignment average scores were strongly correlated with all subscales of academic writing skills, indicating that the overall average score captures the variance in each individual subscale. Thus, it is appropriate to use the baseline and final assignment average scores as a standalone score to assess academic writing skills on a group level (i.e. assessing and reporting on the overall impact of a programme). However, to understand individual pupils' progress, subscale scores should be provided alongside the average score to understand separate constructs of academic writing (e.g. written communication, critical thinking and subject knowledge).

Recommendations
Baseline and final assignment average scores are appropriate for group-level impact reporting.
Subscale scores should be used for feedback to pupils and to inform teaching and learning.

4.3 Moderators of impact

We conducted some additional analyses to further understand whether the impact of the programme is moderated by pupil characteristics, or if the degree of impact varied under different contexts. For example, does the programme have greater impact on girls' writing skills than boys? Is the programme more beneficial for Ever6FSM pupils?

Our preliminary results indicated that the degree of improvement in academic writing skills in the context of The Scholars' Programme was not moderated by gender or Ever6FSM status. However, as the sample size for these analyses were small, this could be due to insufficient statistical power to detect group differences.

Our results revealed a higher proportion of pupils achieving a 1st in STEM subjects compared to Arts/Humanities subjects. Interestingly, this effect of subject stream was not observed in the baseline assignment, leading to a noticeably larger degree of improvement between the baseline and final assignment. As most STEM courses did not have open-ended questions already in place as a baseline assignment, we were restricted to selecting only one STEM PhD Tutor for this project. Therefore, we were unable to verify whether the differential effect of subject stream was a true effect, or was due to variability in the marking process.

Recommendations
Further evaluation work that includes a range of STEM courses is necessary to understand and confirm the subject stream effect.
The degree to which the impact of The Scholars Programme varies by pupils' characteristics (e.g. age, gender and Ever6FSM) should be investigated with a larger sample in the future.

4.4 Challenges and future evaluation work

Results from this project indicate that, overall, pupils show marked positive improvements in all areas of academic writing skills from the beginning to the end of The Scholars Programme. However, it is crucial to interpret the data with context and caveats. This project used a small sample of pupils and PhD tutors on a selective range of courses. The results provide a useful framework for implementation, but more detailed evaluation work on a larger-scale is needed to accurately quantify the impact that The Scholars Programme has on pupils' academic writing skills. The specific challenges and recommendations for future evaluation and implementation of The Scholars Programme are outlined below.

Reliability of the assessment scores

On a group level, pupils showed an increase in final assignment scores – a pattern that is consistent across PhD tutors and across schools, which shows overall support for the positive impact of the programme. Detailed individual-level analyses to test inter-rater reliability or internal consistencies of the assessment scores would have been informative for ensuring an accurate interpretation of subscale scores. However, it was beyond the scope of this project to conduct such detailed analyses, as this would require multiple markers to assess the same piece of work, or the same marker to assess the same piece of work across multiple timepoints.

Validity of the assessment scores

The subscales and content of the mark scheme used in this study was designed based on existing cognitive theories and frameworks as outlined in Sparks et al (2014), therefore in theory it should be an accurate reflection of pupils' ability. However, the validity and reliability of the assessment scores could also be affected due to differences in PhD tutors' interpretation of the mark scheme. This is a well-known problem with using standardised marking criteria and has been discussed extensively elsewhere (see Christodoulou, 2016). Two approaches have been proposed to address this problem, the first is the comparative judgement method in which pupils' work is assessed by an iterative and adaptive algorithm where judges compare pairs of students and choose the one which they think is better, without reference to criteria. The second method is to have benchmark examples to create a common language of standards.

Specificity of impact

An obvious caveat of this project is the lack of a control group. Therefore, any change in academic writing skills could also be due to factors not specific to the programme that were not measured (e.g. maturity, school curriculum). It was not possible for this project to include a control group, as this would have involved tracking a group of pupils who were not taking part in The Scholars Programme.

Unconscious Bias

The written assignments were not anonymous, and were marked by PhD tutors who knew the pupils. One of the reasons why the assignments were not anonymous was so that PhD tutors could tailor feedback to individual pupils. Unconscious bias is a common problem in assessments, where pre-existing assumptions about a student's work and performance influence PhD tutors' judgement. The Scholars Programme has a moderation process in place which can mitigate some of these biases, but due to limited resources only a proportion of scripts are systematically moderated.

Challenges	Recommendations
Reliability	A proportion of the assessment should be externally moderated to test the reliability and consistencies of the scores.
Validity	The validity of the standardised mark scheme should be confirmed by comparing pupils' baseline assignment scores and their performance on a standardised written exam (e.g. KS4 English).
	Other approaches to marking our assignments including comparative judgement or including benchmarking examples should also be systematically tested.
Specificity	Further evaluation work should include a control group through a wait-list design.
Bias	To minimise bias, PhD tutors should be well-trained to follow the marking criteria and to be made aware of the common types of unconscious bias that they should avoid.
	Ensure mechanisms are in place to flag assignments with spurious scores.

4.5 Conclusions

This project rendered some positive preliminary results, indicating a uniform increase in all aspects of pupils' academic writing skills following a series of university-style tutorials. The standardised mark scheme was deemed feasible, as all PhD tutors found this a useful assessment tool. All six areas of academic writing skills were strongly related, but some constructs were more closely related than others. Detailed analysis of the data revealed some variability in patterns across PhD tutors, emphasising the need to train PhD tutors effectively and moderate their marks systematically. A more thorough investigation on the effect of subject stream on pupils' progress in academic writing skills is needed. The qualitative feedback from PhD tutors should be used to inform the practicalities of implementing the baseline assignment and the standardised mark scheme for internal evaluation of the large-scale roll out. Future evaluation work should be carried out to further validate the measures obtained from the mark scheme using a larger sample.

5. Bibliography

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

AQA. (2017). *GCSE English Language: Changes to Assessment Objectives.* Retrieved from http://www.aqa.org.uk/resources/english/gcse/english-language-8700/plan/changes-to-assessment-objectives

Bloom, B. S., Engelhart, M., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives.* New York: David McKay.

Casner-Lotto, J., & Barrington, L. (2006). *Are They Really Ready to Work? Employers' Perspectives on the Basic Knowledge and Applied Skills of New Entrants to the 21st Century U.S. Workforce.* Partnership for 21st Century Skills. Retrieved from https://eric.ed.gov/?id=ED519465

Christodoulou, D. (2016). *Making Good Progress? The future of Assessment for Learning.* Oxford University Press.

Department for Children, Schools and Families. (2009). *Improving writing: a handbook for Key Stage 3.* Retrieved from http://dera.ioe.ac.uk/2403/7/se_iw_handbook0067109_Redacted.pdf

Educational Testing Service. (2013). *Quantitative market research [PowerPoint slides].* Princeton, NJ.

Fowles, M. (2012). *Writing assessment for graduate and professional programs: Lessons learned and a note for the future.* In N. Elliot, & L. Perelman (Eds.),Writing Assessment in the 21st century (pp. 135–146). New York, NY: Hampton Press.

Haberman, S. J. (2008). *Subscores and validity (Research Report No. RR-08-64).* Princeton, NJ: Educational Testing Service.

Haswell, R. H. (2008). *Teaching of writing in higher education. In C. Bazerman (Ed.), Handbook of research on writing (pp. 405–424).* New York, NY: Lawrence Erlbaum.

OECD. (2012). *Education at a Glance 2012.* OECD Publishing. https://doi.org/10.1787/eag-2012-en

Pearson. (2017). *Pearson Edexcel Level 1/ Level 2 GCSE (9-1) English Language (1ENO). Sample Assessment Materials.* Retrieved from https://qualifications.pearson.com/content/dam/pdf/GCSE/English%20Language/2015/specification-and-sample-assesment/GCSE-English-Lang-SAMs.pdf

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment. *ETS Research Report Series, 2014*(2), 1–52. https://doi.org/10.1002/ets2.12035

Standards and Testing Agency. (2016). *2016 Teacher Assessment Exemplification: End of Key Stage 2. English Writing. Working towards the expected standard.* Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/510902/STA-Ex2016-KS2-EW-AlexAnn_PDFA.pdf.

6. Acknowledgements

We would like to thank Dr Loren Parry, Data and Analysis Senior Officer at Teach First and Kate Theobald, Associate Director Research and Evaluation of Ambition School Leadership for reviewing a draft of this work and providing insightful comments.

7. Appendices

Appendix A – Baseline and Final Assignment Examples

Baseline assignment	Final assignment
The October Revolution was inevitable. How far do you agree?	Why did the Bolsheviks use the arts to consolidate their power in Russia during the 1920s to1930s?
Explain what we mean by 'identity' and show some of the ways that reading helps us to develop our identities today.	How did reading shape society and the lives of individuals in the eighteenth century?
How relevant are non-renewable sources of energy today?	Should we allow shale gas extraction in the UK?
Using O'Neill's theory of obligation, explain one obligation that you think is not due to everyone, but only to a specified group of people. Give reasons for you answer	What do you understand by the term 'Paternalism'? Do you think that it is ever justifiable for someone to make our life choices for us? What (if any) restrictions would you place on individuals making choices for themselves?

Appendix B – Mark Scheme Table

Skills	1st (70-100)	2:1 (60-69)	2:2 (50-59)
Addressing the Question	○ All materials used are relevant to the general topic and to the specific question/title ○ Clear justification on how the material used is related to the specific issues that are the focus of the essay	○ Most of the materials used are relevant to the general topic and to the specific question/title ○ Adequate justification on how the material used is related to the specific issues that are the focus of the essay	○ Some of the materials used are relevant to the general topic and to the specific question/title ○ Some justification on how the material used is related to the specific issues that are the focus of the essay
Using Evidence	○ Inclusion of rich sources of research findings, data, quotations or other sourced material as evidence for the claims/ ideas ○ Use evidence to support claims/assertions/ideas , consistently clearly and convincingly	○ Inclusion of adequate sources of research findings, data, quotations or other sourced material as evidence for the claims/ ideas ○ Use evidence to support claims/assertions/ideas, mostly clearly and convincingly	○ Inclusion of some sources of research findings, data, quotations or other sourced material as evidence for the claims/ ideas ○ Use evidence to support claims/assertions/ideas, at times clearly and convincingly
Developing an Argument	○ A point of view or position in relation to the title or question is consistently clear ○ Argument exceptionally well-developed and well-justified ○ A position is clearly established in relation to the question, and is developed effectively and consistently throughout the essay	○ A point of view or position in relation to the title or question is adequately clear ○ Argument clear and well-developed and position justified ○ A position is established in relation to the question, and is well-developed in most of the essay	○ A point of view or position in relation to the title or question is somewhat clear ○ Argument clear but not well-developed ○ A position is established in relation to the question, and is well-developed in parts of the essay

Critical Evaluation	○ Moved beyond description to an assessment of the value or significance of what is described ○ Evaluative points are consistently explicit/systematic/reasoned/justified	○ Mostly description but some assessment of the value or significance of what is described ○ Evaluative points are mostly explicit/systematic/reasoned/justified	○ Only description with minimal assessment of the value or significance of what is described ○ Evaluative points are at times explicit/systematic/reasoned/justified
Structuring	○ Ideas are presented in paragraphs and arranged as a logical sequence of ideas ○ The introduction clearly outlines how the essay will deal with the issues ○ The conclusion summarises all the main points clearly and concisely	○ Ideas are presented in paragraphs with some structure ○ The introduction adequately describes how the essay will deal with the issues ○ The conclusion summarises most of the main points clearly	○ Ideas are presented in paragraphs and are loosely structured ○ The introduction mentions how the essay will deal with the issues ○ The conclusion summarises some of the main points clearly
Use of Language	○ No spelling, grammar or punctuation errors ○ Writing style consistently clear, tone appropriate and easy to follow	○ Minimal spelling, grammar or punctuation errors ○ Writing style mostly clear, tone appropriate and easy to follow	○ Some spelling, grammar or punctuation errors ○ Writing style moderately clear, tone appropriate and easy to follow

Appendix C – Mark Scheme Scoring Sheet

Tutor name: _____ School: _____ Today's date_____

	Marks (out of 100)											
Pupil Names												
Addressing the Question												
Using Evidence												
Developing an Argument												
Critical Evaluation												
Structuring												
Use of Language												

Appendix D. Pupil average baseline assignment (BA) and final assignment (FA) marks (and standard deviation) by subject stream, schools and PhD tutors.

Schools	Tutor	Subject Stream	BA Submission (n)	BA Marks (out of 100)	FA Submission (n)	FA Marks (out of 100)	BA & FA Assignment Submission	Average Change
A	AA	AH	8	61.31 (4.00)	7	68.12 (7.06)	7	7.10 (6.34)
B	AA	AH	7	54.26 (6.83)	7	61.26 (10.10)	7	6.5 (11.11)
C	BB	AH	8	56.15 (4.07)	6	64.31 (6.64)	6	10.41 (5.10)
D	BB	AH	11	60.52 (6.02)	11	64.26 (7.14)	11	3.74 (4.73)
E	CC*	STEM	7	49.88 (3.77)	6	70.62 (8.77)	6	22.11 (10.63)
F	CC	STEM	10	57.42 (6.394)	10	79.21 (11.76)	10	21.55 (9.97)
G	DD	AH	7	61.04 (1.79)	0	-	0	-
H	DD	AH	12	63.86 (2.95)	0	-	0	-
I	EE*	AH	9	58.13 (7.75)	5	62.03 (11.82)	5	6.00 (1.63)
Total/ Average	-	-	83	60.17 (7.00)	61	69.52 (11.48)	56	11.04 (10.18)

*A moderation rule was applied to all the scripts on this placement